Probabilistic Analysis of Various Algorithms

a dissertation presented
by
Anish Sevekari
to
The Department of Mathematics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Algorithms, Combinatorics and Optimization

Carnegie Mellon University
Pittsburgh
Pennsylvania
August 2023

Thesis Committee

Professor Wesley Pegden
Department of Mathematics, Carnegie Mellon University
Thesis Advisor and Committee Chair

Professor Andrej Risteski
Department of Machine Learning, Carnegie Mellon University

Professor Alan Frieze
Department of Mathematics, Carnegie Mellon University

Professor Prasad Tetali
Department of Mathematics, Carnegie Mellon University

Probabilistic Analysis of Various Algorithms

Abstract

In recent year, theory and practice in computer science has steered away from each other in many aspects. Recent improvements in computational capabilities and field of optimization have seen rise to the use of various different heuristics, which work in practice with great success, but have not seen much investigation on the theory side. This has created a need for theoretical investigation to bridge the gap between two branches of computing. More frequently than not, the heuristic choices relies on known empirical observations, and intuitive understanding of trade-off between runtime, memory, quality of approximation and probability of success of these algorithms.

In this thesis, we discuss a few such interesting dilemmas, and try to provide provable justifications for some of them by using various probabilistic and analytical techniques. We will focus on two main topics - average-case approximation quality of various lower bounds used for Euclidean Traveling Salesman Problem and computational versus statistical efficiency of modern generative models, like noise contrastive estimation and score estimation.

Contents

List of Figures

Chapter 1

Introduction

Currently, we are in the era where complexity theory and computational practice of algorithms have outpaced each other. Particularly, in the field of optimization. In variety of applications, even for the problems that are proved to be NP-hard, we do have heuristical approaches which perform significantly better than expectations, and provide either approximate or exact answers even for large instances quite successfully. This gives rise to various questions which would help us fill in these gaps - *which heuristics are likely to work and under what conditions*?

One popular approach to understanding different heuristics is to look at the average-case analysis. The simplest example is (*non-randomized*) quick-sort algorithm, where pivot element is fixed in advance, which has worst case runtime of $O(n^2)$, but average case runtime of $O(n \log n)$. One of the most popular results in this directions would be the *smoothed analysis* of simplex algorithm Spielman and Teng [ST09], which proves that simplex method runs in polynomial time on average[1], explaining why simplex is often preferred over interior-point methods, which have a provably polynomial runtime in worst case[2]. Average-case analysis is an important tool for studying *Euclidean Traveling Salesperson Problem* (TSP). In fact, the value of optimal TSP tour on a typical instance - n uniformly random points in $[0,1]^2$ is highly concentrated around $C\sqrt{n}$ for some absolute constant C [BHH59]. Similar analysis for Karp's Dissection algorithm tells us that it also converges to $C\sqrt{n}$ for the same constant C, justifying success of dissection algorithm in practice. In terms of lower bounds, similar results were proved for min-cost maximum matchings [Pap78] and Held-Karp Linear relaxation of TSP [GB91]. In particular, there was strong empirical evidence suggesting that Held-Karp LP relaxation and TSP converge to the same value, which was recently disproved [FP15]. Although the fact that constants are nearly equal implies that Held-Karp LP relaxation provides a lower bound very close to optimal, the separation of two constants implies that it cannot be used to produce an exact algorithm! In chapters 2 and 3 we will explore extensions of some of the results in this direction.

We observe a similar phenomenon among generative models in machine learning. The overarching theme for generative modeling is to fix a parametric family of distributions $\{p_\theta, \theta \in \Theta\}$ and given samples from a distribution p_*, find the value of θ^* such that p_* and p_{θ^*} are close. The choice of

[1] In fact, average-case runtime for a random perturbation of worst-case instance is polynomial in input size.

[2] Runtime of simplex is exponential in worst-case

the family p_θ and estimation algorithm for θ^* is often made heuristically using previous empirical observations. Apart from runtime complexity, an important parameter for any estimator is statistical efficiency - the rate of convergence of mean-squared error, $\mathbb{E}[\|\hat{\theta}_n - \theta^*\|_2]$, where $\hat{\theta}_n$ is the output of estimator. The gold standard for the estimation algorithm is the *Maxmimum-Likelihood Estimation* (MLE), since MLE has optimal statistical efficiency (Le Cam 1953, see [Vaa98]), and estimator that performs better than MLE only does so on a 0-measure (lebesgue) subset of Θ. But the state-of-art for estimators has diverged from MLE and uses various other estimators, since they are believed to be computationally efficient. Two of the most popular alternatives are *Noise Contrastive Estimation*, introduced in [GH10; GH12] and *score matching*, introduced in [Hyv05]. This gives rise to two questions: How much do noise contrastive estimate and score matching help computationally, and how much do they lose statistically, when compared to MLE? Some facets of these questions have recently been explored [KHR22; Liu+21; Che+22], and we will explore some of the remaining in chapters 5 and 6.

Organization of the thesis:

This thesis groups together 7 fairly different pieces of work, and every chapter corresponds to a different result.

Chapter 2 extends on work of Frieze and Pegden [FP15] and proves that addition of Comb inequalities, which are a special case of cutting plan inequalities to Held-Karp LP relaxation does not suffice to bridge the gap for asymptotic convergence, and prove a separation result. Chapter 3 looks at a partial variant of $\mathtt{TSP}$ problem, where we want a tour through only an ε fraction of points, and prove a separation between behaviors of functionals $\mathtt{TSP}, \mathtt{MST}$ and $\mathtt{MM}, \mathtt{TF}$ as $\varepsilon \to 0$. The per edge cost of maximum-matching or two-factor goes to zero as $\varepsilon \to 0$, but for $\mathtt{TSP}$ it does not, and can be lower bounded by an absolute constant.

Chapter 6 describes a simple case where the score matching estimator performs provably better than MLE. We construct an exponential family over which there is no polynomial time algorithm to compute MLE unless $\mathsf{RP} = \mathsf{NP}$, but on the other hand, score matching estimator can be computed in polynomial time, while only losing a polynomial factor in statistical efficiency over MLE. Chapter 5 describes a scenario where noise-contrastive estimation has exponentially bad complexity as compared to MLE, even when the true distribution and noise distribution have matching first and second moments.

Chapter 4 looks at a problem of sampling nearly shortest self-avoiding walks on a grid graph. We provide a direct dynamic programming algorithm with expected polynomial runtime, while proving that the Markov chain approach in this setting has exponential mixing time. Chapter 7 proves an universal approximation result, showing existence of normalizing flow networks that have a well-conditioned Jacobian, provided that true distribution is log-concave. Finally, Chapter 8 provides an efficient single-pass algorithm to compute a low-dimensional subspace approximation to a given matrix.

Chapter 2

Comb Inequalities for Euclidean Traveling Salesman Problem

Papadimitriou showed that the Euclidean TSP is NP-hard, while Arora [Aro96] and Mitchell [Mit99] described polynomial-time approximation schemes (PTAS) for the Euclidean TSP. On the computational side: efficient implementations of these PTASs have not materialized to supplant the use of heuristics without provable guarantees, while on the other hand, branch-and-cut methods using these heuristics with LP-based lower bounds nevertheless have found (provably) optimal tours in random or real-world (rather than worst-case) problem instances of large size; the current record is a problem instance from an application to integrated circuit design with 85,900 "cities" [App+06].

Underpinning the tension in these developments is the unresolved status (even subject to standard complexity assumptions) of the hardness of finding optimal tours on *typical*—rather than worst-case—instances of the Euclidean TSP:

Question 1. Is there a polynomial-time algorithm for the Euclidean TSP which, given a collection of n independent random points, returns an optimal tour with probability p_n where $p_n \to 1$ as $n \to \infty$?

2.1 Preliminaries

2.1.1 Branch-and-cut for the Euclidean TSP

One of the most successful computational approaches in practice to find optimal tours for the Euclidean TSP is the *branch-and-cut* approach, discussed by Applegate, Bixby, Chvátal and Cook [App+06], and implemented in Cook's software package *Concorde*.

Before discussing branch-and-cut, let us first recall that the more general *branch-and-bound* approach is a combinatorial optimization paradigm based on pruning a branched exhaustive search. In the context of finding optimal TSP tours, the approach combines (sub-optimal) algorithms for finding tours subject to restrictions (e.g., edge inclusions/exclusions), methods to establish lower bounds on tour lengths subject to restrictions, and a branching strategy which recursively partitions the exhaustive search space into complementary sets of restrictions. Efficiency of the approach depends on lower bound methods being strong enough on restricted instances to match the global performance of upper bound (tour-finding) approaches to quickly prune large parts of the search space.

Within this paradigm, *branch-and-cut* algorithms for the TSP specialize by using an LP relaxation lower bound for the TSP, which, for each constrained instance, can be augmented by an adaptive choice of cutting planes. The algorithm *branches*, partitioning a problem instance into a collection of problem instances with complementary restrictions, and then prunes by searching for *cut*ting planes for each.

Frieze and Pegden [FP15] showed that regardless of the tour-finding algorithm used for upper bounds (i.e., even if it actually finds optimal tours), the branch and bound decision tree will inevitably have exponential size if lower bounds are found via the Held-Karp LP-relaxation of the TSP, without any additional cutting planes [HK71].

This *Held-Karp lower bound* on the tour is defined by the linear program:

$$\min \sum_{\{i,j\}\subseteq V} c_{\{ij\}} x_{\{ij\}}$$

$$\text{subject to}$$

$$\begin{array}{llr} \text{(I)} & (\forall i) & \displaystyle\sum_{j\neq i} x_{\{ij\}} = 2 \\ \text{(II)} & (\forall \varnothing \neq S \subsetneq V) & \displaystyle\sum_{\{i,j\}\subseteq S} x_{\{ij\}} \leq |S| - 1 \\ \text{(III)} & (\forall i < j \in V) & x_{\{ij\}} \in [0,1] \end{array} . \tag{2.1}$$

Let $\mathtt{HK}(X)$ denote the value of this LP on a set X. Note that under assumption (I) in (2.1), (II) can be replaced by

$$(\forall \varnothing \neq S \subsetneq V) \sum_{i\in S, j\notin S} x_{\{ij\}} \geq 2 \tag{2.2}$$

as shown in Section 58.5 in [Sch03]; these are known as *subtour-elimination* constraints.

The branch-and-cut approaches used to solve TSP instances of significant size go beyond the branch-and-bound framework considered by Frieze and Pegden, by using additional cutting planes to further prune the TSP search space. Perhaps the most important class of such cutting planes are the so-called *comb-inequalities* (which are valid for any solution x corresponding to a TSP tour [GP86]).

Definition 2 (Comb Inequality)**.** Given sets H and $T_1, \dots, T_t$ for odd t, such that $T_i \cap T_j = \varnothing$

and $T_i \cap H \neq \varnothing$, the *comb inequality* associated to these sets is given by

$$\sum_{\substack{i \in H \\ j \notin H}} x_{\{ij\}} + \sum_{k=1}^{t} \sum_{\substack{i \in T_k \\ j \notin T_k}} x_{\{ij\}} \geq 3t + 1.$$

In this case, we call H to be the *handle* and T_i to be the *teeth* of comb inequality. We refer to $C = H \cup \left(\cup_{k=0}^{t} T_k\right)$ as the comb and we will use the term *size of the comb* to denote $|C|$.

We will obtain in this paper a proof that polynomial-time branch-and-cut algorithms based on comb inequalities of bounded size cannot solve the Euclidean TSP on typical instances. In particular, let $\texttt{Comb}_c(X)$ denote the value of the LP obtained by adding all comb inequalities with combs of size at most c to the Held-Karp LP relaxation of TSP. For a random set $\mathcal{X}_n$ of n points in $[0,1]^d$, we prove:

Theorem 3. *Suppose that we use branch and bound to solve the TSP on $\mathcal{X}_n$, using $\texttt{Comb}_c$ as a lower bound for some fixed constant c. Then the algorithm runs in time $e^{\Omega(n/\operatorname{polylog}(n))}$ almost surely.*

Note that this gives a almost-exponential lower bound on the runtime of any branch and bound strategy. Further, we have a slightly more general version of this result when c is not a constant, but with a slightly weaker but still super-polynomial lower bound on the runtime:

Theorem 4. *Suppose that we use branch and bound to solve the TSP on $\mathcal{X}_n$, using $\texttt{Comb}_c$ as a lower bound for $c = O\left(\frac{\log n}{\log \log n}\right)$. Then the algorithm runs in time $e^{\Omega(n^{0.5})}$ almost surely.*

The set of all combs of size $\frac{\log n}{\log \log n}$ has size at least $n^{\Omega(\log n/(\log \log n))}$, which is super polynomial. It is not clear that there should be a polynomial-time separation algorithm for this set of comb-inequalities. The known results for separation of comb inequalities are for combs with a bounded number of teeth [Car97], and combs that are derived in a specific way [Car04].
The proof of the two theorems above theorem, along with a precise definition of the branch-and-bound paradigm we consider, can be found in Section 2.3. The applicability of Theorems 3 and 4 to branch-and-cut follows from the fact that a branch-and-cut tree using only combs of size $\leq c$ contains as a subtree the corresponding branch-and-bound which uses $\texttt{Comb}_c$ as a lower bound. The proofs of Theorems 3 and 4 depends on a new extension of probabilistic analyses of the Euclidean TSP and its LP relaxations.

2.1.2 Probabilistic analysis of cutting planes for the Euclidean TSP

The proof of Theorem 3 will depend on a probabilistic analysis of the impact of comb-inequality cutting planes on the value of the Held-Karp linear program (2.1). In particular, if $x_1, x_2, \ldots$ is a sequence of random points in $[0,1]^d$ and $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, we aim to show for any constant c that for some $\varepsilon > 0$,

$$\texttt{Comb}_c(\mathcal{X}_n) \leq (1-\varepsilon)\texttt{TSP}(\mathcal{X}_n) \quad \text{almost surely } (a.s.),$$

where $\texttt{TSP}(X)$ denotes the length of a shortest tour through X. The random variable $\texttt{TSP}(\mathcal{X}_n)$ was first studied by Beardwood, Halton and Hammersley [BHH59]. They proved in 1959 that there is an absolute constant β^d_{TSP} such that the length $\texttt{TSP}(\mathcal{X}_n)$ of a minimum length TSP tour through $\mathcal{X}_n$ satisfies

$$\texttt{TSP}(\mathcal{X}_n) \sim \beta^d_{\texttt{TSP}} n^{\frac{d-1}{d}} \qquad a.s.$$

Here $a_n \sim b_n$ indicates that $a_n/b_n \to 1$. This result has since been extended to many structures other than Hamiltonian cycles. Various similar results are also known for problems like Minimum Spanning Tree [BHH59] and Maximum Matching [Pap78], etc. Steele [Ste81] extended this result to a more general framework which proves existence of such asymptotic constants β_F for *subadditive Euclidean functional* F. One peculiar feature of these results is that the true values of the constants are unknown, and even improvements on their estimates are rare. Some results in this direction were proved in [BV90] and [Ste15].

Goemans and Bertsimas established in [GB91] an analogous asymptotic result for the Held-Karp linear program:

$$\texttt{HK}(\mathcal{X}_n) \sim \beta^d_{\texttt{HK}} n^{\frac{d-1}{d}}$$

by proving that $\texttt{HK}(X)$ is a subadditive Euclidean functional. They asked in [GB91] whether $\beta^d_{\texttt{HK}} = \beta^d_{\texttt{TSP}}$; this was answered in the negative in the same paper [FP15] showing that branch-and-bound with $\texttt{HK}(\mathcal{X}_n)$ as a lower bound takes exponential time on typical inputs; Frieze and Pegden proved there that

$$\beta^d_{\texttt{HK}} < \beta^d_{\texttt{TSP}} \qquad \forall d \geq 2. \tag{2.3}$$

Let $\texttt{Comb}_c$ denote the value of the LP obtained by adding all comb inequalities with combs of size at most c to the Held-Karp LP relaxation of TSP. Since $\texttt{Comb}_c(X) \leq \texttt{TSP}(X)$ for all $x \in \mathbb{R}^d$, there is some constant γ such that

$$\limsup_{n\to\infty} \texttt{Comb}_c(\mathcal{X}_n) \cdot n^{-\frac{d}{d-1}} \leq \gamma \quad \text{a.s.} \tag{2.4}$$

Note that $\gamma = \beta^d_{\texttt{TSP}}$ satisfies this equation.

Definition 5. Let Γ denote the set of constants that satisfy (2.4). Define

$$\gamma^{c,d}_{\texttt{Comb}} = \inf_{\gamma\in\Gamma} \gamma.$$

We claim that $\gamma^{c,d}_{\texttt{Comb}} \in \Gamma$. This holds since for all m, we have

$$\mathbb{P}\left[\limsup_{n\to\infty} \texttt{Comb}_c(\mathcal{X}_n) \cdot n^{-\frac{d-1}{d}} > \gamma^{c,d}_{\texttt{Comb}} + \frac{1}{m}\right] = 0.$$

By taking a countable union of all these events, we get that

$$\mathbb{P}\left[\limsup_{n\to\infty} \texttt{Comb}_c(\mathcal{X}_n) \cdot n^{-\frac{d-1}{d}} > \gamma^{c,d}_{\texttt{Comb}}\right] = 0,$$

proving that $\gamma^{c,d}_{\texttt{Comb}}$ satisfies (2.4) and lies in Γ. With these definitions above, we will prove

Theorem 6. *For all constants c and for all $d \geq 2$,*

$$\gamma^{c,d}_{\texttt{Comb}} < \beta^d_{\texttt{TSP}} \tag{2.5}$$

The proof of Theorem 6 appears in Section 2.2. In Section 2.3 we show that this theorem implies Theorem 3.

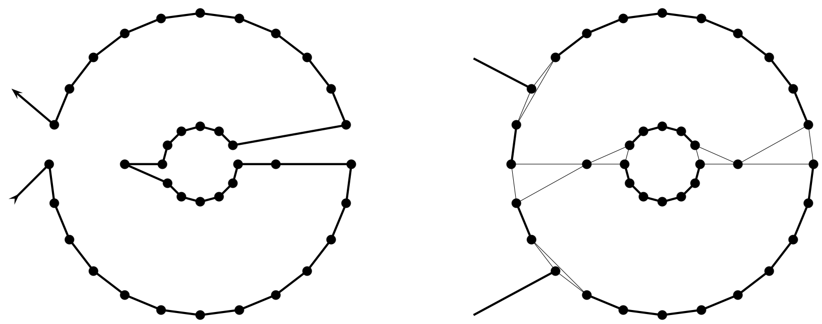


Figure 2.1: Solution when tour enters the gadget only once. Thick edges have value 1 and thin edges have value 0.5.

2.1.3 Notation

Given a graph $G = (V, E)$ and $A, B \subset H$, let $\delta(A)$ denote the set of edges of G, with exactly one vertex inside A. If A, B are disjoint, then let $e(A, B)$ denote the set of edges in G with exactly one vertex in A and one vertex in B.
A weight assignment x is a function $x : E \mapsto \mathbb{R}$. Let $F \subset E$, then

$$x(F) = \sum_{e \in F} x(e)$$

denotes the total weight of edges in F. In particular, $x(\delta(A))$ denotes the total weight leaving the set A, and $x(e(A, B))$ denotes the total weight of edges going from A to B.

2.2 Separating Constant Size Comb LP from TSP

Frieze and Pegden show in [FP15] that for all $d \geq 2$, $\beta^d_{\mathtt{HK}} < \beta^d_{\mathtt{TSP}}$. They prove the result by constructing a gadget such that the length of any tour while passing through the gadget is significantly larger than the total contribution of a solution satisfying subtour elimination constraints. They then prove that suitable approximations to this gadget occur frequently enough in random set to ensure that the an LP solution can be found of length $(1 - \varepsilon)\mathtt{TSP}(\mathcal{X}_n)$. We now define this gadget $S(k)$.

Definition 7. The gadget $S(k)$ consists of $2k$ equally spaced points on the circle of radius 4 and k equally spaced points on the circle of radius 1, along with the points $(2, 0)$ and $(-2, 0)$, which we refer to as the *gap vertices*.

Observation 9 (Observation 3.10 from [FP15]) states that we can enter a copy of this gadget at most twice. Section 2.2 shows the gadget with a TSP (on the left) and corresponding Held-Karp solution (on the right) when the TSP enters/leaves the gadget just once, while Section 2.2 shows the same when the tour enters the same figure at most twice. Note that in both the cases, the tour crossed the gap between smaller and larger circle roughly 3 times, while the half-integral solution

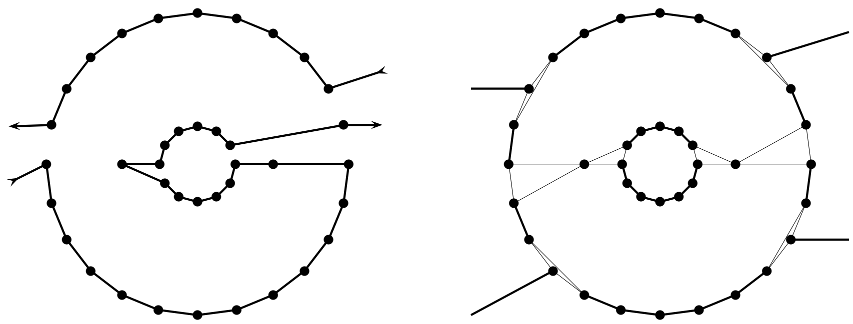


Figure 2.2: Solution when tour enters the gadget exactly twice. Thick edges have value 1 and thin edges have value 0.5.

(on the right) crosses this gap only twice (since the edges crossing the gap have weight 0.5). Thus there is a constant gap between values of these solutions.

The proof in [FP15] of (2.3) incorporates the following two observations. Before stating them, we will recall an important definition from [FP15]:

Definition 8. Consider a set $X \subset \mathbb{R}^n$. A set $T \subset X$ is (ε, D)-copy of $S \subset \mathbb{R}^n$ if there is a set $S' \cong S$[1] and a bijection f between T and S' such that for all $x \in T$, $\|x - f(x)\| < \varepsilon$, and such that T is at distance $> D$ from $X \setminus T$.

Note that when we refer to a scaled (ε, D)-copy of S, by say a factor t, we mean a $(t\varepsilon, tD)$-copy of $t \cdot S$.

Observation 9 (Observation 3.10 from [FP15])**.** Suppose that $S_{\varepsilon,D}$ is an (ε, D) copy of any fixed set S for fixed ε and sufficiently large D. Then there are at most 2 pairs of edges in a shortest TSP tour which join $S_{\varepsilon,D}$ to $V \setminus S_{\varepsilon,D}$.

Observation 10 (Observation 3.1 from [FP15])**.** Let $\{Y_1, Y_2, \ldots\}$ be a sequence of points drawn uniformly at random from $[0,t]^d$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$, where $t = n^{1/d}$. Given any finite point set S, any $\varepsilon > 0$, and any D, $\mathcal{Y}_n$ a.s. contains at least $C^S_{\varepsilon,D} n$ (ε, D)-copies of S, for some constant $C^S_{\varepsilon,D} > 0$.

The structure of the proof of Equation (2.3) from [FP15] is than as follows:

(i) For $\mathcal{Y}_n = t \cdot \mathcal{X}_n$, Observation 10 ensures that we can choose a large constant D and a small constant $\varepsilon > 0$ and find linearly many (ε, D)-copies of the gadget described above.

(ii) By Observation 9, for each (ε, D) copy of the gadget, the shortest tour through $\mathcal{Y}_n$ has either one or two components when restricted to the gadget.

(iii) For both of these two possible cases, in each approximate copy of the gadget, the tour can be locally shortened by relaxing to a (half-integral) LP solution as in Section 2.2.2.

[1]For $S', S \in \mathbb{R}^n$, we write $S' \cong S$ if there is an isometry of $\mathbb{R}^n$ that maps S to S'.

(iv) In total these shorten the tour by $\delta \cdot n$ for some $\delta > 0$, which establishes (2.3) since after rescaling by the factor t we have that $\texttt{TSP}(\mathcal{Y}_n) \sim \beta^d_{\texttt{TSP}} n$.

To extend this approach to prove Theorem 6, we will do the following:

(1) Construct a local half-integral solution on $S = S(k)$ assuming that tour visits S exactly onces, entering and exiting through adjacent vertices (satisfying Property 21).

(2) Prove that this solution satisfies all comb inequalities of size c for $k = O(c)$.

(3) Construct a gadget $\Pi^3(S)$ that contains 12 copies of S and any optimal tour through $\Pi^3(S)$ must go through at least one copy of S while satisfying Property 21.

To begin, we prove some structural lemmas about combs.

2.2.1 Technical lemmas for comb inequalities

For the lemmas in this section, we suppose that x is a half-integral solution to the Held-Karp LP, which has the property that all the edges of weight $1/2$ in x form a graph that can be written as a union of edge disjoint triangles.

Lemma 11. *If C is a comb violated by x with handle H and teeth T_i for $i = 1 \dots t$ for odd t, then following must hold:*

1. $x(\delta(H)) = t$
2. $x(\delta^*(H)) = 0$
3. $x(\delta(T_i)) = 2$ *for all* i.
4. $x(e(A_i, B_i)) = 1$.
5. $x(e(A_i, H \setminus A_i)) = 1$.
6. $x(e(B_i, X \setminus (H \cup T_i))) = 1$.

where $A_i = T_i \cap H$, $B_i = T_i \setminus H$ and $\delta^*(H)$ denotes the edges with exactly one endpoint inside H, and at least one endpoint outside $\bigcup_{i=1}^t T_i$.

Proof. Suppose x violates the comb inequality C with handle H and teeth $T_1, \dots, T_t$ for odd t.

Since this is a comb inequality, we know that T_i intersect H, and are pairwise disjoint. For each i, define $A_i = T_i \cap H$ and $B_i = T_i \setminus H$. For any set two sets S, T, let $e(S,T)$ denote the set of edges with one endpoint in S and another in T, and let $\delta(S)$ denote the set of all edges with exactly one endpoint in S. Let x denote the solution of LP that we are considering. That is, for any edge e, $x(e)$ denote the value associated to that edge. For any set $U \subseteq E$,

$$x(U) = \sum_{e \in U} x(e)$$

is the total weight of the set of edges.
The comb-inequality constraint is given by

$$x(\delta(H)) + \sum_{i=1}^{t} x(\delta(T_i)) \geq 3t + 1.$$

Since the comb inequality is not valid for the solution, we have

$$x(\delta(H)) + \sum_{i=1}^{t} x(\delta(T_i)) < 3t + 1.$$

From subtour elimination, we have $x(\delta(A_i)) \geq 2$ and $x(\delta(B_i)) \geq 2$. Since A_i and B_i partition T_i, we have

$$x(\delta(T_i)) = x(\delta(A_i)) + x(\delta(B_i)) - 2x(e(A_i, B_i)). \tag{2.6}$$

Let $\delta^*(H)$ denote all the edges exiting H that have are not contained inside a single tooth.

$$\delta^*(H) = \delta(H) \setminus \left(\bigcup_{i=1}^{t} e(A_i, B_i) \right)$$

Substituting this into the comb inequality,

$$x(\delta^*(H)) + \sum_{i=1}^{t} \big(x(e(A_i, B_i)) + x(\delta(T_i)) \big) < 3t + 1. \tag{2.7}$$

Because of subtour elimination constraints, we have $x(\delta(T_i)) \geq 2$ for all i, which gives

$$\begin{aligned} x(\delta^*(H)) + \sum_{i=1}^{t} x(e(A_i, B_i)) &< t + 1 \\ \implies \sum_{i=1}^{t} x(e(A_i, B_i)) &< t + 1 - x(\delta^*(H)) \end{aligned} \tag{2.8}$$

on the other hand, (2.6) gives

$$\begin{aligned} x(\delta^*(H)) + \sum_{i=1}^{t} \left(x(e(A_i, B_i)) + x(\delta(A_i)) + x(\delta(B_i)) - 2x(e(A_i, B_i)) \right) &< 3t + 1 \\ \implies \sum_{i=1}^{t} \left(x(\delta(A_i)) + x(\delta(B_i)) \right) - 3t - 1 + x(\delta^*(H)) &< \sum_{i=1}^{t} x(e(A_i, B_i)) \\ \implies t - 1 + x(\delta^*(H)) &< \sum_{i=1}^{t} x(e(A_i, B_i)) \end{aligned} \tag{2.9}$$

Combining the both, we have

$$t - 1 + x(\delta^*(H)) < \sum_{i=1}^{t} x(e(A_i, B_i)) < t + 1 - x(\delta^*(H)). \tag{2.10}$$

This immediately forces only two possible values of $x(\delta^*(H))$, either 0 or 1/2.

Substituting the lower bound (2.9) into (2.7), we get

$$x(\delta^*(H)) + \sum_{i=1}^{t} x(\delta(T_i)) + t - 1 + x(\delta^*(H)) < 3t + 1$$
$$\implies \sum_{i=1}^{t} (x(\delta(T_i)) - 2) < 2 - 2x(\delta^*(H))$$

Recall that edges of weight 1/2 in x form a graph that can be written as union of edge disjoint triangles. For any set S, any triangle can have either exactly 2 edges crossing it, or no edges crossing it. Hence, for any S, a triangle with all edges of weight 1/2 contributes either 1 or 0 to $x(\delta(S))$. Since we can decompose all the edges of weight 1/2 into edge disjoint triangles, no edges are double counted while adding up elements in $\delta(S)$, so for each set S, $x(\delta(S))$ is an integer. Now, observing the equation above, we can note that there is at most one T_i for which $x(\delta(T_i)) = 3$. Further, even this cannot happen if $x(\delta^*(H)) = 1/2$. Now we are left with three cases, namely:

1. $x(\delta^*(H)) = 1/2$ and $x(\delta(T_i)) = 2$ for all i.
2. $x(\delta^*(H)) = 0$, $x(\delta(T_1)) = 3$ and $x(\delta(T_i)) = 2$ for all $i \neq 1$.
3. $x(\delta^*(H)) = 0$ and $x(\delta(T_i)) = 2$ for all i.

we will show that only case (3) can happen.

Case 12. In this case,

$$x(\delta(T_i)) = x(\delta(A_i)) + x(\delta(B_i)) - 2x(e(A_i, B_i)) = 2$$

since $x(\delta(A_i)), x(\delta(B_i)) \geq 2$, this gives $x(e(A_i, B_i)) \geq 1$. Substituting this in (2.8) gives

$$t + 1 - \frac{1}{2} > \sum_{i=1}^{n} x(e(A_i, B_i)) \geq t$$

Since the sum only takes half integral values, this forces the value of the sum to be t. So, $x(e(A_i, B_i)) = 1$ for all i. Now, note that

$$x(\delta^*(H)) = x(\delta(H)) - \sum_{i=1}^{t} x(e(A_i, B_i)) = x(\delta(H)) - t$$

which implies that $x(\delta^*(H))$ is an integer since $x(\delta(H))$ is an integer, and hence must be zero, forcing us to be in case 14 instead.

Case 13. In this case, by the same argument as in case (1), we have $x(e(A_i, B_i)) \geq 1$ for all $i > 1$, and $x(e(A_1, B_1)) \geq 1/2$. Substituting these values in (2.7) gives

$$t - \frac{1}{2} \leq \sum_{i=1}^{t} x(e(A_i, B_i)) < 3t + 1 - \sum_{i=1}^{t} x(\delta(T_i)) = t$$

which forces equality on the left since summation only takes integer values. Therefore, $x(e(A_1, B_1)) = 1/2$, and thus there is exactly one edge of weight 1/2 between A_1 and B_1. This edge is part of a triangle, whose vertex must lie outside T_i. But, then it contributes to $x(\delta^*(H))$, and will contradict the assumption that $x(\delta^*(H)) = 0$. Therefore, case (2) can't hold either, which means we are in case 14

Case 14. Now we have $x(\delta^*(H)) = 0$, and hence if there is an edge of weight $1/2$ in $e(A_i, B_i)$, then the unique triangle containing that edge in the decomposition must also be completely contained in T_i. Therefore, every triangle with edges of weight $1/2$ contributes either 1 or 0 to $x(e(A_i, B_i))$. Further, by the same argument as in analysis in case (1), $x(e(A_i, B_i)) \geq 1$ for all i, and using (2.8) gives

$$t + 1 > \sum_{i=1}^{t} x(e(A_i, B_i)) \geq t.$$

Forcing the following equalities for all i:

1. $x(e(A_i, B_i)) = 1$.
2. $x(e(A_i, H \setminus A_i)) = 1$.
3. $x(e(B_i, X \setminus (H \cup T_i))) = 1$.

and these are the only non empty boundary crossings with respect to x for A_i, B_i. Thus, each of these boundaries is either an edge of weight 1, or a triangle with two edges of weight $1/2$ crossing the boundary. This completes the proof. □

Now we are ready prove a couple of trivial lemmas. But first, we will define *induced subgraphs* with respect to an assignment x.

Definition 15. Given a set $X \in \mathbb{R}^n$, and an assignment x, for every subset $Y \subseteq X$, we define $G[Y]$ to be the graph with vertex set Y and edges e with both endpoints in Y such that $x(e) > 0$.

Lemma 16. *For any comb C violated by x, with teeth T_i, the induced subgraph $G[T_i]$ is connected for all teeth T_i.*

Proof. Suppose not, then applying subtour elimination constraint on each connected component (there are at least two) gives $x(\delta(T_i)) \geq 4$. □

Lemma 17. *Consider an assignment x and a comb C with handle H and teeth T_i such that x violates the comb inequality corresponding to the comb C. If C is the comb with least number of teeth such that x violates C, then the induced subgraph $G[H]$ is connected.*

Proof. Suppose not, and let H_i be the connected components of $G[H]$. Let $\alpha_i = \{j : T_j \cap H_j \neq \varnothing\}$ denote the set of teeth intersecting H_i. Note that by Lemma 11, edges exiting any teeth into the handle must have weight 1. This and the fact that weight $1/2$ edges form a graph that can be decomposed into edge disjoint triangles imply that a tooth can't interest two different connected components of the handle. Then, by the constraints in Lemma 11, $x(\delta(H_i)) = |\alpha_i|$ since edges in H_i can only exit through some teeth T_j with $j \in \alpha_i$, and they must exit with weight 1. Therefore, it follows that

$$x(\delta(H_i)) + \sum_{j \in \alpha_i} x(\delta(T_i)) = 3|\alpha_i|.$$

Since at least one of the α_i must be odd, this gives us a smaller comb on which the solution violates the comb inequality, contradicting minimality of H. □

Lemma 18. *Any comb violated by x must contain an edge of weight $1/2$ inside it.*

Proof. Suppose not. Note that edges exiting the handle exit through a tooth, so all of them must have weight 1 by Lemma 11. Since all the edges intersecting the handle have weight 1, we can split the handle into connected components, which are paths. Note that each path contributes exactly 2 to $x(\delta(H))$, and thus $x(\delta(H))$ must be even, which contradicts that $x(\delta(H)) = t$ is odd. $\square$

Definition 19. For any set S, define $E(S, n)$ to be the size of the smallest set $T \supseteq S$ such that $x(\delta(T)) \leq n$.

Note that $x(\delta(T_i)) = 2$ and $x(\delta(H)) = t$. Hence, a handle can only contain sets that have small $E(S, t)$ values and a tooth can only contain sets that have small $E(S, 2)$ values.

Lemma 20. *Let $S \subset T$ be sets such that for all $u \in S$, $x(e(u, T \setminus S)) \leq 1$. Suppose $x(\delta(S)) = n$ and $x(\delta(T)) = n - 1$. Then there are two vertices $u, v \in S$ such that T contains a path from u to v outside S.*

Proof. For each $u \in S$, define P_u to be the set of vertices in $T \setminus S$ that are connected to u using edges in T but outside S. If $P_u \cap P_v \neq \varnothing$ for some $u \neq v$, then there is a path from u to v strictly contained in $T \setminus S$, and $P_u = P_v$.

Suppose this doesn't happen. Then P_u are disjoint for all u. Let

$$T^* = T \setminus \left(S \cup \bigcup_{u \in S} P_u \right).$$

There are no edges between T^* and S by definition. We have the following:

$$x(\delta(T)) = x(\delta(S)) + x(\delta(T \setminus S)) - 2x(e(S, T \setminus S))$$

T^* along with P_u form a partition of $T \setminus S$. Note that there are no edges between any of these parts by definition. Therefore,

$$x(\delta(T \setminus S)) = x(\delta(T^*)) + \sum_{u \in S} x(\delta(P_u)).$$

and if $\{u, v\} \in e(S, T \setminus S)$ with $u \in S$, then $v \in P_u$ by definition. Therefore,

$$x(e(S, T \setminus S)) = \sum_{u \in S} x(e(u, P_u))$$

Using these identities, we get

$$x(\delta(T)) = x(\delta(S)) + x(\delta(T^*)) + \sum_{u \in S} \big(x(\delta(P_u)) - 2x(e(u, P_u)) \big) \tag{2.11}$$

Observe that

$$x(\delta(P_u \cup u)) = x(\delta(u)) + x(\delta(P_u)) - 2x(e(u, P_u))$$

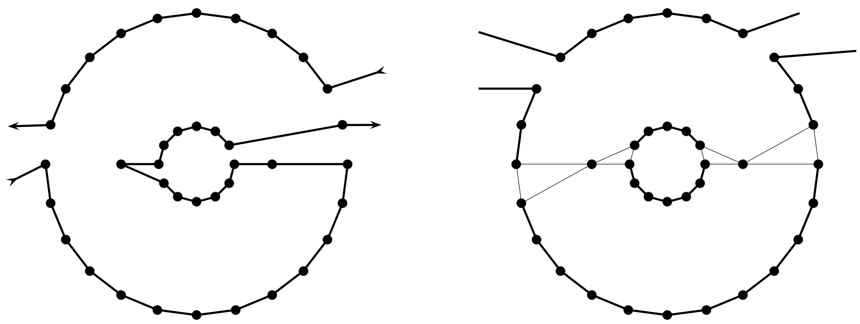


Figure 2.3: The gadget with the tour enters it once. Thick edges have weight 1 and thin edges have weight 0.5.

and therefore that

$$x(\delta(P_u)) - 2x(e(u, P_u)) = x(\delta(P_u \cup u)) - x(\delta(u)) \tag{2.12}$$

Now, we claim that $x(\delta(P_u \cup u)) \geq x(\delta(u))$. We split each of the boundaries into two parts to get

$$x(\delta(u)) = x(e(u, X \setminus (P_u \cup u))) + x(e(u, P_u))$$

$$x(\delta(P_u \cup u)) = x(e(u, X \setminus (P_u \cup u))) + x(e(P_u, X \setminus (P_u \cup u)))$$

Subtracting the equations, we get

$$x(\delta(P_u \cup u))) - x(\delta(u)) = x(e(P_u, X \setminus (P_u \cup u))) - x(e(P_u, u))$$

On the other hand,

$$x(e(P_u, X \setminus (P_u \cup u))) + x(e(P_u, u)) = x(\delta(P_u)) \geq 2.$$

Now, the condition that $x(e(P_u, u)) \leq x(e(u, X \setminus S)) \leq 1$, it must be the case that $x(e(P_u, X \setminus (P_u \cup u))) \geq 1 \geq x(e(P_u, u))$. This implies that

$$x(\delta(P_u \cup u))) - x(\delta(u)) = x(e(P_u, X \setminus (P_u \cup u))) - x(e(P_u, u)) \geq 0.$$

for all u. Substituting this into (2.11) (using (2.12)),

$$x(\delta(T)) \geq x(\delta(S)) + x(\delta(T^*))$$

which is clearly false, since $x(\delta(T)) < x(\delta(S))$ by assumption. This completes the proof of the lemma. □

2.2.2 Construction of Half Integral Solution for the Gadget

We will now describe the construction of a local half-integral modification of the tour at an (ε, D)-copy of the gadget S, which is compatible with all comb inequalities of size c. For any Hamiltonian tour P, this modification can be made on any (ε, D)-copy of S that has the following property with respect to P:

Property 21. We say that an (ε, D)-copy S_1 of S has Property 21 with respect to a Hamiltonian path or tour P if and only if

1. P visits S_1 exactly once, and enters and leaves through consecutive vertices on the outer circle vertices.
2. If x, y are the points adjacent to S_1 in P, then the points x, y are respectively connected to points of S_1 which are closest to them.

We will construct another gadget $\Pi^3_S(k)$ in Section 2.2.3 that contains multiple copies of $S = S(k)$, such that given any optimal Hamiltonian tour P, at least one (ε, D)-copy of S contained in an approximate copy of Π^3_S must satisfy Property 21 with respect to P.

The local half-integral solution mentioned above on S (see Section 2.2.2) consists of

(i) Four edge-disjoint triangles of edges of weight $\frac{1}{2}$—for each gap vertex, one such triangle joins that point to the closest two points on the outer and inner circles, respectively;

(ii) Weight-1 edges joining the remaining consecutive pairs of points on the inner ring of the gadget;

(iii) Weight-1 edges joining the remaining consecutive pairs of points on the outer ring of the gadget, except between entry/exit edges.

Moreover, we require that the entry/exit edges are separated by at least $c-1$ points on the circle from the weight $\frac{1}{2}$ edges. Section 2.2.2 shows the local solutions when $c = 2$ and $k = 12$, under Property 21. Now, we have the following lemma:

Lemma 22. *Consider gadget $S = S(k)$. Let x, y be points outside S and let P be a Hamiltonian path P from x to y in $\{x, y\} \cup S$ satisfies Property 21. Then length of Hamiltonian path P is at least*

$$\operatorname{dist}(x, S) + \operatorname{dist}(y, S) + 10\pi + 8 - O\left(\frac{1}{k}\right).$$

On the other hand, cost of the half-integral solution described above is at most

$$\operatorname{dist}(x, S) + \operatorname{dist}(y, S) + 10\pi + 6 + O\left(\frac{c}{k}\right).$$

Proof. Proof of the first lower bound is given in [FP15]. We include a discussion about the lower bound in Section A.2.4 for sake of completeness.

For the second bound, observe that in the half-integral solution, the total length of half-integral edges is $12 + 8\pi\frac{4}{2k} + 2\pi\frac{4}{k}$. On the other hand, the total length of integral edges contained in S is $10\pi - 8\pi\frac{3}{2k} - 2\pi\frac{2}{k}$, since we are missing 3 edges on bigger circle and 2 edges on smaller circle. Further, length of entry and exit segments is at most $\operatorname{dist}(x, S) + \operatorname{dist}(y, S) + 2\frac{8\pi c}{2k}$ since these are the original entry / exit points moved by length at most c points. Therefore, we get the total length of at most

$$\operatorname{dist}(x, S) + \operatorname{dist}(y, S) + \frac{8\pi c}{k} + 10\pi - \frac{16}{k} + 6 + \frac{12}{k}$$

which gives the required bound. □

Corollary 23. *There exists a constant γ such that if $k = \gamma c$ and $S = S(k)$, then for any points points x, y, and an Hamiltonian path from x to y on $S \cup \{x, y\}$ such that S satisfies Property 21 with respect to P, the half-integral solution described above has total value at least* 1 *smaller than length of P.*

In particular, $\gamma = 16\pi$ ensures that for $c \geq 3$, $k \geq 48\pi$, the total cost of the half integral solution is at most $L + 6.5$. Following the computations in Section A.2.4, total length of any Hamiltonian path P from x to y on S_1 is at least $L + 7.5$, which implies the result.

The two results above, namely Corollary 23 and lemma 22 show that the proposed half-integral solution is much smaller than the shortest tour. What remains is to show that this half-integral solution also satisfies all comb inequalities of bounded size.

Satisfying combs with 3 teeth

Now we prove that the half-integral solution described above in the gadget S satisfies all 3-combs of size at most c, assuming $c < k$. Note that we will pick $k = \gamma c$ where $\gamma > 2$, and hence this condition is always satisfied.

Lemma 24. *If a gap vertex is contained in a* 3*-comb such that the comb inequality corresponding to the* 3*-comb is violated, then $|H \cup T_1 \cup T_2 \cup T_3| \geq 2c$.*

Proof. The gap vertex is contained in two triangles of weight-$\frac{1}{2}$ edges.

Case 25. If any triangle P is contained in some tooth T, then note that $x(\delta(P)) = 3$ and $x(\delta(T)) = 2$. Further, each vertex of triangle has exactly 1 weight going outside the triangle. Therefore, by Lemma 20, T contains a path between two vertices of the triangle that lies completely outside the triangle. Any such path either must go along entire inner circle, or entire outer circle or it exits the gadget and enters again. In first two cases, this path as length at least $2k$ and in second case, the path has length at least $2c$, implying that $|T| \geq 2c$.

Case 26. If some tooth contains exactly two vertices of one of the two triangles, then since it doesn't contain the third vertex, all the conditions of Lemma 20 are satisfied. This again implies that T contains a path between the two vertices that does not use any edges in the triangle, and hence must have size at least $2c$.

Therefore, a tooth can contain at most one vertex of the triangles that contain the gap vertex. Hence, the handle must contain the gap vertex, and both the triangles containing the gap vertex. Let Q denote the union of both the triangles. Then $x(\delta(Q)) = 4$, and every vertex has edges of weight exactly 1 going out of Q. By Lemma 11 and the fact that this is a 3-comb, we have $x(\delta(H)) = 3$. Thus, it satisfies the conditions of Lemma 20, and hence, $H \setminus Q$ contains a path between two vertices in Q which lies completely outside Q. This path must have length at least $2c$ by exactly the same argument as above.

Hence, in all cases, either a tooth or the handle must contain at least c vertices, proving the lemma that we want. □

Since all the half weight edges in the Gadget are contain at least one gap vertex, every 3-comb must have a gap vertex in it, and hence must have large size.

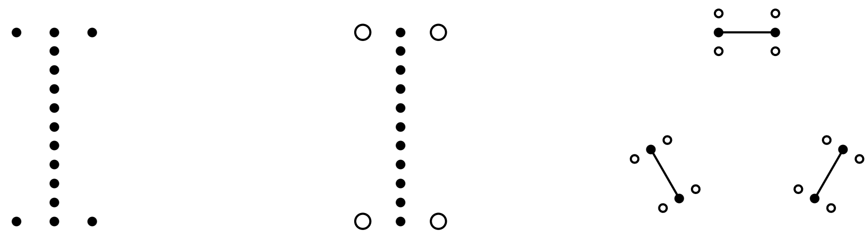


Figure 2.4: The figure shows (a)$\Pi(t, h, w)$, (b)$\Pi(S, t, h, w)$ and (c)$\Delta(S, D)$ from left to right.

Satisfying combs with 5 or more teeth

Lemma 27. *If a gap vertex is contained in a t-comb, with $t \geq 5$, then*

$$\left| H \cup \bigcup_{i=1}^{t} T_i \right| \geq c.$$

Proof. For this case, we will assume that the given comb in minimal, in particular, we have a comb with minimum value of t. If not, then we can always show the result for a smaller comb contained in this comb. From Lemma 17, if H is not connected, then we can always find a smaller comb that invalidates the solution. Hence, we will assume that H is connected for rest of the proof.

First, note that case (1) of Lemma 24 does not use the assumption that the comb has only 3 teeth. Therefore, we can conclude that teeth of comb of any size cannot contain the gap vertex. Hence, we only need to handle the case that the handle of the comb contains a gap vertex.

Note that any edge leaving the gadget is at least c distance away from the gap vertex. Since the comb, that is $H \cup \bigcup_{i=1}^{t} T_i$ is connected, if the comb contains any vertex outside the gadget, then it must have at least c vertices. Thus, we can assume that the comb is completely inside the gadget.

Further, any tooth can't have an edge of weight 1/2, since that would mean it contains two vertices of a gap triangle, and then by case (1) of Lemma 24 the tooth must contains a large cycle. Hence, all the edges strictly inside a tooth have weight exactly 1. Hence, each tooth is completely contained inside one of the 4 paths left after deleting all the edges of weight 1/2 in the gadget. Note that there are only 4 paths and at least 5 teeth. Let L_1, L_2, L_3, L_4 be the 4 paths.

Hence, one of the paths, say P, contains at least 2 teeth. Let these be T_1, T_2 such that the closest point in T_1 is closer to the gap vertex than the closest point in T_2. Now, since T_1, T_2 are connected, this implies that every vertex in T_1 is closer to the gap vertex than every vertex in T_2. But, since T_2 intersects the handle, there is a path from the gap vertex to T_2, say Q, which is completely contained in the handle. If Q is completely contained in P, then it contains entire T_1 implying that T_1 is contained in the handle, which contradicts definition of the comb. Otherwise, the path Q must wrap around using one of the other paths, $L_i \neq P$. Since it must contain the whole path, that implies that the handle has size at least k. This completes the proof. $\square$

2.2.3 Expanding the gadget

We will now describe the gadget $\Pi_S^3 = \Pi_S^3(k)$ that contains 12 copies of $S(k)$, such that there is an almost optimal Hamiltonian tour that satisfies Property 21 in at least one copy of S in all (ε, D) copies of Π_S^3. This gadget is obtained by combing two more gadgets with S, namely $\Pi_S = \Pi(S, t, h, w)$ and $\Pi_S^3 = \Delta(\Pi_S, D)$. Section 2.2.3 illustrates these gadgets. We will provide the formal definitions and state few lemmas below, proofs of which are given in Section A.2.

Definition 28. We define the gadget $\Pi(t, h, w)$ for $t \in \mathbb{Z}_{\geq 0}$ and $h, w \in \mathbb{R}_{\geq 0}$, given by points $\pi_1 = \left(-\frac{w}{2}, 0\right)$, $\pi_2 = \left(\frac{w}{2}, 0\right)$, $\pi_3 = \left(-\frac{w}{2}, h\right)$, $\pi_4 = \left(\frac{w}{2}, h\right)$ and points $v_1, \ldots, v_t$ which are evenly spaced along $(0,0), (0,h)$, with $v_1 = (0,0)$ and $v_t = (0,h)$. We will refer to sets $\{\pi_1\pi_2\}$ and $\{\pi_3\pi_4\}$ as *shorter sides* of the gadget, and sets $\{\pi_1\pi_3\}$ and $\{\pi_2\pi_4\}$ as *longer sides* of the gadget.

Lemma 29. *Let p, q be two points on the opposite sides of the horizontal line $y = \frac{h}{2}$ such that*

$$\mathrm{dist}(\{x, y\}, \Pi(t, h, w)) \geq D.$$

Let P be a shortest Hamiltonian path from p to q in $\Pi(t, h, w) \cup \{p, q\}$. Suppose all of the following inequalities hold:

$$D \geq \tfrac{h^2+w^2}{4w} \qquad h \geq 2w \qquad t \geq \tfrac{16h}{w}$$

Then for at least two $i \in 1, 2, 3, 4$ we have that neither neighbor v_i^1, v_i^2 of π_i on P is not in $\{p, q\}$ and moreover, v_i^1, v_i^2 are two points in $\{v_1, \ldots, v_t\}$ closest to π_i.

Intuitively, this lemma holds since the shortest path through $\Pi(t, h, w)$ must travel through both the shorter sides, and connect them using the middle segment. The condition on positions of p, q ensures that it is beneficial to enter the gadget on one of the shorter sides and exit from the other shorter side. A formal proof is given in Section A.2.1.

Now, we extend this gadget to the gadget Π_S, which is constructed by replacing each of the four corner points in $C = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ by a copy of gadget $S(k)$ defined in Definition 7.

Definition 30. We construct the gadget $\Pi(S(k), t, h, w)$ by replacing points in C by copies of $S(k)$ centered at each point $\pi_i \in C$. We let S_i denote the copy centered at π_i.

Lemma 31. *Let p, q be two points on the opposite sides of the line $y = \frac{h}{2}$ such that*

$$dist(\{p, q\}, \Pi(t, h, w)) \geq D.$$

Let P be a shortest Hamiltonian path from p to q in $\Pi(S(k), t, h, w) \cup \{p, q\}$. Suppose all of the following inequalities hold:

$$D \geq \tfrac{h^2+w^2}{4w} \qquad h \geq 2w \qquad w \geq 100 \qquad t \geq 2h \qquad \frac{h}{t} \leq \frac{4\pi}{k}$$

Then there is a Hamiltonian path Q from p to q in $\Pi(S(k), t, h, w) \cup \{p, q\}$ such that Q visits each S_i at most once, $\ell(Q) \leq \ell(P) + O(1/k)$ and for at least two $i \in 1, 2, 3, 4$ we have that neither neighbor v_i^1, v_i^2 of S_i on Q is not in $\{p, q\}$ and moreover, v_i^1, v_i^2 are two points in $\{v_1, \ldots, v_t\}$ closest to S_i.

In particular, $\Pi_S(k) = \Pi(S(k), \frac{200k}{4\pi}, 200, 100)$ satisfies this lemma for $D = 125$. A complete proof of Lemma 31 is given in Section A.2.2.

Now, we introduce the final piece of the puzzle, the gadget $\Delta(D, \Pi_S(k))$ which contains three copies of the gadget $\Pi_S(k)$. This gadget is designed in a way that at least one of the copy of $\Pi_S(k)$ is visited exactly once in any optimal tour.

Definition 32. For any gadget $T \in \mathbb{R}^2$ with diameter d and $D \in \mathbb{R}_{\geq 0}$, we define the gadget $\Delta(D, T)$ containing three copies of T, T_1, T_2, T_3 centered at points $C_1 = \left(R, \frac{\pi}{2}\right)$, $C_2 = \left(R, \frac{7\pi}{6}\right)$ and $C_3 = \left(R, \frac{11\pi}{6}\right)$ for $R = \frac{D+2d}{\sqrt{3}}$ and rotated clockwise in angles of $\frac{\pi}{2}$, $-\frac{\pi}{6}$, and $\frac{\pi}{6}$ respectively.

An illustration for $\Delta(D, \Pi_S)$ is given in Section 2.2.3. The rotations are to ensure that gadgets are symmetrically situated around rays along OC_1, OC_2 and OC_3 respectively. Further, distance between T_i and T_j is at least D for any $i \neq j$. Now, we have following lemma about this gadget:

Lemma 33. *Let $\varepsilon > 0$ be positive real. Then there exists constants $D_1, D_2 \geq 0$ such that if P is an optimal Hamiltonian tour over V, and if Δ_1 is any (ε, D_2) copy of $\Delta(D_1, \Pi_S(k))$, then there exists an $i \in \{1, 2, 3\}$ such that P visits Π_i exactly once, where Π_1, Π_2, Π_3 are (ε, D_1)-copies of $\Pi_S(k)$ contained in Δ_1, with centers C_1, C_2, C_3 respectively. Further if p, q are neighbors of T_i in P, then p, q lie on the opposite side of $\overleftrightarrow{OC_i}$, where O is the center of Δ_1. In particular, the values*

$$D_1 = \frac{2000}{1 - \cos \frac{\pi}{10}} \qquad \text{and} \qquad D_2 = \frac{30000}{\left(1 - \cos \frac{\pi}{10}\right)^2} \tag{2.13}$$

suffice.

Proof of this lemma is a repeated application of Observation 9, in particular, the condition that if an (ε, D)-copy of a gadget T is visited multiple times, then the entry and exit rays must be parallel. The details are given in Section A.2.3.

We define $\Pi_S^3 = \Pi_S^3(k) = \Delta(D_1, \Pi_s(k))$ where $D_1 = \frac{2000}{1-\cos(\pi/10)}$ is as defined in Lemma 33. Combining Lemmas 31 and 33, we get the following lemma which shows the existence of (ε, D)-copies of S which have Property 21.

Lemma 34. *For any $k \geq 4$ and $\varepsilon > 0$, there is are constants $D_1, D_2 > 0$ such that any (ε, D_2)-copy Π_1^3 of $\Pi_S^3(k)$ contained in V, and for any optimal Hamiltonian tour P on V, there is a Hamiltonian tour Q such that $\ell(Q) \leq \ell(P) + 40k\varepsilon$ and (ε, D_1)-copy S_1 of $S(k)$ such that $S_1 \subseteq \Pi_1$ and S_1 has Property 21 with respect to path Q.*

Further, if $k \geq \gamma c$ where γ is given by Corollary 23 and if $\bar{y}$ is the half-integral solution described in Section 2.2.2, then replacing Q by S_1 gives an half-solution satisfying $\texttt{Comb}_c$ inequalities. This replacement can be made in all disjoint (ε, D_2)-copies of $\Pi_S^3(k)$ in V simultaneously.

Proof. We pick D_1, D_2 as defined in Lemma 33, namely

$$D_1 = \frac{2000}{1 - \cos \frac{\pi}{10}} \qquad \text{and} \qquad D_2 = \frac{30000}{\left(1 - \cos \frac{\pi}{10}\right)^2}.$$

Recall that $\Pi_S(k) = \Pi(S(k), \frac{200k}{4\pi}, 200, 100)$. It follows from Definitions 7, 30 and 32 that $\Pi_S(k)$ has at most $40k$ points for $k \geq 4$. Since Π_1^3 is an (ε, D_2)-copy of $\Pi_S^3(k)$, there exists a translation T_1 of $\Pi_S^3(k)$ and a bijection $f : T_1 \to \Pi_1^3$ such that $\|x - f(x)\| \leq \varepsilon$.

Using Lemma 33, there is an (ε, D_1)-copy Π_1 of $\Pi_S(k)$ that is visited by P exactly once. Further, if p and q are the points adjacent to Π_1 in P, then p, q are on opposite side of OC_1 where O is center of T_1 and C_1 is the center of $T_2 = f^{-1}(\Pi_1)$. Let $\bar{P}$ be an optimal Hamiltonian path from p to q in $T_1 \cup \{p, q\}$. Then by Lemma 31, there is an Hamiltonian path $\bar{Q}$ from p to q in $T_1 \cup \{p, q\}$ such that $\ell(\bar{Q}) \leq \ell(\bar{P}) + O(1/k)$ and a copy T_2 of $S(k)$ such that $S(k)$ has Property 21 with respect to $\bar{Q}$. Let Q be the Hamiltonian tour that equals P outside $f(T_1) \cup \{p, q\}$ and $f(\bar{Q})$ inside $f(T_1) \cup \{p, q\}$. Then $f(T_2)$ has Property 21 with respect to Q. Since $\bar{P}$ was optimal on $T_1 \cup \{p, q\}$ and T_1 has at most $40k$ points, we must have $\ell(\bar{P}) \leq \ell(P \cap (T_1 \cup \{x, y\})) + 40k\varepsilon$. It follows that $\ell(Q) \leq \ell(P) + 40k\varepsilon + O(1/k)$.

Using Corollary 23 for $k = \gamma c$, and $\varepsilon = O(\frac{1}{k})$, we can ensure that the cost of half-integral solution $\bar{y}$ is at least 1 smaller than length of Q, and at least 0.5 smaller than length of P. Further, Section 2.2.2 implies that we can make this replacement in a single gadget without violating $\texttt{Comb}_c$ inequalities.

If we do the replacement simultaneously in multiple disjoint (ε, D_2)-copies of $\Pi_S^3(k)$, then any comb of size at most c containing an half-integral edge must be completely contained in an $(\varepsilon, 0)$-copy $S(k)$. And hence again by Section 2.2.2, we do not violate any $\texttt{Comb}_c$ inequalities. $\square$

Now, we are in a position to complete proof of Theorem 6. Let $c > 0$ be a constant. Using Lemma 171 (which is a tighter version of Observation 10), we can find $C_\Pi n$ disjoint (ε, D_2)-copies of $\Pi_S^3(k)$ in $\mathcal{Y}_n$ with probability at least $1 - \frac{1}{n^2}$, where $k = \gamma c$, $\varepsilon = O(1/k)$ and $D_2 = O(1)$ is an absolute constant, where $C_\Pi = C_\Pi(k, \varepsilon, D_2)$. Note that when c is a constant, then so is C_Π.

Using Lemma 34, given any optimal tour in $\mathcal{Y}_n$, we can find a half-integral solution $\bar{y}$ on the edges of $\mathcal{Y}_n$ which is at least $\frac{1}{2}C_\Pi n$ smaller than the length of optimal tour. Therefore,

$$\texttt{Comb}_c(\mathcal{Y}_n) \leq \texttt{TSP}(\mathcal{Y}_n) - C_\Pi^* n$$

with probability $1 - \frac{1}{n}$ where $C_\Pi^* = \frac{1}{2}C_\Pi$ is an absolute constant. Therefore, we have

$$\sum_n \mathbb{P}\Big[\texttt{Comb}_c(\mathcal{Y}_n) \geq \texttt{TSP}(\mathcal{Y}_n) - C_\Pi^* n\Big] \leq \sum_n \frac{1}{n^2} < \infty$$

Therefore, by Borel-Cantelli Lemma,

$$\limsup_{n \to \infty} \texttt{Comb}_c(\mathcal{Y}_n) \leq \lim_{n \to \infty} \texttt{TSP}(\mathcal{Y}_n) - C_\Pi^* n = \big(\beta_{\texttt{TSP}}^2 - C_\Pi^*\big) n$$

almost surely, since $\lim_{n\to\infty} \texttt{TSP}(\mathcal{Y}_n) = \beta_{\texttt{TSP}}^2 n$ almost surely. This implies that

$$\gamma_{\texttt{Comb}}^{c,2} \leq \beta_{\texttt{TSP}}^2 - C_\Pi^*$$

completing the proof of **Theorem 6** in two dimensions.

2.2.4 Higher Dimensions

Note that the construction above only works for $d = 2$. For higher dimensions, we construct gadget $T_d(k)$ which contains 5 copies of $\Pi_S^3(k)$ which are at least D_2 distance apart from each other and lie in the same 2-dimensional plane, where D_2 is as defined in Lemma 34.

Since an optimal tour can enter $T_d(k)$ at most twice (Lemma 157), at least one of the 5 copies of $\Pi_S^3(k)$ must only be connected to points in $T_d(k)$. This reduces problem to two dimensional case, and we can then use Lemma 34 to conclude higher dimensional version of Lemma 34!

Lemma 35. *For any $k \geq 4$ and $\varepsilon > 0$, there is are constants $D_1, D_2 > 0$ such that any (ε, D_2)-copy T_d of $T_d(k)$ contained in V, and for any optimal Hamiltonian tour P on V, there is a Hamiltonian tour Q such that $l(Q) \leq l(P) + 200k\varepsilon$ and (ε, D_1)-copy S_1 of $S(k)$ such that $S_1 \subseteq \Pi_1$ and S_1 has Property 21 with respect to path Q.*

Further, if $k \geq \gamma c$ where γ is given by Corollary 23 and if $\bar{y}$ is the half-integral solution described in Section 2.2.2, then replacing Q by S_1 gives an half-solution satisfying $\texttt{Comb}_c$ inequalities. This replacement can be made in all disjoint (ε, D_2)-copies of $T_d(k)$ in V simultaneously.

In particular, following the Borel-Cantelli argument in 2 dimensional case, this gives us the separation in higher dimensions, namely

$$\gamma^d_{\texttt{Comb}} \leq \beta^d_{\texttt{TSP}} - C$$

for some constant C.

2.3 Branch and Bound Algorithms

In this section, we will prove Theorem 3. For this section, we will assume that we are working in some fixed dimension d. Further, throughout this section, O notation will hide constants dependent on d.

As considered here, a branch and bound algorithm depends on three choices:

(1) A choice of heuristic to find (not always optimal) TSP tours.
(2) A choice of lower bound for TSP (such as $\texttt{Comb}_c$ or HK).
(3) A branching strategy.

The result of a branch-and-bound approach is a branch-and-bound tree, which is a rooted tree such that to each vertex v of this tree, we associate two sets I_v and O_v such that

(1) When v is the child of u, $I_v \supseteq I_u$ and $O_v \supseteq O_u$
(2) If u has children $v_1, \ldots, v_k$, then we have $\Lambda_u = \bigcup_{i=1}^k \Lambda_{v_i}$, where Λ_u denotes the set of TSP tours which include all the edges in I_u and exclude all the edges in O_u.
(3) The leaves of the (unpruned) branch and bound tree satisfy $|\Lambda_v| = 1$.

For any node v of the branching tree, let b_v denote the value of the lower bound, which in our case is the value of $\texttt{Comb}_c$ under the additional constraints given by I_v and O_v (that is, the solution for $\texttt{Comb}_c$ must include all the edges in I_v with weight 1 and must exclude all the edges in O_v). Let B be the value of the tour given by our heuristic. For each vertex v, we find a tour using the some heuristic that includes the edges in I_v and excludes the edges in O_v, and whenever we find a tour smaller than B, we update B. For every vertex v such that $b_v \geq B$, we know that we have already found a tour as good as any in Λ_v, and we prune the tree at v. The process ends when the set L of leaves of the pruned tree satisfies $v \in L \Rightarrow b_v \geq B$. Note that such a tree in fact gives a proof that B is an optimal tour.

Note that following any branching strategy to generate the tree will give us an optimal tour and proof of its optimality. For a branch and bound to be efficient, we want to prune the tree such that only polynomially many leaves remain.

We can now state a more precise version of Theorem 3 as follows:

Theorem 36 (Theorem 3 restated)**.** *For any TSP heuristic, any branching strategy and a lower bound heuristic which is* $\texttt{Comb}_c$ *for some constant* c*, the pruned branch and bound tree will have* $e^{\Omega(n/\log^5 n)}$ *leaves almost surely.*

Further, we state a generalization of above result when c is not a constant as follows:

Theorem 37 (Theorem 4 restated)**.** *Given any* $\varepsilon > 0$*, For any TSP heuristic, any branching strategy and a lower bound heuristic which is* $\texttt{Comb}_c$ *for* $c = O\left(\frac{\varepsilon \log n}{\log \log n}\right)$*, the pruned branch and bound tree will have* $e^{\Omega(n^{1-6\varepsilon})}$ *leaves almost surely.*

Note that setting $\varepsilon = 0.08$ gives us Theorem 4

Any branch-and-bound approach should produce not only an optimal tour, but, via the pruned tree and computed bounds, a certificate verifying that the returned tour is optimal. Theorem 36 shows that even just the size of this certificate is exponential. Our general strategy to prove Theorem 36 will be to show that when $\texttt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \geq \texttt{TSP}(\mathcal{X}_n)$ then either I_v or O_v is must be large, and hence Λ_v is in fact small. Since $\Lambda = \bigcup \Lambda_v$, this would imply that there are a lot of leaves in any pruned tree.

Following [FP15], we will further modify this approach by looking at a special set of tours $\overline{\Lambda}$. Given the point set $\mathcal{X}_n$, we will consider the division of $[0,1]^d$ into $s = \frac{n}{\sigma}$ boxes of side-length $s^{-\frac{1}{d}}$. We will eventually $\sigma = \Omega(\log n)$ as required for the runtime bounds. Let $B_1, \ldots, B_s$ denote these boxes, taken in some order such that consecutive terms share a $(d-1)$ dimensional face. Note that

$$|x - y| \leq \sqrt{d} \cdot s^{-\frac{1}{d}} = O\left(s^{-\frac{1}{d}}\right)$$

if x, y lie in the same box. We consider $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, and for each $2 \leq j \leq s-1$, we let $x_j^1, x_j^2, x_j^3, x_j^4$ denote the four points $x_i \in \mathcal{X}_n \cap B_j$ of smallest index (this choice can be arbitrary, and is just for definiteness). We also chose points $x_1^3, x_1^4 \in \mathcal{X}_n \cap B_1$ and $x_s^1, x_s^2 \in \mathcal{X} \cap B_s$, again by simply choosing points of minimum index. These points chosen as above can be viewed as preselected interface points between boxes B_j. In particular, we let $\overline{\Lambda}$ denote the set of TSP tours in $\mathcal{X}_n$ with the properties that, in that tour,

1. x_1^4 is joined to x_1^3 by a path lying entirely in B_1;
2. for $1 \leq j \leq s-1$, x_j^3 and x_{j+1}^1 are adjacent;
3. for $2 \leq j \leq s-1$, x_j^1 is joined to x_j^3 by a path lying completely in B_j;
4. x_s^1 is joined to x_s^2 by a path lying entirely in B_s;

5. for $s \geq j \geq 2$, x_j^2 and x_{j-1}^4 are adjacent; and
6. for $s-1 \geq j \geq 2$, x_j^2 is joined to x_j^4 by a path lying completely in B_j.

We will only restrict our attention to these special tours. Note that we are now only looking at a smaller subset of tours. We claim that these tours have asymptotically almost the same length as the TSP tour *almost surely.* These tours are similar to those produced by Karp's *fixed dissection* heuristic [Kar77], which divides the square into s boxes like we have, finds optimal tours through each, and then joins into a closed walk by means of an optimal tour through a set of representatives.

For the sake of notation, let $\overline{\texttt{TSP}}(\mathcal{X}_n)$ denote the best tour in $\overline{\Lambda}$. Let $\texttt{TSP}^F(\mathcal{X}_n)$ denote the tour given by fixed dissection heuristic. We claim that asymptotically

$$\texttt{TSP}(\mathcal{X}_n) \sim \overline{\texttt{TSP}}(\mathcal{X}_n)$$

The proof is based off the techniques used to show $\texttt{TSP}(\mathcal{X}_n) \sim \texttt{TSP}^F(\mathcal{X}_n)$. We will leverage parts of Lemma 4 in Chapter 6 from [Law85], in particular,

Lemma 38. *Let $\texttt{TSP}(B_j)$ denote the best tour in $\mathcal{X}_n \cap B_j$. Then we have the following bound:*

$$\sum_{j=1}^{s} \texttt{TSP}(B_j) \leq \texttt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-2}{d-1}} s^{\frac{1}{d(d-1)}}\right) + O\left(s^{\frac{d-1}{d}}\right) = \texttt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}} \sigma^{-\frac{1}{d(d-1)}}\right) \tag{2.14}$$

where s is the number of boxes B_j. Recall that $s = o(n)$.

Apart from finding the best tour in each cube B_j, the cost of modifying this solution into a path that starts at x_j^1 and ends at x_j^3 is at most $2d^{1/2}s^{-1/d}$. The cost of patching edges between B_j and B_{j+1} also at most $2d^{1/2}s^{-1/d}$. This gives us the upper bound:

$$\overline{\texttt{TSP}}(\mathcal{X}_n) \leq \texttt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-2}{d-1}} s^{\frac{1}{d(d-1)}}\right) + O\left(s^{\frac{d-1}{d}}\right) + O\left(s \cdot s^{-\frac{1}{d}}\right)$$

Since we choose $s = \frac{n}{\sigma} = o(n)$, we get

$$\texttt{TSP}(\mathcal{X}_n) \leq \overline{\texttt{TSP}}(\mathcal{X}_n) \leq \texttt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}}\left(\sigma^{-\frac{1}{d(d-1)}} + \sigma^{-\frac{d-1}{d}}\right)\right)$$
$$\therefore \overline{\texttt{TSP}}(\mathcal{X}_n) \leq \texttt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}} \sigma^{-\frac{1}{d(d-1)}}\right)$$

where the O-notation hides constants dependent only on d. Note that this statement holds true deterministically.

Now we use the bounds on sizes of $\overline{\Lambda}$ and $\overline{\Lambda}_v = \Lambda_v \cap \overline{\Lambda}$ proved in [FP15] (equations $29-32$). Let $\beta_j = |\mathcal{X}_n \cap B_j|$, let O_v^j denote the set of edges in O_v that have both the endpoints in B_j and let I_v^j be the set of edges in I_v that have both the endpoints in B_j. Let $I_v' \subseteq I_v$ denotes edges in I_v of the form $\{x_j^3, x_{j+1}^1\}$ or $\{x_j^2, x_{j-1}^4\}$. We will provide short proofs of these bounds again for sake of completeness.

$$\left|\overline{\Lambda}\right| = (\beta_1 - 2)!\left(\prod_{j=2}^{s-1}(\beta_j - 3)!\right)(\beta_s - 2)! \tag{2.15}$$

This bound follows since we can choose tour in every box B_j by choosing the path from x_j^1 to x_j^3 and the path from x_j^2 to x_j^4, by choosing a permutation of $(\beta_j - 4)$ vertices ($(\beta_j-4)!$ choices) and breaking it up into 2 parts ($\beta_j - 3$ choices). First and last terms follow from a similar logic on box B_1 and B_s, which only have 2 special vertices instead of 4. Now, observe that $\Lambda_v = \varnothing$ unless $I_v = I'_v \cup \bigcup_{j=1}^s I_v^j$. To get an upper bound on $\overline{\Lambda}_v$, we look at the portion of tour in B_j, which can be represented as a permutation of $(\beta_j - 3)$ symbols. Given an orientation of edges in I_v^j, each edge reduces the number of free symbols in the permutation by at 1, giving us an upper bound of

$$\left|\overline{\Lambda}_v\right| \leq (\beta_1 - 2 - \left|I_v^1\right|)!2^{\left|I_v^1\right|}\left(\prod_{j=2}^{s-1}(\beta_j - 3 - \left|I_v^j\right|)!2^{|I_v|}\right)(\beta_s - 2 - |I_v^s|)!2^{|I_v|}$$

Let $\overline{I}_v = \bigcup_{j=1}^s I_v^j$. Using Sterling's Approximation, we get

$$\left|\overline{\Lambda}_v\right| \leq \left|\overline{\Lambda}\right| \cdot \prod_{j=1}^s \left(\frac{2e}{\beta_j - 3}\right)^{\left|I_v^j\right|} \leq \left|\overline{\Lambda}\right| \cdot e^{-\left|\overline{I}_v\right|} \tag{2.16}$$

assuming that $\beta_j \geq 2e^2 + 3$ for all j.

Note that a crude application of the Chernoff bound gives that for each j,

$$\beta_j \in (1 \pm 0.5)\sigma$$

with probability at least $1 - e^{-\sigma}$, where $s = \frac{n}{\sigma}$. Then by union bound, we get that the same expression holds for all j simultaneously with probability at least $1 - ne^{-\sigma}$. This implies that Equation (2.16) holds with probability at least $1 - \frac{1}{n^2}$ provided that $\sigma = \Omega(\log n)$. In particular, Equation (2.16) holds with high probability.

On the other hand, observe that number of permutations on $\beta_j - 3$ symbols that avoid one particular edge is at most

$$(\beta_j - 3)! - (\beta_j - 4)! \leq \left(1 - \frac{1}{\beta_j}\right)(\beta_j - 3)!$$

simply by subtracting number of permutations that include this particular edge. Define $\delta_A = 1$ if $|A| \geq 1$ and $\delta_A = 0$ otherwise. Then we have an upper bound

$$\left|\overline{\Lambda}_v\right| \leq \left|\overline{\Lambda}\right| \cdot \prod_{j=1}^s \left(1 - \frac{\delta_{O_v^j}}{\beta_j}\right)$$

Let $\overline{O}_v = \bigcup_{j=1}^s \overline{O}_v^j$. Since $\beta_j \leq 2\sigma$ for all j with probability at least $1 - \frac{1}{n^2}$, there must be at least $\left|\overline{O}_v^j\right|\left(\frac{1}{2\sigma}\right)^2$ integers j such that $|O_v^j| \geq 1$. Therefore, we get the upper bound:

$$\overline{\Lambda}_v \leq \overline{\Lambda} \cdot \left(1 - \frac{1}{2\sigma}\right)^{\left|\overline{O}_v\right|\left(\frac{1}{2\sigma}\right)^2} \leq \overline{\Lambda} \cdot e^{-\left|\overline{O}_v\right|/(2\sigma)^3} \tag{2.17}$$

Now that we have established these bounds, we know that a large $\overline{I}_v$ or large $\overline{O}_v$ forces $\overline{\Lambda}_v$ to be small. Define $\overline{L} = \left\{v \in L \,\middle|\, \overline{\Lambda}_v \neq \varnothing\right\}$. Note that $\overline{\Lambda} = \bigcup_{v \in \overline{L}} \overline{\Lambda}_v$. Now, since $\overline{\Lambda}$ itself is large, it suffices to show that $v \in \overline{L}$ implies that either $\overline{I}_v$ or $\overline{O}_v$ is large. Indeed, we have

Lemma 39. *Let d be a fixed integer. If following conditions hold with correct constants (dependent on d),*

$$\sigma = \Omega(\log n) \qquad \tau = \Omega\left(\sigma^{\frac{d}{d-1}}\right) \qquad c = O\left(\frac{\log \sigma}{\log\log \sigma}\right)$$

Then with probability at least $1 - O\left(\frac{1}{n^2}\right)$, either

$$\left|\overline{I}_v\right| + \left|\overline{O}_v\right| \geq t = \frac{n}{\tau} \qquad \forall v \in \overline{L}$$

or else that

$$\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \leq \mathtt{TSP}(\mathcal{X}_n)$$

for large enough n.

Proof. We will show that if $\left|\overline{I}_v\right| + \left|\overline{O}_b\right| \leq t = \frac{n}{\tau}$, then $\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \leq \mathtt{TSP}(\mathcal{X}_n)$. This proof has two components. First, we upper bound $\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v)$ given that $\overline{I}_v, \overline{O}_v$ are small. More precisely, we will show that

$$\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v) \leq \overline{\mathtt{TSP}}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}}\sigma^{\frac{d+1}{d}}\tau^{-1}\right) \tag{2.18}$$

which follows from making local modifications to the optimal tour in $\overline{\Lambda}$. In the second part, we will bound the value of $\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v)$ given that $\overline{I}_v, \overline{O}_v$ are small. In particular, we have:

$$\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \leq \overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v) - O\left(\left(e^{-O(c\log c)} - \frac{1}{\tau} - \frac{1}{\sigma}\right)n^{\frac{d-1}{d}}\right) \tag{2.19}$$

The proof of the second part it similar to that of Theorem 6.

For the first part, notice that there are at most t integers j such that $|I_v^j| + |O_v^j| > 0$. We shall use the term *restricted boxes* to denote all such boxes B_j. We construct a tour by modifying the optimal tour in $\overline{\Lambda}$, by replacing the portion of tour by any feasible tour in all the restricted boxes. Note that if there are no feasible tour in any of the boxes, then $\overline{\Lambda}_v = \varnothing$, and hence $v \notin \overline{L}$, which is a contradiction. Therefore, such a patching always exists.

In the restricted boxes, the total length of the tour can be the worst case length of the tour, which is $\beta_j s^{-1/d}\sqrt{d}$. Since with high probability, $\beta_j \leq 2\sigma$ for all j, we can conclude that

$$\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v) \leq \overline{\mathtt{TSP}}(\mathcal{X}_n) + 2\sigma s^{-\frac{1}{d}}\frac{n}{\tau} = \overline{\mathtt{TSP}}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}}\sigma^{\frac{d+1}{d}}\tau^{-1}\right)$$

On the other hand, following the proof of Lemma 171, where we ensure that the smaller boxes of side-length $3D$ are contained in the boxes of side-length $s^{-1/d}$, we can find $e^{-O(c\log c)}n$ (ε, D)-copies of the gadget $\Pi_S^3(k)$, scaled by $n^{-1/d}$ (Note that ε and D also gets scaled by a factor $n^{-1/d}$). Here $\varepsilon = \Omega(1/c)$ and $D = D_2$ is the absolute constant specified in Lemma 33.

Observe that Lemma 171 holds only when $\exp(O(c\log c)) = o(n)$. When the third hypothesis condition holds with correct constant, that is

$$c = O\left(\frac{\log \sigma}{\log\log \sigma}\right)$$

ensures that $c \log c \leq K \log \sigma$ for some constant $K \leq 1$. This implies

$$\exp(O(c \log c)) = O(\sigma^K) = o\left(\frac{n}{2 \log n}\right)$$

Therefore, Lemma 171 holds not just with high probability, but with probability at least $1 - \frac{1}{n^2}$.

We look at the optimal TSP tour which has length $\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v)$. We cannot directly use Lemmas 29, 31, 33, 34, 155 and 157 on this tour to construct a solution that satisfies $\mathtt{Comb}_c$, since the optimal tour in $\overline{\Lambda}$ might not by an optimal TSP tour.

But, for any (ε, D)-copy of the gadget S_1 which is contained in box B_j, all the results will go through as long as we can perform the modification used in the proofs of Lemmas 29, 31, 33, 34, 155 and 157 and get a tour that is contained in $\overline{\Lambda}_v$. These modifications can be made as long as any of the points in the (ε, D)-copy of the gadget or the points adjacent to these gadget are not contained in an edge in $\overline{I}_v$ or $\overline{O}_v$, and are not one of the special points $x_j^{\{1,2,3,4\}}$ used in definition of $\overline{\Lambda}$. Therefore, we can make these modifications on all but $O(s+t)$ gadgets! Therefore, we can construct a half-integral solution which satisfied $\mathtt{Comb}_c$ constraints and respects the sets I_v and O_v of value at most

$$\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v) - O\left(\left(e^{-O(c \log c)} - \frac{1}{\tau} - \frac{1}{\sigma}\right) n^{\frac{d-1}{d}}\right)$$

In particular, this proves Equation (2.19).

The condition $\tau = \Omega(\sigma^{d/(d-1)})$ along with Equation (2.18) and lemma 38 implies that

$$\overline{\mathtt{TSP}}(\mathcal{X}_n \mid I_v, O_v) \leq \mathtt{TSP}(\mathcal{X}_n) + O\left(n^{\frac{d-1}{d}} \sigma^{-\frac{1}{d(d-1)}}\right)$$

which again along with $\tau = \Omega(\sigma^{d/(d-1)})$ and $\sigma = \omega(1)$ gives us

$$\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \leq \mathtt{TSP}(\mathcal{X}_n) + n^{\frac{d-1}{d}} O\left(\sigma^{-\frac{1}{d(d-1)}} - e^{-O(c \log c)}\right)$$

We can now choose

$$c = O\left(\frac{\log \sigma}{d(d-1) \log \log \sigma}\right) = O\left(\frac{\log \sigma}{\log \log \sigma}\right)$$

to get that $\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \leq \mathtt{TSP}(\mathcal{X}_n)$ holds for large enough n.

The result is conditioned on two probabilistic events happening, first one is the event that $\beta_j \in (1 \pm 0.5)\sigma$, which happens with probability $1 - \frac{1}{n^2}$ and the second is that Lemma 171 holds for $\mathcal{X}_n$, which also happens with probability $1 - \frac{1}{n^2}$. Therefore, overall, this results holds with probability $1 - O\left(\frac{1}{n^2}\right)$ for large enough n. $\square$

Proof of Theorems 36 and 37: Now, we are in a position to complete the proofs of these results by choosing σ and τ appropriately.

If $v \in \overline{L}$, then $\mathtt{Comb}_c(\mathcal{X}_n \mid I_v, O_v) \geq \mathtt{TSP}(\mathcal{X}_n)$ and hence, by the result above, we must have that

$$|\overline{I}_v| + |\overline{O_v}| \geq \frac{n}{\tau}$$

Then by Equation (2.16) and Equation (2.17) gives

$$\overline{\Lambda}_v \leq \overline{\Lambda} e^{-\Omega\left(\frac{n}{\sigma^3 \tau}\right)}$$

which implies that

$$\left|\overline{L}\right| \geq e^{\Omega\left(\frac{n}{\sigma^3 \tau}\right)}$$

Observe that for a constant c, choosing $\sigma = K \log n$ and $\tau = \sigma^{d/(d-1)}$ gives us that

$$\left|\overline{L}\right| \geq \exp\left(\Omega\left(\frac{n}{\log n^{4+\frac{d}{d-1}}}\right)\right) = e^{\Omega\left(\frac{n}{\log^5 n}\right)}$$

Therefore, with probability $1 - O(n^{-2})$, the pruned branch and bound tree will have $e^{\Omega(n/\log^5 n)}$ leaves for large enough n. This implies that

$$\sum_{n=1}^{\infty} \mathbb{P}\left[\left|\overline{L}\right| \leq e^{\Omega(n/\log^5 n)}\right] < \infty$$

Therefore, by Borell-Cantelli Lemma,

$$\mathbb{P}\left[\limsup_{n\to\infty}\left|\overline{L}\right| \leq e^{\Omega(n/\log^5 n)}\right] = 0$$

which recovers Theorem 36, that is

Theorem 36 (Theorem 3 restated)**.** *For any TSP heuristic, any branching strategy and a lower bound heuristic which is* $\texttt{Comb}_c$ *for some constant c, the pruned branch and bound tree will have* $e^{\Omega(n/\log^5 n)}$ *leaves almost surely.*

Further, we get the exact same bound on the number of leaves when $c = O\left(\frac{\log\log n}{\log\log\log n}\right)$. Similarly, for any $\varepsilon > 0$ we can set $\sigma = n^{\varepsilon}$ and $\tau = n^{\varepsilon d/(d-1))}$ to get that for any $c = O\left(\frac{\varepsilon \log n}{\log\log n}\right)$ we have

$$\left|\overline{L}\right| \geq \exp\left(\Omega\left(\frac{n^{1-\varepsilon}}{n^{\varepsilon\left(4+\frac{d}{d-1}\right)}}\right)\right) = e^{\Omega(n^{1-6\varepsilon})}$$

with probability at least $1 - O(n^{-2})$. Now, a similar Borell-Cantelli argument recovers Theorem 37, that is

Theorem 37 (Theorem 4 restated)**.** *Given any* $\varepsilon > 0$*, For any TSP heuristic, any branching strategy and a lower bound heuristic which is* $\texttt{Comb}_c$ *for* $c = O\left(\frac{\varepsilon \log n}{\log\log n}\right)$*, the pruned branch and bound tree will have* $e^{\Omega(n^{1-6\varepsilon})}$ *leaves almost surely.*

Chapter 3

Separation for Partial Euclidean Functionals

The classic problem of average case analysis on various Euclidean functionals often shows that the functionals almost surely converge to a value, when the region of plan is appropriately scaled. The typical formal setting to look at is when you have n points in $[0, t]^d$ where $t = n^{1/d}$ or when you have a Poisson point process at the rate of 1, on the same region, which ensures the expected number of points to be n. There are various results analysing functional like the Travelling Salesman Tour, Minimal Spanning Tree, Min Cost Maximum Matching, proving that their value almost surely converges to various different constants. A general framework for was established by proving that *Subadditive Euclidean Functionals* almost surely converge to a constant value in the first setting.

Establishing a separation between these constants has been an problem of interest, especially since it has implications on the running time of exact algorithms that use solution of one of the problems as a proxy to the other. The separation between constant for TSP and the Held-Karp lower bound was of perticular interest, as there was empirical evidence which suggested that the constants might have the same value. This problem was resolved by who proved that two constants are infact different, also implying that any exact branch and bound algorithm for TSP that uses Held-Karp inequalities as a lower bound must in fact have exponential running time, even in the average case.

3.1 Lower bound on length of large cycles

In this section, we will show a lower bound on length of large cycles which is linear in size of the cycle. Suppose that there is a large cycle consisting of k points, of length at most δk. First, we will use Gasoline Lemma [Lov79, Problem 3.21], which we formalize as follows:

Lemma 40 (Gasoline Lemma). *Consider a path $(x_1, \ldots, x_k)$ with k vertices. For simplicity of notation, assume that $x_i \equiv x_j$ if $i = j \mod k$. Suppose the total length of the path is given by*

$$\ell = \sum_{i=1}^{k} d(x_i, x_{i+1})$$

Suppose there are reals $a_1, \dots a_k \in \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^k a_i \geq l$. Then there exists an index s such that

$$\sum_{i=1}^{j} d(x_{s+i}, x_{s+i+1}) \leq \sum_{i=1}^{j} a_{s+i} \qquad \forall\, j, 1 \leq j \leq k$$

Suppose that $x_1, \dots, x_k$ is a cycle of total length δk. Then by choosing $a_i = \delta$ for $1 \leq i \leq k$, and Lemma 40, there exists an index s such that

$$\sum_{i=1}^{j} d(x_{s+i}, x_{s+i+1}) \leq j\delta \qquad \forall\, j, 1 \leq j \leq k. \tag{3.1}$$

Without loss of generality, we may assume that $s = 1$. Hence, if such a cycle exists, then there is a sequence of points $x_1, \dots, x_k$ such that Equation (3.1) holds for all $j \leq k$. We will compute the expected number of such sequences. Define the region $\mathcal{R} \in (R^2)^k$ containing of points $(x_1, \dots, x_k)$ such that $x_1 \in [0, \sqrt{n}]^2$, and Equation (3.1) holds for all $j \leq k$. Using Theorem 198, the expected number of such sequences is given precisely by computing volume $V(\mathcal{R})$. We can evaluate this volume using the following expression:

$$V(\mathcal{R}) = n \cdot (2\pi)^k \int_{r_1 \leq \delta} \int_{r_1 + r_2 \leq 2\delta} \cdots \int_{r_1 + \cdots + r_k \leq k\delta} r_1 \cdots r_k dr_k \cdots dr_1 \tag{3.2}$$

By AM-GM inequality, observe that

$$\prod_{i=1}^{k} r_i \leq \left(\frac{\sum_{i=1}^k r_i}{k} \right)^k \leq \left(\frac{k\delta}{k} \right)^k = \delta^k \tag{3.3}$$

Therefore, we get the upper bound

$$V(\mathcal{R}) \leq n \cdot (2\pi\delta)^k \cdot V(\mathcal{P}_k(\delta)) \tag{3.4}$$

where $V(S)$ denotes volume of the set S, and $\mathcal{P}_k(\delta)$ is the set of sequences $r = (r_1, \dots, r_k)$ such that $\sum_{i=1}^j r_i \leq \delta j$. We define $s = (s_1, \dots, s_k)$ such that $s_j = \sum_{i=1}^j r_i$. Let

$$\mathcal{Q}_k(\delta) = \{s = (s_1, \dots, s_k), s_1 \leq \cdots \leq s_k, 0 \leq s_i \leq i\delta\}.$$

The relation between s and r provides a diffeomorphism between $\mathcal{P}_k(\delta)$ and $\mathcal{Q}_k(\delta)$. Hence it suffices to compute volume of $\mathcal{Q}_k(\delta)$. Given any permutation $\pi \in S_k$, and a point $s \in \mathbb{R}^k$, we denote by $\pi(s)$ the action of π on coordinates of s. Since points in $\mathcal{Q}_k(\delta)$ have non-decreasing coordinates, for any $\pi_1, \pi_2 \in S_k$, $\pi_1 \mathcal{Q}_k(\delta)$ and $\pi_2 \mathcal{Q}_k(\delta)$ intersect on region of measure zero, since they can only intersect on the points that have at least two coordinates which are equal. Therefore,

$$V(S_k \mathcal{Q}_k(\delta)) = k! \cdot V(\mathcal{Q}_k(\delta))$$

where $S_k \mathcal{Q}_k(\delta) = \bigcup_{\pi \in S_k} \mathcal{Q}_k(\delta)$. For any point $s \in \mathbb{R}^k_{\geq 0}$, define $\phi : \mathbb{R}^k_{\geq 0} \to \mathbb{Z}^k_{\geq 0}$ such that $\phi(s)_i = \left\lfloor \frac{s_i}{\delta} \right\rfloor$. If $s \in S_k \mathcal{Q}_k(\delta)$, observe that

$$|\{i : \phi(s)_i \geq k - j\}| \leq j \tag{3.5}$$

Therefore, $\phi(s)$, thought of as a function of coordinate i, that is $i \mapsto \phi(s)_i$ is a parking function on set $\{0, \ldots, k-1\}$ for all $s \in S_k\mathcal{Q}_k(\delta)$, since eq. (3.5) is exactly the definition of parking functions. While it will be sufficient for our purposes to bound the number of choices for $\phi(s)$ by the number of all functions on $\{0, 1, \ldots, k-1\}$, namely k^k, we recall the following theorem due to Pyke [Pyk59] and Konheim and Weiss [KW66]:

Theorem 41. *Number of parking functions on set* $\{0, \ldots, k-1\}$ *is precisely* $(k+1)^{k-1}$.

Proof. (Due to Pollack (1974)[FR74; Staa; Stab]) We look at the set of all functions from $\{0, \ldots, k-1\}$ to $\{0, \ldots, k\}$. Given any such function f, we can define an injective function g, by doing the following: for $i = 0, \ldots, k-1$, find the smallest $m \geq 0$ such that $f(i) + m \mod (k+1)$ is free, that is there is no $j < i$ such that $g(j) = f(i) + m \mod (k+1)$. This is always possible since there are $k+1$ possible choice of m, and number of occupied positions can be at most k. Observe that for each f, there is exactly one index $\tau(f)$ such that $g^{-1}(\tau(f)) = \varnothing$, and f is a parking function if and only if $\tau(f) = k$. Now, define function f_i such that

$$f_i(x) = f(x) + i \mod (k+1)$$

Then $\tau(f_i) = \tau(f) + i \mod (k+1)$. Note that these translations partition the set of all functions, and each partition contains a unique parking function, therefore, total number of parking functions is $(k+1)^k/(k+1) = (k+1)^{k-1}$. $\square$

Now, we are in a position to compute $S_k\mathcal{Q}_k(\delta)$. The mapping ϕ associates exactly a set of size δ^k to each parking function. Therefore,

$$V(S_k\mathcal{Q}_k(\delta)) = (k+1)^{k-1}\delta^k$$

And therefore, we have

$$V(\mathcal{P}_k(\delta)) = V(\mathcal{Q}_k(\delta)) = \frac{(k+1)^{k-1}\delta^k}{k!} \leq \frac{k^k\delta^k}{k!} \leq (3\delta)^k \tag{3.6}$$

Where second last inequality follows from number of parking functions ($(k+1)^{k-1}$) being smaller than number of all functions (k^k), and the last inequality follows from a loose version of sterling's formula, $k! \geq \left(\frac{k}{3}\right)^k$. Putting this together with Equation (3.4), we get

$$V(\mathcal{R}) \leq n(6\pi\delta^2)^k \tag{3.7}$$

Let P denote the expected number of paths $(x_1, \ldots, x_k)$ of distinct points such that $(x_1, \ldots, x_k)$ satisfy Equation (3.1). Then we have

$$\mathbb{E}[P] \leq V(\mathcal{R}) \leq n(6\pi\delta^2)^k$$

Using Markov's inequality, we have

$$\mathbb{P}[P \geq 1] \leq \mathbb{E}[P] \leq n\left(6\pi\delta^2\right)^k. \tag{3.8}$$

Therefore, if $6\pi\delta^2 \leq \frac{3}{4}$, then $\mathbb{P}[P \geq 1] \leq ne^{-k/4}$. Hence, we get the following theorem:

Theorem 42. *For any $\epsilon > 0$, suppose the cost of optimal* $\texttt{TSP}$ *tour that visits at least $k = \varepsilon n$ points in $\mathcal{X}_n$ be denoted by $\texttt{TSP}_\varepsilon(\mathcal{X}_n)$. Then $\forall \varepsilon \in \mathbb{R}_{\geq 0}$*

$$\frac{\texttt{TSP}_\varepsilon(\mathcal{X}_n)}{k} \geq C \qquad \textit{with probability } 1 - e^{-\Omega(\varepsilon n)} \tag{3.9}$$

for some absolute constant $C \geq (6\pi)^{-1/2}$.

3.2 Upper bound on Average value of

In this section, we will upper bound the per point cost of min-cost ε-matching and ε-two-factor problems, where are are looking for matching (two-factor) that spans $k = \varepsilon n$ points among given n points. In particular, we will prove the following theorem:

Theorem 43. *For any $\varepsilon > 0$, suppose the cost of optimal matching and two-factor that covers at least $k = \varepsilon n$ points in $\mathcal{X}_n$ be denoted by $\texttt{MM}_\varepsilon(n)$ and $\texttt{TF}_\varepsilon(n)$ respectively. Then*

$$\lim_{\varepsilon \to 0} \frac{\texttt{MM}_\varepsilon(\mathcal{X}_n)}{k} = 0 \tag{3.10}$$

$$\lim_{\varepsilon \to 0} \frac{\texttt{TF}_\varepsilon(\mathcal{X}_n)}{k} = 0 \tag{3.11}$$

almost surely.

Proof. In order to construct matching or two factor with small cost, we break region $[0, \sqrt{n}]^2$ into squares with side length d. Then we will count the number of squares that contain either 2 points for matching or 3 points for a two factor.

Suppose there are s squares, $Q_1, \ldots, Q_s$ each of side length d, which cover $[0, \sqrt{n}]^2$. Then we have $s = n/d^2$. Let v denote the volume of each square, that is, $v = d^2$. Let X_i denote the event that square Q_i contains at least 2 points. Since $\mathcal{X}$ is generated using a poisson process,

$$\mathbb{P}[X_i = 1] = 1 - (1 + v)e^{-v} = \int_0^v xe^{-x}\, dx \geq \int_0^v x(1-x)\, dx = \frac{v^2}{2} - \frac{v^3}{3}.$$

For $v \leq 1$, we can further lower bound this by $\frac{v^2}{6}$, getting the following lower bound:

$$\mathbb{P}[X_i = 1] \geq \frac{v^2}{6} \tag{3.12}$$

Let $X = \sum_{i=1}^s X_i$. Then by Chernoff's inequality,

$$\mathbb{P}\big[X \leq (1-\delta)\mathbb{E}[X]\big] \leq \exp\left(-\frac{\delta^2 \mathbb{E}[X]}{2}\right). \tag{3.13}$$

By linearity of expectation,

$$\mathbb{E}[X] \geq s \cdot \frac{v^2}{6} = \frac{vn}{6}.$$

Choosing $v = 6\varepsilon$ and $\delta = \frac{1}{2}$, we get that $X \geq \frac{\varepsilon n}{2}$ with probability at least $1 - e^{-\Omega(\varepsilon n)}$. Since X counts the number of squares with two points, we can pick any two points and pick the edge between them in the matching. This provides a matching of total weight at most $2^{-0.5}\varepsilon nd$. Since we choose $v = 6\epsilon, d = \sqrt{6\varepsilon}$. Therefore, we have

$$\frac{\mathtt{MM}_{\varepsilon}(\mathcal{X}_n)}{k} = \frac{3^{0.5} \cdot \varepsilon^{1.5} n}{\varepsilon n} = O(\sqrt{\varepsilon}) \qquad \text{with probability } 1 - e^{-\Omega(\varepsilon n)}.$$

Where both O and Ω notation hide absolute constants. This implies the limit in eq. (3.10).

To upper bound the value of optimal two-factor, we follow the same argument but look at the squares that contain at least 3 points. We can then construct a triangle in each of these squares which gives us a valid two factor. Let Y_i denote the event that Q_i contains at least 2 points. Then

$$\mathbb{P}[Y_i = 1] = 1 - \left(1 + v + \frac{v^2}{2}\right)e^{-v} = \int_0^v \frac{x^2}{2} e^{-x}\, dx \geq \int_{0^v} \frac{x^2}{2}(1 - x)\, dx = \frac{v^3}{6} - \frac{v^4}{8}.$$

For $v \leq 1$, we have the lower bound

$$\mathbb{P}[Y_i = 1] \geq \frac{v^3}{24} \tag{3.14}$$

Defining $Y = \sum_{i=1}^{s} Y_i$, linearity of expectation gives us

$$\mathbb{E}[Y] \geq s \cdot \frac{v^3}{24} = \frac{v^2 n}{24}$$

Choosing $v^2 = 16\varepsilon$ and $\delta = \frac{1}{2}$, we get that $Y \geq \frac{\varepsilon n}{3}$ with probability at least $1 - e^{-\Omega(\varepsilon n)}$. Picking a triangle from each of these squares provides us a two factor with εn points, with total weight at most $2^{0.5}\varepsilon nd$. Since $v^2 = 16\varepsilon$, $d = 2\varepsilon^{0.25}$. Therefore, we have

$$\frac{\mathtt{TF}_{\varepsilon}(\mathcal{X}_n)}{k} = \frac{2^{0.5}\varepsilon^{1.25} n}{\varepsilon n} = O(\sqrt[4]{\varepsilon}) \qquad \text{with probability } 1 - \varepsilon^{-\Omega(\varepsilon n)}.$$

Again, both O and Ω notation hide absolute constants. This gives us the limit in eq. (3.11). $\square$

Chapter 4

Direct sampling for paths on grid

Analysis of political redistrictings has created a significant impetus for the problem of random sampling of graph partitions into connected pieces—e.g., into districtings.

The most common approach to this problem in practice is to use a Markov Chain; e.g., Glauber dynamics, or chains based on cutting spanning trees (e.g., [DDS19; Aut+19; Aut+21]). Rigorous understanding of mixing behavior is the exception rather than the rule; for example, [MP15] established rapid mixing of a Markov chain for the special case where both partition classes are unions of horizontal bars, which in each case meet a common side. No rigorous approach is known, for example, which can approximately uniformly sample from contiguous 2-partitions even of lattice graphs like the $n \times n$ grid in polynomial time

In this paper we consider a direct approach, where instead of leveraging a Markov chain with unknown mixing time to generate approximate uniform samples, we use a dynamic programming algorithm and rejection sampling to exactly sample from self-avoiding walks in the lattice $\mathbb{Z}^2$ (which correspond to partition boundaries) in polynomial expected time. Counting self-avoiding lattice walks is a significant long-standing challenge; the connective constant—the base of the exponent in the asymptotic formula for the number of such walks—is not even known for $\mathbb{Z}^2$. But we will be interested in sampling *nearly-shortest* self avoiding walks, motivated by districting constraints which discourage the use of large district perimeters relative to area. In particular, we will prove:

Theorem 44. *For any C and $\varepsilon > 0$ and for any n_1, n_2, and $n = n_1 + n_2$, there is a randomized algorithm which runs w.h.p in polynomial time, and produces a uniform sample from the set of self-avoiding walks in $\mathbb{Z}^2$ from $(0,0)$ to (n_1, n_2) of length at most*

$$n + Cn^{1-\varepsilon}.$$

A variant of this algorithm can be used to sample from contiguous 2-partitions of the Aztec diamond with restricted partition-class perimeter, by sampling short paths between nearly-antipodal points on the dual of the Aztec diamond. These paths are in bijection with the contiguous 2-partitions of the Aztec diamond, by mapping a partition to it's boundary which gives us a path. This approach generates samples in polynomial time w.h.p. In contrast, we show that the traditional approach using Markov chains is inefficient:

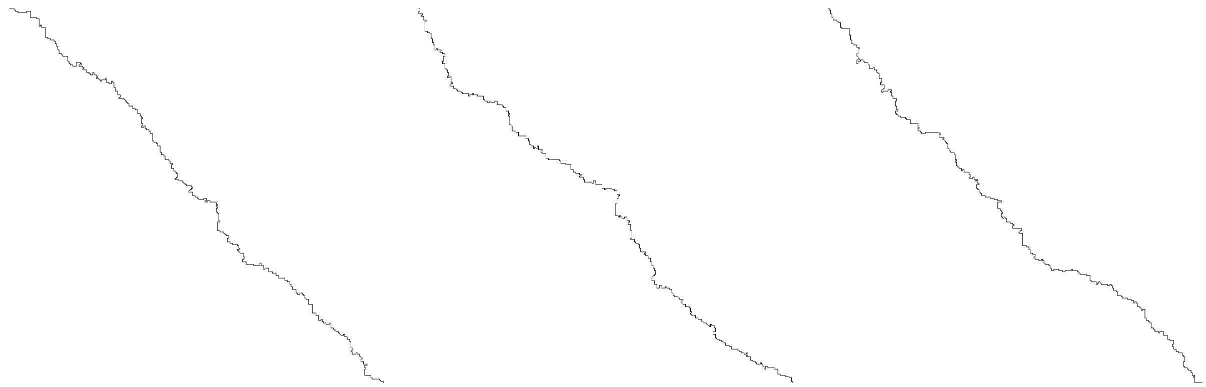


Figure 4.1: Uniformly random self-avoiding walks of length 700 between corners of a 300×300 grid, generated with the algorithm from Theorem 44.

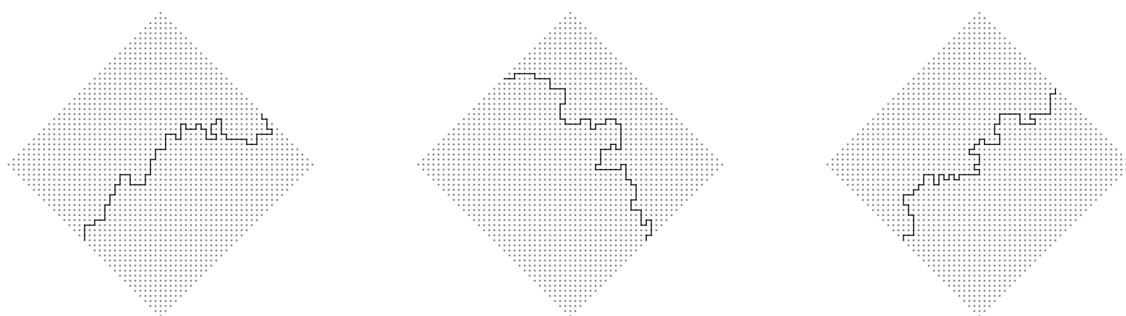


Figure 4.2: Uniformly random self-avoiding walks on A_{30} such that both sides have perimeter of at most 220.

Theorem 45. *For any C and $\varepsilon > 0$, Glauber dynamics has exponential mixing time on contiguous 2-partitions of the Aztec diamond A_k when constrained by perimeter slack $Ck^{1-\varepsilon}$.*

Organization of the Paper: The paper is organized in the following manner: Section 4.1 describes a dynamic programming algorithm (Algorithm 1) to sample walks without short cycles and proves its correctness. Sections 4.2 and 4.3 show that the algorithm actually returns a self-avoiding path from $(0,0)$ to (n_1, n_2) in the unbounded lattice graph $\mathbb{Z}^2$ in polynomial time with high probability, enabling the random sampling of paths for rejection sampling. Section 4.4 provides the same result for *wide* subgraphs of the lattice, the notion of *wide* subgraph is also defined in this section. The last section, Section 4.5 is dedicated to proving Theorem 45, and showing that Aztec diamond is a *wide* subgraph of the lattice.

Notation: For the rest of the paper, we will typically use letters $A, B, \ldots$ for denoting paths from $O = (0,0)$ to $P = (n_1, n_2)$. We will use letters $Q, R, \ldots$ to denote points on the grid. Each path A from O to P of length $n + 2k$ has two representations, we can describe A by the sequence of moves $a_1, \ldots, a_{n+2k}$ where $a_i \in L, R, U, D$ denotes the direction of next step in the path. On the other hand, we can also denote path A by the sequence of points that it visits, namely, $O = A_0, \ldots, A_{n+2k} = P$. Typically, we will also use B to denote a shortest path, and A to denote a larger path.

We will further let P_k, W_k, W_k^l denote the number of paths (self-avoiding walks), number of walks, and number of walks without cycles smaller than $2l$ from O to P of length $n+2k$ respectively.

4.1 Dynamic Programming Algorithm

In this section, we will describe the dynamic programming algorithms that counts W_k^l, the number of walks of length $n+2k$ without short cycles, that is, without cycles of length smaller than $2l$ from $O = (0,0)$ to $P = (n_1, n_2)$ in a subgraph S of the grid $\mathbb{Z}^2$. The algorithm memorizes the number of paths from every point $Q \in S$ to P, along with previous $2l$ steps, which is given by a walk w of length $2l$ ending at Q. Let $\Phi_l(Q)$ denote the set of paths ending at Q of length at most $2l$.

Algorithm 1 Counting Low Girth Walks

$\mathrm{DP}(Q, P, w, t) = 0$ for $Q \in S$, $w \in \Phi_l(Q)$, $0 \le t \le n+2k$
function WALKS(Q, P, w, t)
  if $t = 0$ **then**
    if $Q = P$ **then**
      return $\mathrm{DP}(Q, P, w, t) = 1$
    else
      return $\mathrm{DP}(Q, P, w, t) = 0$
    end if
  end if
  if $\mathrm{DP}(Q, P, w, t) \neq 0$ **then**
    return $\mathrm{DP}(Q, P, w, t)$
  end if
  for $d \in \{(1,0), (0,1), (-1,0), (0,-1)\}$ **do**
    if $Q + d \in S$ **and** $d \notin w$ **then**
      $R = Q + d$
      w' is the path obtained by appending R to w and trimming down to length $2l$.
      $\mathrm{DP}(Q, P, w, t)$ += WALKS$(R, P, w', t-1)$
    end if
  end for
  return $\mathrm{DP}(Q, P, w, t)$
end function

Algorithm 2 Sampling Low Girth Walks

function SAMPLE WALKS(k)
  $w = O$
  for $i = 0$ to $n + 2k$ **do**
    for $Q \sim w[i]$ **do**
      $p_Q = \mathrm{DP}(Q, w', n + 2k - 1 - i)$
      where $w' = w[i - 2l + 1] \cdots w[i]Q$ is path of length $2l$ ending at Q
    end for
    Sample $w[i+1]$ from $Q \sim w[i]$ proportional to p_Q.
  end for
  return w
end function

Algorithm 3 Sampling Paths

function SAMPLE PATHS(k)
  while w is not a path **do**
    $w =$ SAMPLE WALKS(k)
  end while
  return w
end function

Once we have number of these paths, we can sample a walk of length $n + 2k$ without cycles of length smaller than $2l$ by starting at O and sampling points in the walk with correct probability using memoized values obtained by algorithm 1.

Since there are at most 4^{2l} paths of length $2l$, $|\Phi_l(Q)| \leq \sum_{i=0}^{l} 16^i = 2 \cdot 16^l$ for any point Q. Therefore, size of the DP table in Algorithm 1 is $|S| \cdot 16^l$, and each entry in this table takes $O(l)$ time to compute, since deg of each vertex in S is at most 4. Therefore, Algorithm 1 takes $O\big(|S| \cdot l \cdot 16^l\big) = O(|S|)$ time for constant l. Note that these paths are restricted to the set of points $\mathcal{R} = \{Q \mid O - (k,k) \leq Q \leq P + (k,k)\}$. Thus, for large S (in particular for $S = \mathbb{Z}^2$), we can restrict the algorithm to $S' = \mathcal{R} \cap S$.

Further, once the DP table is computed, Algorithm 2 runs in $O(n+2k)$ time. We will prove in Theorem 60 that for $k \leq Cn^{1-\varepsilon}$ and $S = \mathbb{Z}^2$, Algorithm 2 actually returns a path with probability $1 - o(1)$ for $l > \frac{1}{\varepsilon}$. This implies that Algorithm 3 runs in $O(n+2k)$ time with high probability, completing the proof of Theorem 44. We will provide a sufficient condition for subgraphs $S \subseteq \mathbb{Z}^2$ in Theorem 65 which implies the same probability bound for these specific subgraphs S.

4.2 Number of Paths in a Grid

This section focuses on getting bounds on the number of paths from $O = (0,0)$ to $P = (n_1, n_2)$ in the grid. Recall that paths are in fact *self-avoiding walks*. Let $n = n_1 + n_2$ be the length of a shortest path from O to P. We will provide some upper and lower bounds on the number of paths of length $n + 2k$ from O to P in terms of number of shortest paths from O to P. These upper and lower bounds are based on constructing extensions of shortest paths.

In general, we will associate a shortest *base path* to every path from O to P. This association is described in Definition 52. We will also provide procedures for extending shortest paths to larger paths, which respects the base path mapping. Then the lower bound on paths of length will follow by bounding the number of extensions of each shortest path, and upper bound will follow from bounding the number of paths of length $n+2k$ that have a specific given path as the associated *base path*.

Let a shortest path B be described by sequence of moves $b_1, \ldots, b_n$ where $n = n_1 + n_2$, where each $b_i \in \{U, R\}$ describes the direction of move at i^{th} step. Then we have the following procedure to extend the path B to a path A from O to P of length $n + 2k$.

Definition 46. Given a shortest path B represented by $b_1, \ldots, b_n$ from $O = (0,0)$ to $P = (n_1, n_2)$ where $n = n_1 + n_2$, and a set $M = \{i_1, \ldots, i_k\}$ of indices, we define the extended path $A = \mathcal{A}(B, M)$ obtained by performing following replacements for all $j = 1, \ldots, k$:

1. If $b_{i_j} = R$, replace it by DRU.
2. If $b_{i_j} = U$, replace it by LUR.

For an edge b_i, we will also refer to the operation above as *bumping the edge*. Further, we will say that an edge b_i can be *bumped* if bumping the edge b_i gives us a path.

Figure 4.3 illustrates how Definition 46 behaves when extending shortest paths. It is not true that for all choices of M the map $\mathcal{A}(B, M)$ is a path. But, we will show that for a large choice of set M, it is a path.

Figure 4.3: Bumping a shortest path at indices $2, 6, 9$

Lemma 47. *For any choice of M such that $b_{i_j-1} = b_{i_j}$ for all j, the map $\mathcal{A}(B, M)$ gives us a path.*

Proof. Let path B be go through the points $O = B_0 \dots B_n = P$. Then for any point $B_i = (x_i, y_i)$ if the point $X = (x_i - 1, y_i)$ is also in the path B then X must be connected to B_i, and hence $B_{i-1} = X$ since otherwise there is a subpath from (x_i, y_i) to $(x_i - 1, y_i + 1)$ (or the other way around) in B, which implies that B is not a shortest path.

In particular, if $b_{i-1} = b_i = U$ then the points to the left of B_{i-1} and B_i, that is, the points $(x_{i-1} - 1, y_{i-1})$ and $(x_i - 1, y_i)$ are not in B. Therefore, if we replace b_i by LUR, we change the portion of path from B_{i-1} to B_i to look like

$$B_{i-1} = (x_{i-1}, y_{i-1}) \to (x_{i-1} - 1, y_{i-1}) \to (x_{i-1} - 1, y_{i-1} + 1) = (x_i - 1, y_i) \to (x_i, y_i) = B_i$$

which is a path since newly added points were not in B initially. Similar argument works for DRU modifications. The modifications of type $U \to LUR$ and $R \to DRU$ happen on opposite side of the path B, and hence don't intersect. Further, all the modifications of type $U \to LUR$ don't intersect unless they are adjacent to each other. Therefore, if the set M contains non-adjacent indices, then we can perform all the modifications simultaneously without creating any loops. Further, observe that these modifications do not intersect each other if the set M contains non-adjacent indices, and can be performed simultaneously. $\square$

We will use this procedure described in Definition 46 to generate a family of paths of length $n + 2k$. To ensure that there are a lot of choices for M, we need to argue that most shortest paths from O to P have $\frac{n}{2} - o(n)$ many places where the hypothesis of Lemma 47 is satisfied. This is formalized in the next lemma.

Lemma 48. *For any point $P = (n_1, n_2)$ with $n_1 + n_2 = n$, a shortest path from $O = (0, 0)$ to P drawn uniformly at random has at least $\frac{n}{2} - O(\sqrt{-n \log \varepsilon})$ places with two consecutive moves in the same direction with probability $1 - \varepsilon$.*

Proof. Let B be a shortest path from O to P. B can be denoted as a sequence of exactly n_1 right moves and exactly n_2 up moves. Let us denote this path by $b_1, \dots, b_n$ where $b_i \in R, U$. We can draw a path uniformly at random by picking uniformly at random from a bag with n_1 R symbols and n_2 U symbols without replacement. Let X_i be the indicator random variable for the event that $b_i = b_{i+1}$. Now, we first observe that

$$\mathbb{P}[X_i \mid b_1, \dots, b_{i-1}] = \frac{p(p-1)}{r(r-1)} + \frac{q(q-1)}{r(r-1)} = \frac{p^2 + q^2 - r}{r(r-1)} \geq \frac{1}{2} - \frac{1}{2(r-1)}$$

where p is number of U symbols left in the bag, q is number of R symbols left in the bag and $r = p + q$. Now, we will show that $\mathbb{P}[X_i \mid X_1, \ldots, X_{i-1}] \geq \frac{1}{2} - \frac{1}{n-i-1}$. It suffices to show that $\mathbb{P}[X_i = 1 \mid b_1, \ldots, b_{i-2}, X_{i-1}] \geq \frac{1}{2} - \frac{1}{n-i-1}$. We will show this by doing two cases: $X_{i-1} = 0$ and $X_{i-1} = 1$. In the first case, $X_{i-1} = 0$,

$$\begin{aligned}
\mathbb{P}[X_i = 1 \mid b_1, \ldots, b_{i-2}, X_{i-1} = 0] &= \frac{\mathbb{P}[X_i = 1, X_{i-1} = 0 \mid b_1, \ldots, b_{i-2}]}{\mathbb{P}[X_{i-1} = 0 \mid b_1, \ldots, b_{i-2}]} \\
&= \frac{\frac{pq(q-1)+qp(p-1)}{r(r-1)(r-2)}}{\frac{pq+qp}{r(r-1)}} \\
&= \frac{pq(p+q-2)}{2pq(r-2)} = \frac{1}{2}
\end{aligned}$$

where p is number of U symbols left, q is the symbol of R symbol left, and $r = p + q$. In the second case, using the same notation, we have

$$\begin{aligned}
\mathbb{P}[X_i = 1 \mid b_1, \ldots, b_{i-2}, X_{i-1} = 1] &= \frac{\mathbb{P}[X_i = 1, X_{i-1} = 1 \mid b_1, \ldots, b_{i-2}]}{\mathbb{P}[X_{i-1} = 1 \mid b_1, \ldots, b_{i-2}]} \\
&= \frac{\frac{p(p-1)(p-2)+q(q-1)(q-2)}{r(r-1)(r-2)}}{\frac{p(p-1)+q(q-1)}{r(r-1)}} \\
&= \frac{p(p-1)(p-2) + q(q-1)(q-2)}{(p(p-1) + q(q-1))(r-2)} \\
&= \frac{p^3 + q^3 - 3(p^2 + q^2) + 2(p+q)}{(p^2 + q^2 - (p+q))(r-2)} \\
&= \frac{r^3 - 3pqr - 3(r^2 - 2pq) + 2r}{(r^2 - 2pq - r)(r-2)} \\
&= \frac{r^3 - 3r^2 + 2r - 3pq(r-2)}{(r^2 - r - 2pq)(r-2)} \\
&= \frac{r(r-1)(r-2) - 3pq(r-2)}{(r^2 - r - 2pq)(r-2)} \\
&= \frac{r(r-1) - 3pq}{r(r-1) - 2pq}
\end{aligned}$$

Note that this term is maximized when pq is minimized, and is minimized when pq is maximized. Constrained to the fact that $p + q = r$ and $p, q \geq 0$, we get

$$1 \geq \frac{r(r-1) - 3pq}{r(r-1) - 2pq} \geq \frac{4r^2 - 4r - 3r^2}{4r^2 - 4r - 2r^2} = \frac{r-4}{2(r-2)} = \frac{1}{2} - \frac{1}{r-2}$$

Therefore, in both cases, we have

$$\mathbb{P}[X_i = 1 \mid X_1, \ldots, X_{i-1}] \geq \frac{1}{2} - \frac{1}{n-i-1}$$

Now, we couple variables X_i with variables e_i, drawn independently such that $\mathbb{P}[e_i = 1] = \frac{1}{2} - \frac{1}{n-i-1}$. To begin with, we draw b_1 with correct probabilities. Then for each i, we draw f_i uniformly at random from $[0, 1]$. We set $e_i = 1$ if $f_i \leq \mathbb{P}[e_i = 1]$ and we set $e_i = 0$ otherwise. Further, if $f_i \leq \mathbb{P}[X_i = 1 \mid X_1, X_2, \ldots, X_{i-1}]$, then we set a_{i+1} such that $X_i = 1$, otherwise we set a_{i+1} such that $X_i = 0$; note that the status of X_{i+1} uniquely determines the choice of a_{i+1}. Therefore, $e_i = 1 \implies X_i = 1$, and hence $\sum_{i=1}^{n-1} e_i \leq \sum_{i=1}^{n-1} X_i$. Notice that e_i are still independent random variables. Therefore,

$$\mathbb{P}\Big[\sum X_i \leq \mathbb{E}\Big[\sum e_i\Big] - t\Big] \leq \mathbb{P}\Big[\sum e_i \leq \mathbb{E}\Big[\sum e_i\Big] - t\Big] \leq \exp\Big(-\frac{2t^2}{n}\Big)$$

Where the last inequality follows from Hoeffding's inequality. Note that

$$\mathbb{E}\Big[\sum e_i\Big] = \sum \frac{1}{2} - \frac{1}{n-i+1} \geq \frac{n}{2} - 2\log n$$

Given any $\varepsilon > 0$, and $t = \sqrt{-n \log \varepsilon}$, we get that

$$\mathbb{P}\Big[\sum X_i \leq \frac{n}{2} - 2\log n - \sqrt{-n\log\varepsilon}\Big] \leq \varepsilon$$

This proves the required result. $\square$

This allows us to lower bound the number of paths of length $n + 2k$ from $O = (0, 0)$ to $P = (n_1, n_2)$ where $n_1 + n_2 = n$. Recall that P_k denotes the number of these paths.

Lemma 49. *For any $k \leq 0.1n$ and $1 > \varepsilon \geq 0$, we have the lower bound*

$$P_k \geq (1-\varepsilon)P_0\binom{t-2k}{k}$$

where $t = \frac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon)}$. Further, there is $n_0 = n_0(\varepsilon)$, such that for all $n \geq n_0$,

$$P_k \geq (1-\varepsilon)P_0\frac{(0.49)^k n^k}{k!}\exp\Big(-O\Big(\frac{k^2}{n}\Big)\Big) \tag{4.1}$$

Proof. Consider a path B of length n from O to P. Let B be represented by $b_1, \ldots, b_n$ where $b_i \in \{R, U\}$. Then using Definition 46 and lemma 47, we can extend B to a path $A = \mathcal{A}(B, M)$ of lenght $n + 2k$ if we choose M to be a set such that there are no adjacent indices in M and further, for each $i \in M$, $b_{i-1} = b_i$. There are at least

$$t = \frac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon)}$$

such indices, for at least $(1-\varepsilon)P_0$ many paths. For each of these paths, we need to choose a set of k non-adjacent indices. This can be done in at least

$$\frac{t(t-3)(t-6)\ldots(t-3(k-1))}{k!} \geq \frac{(t-2k)(t-2k-1)\ldots(t-3k+1)}{k!} = \binom{t-2k}{k} \tag{4.2}$$

many ways, since after picking first index, we lost 3 possible choices for rest of the indices. Further, observe that any longer path A that is obtained in this way corresponds to exactly one shortest path B. We can find this path B by looking at patterns LUR and DRU and replacing them by U and R respectively. If M is choosen satisfying conditions of Lemma 47, then it is clear that every L in the extended path A is followed by UR and every D in A is followed by RU. Hence, these replacements can be made unambiguously. Since we can do this for all $(1-\varepsilon)P_0$ paths, we get the lower bound.

$$P_k \geq (1-\varepsilon)P_0\binom{t-2k}{k}$$

Since $2\log n + \sqrt{n\log(1/\varepsilon)} = o(n)$, there is $n = n(\varepsilon)$ such that for all $n \geq n(\varepsilon)$, $2\log n + \sqrt{n\log(1/\varepsilon)} \leq 0.01n$, and hence $t \geq 0.49n$. This gives us the lower bound

$$P_k \geq (1-\varepsilon)P_0\binom{0.49n-2k}{k}$$

Using Equation (D.6), we have

$$\begin{aligned}
P_k &\geq (1-\varepsilon)P_0\frac{(0.49)^k n^k}{k!}\exp\left(\frac{-4k^2-k^2+k}{0.49n} - \frac{2k(2k+k)}{0.49n}\right) \\
&\geq (1-\varepsilon)P_0\frac{(0.49)^k n^k}{k!}\exp\left(-\frac{25k^2}{n}\right) \\
\implies P_k &\geq (1-\varepsilon)P_0\frac{(0.49)^k n^k}{k!}\exp\left(-O\left(\frac{k^2}{n}\right)\right)
\end{aligned}$$

completing the proof of the lemma. □

The next task is to extend this result to get similar bounds for extending paths of length $n+2k$ to paths of length $n+2k+2l$. We will prove the following:

Lemma 50. *For any $k, l \leq 0.1n$ and $1 > \varepsilon \geq 0$, there is $n_0 = n_0(\varepsilon)$ such that for all $n \geq n_0(\varepsilon)$,*

$$P_{k+l} \geq (1-\varepsilon)P_k\binom{t-8k-3l}{l}\binom{k+l}{l}^{-1}$$

where $t = \frac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon) + 2kn\log n + 30k^2}$. Further, there is $n_1 = n_1(\varepsilon)$, such that for all $n \geq n_1$,

$$P_{k+1} \geq (1-\varepsilon)P_k\frac{(0.49)^l n^l k!}{(k+l)!}\exp\left(-O\left(\frac{k(k+l)}{n}\right)\right) \tag{4.3}$$

The outline of proof of this lemma will be similar to Lemma 49. Consider a path A of length $n+2k$ from O to P. We want to show that for a large number of sets $M = \{i_1, \ldots, i_k\}$, we can construct the extended path $C = \mathcal{A}(A, M)$. To ensure we can find a large number of candidates for M, we will associate a shortest path to each path A. We define a map $\mathcal{B}$ in Definition 52 such that $\mathcal{B}(A)$ gives us such a shortest path. We further associate each the edges of $B = \mathcal{B}(A)$ to some of the edges of A, and we call these the *good edges* of A and all other edges of A as *bad edges* of A. This mapping is defined in Definition 56. We claim that the set of indices where we cannot do modifications in the extension procedure defined in Definition 46 corresponds to either a corner of B or a *bad edge* of A. Then we can bound the number of corners and bad edges to get the bound required.

We begin the proof begin by defining *lattice boxes* to make notation easier, and then use those to define the map $\mathcal{B}$.

Definition 51. Given points $P_1, P_2 \in \mathbb{Z}^2$, such that $P_1 \leq P_2$, we define the *lattice box* $\mathcal{R}(P_1, P_2)$ with left bottom corner P_1 and right top corner P_2 to be the rectangle with sides parallel to the axis with P_1 and P_2 as diagonally opposite corners. To be precise,

$$\mathcal{R}(P_1, P_2) = \left\{x \in \mathbb{Z}^2 \,\middle|\, P_1 \leq x \leq P_2\right\}$$

We further define *boundary* of a lattice box (and more generally of any set $S \subseteq \mathbb{Z}^2$) to be the set of vertices $v \in S$ such that v has at least one neighbor outside S in the infinite grid graph.

Definition 52. We define the map $\mathcal{B}$ as follows. Consider a path A given by points $O = A_0, \ldots, A_{n+2k} = P$ from $O = (0,0)$ to $P = (n_1, n_2)$ with $n_1, n_2 \geq 0$ and $n = n_1 + n_2$. We will build $\mathcal{B}(A) = B$ inductively, starting at $O = (0,0)$. We will do this by constructing a sequence of points R_i which will all lie in the intersection $A \cap B$. Let $R_0 = O$. Suppose we have constructed $R_0, \ldots, R_i$.

1. Construct a box $\mathcal{R}_i = \mathcal{R}(R_i, P)$ with R_i as the bottom left corner and P as the top right corner.
2. Find the next point R_{i+1} on A, after R_i such that $R_{i+1} \in \mathcal{R}_i$.
3. Extend B to R_{i+1} using the shortest path along the boundary of $\mathcal{R}_i$ if $R_{i+1} \neq P$.
4. If $R_{i+1} = P$, then let $\bar{A}$ be part of A between $R_i = (R_i(x), R_i(y))$ and P.
   - If $\bar{A}$ intersects $y = n_2$ before $x = n_1$, define $\bar{R} = (R_i(x), n_2)$
   - Otherwise define $\bar{R} = (n_1, R_i(y))$.

   Extend B from R_i to $\bar{R}$ to P.

Lemma 53. *The map $\mathcal{B}$ in Definition 52 is well defined.*

Proof. Given $R_i \neq P$, we can always find R_{i+1} since $P \in \mathcal{R}_i$ and $P \in A$, so A eventually intersects $\mathcal{R}_i$. Therefore, steps $(1,2)$ in Definition 52 are well defined. For step (3), observe that P is the only point on boundary of $\mathcal{R}_i$ that has two shortest paths from R_i along the boundary. Therefore, (3) is well defined as long as $R_{i+1} \neq P$.

For step (4), observe that if $\mathcal{R}_i$ is degenerate, then there is a unique path from R_i to P, and this step is well defined. Suppose $\mathcal{R}_i$ is non-degenerate. That is, R_i and P differ at both x and y coordinates. In this case, $\bar{A}$ cannot intersect both the lines $y = n_2$ and $x = n_1$ simultaneously, and it must intersect both of them eventually. Hence, step (4) is well defined as well. $\square$

We now define the *good edge mapping*. First, we will start by making a few notational definitions.

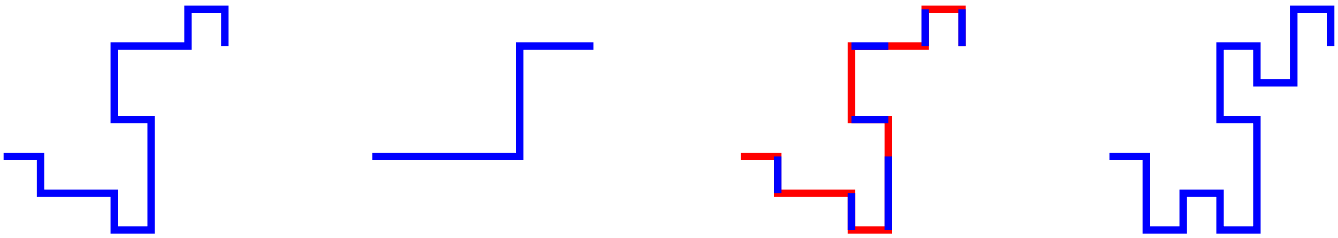


Figure 4.4: Illustrations for Definitions 52 and 56. First image shows a non-shortest path A, second image is the *base path* $\mathcal{B}(A)$, third image indicates good forward edges in green, and forth image is the path obtained by bumping at indices $3, 14$.

Definition 54. Given a path A from O to P with points $O = A_0, \ldots, A_{n+2k} = P$, we can represent it as a sequence of moves, $a_1 \ldots a_{n+2k}$, where each move is one of the four directions (U, D, L, R). We say that i^{th} point (A_i) on this path is a *corner* if $a_i \neq a_{i+1}$. We further include O and P to be corner points.

We define *last corner point* to be the corner point $Q \neq P$ with highest index. We will also refer to O as the *starting point* and to P as the *ending point*.

Definition 55. Let A be a path of length $n+2k$ from $O = (0,0)$ to $P = (n_1, n_2)$, where $n = n_1 + n_2$. Let A be given by point $O = A_0, \ldots, A_{n+2k} = P$. Then, we divide edges of A into two categories. Any edge going in the directions D or L will be reffered to as a *reverse edge*, and any edge going in the direction U and R will be reffered to as a *forward edge*.

Definition 56. In the setting described in the previous definition, let $B = \mathcal{B}(A)$, where $\mathcal{B}$ is defined in Definition 52. Let B be given by $O = B_1, \ldots, B_n = P$. We define a *good edge mapping* to be any function $\mathcal{F}_A : \mathbb{Z}_{[0,n-1]} \to \mathbb{Z}_{[0,n+2k-1]}$, where $\mathbb{Z}_{[0,t]} = \mathbb{Z} \cap [0,t]$ satisfying

1. $\mathcal{F}_A$ is injective.
2. For $i < j$, $\mathcal{F}_A(i) < \mathcal{F}_A(j)$.
3. The edges $A_{\mathcal{F}_A(i)} A_{\mathcal{F}_A(i)+1}$ and $B_i B_{i+1}$ are *super-parallel*, that is
   - If edge $B_i B_{i+1} = (x,y) \to (x, y+1)$, the edge $A_{\mathcal{F}_A(i)} A_{\mathcal{F}_A(i)+1} = (\bar{x}, y) \to (\bar{x}, y+1)$ for some $\bar{x}$.
   - If edge $B_i B_{i+1} = (x,y) \to (x+1, y)$, the edge $A_{\mathcal{F}_A(i)} A_{\mathcal{F}_A(i)+1} = (x, \bar{y}) \to (x+1, \bar{y})$ for some $\bar{y}$.

Given such a mapping $\mathcal{F}$, we will refer to any edge of form $A_{\mathcal{F}(i)} A_{\mathcal{F}(i)+1}$ to be a *good forward edge*, and any edge that is not a good forward edge as a *bad forward edge*.

Figure 4.4 illustrates the definitions above. We show that such a mapping exists in the lemma below.

Lemma 57. *Given a map A of length $n + 2k$ and let $B = \mathcal{B}(A)$. Using notation in Definitions 55 and 56, there exists a good edge mapping $\mathcal{F}$ satisfying conditions in Definition 56.*

Proof. First, it immediately follows from definitions 52 and 54 that all the corners of path B are contained in the set $\{O = R_0, R_1, \ldots, R_m = P, \bar{R}\}$, since the portions of B in between these points are straight lines. Now, we define the mapping $\mathcal{F} = \mathcal{F}_A$ for parts of B between R_i and R_{i+1} for $0 \leq i \leq m-2$, for each edge B_jB_{j+1} between R_iR_{i+1} in B, we define $\mathcal{F}(j) = k$ to be the least index such that A_kA_{k+1} and B_jB_{j+1} are super-parallel, that is, they satisfy the condition (3) in Definition 56.

We claim that this is strictly monotonic for each i. Suppose not, then there is an index j such that such that $\mathcal{F}(j+1) \leq \mathcal{F}(j)$. If $\mathcal{F}(j+1) = \mathcal{F}(j)$, then edges B_jB_{j+1} and $B_{j+1}B_{j+2}$ are super-parallel, which is a contradiction. Without loss of generality, let the points $R_i, B_j, B_{j+1}, R_{i+1}$ share the same x coordinate, that is, let $R_i = (x_0, y_0)$, $B_j = (x_0, y_1)$, $B_{j+1} = (x_0, y_1+1)$ and $R_{i+1} = (x_0, y_2)$. Then $A_{\mathcal{F}(j+1)} = (x_1, y_1+1)$ for some x_1. Then the path from $R_i = (x_0, y_0)$ to (x_1, y_1+1) must have an edge of the form $(x_2, y_1) \to (x_2, y_1+1)$ since $y_0 \leq y_1$. Therefore, there is an index $k < \mathcal{F}(j+1)$ such that A_kA_{k+1} is super-parallel to the edge B_jB_{j+1}, which implies $\mathcal{F}(j) < \mathcal{F}(j+1)$, a contradiction!

If the path between R_{m-1} and $R_m = P$ is straight line, we can extend the definition above when $i = A-1$. Otherwise, the point $\bar{R}$ is well defined. Let $\bar{A}$ be portion of A between R_{m-1} and P. Without loss of generality, let $\bar{A}$ intersect the line $y = n_2$ before the line $x = n_1$ at a point Q. Suppose $Q = (x_0, n_2)$, then $x_0 < n_1$, otherwise the path from R_{m-1} to Q will intersect the line $x = n_1$. Since Q is also outside $\mathcal{R}(R_{m-1}, P)$, it follows that $x_0 < x_1$ where $R_i = (x_1, y_1)$.

Now, for all B_j between R_{m-1} and $\bar{R}$, we define $\mathcal{F}(j) = k$ where k is the smallest index such that A_k is between R_{m_1} and Q such that B_jB_{j+1} and A_kA_{k+1} are super-parallel and for all B_j between $\bar{R}$ and P, we define $\mathcal{F}(j) = k$ where k is the smallest index such that A_k is between Q and P such that B_jB_{j+1} and A_kA_{k+1} are super-parallel.

This map is well defined and monotonic since $\bar{A}$ must go from $y = y_1$ to $y = n_2$, and then from $x = x_0$ to $x = n_1$, and hence edges super parallel to B_jB_{j+1} exists for all B_j between R_{m-1} and P. Further, the map is strictly monotonic by an argument earlier in the proof. This gives us the *good edge mapping* that we want. $\square$

The next lemma proves that a large number of *good edges* can be bumped.

Lemma 58. *Consider a path A of length $n+2k$. Let $B = \mathcal{B}$ be the base path associated with it. Suppose B has c corners. Then there is a set G of indices of at least $n-c-8k$ good edges in A which can be* bumped*.*

Proof. Note that A has exactly n good forward edges, k bad forward edges and k reverse edges. Now, we transverse A, and for each good forward edge, we check if we can *bump* the good forward edge. To be presice, consider a good forward edge S_1S_2. Without loss of generality, we will assume that the edge goes in U direction, and is given by $(x_0, y_0) \to (x_0, y_0+1)$.

Suppose S_1S_2 is a good forward edge that cannot be bumped. We will associate either

1. a reverse edge
2. a bad forward edge
3. or a corner of B

as the reason why bumping at S_1 is blocked. Since S_1S_2 cannot be bumped, either $S_3 = (x_0-1, y_0)$ is in A or $S_4 = (x_0-1, y_0+1)$ is in A.

First, consider the case when S_3 is contained in A. Look at the edge e going out of S_3 in A. We have following cases:

1. If there is no such edge, then $S_3 = P$. In this case, we say that P blocks bumping at S_1.
2. If the edge e is either a reverse edge or a bad forward edge, then we say that this edge blocks bumping at S_1.
3. If the edge e is going in U direction and is a good forward edge, then there is an unique edge $f \in B$ that is obtained by moving e and S_1S_2 perpendicular to their respective directions. This contradicts the definition of $\mathcal{F}$.
4. If the edge is going in R direction and is a good forward edge, $S_3S_1S_2$ are consecutive in A. Let j be such that $A_j = S_3, A_{j+1} = S_1$ and $A_{j+2} = S_2$. Since these are good forward edges, there is i such that $\mathcal{F}(i) = j$. Since $\mathcal{F}$ is strictly monotonic, $(i+1) = j+1$. Therefore, B_{i+1} is a corner point in B. In this case, we say that the corner point B_{i+1} is blocking the bumping at S_1.

Now, suppose S_4 is contained in A. Look at the edge e going into S_4 in A. We again that 4 cases:

1. If there is no such edge, then $S_4 = O$. In this case, we say that O is blocking bumping at S_1.
2. If the edge e is either a reverse edge or a bad forward edge, then we say that this edge is blocking the bump at S_1.
3. If the edge e is going in U direction and is a good forward edge, then it is exactly the same edge as the one considered in case (3) above.
4. If the edge e is going in R direction, then both e and S_1S_2 end at S_2, which cannot happen as A is a path.

Each reverse forward edge or backward edge can block at most 4 good forward edges from bumping, two in each direction, one where it is blocking S_3 and one where it is blocking S_4. On the other hand, each corner including O and P can block at most one edge. Therefore, there are at least $n - c - 8k$ good forward edges which can be bumped, completing the proof. $\square$

In order to finish the proof of Lemma 50, we need a bound on number of paths A of length $n + 2k$ such that the base path $B = \mathcal{B}(A)$ has a large number of corners. We will do this by bounding the number of paths A such that $\mathcal{B}(A) = B$, and then using Lemma 48 to bound number of paths B with a large number of corners. We will give a rather trivial bound that suffices.

Lemma 59. *Given a shortest path B and $k \leq 0.1n$, the number of paths A of length $n + 2k$ such that $\mathcal{B}(A) = B$ is at most*

$$2 \cdot 3^{2k} \binom{n+2k}{2k}.$$

Proof. First, we express B as a sequence of directions of length n. Now, from $n + 2k$ positions, we choose $2k$ positions, and fill up the rest with the sequence of directions used in B. For the remaining $2k$ places, we have at most 3 choices each since we canot leave in the direction we came from, unless we are picking the starting direction, in which case we might have 4 choices. This gives an upper bound of

$$3^{2k-1}\left(4\binom{n+2k-1}{2k-1} + 3\binom{n+2k-1}{2k}\right) = 3^{2k}\binom{n+2k}{2k} + 3^{2k-1}\binom{n+2k-1}{2k-1}$$

since $\binom{n+2k-1}{2k-1} = \frac{2k}{n+2k}\binom{n+2k}{2k} \leq 3\binom{n+2k}{2k}$ for $k \leq 0.1n$, we get the result. $\square$

Now we are in a position to finish the proof of Lemma 50.

Proof. Recall that by Lemma 49, there is $n_0 = n_0(\varepsilon)$ such that for all $n \geq n_0$,

$$P_k \geq \frac{1}{2} P_0 \frac{(0.49)^k n^k}{k!} \exp\left(-\frac{25k^2}{n}\right)$$

On the other hand, for any given ε_1, we have that the number of paths A such that the base path $B = \mathcal{B}(A)$ has at least $\frac{n}{2} + 2\log n + \sqrt{n \log(1/\varepsilon_1)}$ corners is upper bounded by

$$2\varepsilon_1 P_0 3^{2k} \binom{n+2k}{2k} \leq 2\varepsilon_1 P_0 \frac{3^{2k} n^{2k}}{(2k)!} \exp\left(\frac{8k^2 - 4k^2 + 2k}{n}\right) \leq 2\varepsilon_1 P_0 \frac{3^{2k} n^{2k}}{(2k)!} \exp\left(\frac{5k^2}{n}\right) = T$$

Hence, if we choose ε_1 such that

$$\varepsilon_1 \leq \frac{\varepsilon}{4} \cdot \frac{(0.49)^k (2k)!}{3^{2k} n^k k!} \exp\left(\frac{-30k^2}{n}\right)$$

or equivalently, if

$$\log(1/\varepsilon_1) \geq \log(1/\varepsilon) + \log 4 + k \log n + 4k \log 3 - k \log k + \frac{30k^2}{n}$$

It follows that there are at most εP_k paths A of length $n + 2k$ such that B has at most

$$\frac{n}{2} + 2\log n + \sqrt{n \log(1/\varepsilon) + 2nk \log n + 30k^2}$$

corners, when $k \leq 0.1n$ and $n \geq 81$. Therefore, in this setting, every path A has at least $t - 8k$ good edges which can be bumped where

$$t = \frac{n}{2} - 2\log n - \sqrt{n \log(1/\varepsilon) + 2kn \log n + 30k^2}$$

Note that every edge that is bumped can prevent at most 3 new edges from being bumped. For example, if we bump and edge that looks like $(x_0, y_0) \to (x_0, y_0 + 1)$ it can stop the edges $(x_0, y_0 - 1) \to (x_0, y_0)$, $(x_0, y_0 + 1) \to (x_0, y_0 + 2)$ and $(x_0 - 2, y_0 + 2) \to (x_0 - 1, y_0 + 2)$ from bumping, which it initially did not. Therefore, we can choose set M of l edges which can be bumped simultaneously in

$$\frac{t(t-4)(t-8)\cdots(t-4(l-1))}{l!} \geq \frac{(t-3l)\cdots(t-4l+1)}{l!} = \binom{t-3l}{l}$$

many ways. Further, each path of length $n + 2k + 2l$ can have $k + l$ bumps, and can potentially be obtained in $\binom{k+l}{k}$ many different paths of length l. This gives us the lower bound

$$P_{k+l} \geq (1-\varepsilon) P_k \binom{t - 8k - 3l}{l} \binom{k+l}{k}^{-1}$$

as required. Note that for $k \leq \frac{n}{(\log n)^2}$, there exists $n(\varepsilon)$ such that for all $n \geq n(\varepsilon)$, $t \geq 0.49n$. Using Equation (D.6), we get the simplified lower bound:

$$\begin{aligned}
P_{k+l} &\geq (1-\varepsilon) P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-\frac{2(8k+3l)l - l^2 + l}{0.49n} - \frac{2l(8k+3l)}{0.49n}\right) \\
&\geq (1-\varepsilon) P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-\frac{32(kl+l^2)}{0.49n}\right) \\
&\geq (1-\varepsilon) P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-\frac{70(kl+l^2)}{n}\right) \\
&\geq (1-\varepsilon) P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-O\left(\frac{(kl+l^2)}{n}\right)\right)
\end{aligned}$$

□

4.3 Number of Low Girth Walks in the Grid

In this section, we will use the bounds obtained in the section above to compare the number of paths from $O = (0,0)$ to $P = (n_1, n_2)$ to the number of walks from O to P that do not have cycles of length less than $2l$. For the sake of notation, let W_k^l denote the number of walks from O to P that do not have cycles of length less than $2l$. Then we have the following:

Theorem 60. *Given constants $C, \delta, \alpha \geq 0$, there exists $n(C, \delta, \alpha)$, such that for all $n \geq n(C, \delta, \alpha)$, and for all k, l such that $k \leq Cn^{1-\delta}$ and $l\delta > 1 + 2\alpha$,*

$$P_k \leq W_k^l \leq \left(1 + 16n^{-\alpha}\right) P_k. \tag{4.4}$$

Proof. We will show this by induction on k. Note that result holds for $0 \leq k < l$ since in this setting, $W_k^l = P_k$. Suppose by induction hypothesis, $W_{\bar{k}}^l \leq (1 + 8n^{-\alpha}) P_{\bar{k}}$ for $0 \leq \bar{k} < k$. Since every walk of length $n + 2k$ with no cycles of length smaller than $2l$ is either or path or can be decomposed into a cycle of length $t \geq 2\ell$ and a walk of length $n + 2k - t$ with no cycles of length smaller than $2l$, we get the following bound:

$$W_k^l \leq P_k + \sum_{t=0}^{k-l} W_t^l 16^{k-t} (n + 2t) \leq P_k \sum_{t=0}^{k-l} \left(1 + 8n^{-\alpha}\right) \cdot 2 \cdot P_t 16^{k-t} n.$$

Here 16^t is a simple upper bound on the number of cycles of length 16^t through a fixed point. Note that for $t \leq Cn^{1-\delta}$, $n + 2t \leq 2n$. Now, using Lemma 50 with $\varepsilon = 0.5$, we have

$$\begin{aligned}
\frac{P_t 16^{k-t} n}{P_k} &\leq 2 \cdot \frac{k!}{t!} \cdot \frac{16^{k-t} n}{n^{k-t} (0.49)^{k-t}} \exp\left(\frac{70(k-t)(k-t+t)}{n}\right) \\
&\leq 2 \exp\left((k-t)(\log k + \log 16 - \log n - \log(0.49)) + \frac{70(k-t)k}{n} + \log n\right) \\
&\leq 2 \exp\left((k-t)((1-\delta)\log n + \log C - \log n + \log 40) + \frac{70k(k-t)}{n} + \log n\right).
\end{aligned}$$

Let $k-t=l+r$, and let l be an integer constant such that $l\delta > 1$, then we can upper bound the summation as below:

$$\frac{W_k^l}{P_k} \le 1 + \sum_{r=0}^{k-l} \big(1+8n^{-\alpha}\big)\cdot 4 \cdot \exp\bigg((1-l\delta)\log n + C_1 l + -r\delta \log n + C_1 r + \frac{50k(l+r)}{n}\bigg)$$
$$\le 1 + 4\big(1+8n^{-\alpha}\big)\exp\bigg((1-l\delta)\log n + C_1 l + 50lCn^{-\delta}\bigg)\Bigg(\sum_{r=0}^{k-l}\exp\big(r\big(-\delta\log n + C_1 + 50Cn^{-\delta}\big)\big)\Bigg)$$
$$\le 1 + 4\big(1+8n^{-\alpha}\big)\exp\bigg(-\alpha\log n\bigg)\Bigg(\sum_{r=0}^{\infty}\exp(-rC_2\log n)\Bigg),$$

where these equations hold with constants $C_1 = \log 40C$ and $C_2 = \frac{\delta}{2}$ for $n \ge n_1(C,\delta)$. Simplifying, we get the upper bound:

$$\frac{W_k^l}{P_k} \le 1 + 4\big(1+8n^{-\alpha}\big)n^{-\alpha}\frac{1}{1-n^{-C_2}}$$
$$\le 1 + 16\big(n^{-\alpha}\big),$$

where the last inequality holds for $n \ge n_2(\alpha)$, so that $8n^{-\alpha}, n^{-C_2} \le 0.5$. Therefore, for $n \ge n(C,\delta,\alpha) = \max(n_1(C,\delta), n_2(\alpha))$, we get the result. □

4.4 Subgraphs of the Lattice

In this section, we do the same analysis for number of paths in induced subgraphs of the lattice $\mathbb{Z}^2$. To ensure that the sampling procedure works efficiently, we will prove the analogues of Lemmas 49 and 50 and theorem 60 where we restrict ourselves to paths bounded in some set $S \subseteq \mathbb{Z}^2$. First, let us setup some notation:

Notation. For this section, let $S \subseteq \mathbb{Z}^2$ be an induced subset of lattice. Let O, P be two points in S. Without loss of generality, we will assume that $O = (0,0)$ and $P = (n_1, n_2) \ge O$. Let $n = n_1 + n_2$ denote the length of shortest path from n_1 to n_2 in $\mathbb{Z}^2$. Let P_k denote the number of paths (*self avoiding walk*) from O to P of length $n+2k$ that are contained in S Let W_k^l denote the number of walks from O to P of length $n+2k$ that do not have cycles of length smaller than l and are contained in S.

Now, we make a few definitions which are helpful in the analysis

Definition 61. Given set $S \subseteq \mathbb{Z}^2$, we define the boundary of S, denoted by ∂S as the set of points $Q \in S$ such that at least on neighbor of Q is outside S.

Definition 62. Given an induced subgraph $S \subseteq \mathbb{Z}^2$ and points $O, P \in S$, we say that S is (k, s, β)-wide if at least $(1-\beta)$ fraction of paths of length $n+2k$ from O to P contained in S intersect the boundary ∂S of S in at most s points.

To give some trivial examples, every set S is $(k, s, 1)$ wide for all k, s and on the other hand, every set S is $(k, n+2k, \beta)$-wide for all k, β. We are now ready to state and prove variants of Lemmas 49 and 50 that hold for bounded subgraphs of the lattice $\mathbb{Z}^2$.

Lemma 63. *Given an induced subgraph $S \subseteq \mathbb{Z}^2$ and points O, P in S such that S is $(0, s, \beta)$-wide, and numbers $k \in \mathbb{Z}$ and $\varepsilon \in \mathbb{R}$, $\varepsilon, k > 0$, we have the lower bound on number of paths from O to P contained in S:*

$$P_k \geq (1 - \varepsilon - \beta)\binom{t - 2s - 2k}{k}$$

where $t = \frac{n}{2} - 2\log n - \sqrt{n \log(1/\varepsilon)}$. Further, there is $n_0 = n_0(\varepsilon)$ such that for all $n \geq n_0$,

$$P_k \geq (1 - \varepsilon - \beta)P_0 \frac{(0.49)^k n^k}{k!} \exp\left(-O\left(\frac{k(k+s)}{n}\right)\right) \tag{4.5}$$

Proof. The proof is almost the same as Lemma 49, except one major change, we need to ensure that the constructed paths $\mathcal{A}(B, M)$ using Definition 46 stays inside set S. We can bump a path B at index i if the point B_i and $B_i + 1$ are not on the boundary ∂S. Further, there are at least $(1 - \varepsilon - \beta)P_0$ shortest paths that have at most $\frac{n}{2} + 2\log n + \sqrt{n \log(1/\varepsilon)}$ corners and at most s points that are on the boundary. For these paths, there are at least $\frac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon)} - 2s$ indices which can be bumped while keeping the path inside set S. Using Equation (4.2), we get the lower bound:

$$P_k \geq (1 - \varepsilon - \beta)\binom{t - 2s - 2k}{k}$$

for

$$t = \tfrac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon)}.$$

Since $t = \frac{n}{2} - o(n)$ there is $n_0 = n_0(\varepsilon)$ such that for all $n \geq n_0$, $t \geq 0.49n$. This gives us the lower bound, due to computation similar to Lemma 49.

$$\begin{aligned}
P_k &\geq (1 - \varepsilon - \beta)P_0 \frac{(0.49)^k n^k}{k!} \exp\left(\frac{-2(2s+2k)k - k^2 + k}{0.49n} - \frac{2k(2k+k+2s)}{0.49n}\right) \\
&\geq (1 - \varepsilon)P_0 \frac{(0.49)^k n^k}{k!} \exp\left(-\frac{25k(k+s)}{n}\right) \\
\implies P_k &\geq (1 - \varepsilon)P_0 \frac{(0.49)^k n^k}{k!} \exp\left(-O\left(\frac{k(k+s)}{n}\right)\right)
\end{aligned}$$

completing the proof of the lemma. □

Lemma 64. *Given an induced subgraph $S \subseteq \mathbb{Z}^2$ and points O, P in S such that S is (k, s, β)-wide and $(0, s, \beta)$-wide, and numbers $k \in \mathbb{Z}$ and $\varepsilon \in \mathbb{R}$, $\varepsilon, k > 0$, then there is $n_0 = n_0(\varepsilon)$ such that we have the lower bound on number of paths from O to P contained in S for $n \geq n_0(\varepsilon)$:*

$$P_{k+l} \geq (1 - \varepsilon - \beta)\binom{t - 2s - 8k - 3l}{l}$$

where $t = \frac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon) + 2kn\log n + 30k^2}$. *Further, if* $k, s \leq \frac{n}{(\log n)^2}$, *there is* $n_1 = n_1(\varepsilon)$ *such that for all* $n \geq n_1$,

$$P_{k+1} \geq (1-\varepsilon-\beta)P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-O\left(\frac{l(k+s+l)}{n}\right)\right) \tag{4.6}$$

Proof. The proof of this lemma is similar to Lemma 50, and we will only mention the key differenecs. First, observe that if $B = \mathcal{B}(A)$ has c corners, then there are at least $n - c - 8k$ indices in A that can be bumped. Among these, there are at most $2s$ indices where the points A_i or A_{i+1} are on boundary. Further, choice of ε_1 in the proof of Lemma 50 changes to satsify

$$\log(1/\varepsilon_1) \geq \log(1/\varepsilon) + \log 4 + k\log n + 4k\log 3 - k\log k + \frac{30k(k+s)}{n}$$

Therefore, there are at most εP_k paths A of length $n+2k$ such that B has at most

$$\tfrac{n}{2} + 2\log n + \sqrt{n\log(1/\varepsilon) + 2nk\log n + 30k(k+s)}$$

corners, there are at most βP_k paths A of length $n+2k$ that may have more that s points on the boundary ∂S. This gives us that at least $(1-\varepsilon-\beta)P_k$ paths of length $n+2k$ can be bumped at $t - 2s - 8k$ positions for

$$t = \tfrac{n}{2} - 2\log n - \sqrt{n\log(1/\varepsilon) + 2nk\log n + 30k(k+s)}$$

For $k, s \leq \frac{n}{(\log n)^2}$, $t = \frac{n}{2} - o(n)$, implying that there is $n_1 = n_1(\varepsilon)$ such that $t \geq 0.49n$. Using Equation (D.6) and computations similar to Lemma 50, we get the lower bound:

$$\begin{aligned} P_{k+l} &\geq (1-\varepsilon-\beta)P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-\frac{2(8k+3l+2s)l - l^2 + l}{0.49n} - \frac{2l(8k+3l+2s)}{0.49n}\right) \\ &\geq (1-\varepsilon-\beta)P_k \frac{(0.49)^l n^l k!}{(k+l)!} \exp\left(-\frac{70l(k+l+s)}{n}\right) \end{aligned}$$

This gives us the proposed bound, finishing the proof. □

Next step is to prove that variant of Theorem 60 holds for induced subgraph S of the lattice provided that the set S is satisfies certain properties.

Theorem 65. *Given constants* $C, \delta, \alpha \geq 0$, *a subgraph* $S \subseteq \mathbb{Z}^2$, *and a function* $s = s(k)$ *such that* S *is* $(k, s(k), \beta)$*-wide where* $\beta \leq 0.25$ *and* $s(k) \leq Cn^{(1-\delta)}$ *for all* $k \leq Cn^{(1-\delta)}$, *there exists* $n_0 = n_0(C, \delta, \alpha)$ *such that for all* $n \geq n_0$, $k \leq Cn^{(1-\delta)}$ *and* $l \geq 0$ *such that* $l\delta > 1 + 2\alpha$,

$$P_k \leq W_k^l \leq (1 + 32n^{-\alpha})P_k \tag{4.7}$$

Proof. The proof is similar to the proof of Theorem 60. The recursive bound still holds, that is,

$$W_k^l \leq P_k + \sum_{t=0}^{k-l} W_t^l 16^{k-t}(n+2t) \leq P_k \sum_{t=0}^{k-l} \left(1 + 8n^{-\alpha}\right) \cdot 2 \cdot P_t 16^{k-t} n$$

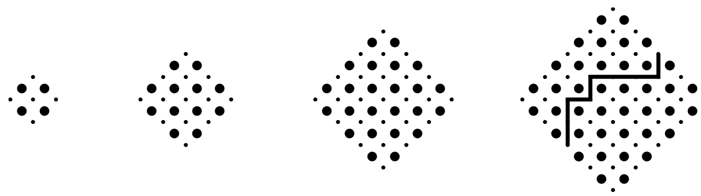


Figure 4.5: The large dots show the vertex sets of the the Aztec diamonds A_1, A_2, A_3, and A_4, which are subsets of the dual lattice $\mathbf{Z}'$. The small dots show the vertex sets of the corresponding A'_1, A'_2, A'_3 and A'_4, which are subsets of the integer lattice $\mathbf{Z} = \mathbb{Z}^2$. In the last case, a path $\mathcal{P}_\omega$ in A'_4 corresponding to a partition ω of A_4 is shown.

since we are restricting all the paths and walks to be restricted to set S. Using Lemma 64 with $\varepsilon = 0.5$, we get

$$\begin{aligned}\frac{P_t 16^{k-t} n}{P_k} &\leq 4 \cdot \frac{k!}{t!} \cdot \frac{16^{k-t} n}{n^{k-t}(0.49)^{k-t}} \exp\left(\frac{70(k-t)(k-t+t+s)}{n}\right)\\ &\leq 4 \exp\left((k-t)((1-\delta)\log n + \log C - \log n + \log 40) + \frac{70(k+s)(k-t)}{n} + \log n\right),\end{aligned}$$

which follows from computations in Theorem 60. The last expression holds for $C_1 = \log 40C$ and $C_2 = \frac{\delta}{2}$ for $n \geq n_1(C,\delta)$. Following the steps in Theorem 60 to evaluate the summation, we get the upper bound

$$W_k^l \leq 1 + 8\big(1 + 32n^{-\alpha}\big)n^{-\alpha}\frac{1}{1 - n^{-C_2}} \leq 1 + 32n^{-\alpha},$$

where the last inequality holds for $n \geq n_2(\alpha)$ chosen such that $16n^{-\alpha}, n^{-C_2} \leq 0.5$. Therefore, for $n \geq n(C,\delta,\alpha) = \max(n_1(C,\delta), n_2(\alpha))$, we get the result. □

4.5 The Aztec Diamond

We let $\mathbf{Z}$ denote the planar graph of the integer lattice $\mathbb{Z}^2$ and let $\mathbf{Z}'$ be its planar dual, with vertices using half-integer coordinates.

We define the Aztec Diamond graph A_k to be the subgraph of $\mathbf{Z}'$ induced by the set

$$V(A_k) = \{(x,y) \in \mathbb{Z}^2 + (\tfrac{1}{2}, \tfrac{1}{2}) \mid |x| + |y| \leq k\}, \tag{4.8}$$

and define A'_k to be the subgraph of $\mathbf{Z}$ induced by the set

$$V(A'_k) = \{(x,y) \in \mathbb{Z}^2 \mid |x| + |y| \leq k\}; \tag{4.9}$$

see Figure 4.5. We define the boundary $\partial A'_k$ to be those vertices of A'_k (x,y) with $|x| + |y| = k$.

We consider as a toy example the problem of randomly dividing the Aztec diamond into two contiguous pieces S_1, S_2, whose boundaries are both nearly as small as possible. Here we use the *edge-boundary* of S_i, which is the number of edges between S_i and $\mathbf{Z}' \setminus S_i$. Note that this is the same has the length of the closed walk in $\mathbf{Z}$ enclosing S_i. We collect the following simple observations about these sets and their boundaries:

Observation 66. A_k has $8k$ boundary edges. □

Observation 67. Every shortest path in A'_k between antipodal points on $\partial A'_k$ has length $2k$. □

Observation 68. For $x \geq 0$, the (unique) shortest path between points (x, y_1) and $(x, -y_1)$ of $\partial A'_k$ has length $2k - 2x$. □

In particular, there is no partition of A_k into two contiguous partition classes such that both have boundary size less than $6k$. With this motivation, we define $\Omega = \Omega_{C,\varepsilon,k}$ to be the partitions of A_k into two contiguous pieces, each with boundary sizes at most $6k + Ck^{1-\varepsilon}$, and consider the problem of uniform sampling from Ω. We will show that this problem can be solved in polynomial time with our approach, but also that Glauber dynamics on this state space has exponential mixing time. Observe that we can equivalently view Ω as set of paths in A'_k between points of $\partial A'_k$, and for any partition $\omega \in \Omega$ we write $\mathcal{P}_\omega$ for this corresponding path.

Writing $\omega \sim \omega'$ for $\omega, \omega' \in \Omega$ whenever (viewed as partitions) ω, ω' agree except on a single vertex of A_k, we define the Glauber dynamics for Ω to be the Markov chain which transitions from ω to a uniformly randomly chosen neighbor ω'. Recall that we define the *conductance* by

$$\Phi = \min_{\pi(S) \leq \frac{1}{2}} \frac{Q(S, \bar{S})}{\pi(S)} \tag{4.10}$$

where

$$Q(S, \bar{S}) = \sum_{\substack{\omega \in S \\ \omega' \in \bar{S}}} \pi(\omega) P(\omega, \omega') \leq \pi(\partial S),$$

where ∂S is the set of all $\omega \in S$ for which there exists an $\omega' \in \bar{S}$ for which $P(\omega, \omega') > 0$.

The mixing time t_{mix} of the Markov chain with transition matrix P is defined as the minimum t such that the total variation distance between vP^t and the stationary distribution π is $\leq \frac{1}{4}$, for all initial probability vectors v. With these definitions we have

$$t_{\text{mix}} \geq \frac{1}{4\Phi} \tag{4.11}$$

(e.g. see [LPW06], Chapter 7) and so to show the mixing time is exponentially large it suffices to show that the conductance Φ is exponentially small.

To this end, we define $S \subseteq \Omega$ to be the set of ω for which the endpoints (x_1, y_1) and (x_2, y_2) of $\mathcal{P}_\omega$ satisfy

$$x_1 \leq x_2 \quad y_1 \leq y_2. \tag{4.12}$$

Our goal is now to show that $|S|$ is large while $|\partial S|$ is small. For simplicity we consider the case where k is even but the odd case can be analyzed similarly.

To bound S from below it will suffice to consider just the partitions whose boundary path in A'_k is a shortest path from the point $(-\frac{k}{2}, -\frac{k}{2})$ to the point $(\frac{k}{2}, \frac{k}{2})$; note that such a path for the case where $k = 4$ is shown in Figure 4.5 There are $\binom{2k}{k}$ such paths and so we have lower bound

$$|S| \geq \binom{2k}{k} = \Omega\left(\frac{2^{2k}}{\sqrt{k}}\right). \tag{4.13}$$

To bound $|\partial S|$ from above We will make use of the following count of walks in the lattice:

Lemma 69. *For any point $P = (n_1, n_2)$ such that $n_1 + n_2 = n$, the number of walks from $O = (0, 0)$ to P of length $n + 2t$ is given by*

$$\binom{n+2t}{t}\binom{n+2t}{n_1+t}$$

Proof. Let W_t denote number of such walks. Note that any such path can be denoted as a sequence of symbols U, D, L, R which denote moves in the corresponding directions. For a direction $Z \in \{U, D, L, R\}$, let n_Z denote number of symbols signifying the direction that appear in the walk; then the walks from O to P are in bijection with the sequences over $\{U, D, L, R\}$ of length $n + 2t$ for which $n_U - n_D = n_1$ and $n_R - n_L = n_2$. Note then that $n_L + n_D = t$, $n_U + n_L = n_1 + t$, and $n_R + n_D = n_2 + t$. There is a bijection from the set of these sequences s to pairs of subsets $(X_s, Y_s) \subseteq [n + 2t]$ where $|X_s| = t$ and $|Y_s| = n_1 + t$ as follows. Given such a sequence s, we can let X_s be the set of indices with symbols L or D, while Y_s is the set of indices with symbols U or L. The sequence s is recovered from the sets X_s and Y_s by assigning the symbol U to indices in $X_s \setminus Y_s$, the symbol L to indices in $X_s \cap Y_s$, the symbol D to those in $Y_s \setminus X_s$, and the symbol R to indices in neither X_s nor Y_s. ☐

Now the boundary ∂S of S thus consists of paths which satisfy either $x_1 = x_2$ or $y_1 = y_2$. Observation 68, together with the condition that the total length of a closed walk enclosing each partition class is at most $6k + O(k^{1-\varepsilon})$, implies that in these cases, we must have $|y_i| = O(k^{1-\varepsilon})$ in the case where $x_1 = x_2$ or $|x_i| = O(k^{1-\varepsilon})$ in the case where $y_1 = y_2$. In particular, we have without loss of generality that $x_1 = x_2$, and $y_2 = y_1 + 2k - O(k^{1-\varepsilon})$. In particular, letting $\ell_\omega = y_2 - y_1$, we have that the path $\mathcal{P}_\omega$ has length $\ell_\omega + O(\ell_\omega^{1-\varepsilon})$. Now by Lemma 69, the number of choices for such walks (for fixed x_i, y_i, for which there are only polynomially many choices) is

$$\binom{\ell_\omega + O(\ell_\omega^{1-\varepsilon})}{O(\ell_\omega^{1-\varepsilon})}^2 \leq 2^{O(\ell_\omega^{1-\varepsilon})} \tag{4.14}$$

for $0 \leq \varepsilon \leq 1$. Together, (4.14) and (4.13) imply that

$$\Phi = \frac{2^{O(\ell_\omega^{1-\varepsilon})}}{2^{2k}/\sqrt{k}} \lesssim \frac{1/4}{2^{\varepsilon k}},$$

and so the mixing time t_{mix} satisfies

$$t_{\text{mix}} \geq 2^{\varepsilon k}, \tag{4.15}$$

with respect to the fixed parameter $\varepsilon > 0$. This gives the following theorem:

Theorem 70 (Theorem 45 restated)**.** *Glauber Dynamics on contiguous* 2*-partitions of* A_k *with boundary of length at most* $6k + Ck^{(1-\epsilon)}$ *has exponential mixing time.*

On the other hand, we claim that we can sample the partitions $\omega \in \Omega$ efficiently using Algorithm 3, by applying it to each pair of points on the boundary A'_k, to generate the path P_ω. To show this, we will argue that the set A'_k has the correct width property with endpoints of P_ω. Formally,

Lemma 71. *Let* $\omega \in \Omega$ *be a partition of* A_k*. Let* P_ω *be corresponding path in* A'_k *with endpoints* P_1, P_2*. Then* A'_k *is* $(\ell, 16\ell + 4Ck^{(1-\epsilon)}, 0)$*-wide with respect to points* P_1, P_2 *for all* ℓ*.*

Proof. Let $P_i = (x_i, y_i)$ for $i = 1, 2$. Without loss of generality, let $(x_2, y_2) \geq (0, 0)$. Let $Q = (-x_2, -y_2)$ be the point anti-podal to P_2 in $\partial A'_k$. We will break the proof into three cases, based on which quadrant P_1 is in.

Suppose P_1 is in third quadrant. Then the distance between P_1 and P_2 is exactly $2k$. Therefore, P_2 is at most $Ck^{(1-\epsilon)}$ distance from Q. The lattice box $\mathcal{R}(P_1, P_2)$ has at most

$$2|x_1 + x_2| + 2|y_1 + y_2|$$

points in $\partial A'_k$. This is exactly the distance between P_1 and Q. Therefore, a shortest path from P_1 to P_2 can intersect $\partial A'_k$ at at most $2Ck^{(1-\epsilon)}$ many points. It follows that a path of length $2k + 2\ell$ is contained in $\mathcal{R}(P_1 - (\ell, \ell), P_2 + (\ell, \ell))$, which contains at most $16\ell + 2Ck^{(1-\epsilon)}$ points in $\partial A'_k$, implying that any path of length $2k + 2\ell$ can intersect $\partial A'_k$ in at most $16\ell + 4Ck^{(1-\epsilon)}$.

Suppose P_1 is in the second quadrant. Then the distance between P_1 and P_2 is

$$x_2 - x_1 + |y_2 - y_1| = x_2 - x_1 + \max(y_1, y_2) - \min(y_1, y_2) \geq 2k - 2\min(y_1, y_2)$$

Further, length of the lower boundary of A_k between P_1 and P_2 is at least $4k + y_1 + y_2$, and hence boundary of the lower partition is at least $6k + |y_2 - y_1|$, which implies that

$$|y_2 - y_1| \leq Ck^{(1-\epsilon)}$$

The lattice box $\mathcal{R}(P_1, P_2)$ contains at most $2|y_2 - y_1| + 4$ points on the boundary $\partial A'_k$. By similar argument to above, we can conclude that any path of length 2ℓ larger than the shortest path is contained in a slightly bigger lattice box, and can intersect the boundary $\partial A'_k$ in at most

$$2|y_2 - y_1| + 16\ell + 4 \leq 16\ell + 4Ck^{(1-\epsilon)}$$

points.

The case when P_1 is in the fourth quadrant is handled similarly to the case when P_1 is in the second quadrant. This proves that in all cases, the Aztec Diamond is $(\ell, 16\ell + 4Ck^{(1-\epsilon)}, 0)$-wide. □

This lemma implies that for $\ell \leq Ck^{(1-\epsilon)}$, and $s(l) = 20Ck^{(1-\epsilon)}$, the set A'_k satisfies the hypothesis of Theorem 65 for all points P_1, P_2 that are endpoints of P_ω for some $\omega \in \Omega$. Hence, for each pair of points $P_1, P_2 \in \partial A'_k$, we can compute $W^\lambda_\ell(P_1, P_2)$ for all $\ell \leq Ck^{(1-\epsilon)}$, where $\lambda\epsilon > 1$. This allows us to uniformly sample P_ω, for $\omega \in \Omega$, with rejection sampling, using the following algorithm:

Algorithm 4 Partition Sampling

1: Compute $\mathrm{DP}(Q, P, w, t)$ for all $Q \in A'_k$, $P \in \partial A'_k$, $w \in \Phi_\lambda$, $0 \leq t \leq 2k + Ck^{1-\varepsilon}$ using Algorithm 1
2: **while** P_ω is not a path **do**
3: Sample P_1, P_2, t proportional to $\mathrm{DP}(Q, P, \{O\}, t)$
4: Sample P_ω from P_1 to P_2 of length t using Algorithm 2
5: **end while**
6: **return** P_ω

Chapter 5

Pitfalls of using Gaussian as a noise distribution in NCE

Noise contrastive estimation (NCE), introduced in [GH10; GH12], is one of several popular approaches for learning probability density functions parameterized up to a constant of proportionality, i.e. $p(x) \propto \exp(E_\theta(x))$, for some parametric family $\{E_\theta\}_\theta$. A recent incarnation of this paradigm is, for example, energy-based models (EBMs), which have achieved near-state-of-the-art results on many image generation tasks [DM19; SE19]. The main idea in NCE is to set up a self-supervised learning (SSL) task, in which we train a classifier to distinguish between samples from the data distribution P_* and a known, easy-to-sample distribution Q, often called the "noise" or "contrast" distribution. It can be shown that for a large choice of losses for the classification problem, the optimal classifier model is a (simple) function of the density ratio p_*/q, so an estimate for p_* can be extracted from a good classifier. Moreover, this strategy can be implemented *while avoiding* calculation of the partition function, which is necessary when using maximum likelihood to learn p^*.

The noise distribution q is the most significant "hyperparameter" in NCE training, with both strong empirical [RXG20] and theoretical [Liu+21] evidence that a poor choice of q can result in poor algorithmic behavior. [CGH22] show that even the optimal q for finite number of samples can have an unexpected form (e.g., it is not equal to the true data distribution p_*). Since q needs to be a distribution that one can efficiently draw samples from, as well as write an expression for the probability density function, the choices are somewhat limited.

A particularly common way to pick q is as a Gaussian that matches the mean and covariance of the input data [GH12; RXG20]. Our main contribution in this paper is to formally show that such a choice can result in an objective that is statistically poorly behaved, even for relatively simple data distributions. We show that even if p^* is a *product distribution* and a member of a very simple *exponential family*, the Hessian of the NCE loss, when using a Gaussian noise distribution q with matching mean and covariance has exponentially small (in the ambient dimension) spectral norm. As a consequence, the optimization landscape around the optimum will be exponentially flat, making gradient-based optimization challenging. As the main result of the paper, we show the asymptotic sample efficiency of the NCE objective will be *exponentially bad* in the ambient dimension.

5.1 Overview of Results

Let P_* be a distribution in a parametric family $\{P_\theta\}_{\theta\in\Theta}$. We wish to estimate P_* via P_θ for some $\theta_* \in \Theta$ by solving a noise contrastive estimation task. To set up the task, we also need to choose a noise distribution Q, with the constraint that we can draw samples from it efficiently, and we can evaluate the probability density function efficiently. We will use p_θ, p_*, q to denote the probability density functions (pdfs) of P_θ, P_*, and Q. For a data distribution P_* and noise distribution Q, the NCE loss of a distribution P_θ is defined as follows:

Definition 72 (NCE Loss)**.** The NCE loss of P_θ w.r.t. data distribution P_* and noise Q is

$$L(P_\theta) = -\frac{1}{2}\mathbb{E}_{P_*} \log \frac{p_\theta}{p_\theta + q} - \frac{1}{2}\mathbb{E}_Q \log \frac{q}{p_\theta + q}. \tag{5.1}$$

Moreover, the empirical version of the NCE loss when given i.i.d. samples $(x_1, \dots, x_n) \sim P_*^n$ and $(y_1, \dots, y_n) \sim Q^n$ is given by

$$L^n(\theta) = \frac{1}{n}\sum_{i=1}^n -\frac{1}{2}\log \frac{p_\theta(x_i)}{p_\theta(x_i) + q(x_i)} + \frac{1}{n}\sum_{i=1}^n -\frac{1}{2}\log \frac{q(y_i)}{p_\theta(y_i) + q(y_i)}. \tag{5.2}$$

By a slight abuse of notation, we will use $L(\theta), L(p_\theta)$ and $L(P_\theta)$ interchangeably.

The NCE loss can be interpreted as the binary cross-entropy loss for the classification task of distinguishing the data samples from the noise samples. To avoid calculating the partition function, one considers it as an additional parameter, namely we consider an augmented vector of parameters $\tilde{\theta} = (\theta, c)$ and let $p_{\tilde{\theta}}(x) = \exp(E_\theta(x) - c)$. The crucial property of the NCE loss is that it has a unique minimizer:

Lemma 73 ([GH12])**.** *The NCE objective in Definition 72 is uniquely minimized at $\theta = \theta_*$ and $c = \log(\int_x \exp(E_{\theta^*}(x))dx)$ provided that the support of Q contains that of P_*.*

We will be focusing on the Hessian of the loss L, as the crucial object governing both the algorithmic and statistical difficulty of the resulting objective. We will show the following two main results:

Theorem 74 (Exponentially flat Hessian)**.** *For $d > 0$ large enough, there exists a distribution $P_* = P_{\theta_*}$ over $\mathbb{R}^d$ such that*

- $\mathbb{E}_{P_*}[x] = 0$ *and* $\mathbb{E}_{P_*}[xx^\top] = I_d$.
- P_* *is a product distribution, namely* $p_*(x_1, x_2, \dots, x_d) = \prod_{i=1}^d p^*(x_i)$.
- *The NCE loss when using $q = \mathcal{N}(0, I_d)$ as the noise distribution has the property that*

$$\|\nabla^2 L(\theta_*)\|_2 \le \exp\left(-\Omega(d)\right).$$

We remark the above example of a problematic distribution P^* is extremely simple. Namely, P^* is a product distribution, with 0 mean and identity covariance. It actually is also the case that

P^* is log-concave—which is typically thought of as an "easy" class of distributions to learn due to the fact that log-concave distributions are unimodal.

The fact that the Hessian is exponentially flat near the optimum means that gradient-descent based optimization without additional tricks (e.g., gradient normalization, second order methods like Newton's method) will fail. (See, e.g., Theorem 4.1 and 4.2 in Liu et al. [Liu+21].) For us, this will be merely an intermediate result. We will address a more fundamental issue of the sample complexity of NCE, which is independent of the optimization algorithm used. Namely, we will show that without a large number of samples, the best minimizer of the empirical NCE might not be close to the target distribution. Proving this will require the development of some technical machinery.

More precisely, we use the result above to show that the asymptotic statistical complexity, using the above choice of P^*, Q, is exponentially bad in the dimension. This substantially clarifies results in Gutmann and Hyvärinen [GH12], who provide an expression for the asymptotic statistical complexity in terms of P^*, Q (Theorem 3, Gutmann and Hyvärinen [GH12]), but from which it's very difficult to glean quantitatively how bad the dependence on dimension can be for a particular choice of P^*, Q. Unlike the landscape issues that [Liu+21] point out, the statistical issues are impossible to fix with a better optimization algorithm: they are fundamental limitations of the NCE loss.

Theorem 75 (Asymptotic Statistical Complexity). *Let $d > 0$ be sufficiently large and $Q = \mathcal{N}(0, I_d)$. Let $\hat{\theta}_n$ be the optimizer for the empirical NCE loss $L^n(\theta)$ with the data distribution P_* given by Theorem 74 above and noise distribution Q. Then, as $n \to \infty$, the mean-squared error satisfies*

$$\mathbb{E}\left[\left\|\hat{\theta}_n - \theta_*\right\|_2^2\right] = \frac{\exp(\Omega(d))}{n}.$$

5.2 Exponentially flat Hessian: Proof of Theorem 74

The proof of Theorem 74 consists of three ingredients. First, in Section 5.2.1, we will compute an algebraically convenient upper bound for the spectral norm of the Hessian of the loss (eq. (5.1)). We will restrict our attention to the case when $\{P_\theta\}$ belongs to an exponential family. The upper bound will be in terms of the total variation distance $\mathrm{TV}(P_*, Q)$ and the Fisher information matrix of the sufficient statistics at θ_*. Here, P_* denotes the true data distribution and Q denotes the noise distribution.

Then, in Section 5.2.2, we construct a distribution P^* for which the TV distance between P^* and Q is large. We do this by "tensorizing" a univariate distribution. Namely, we construct a univariate distribution with mean 0 and variance 1 that is at a constant TV distance from a standard univariate Gaussian. Then, we use the fact that the *Hellinger* distance tensorizes, along with the relationship between TV and Hellinger distance, to show that $TV(P^*, Q) \geq 1 - \delta^d$ for some constant $\delta < 1$. (See [Was20] for a detailed review of distance measures.) Section 5.2.3 bounds the Fisher information matrix term, completing all the components required to establish Theorem 74.

5.2.1 Bounding the Hessian in terms of TV distance

Suppose $\{P_\theta\}$ is an exponential family of distributions, that is $p_\theta(x) = \exp(\theta^\top T(x))$, where $T(x)$ is a known function. Then, a straightforward calculation (see e.g., Appendix A in [Liu+21]) shows that the gradient and the Hessian of the NCE loss (eq. (5.1)) with respect to θ have the following forms:

$$\nabla_\theta p_\theta(x) = p_\theta(x) \cdot T(x), \tag{5.3}$$

$$\nabla_\theta L(p_\theta) = \frac{1}{2}\int_x \frac{q}{p_\theta + q}(p_\theta - p_*)T(x)dx, \tag{5.4}$$

$$\nabla_\theta^2 L(p_\theta) = \frac{1}{2}\int_x \frac{(p_* + q)p_\theta q}{(p_\theta + q)^2} T(x)T(x)^\top dx. \tag{5.5}$$

For $\theta = \theta_*$ and $p_\theta = p_*$, we have

$$\nabla_\theta^2 L(p_{\theta_*}) = \frac{1}{2}\int_x \frac{p_* q}{p_* + q} T(x)T(x)^\top dx \preceq \frac{1}{2}\int_x \min(p_*, q) T(x)T(x)^\top dx \tag{5.6}$$

The second line holds since $\frac{p_* q}{p_* + q} = \min(p_*, q) \cdot \frac{\max(p_*, q)}{p_* + q} \leq \min(p_*, q)$. Applying the matrix version of the Cauchy-Schwarz inequality (Lemma 199, Section D.3) to eq. (5.6) with two parts $\frac{\min(p_*(x), q(x))}{\sqrt{p_*(x)}}$ and $T(x)T(x)^\top \sqrt{p_*(x)}$, we obtain

$$\begin{aligned}
\|\nabla_\theta^2 L(P_*)\|_2 \leq \left\|\nabla_\theta^2 L(P_*)\right\|_F &\leq \frac{1}{2}\left(\int_x \frac{\min(p_*, q)^2}{p_*}\right)^{\frac{1}{2}} \left(\int_x \left\|T(x)T(x)^\top\right\|_F^2 p_*(x)dx\right)^{\frac{1}{2}} \\
&\leq \frac{1}{2}\left(\int_x \min(p_*, q)dx\right)^{\frac{1}{2}} \left(\int_x \left\|T(x)T(x)^\top\right\|_F^2 p_*(x)dx\right)^{\frac{1}{2}} \\
\implies \left\|\nabla_\theta^2 L(P_*)\right\|_2 &\leq \frac{1}{2}\Big(1 - \mathrm{TV}(P_*, Q)\Big)^{\frac{1}{2}} \left(\int_x \left\|T(x)T(x)^\top\right\|_F^2 p_*(x)dx\right)^{\frac{1}{2}}.
\end{aligned} \tag{5.7}$$

We bound the two terms in the product above separately. The first term is small when P_* and Q are significantly different. The second term is an upper bound of the Frobenius norm of the Fisher matrix at P_*. We will construct P_* such that the first term dominates, giving us the upper bound required.

5.2.2 Constructing the hard distribution P_*

The hard distribution P_* over $\mathbb{R}^d$ will have the property that $\mathbb{E}_{P_*}[x] = 0$, $\mathbb{E}_{P_*}[xx^\top] = I_d$, but will still have large TV distance from the standard Gaussian $Q = \mathcal{N}(0, I_d)$. This distribution will simply be a product distribution—the following lemma formalizes our main trick of tensorization to construct a distribution having large TV distance with the Gaussian.

Lemma 76. *Let $d > 0$ be given. Let $Q = \mathcal{N}(0, I_d)$ be the standard Gaussian in $\mathbb{R}^d$. Then, for some $\delta < 1$, there exists a log-concave distribution P (also over $\mathbb{R}^d$) with mean 0 and covariance I_d satisfying* $\mathrm{TV}(P, Q) \geq 1 - \delta^d$.

Proof. Let $\hat{Q}$ denote the standard normal distribution over $\mathbb{R}$. Let $\hat{P}$ be any other distribution over $\mathbb{R}$ with mean 0 and variance 1 that satisfies $\rho(\hat{P}, \hat{Q}) = \delta < 1$, where $\rho(\hat{P}, \hat{Q}) = \int_x \sqrt{\hat{p}\hat{q}}\, dx$ is the Bhattacharya coefficient. Since ρ tensorizes [Was20], we have that $\rho(\hat{P}^d, \hat{Q}^d) = \rho(\hat{P}, \hat{Q})^d$ for any $d > 1$. We can then write the Hellinger distance between P, Q as

$$H^2(P, Q) := 1 - \int_x \sqrt{pq}\, dx = 2(1 - \rho(\hat{P}, \hat{Q})^d). \tag{5.8}$$

Further, we also know that

$$\frac{1}{2} H^2(\hat{P}^d, \hat{Q}^d) \leq \mathrm{TV}(\hat{P}^d, \hat{Q}^d) \implies 1 - \rho(\hat{P}, \hat{Q})^d \leq \mathrm{TV}(\hat{P}^d, \hat{Q}^d) \implies 1 - \delta^d \leq \mathrm{TV}(\hat{P}^d, \hat{Q}^d).$$

Setting $P = \hat{P}^d$ and noting that $\hat{Q}^d = Q = \mathcal{N}(0, I_d)$, we have $\mathrm{TV}(P, Q) \geq 1 - \delta^d$. Finally, if the chosen $\hat{P}$ is a log-concave distribution, then so is $\hat{P}^d$, since the product of log-concave distributions is log-concave, which completes the proof. □

We will now explicitly define the distribution P_* that we will work with for rest of the paper.

Definition 77. Consider the exponential family $\left\{p_\theta(x) = \exp\left(\theta^\top T(x)\right)\right\}_{\theta \in \mathbb{R}^{d+1}}$ given by the sufficient statistics $T(x) = (x_1^4, \ldots, x_d^4, 1)$. Let $P_* = \hat{P}^d$ where $\hat{P}$ is the distribution on $\mathbb{R}$ with density function $\hat{p}$ given by

$$\hat{p}(x) \propto \exp\left(-\frac{x^4}{\sigma^4}\right).$$

We will set the constant of proportionality C and σ appropriately to ensure that $\hat{P}$ has mean 0 and variance 1. Note that $P_* = P_{\theta_*}$ for $\theta_* = -\left(\frac{1}{\sigma^4}, \ldots, \frac{1}{\sigma^4}, \log C\right)$.

Since $\frac{d^2 \log \hat{p}}{dx^2} = -\frac{12x^2}{\sigma^4} \leq 0$, $\hat{p}$ is log-concave. Further, symmetry of $\hat{p}$ around the origin gives $\mathbb{E}[\hat{P}] = 0$, and the choice of σ ensures that $\mathrm{Var}[\hat{P}] = 1$. The normalizing constant C satisfies

$$C = \int_{-\infty}^{\infty} e^{-\frac{x^4}{\sigma^4}} dx = 2 \int_0^{\infty} e^{-\frac{x^4}{\sigma^4}} dx.$$

Substituting $t = \frac{x^4}{\sigma^4}$, $dt = \frac{4x^3}{\sigma^4} dx = \frac{4t^{3/4}}{\sigma} dx$ gives

$$C = \frac{\sigma}{2} \int_0^{\infty} t^{-3/4} e^{-t} dt = \frac{\sigma}{2} \Gamma\left(\frac{1}{4}\right) = 2\sigma \Gamma\left(\frac{5}{4}\right).$$

where $\Gamma(z)$ is the gamma function defined as $\Gamma(z) \int_0^{\infty} x^{z-1} e^{-x} dx$. The variance is given by

$$\mathrm{Var}\left[\hat{P}\right] = \frac{1}{C} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^4}{\sigma^4}} dx = \frac{2}{C} \int_0^{\infty} x^2 e^{-\frac{x^4}{\sigma^4}} dx.$$

The same substitution as above gives

$$\mathrm{Var}(\hat{P}) = \frac{1}{2C} \int_0^{\infty} t^{1/2} t^{-3/4} \sigma^3 e^{-t} dt = \frac{\sigma^3}{2C} \int_0^{\infty} t^{-1/4} e^{-t} dt = \frac{\sigma^3}{2C} \Gamma\left(\frac{3}{4}\right) = \frac{\sigma^2}{4} \frac{\Gamma(3/4)}{\Gamma(5/4)}.$$

Thus, setting $\sigma = \sqrt{\frac{4\Gamma(5/4)}{\Gamma(3/4)}}$ results in $\mathrm{Var}[\hat{P}] = 1$. Correspondingly, we have $C = \frac{4\Gamma(5/4)^{3/2}}{\sqrt{\Gamma(3/4)}}$.

For this choice of $\hat{P}$, the Bhattacharya coefficient $\rho(\hat{P}, \hat{Q})$ is given by:

$$\rho(\hat{P}, \hat{Q}) = \int_{-\infty}^{\infty} \sqrt{\hat{p}(x)\hat{q}(x)}dx = \frac{1}{\sqrt{C\sqrt{2\pi}}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{4} - \frac{x^4}{2\sigma^4}\right) dx \approx 0.9905 \leq 0.991 < 1.$$

Thus, in the proof of Lemma 76, we can use this choice of $\hat{P}$, and we have that for $\delta = 0.991$ and $P_* = \hat{P}^d$, $\mathrm{TV}(P_*, Q) \geq 1 - \delta^d$, as required.

5.2.3 Bounding the Fisher information matrix

In this subsection, we bound the second factor in eq. (5.7), which is an upper bound on the Frobenius norm of the Fisher information matrix at θ_*.

Lemma 78. *For some constant $M > 0$, we have*

$$\int_x \left\|T(x)T(x)^\top\right\|_F^2 p_*(x)dx \leq d^2 M, \tag{5.9}$$

Proof. Recall that $T(x) = (x_1^4, \ldots, x_d^4, 1)$. Then,

$$\left\|T(x)T(x)^\top\right\|_F^2 = \sum_i x_i^{16} + \sum_{i \neq j} x_i^8 x_j^8 + 2\sum_i x_i^4 + 1. \tag{5.10}$$

Therefore, by linearity of expectation, and using the fact that P_* is a product distribution,

$$\int_x \left\|T(x)T(x)^\top\right\|_F^2 p_*(x)dx = d \cdot \mathbb{E}_{\hat{P}}\left[x^{16}\right] + d(d-1) \cdot \left(\mathbb{E}_{\hat{P}}\left[x^8\right]\right)^2 + 2d \cdot \mathbb{E}_{\hat{P}}\left[x^4\right] + 1 \leq d^2 M,$$

for an appropriate choice of constant M. This constant exists since all the expectations above are bounded owing to the fact that the exponential density $\hat{p}$ dominates in the integrals. $\square$

5.2.4 Putting things together

For P_* defined as above, and $Q = \mathcal{N}(0, I_d)$, Lemma 76 ensures that $1 - \mathrm{TV}(P_*, Q) \leq \delta^d$, for $\delta = 0.991$. From Lemma 78, we have that

$$\int_x \left\|T(x)T(x)^T\right\|_F^2 p_*(x)dx \leq d^2 M.$$

Substituting these bounds in eq. (5.7), we get that

$$\left\|\nabla_\theta^2 L(P_*)\right\|_2 \leq \frac{1}{2}\delta^{d/2} d\sqrt{M} = \exp(-\Omega(d)).$$

By construction, p_* is a product distribution with $\mathbb{E}_{p_*}[x] = 0$ and $\mathbb{E}_{p_*}\left[xx^\top\right] = I_d$, which completes the proof of the theorem.

5.3 Proof of Theorem 75

We will bound the error of the optimizer $\hat{\theta}_n$ of the empirical NCE loss (eq. (5.2)) using the bias-variance decomposition of MSE. To do this, we will reason about the random variable $\sqrt{n}(\hat{\theta}_n - \theta_*)$; let Σ be its covariance matrix. Since $\hat{\theta}_n$ is an unbiased estimate of θ_*, the MSE decomposes as

$$\mathbb{E}\left[\left\|\hat{\theta}_n - \theta_*\right\|_2^2\right] = \frac{1}{n}\operatorname{Tr}(\Sigma). \tag{5.11}$$

The proof of Theorem 75 proceeds as follows. In Section 5.3.1, we show that the random variable $\sqrt{n}(\hat{\theta}_n - \theta_*)$ is asymptotically normal with mean 0 and covariance matrix Σ given by

$$\Sigma = \nabla_\theta^2 L(\theta_*)^{-1}\operatorname{Var}\left[\sqrt{n}\nabla_\theta L^n(\theta_*)\right]\nabla_\theta^2 L(\theta_*)^{-1}. \tag{5.12}$$

We prove that the Hessian $\nabla_\theta^2 L(\theta_*)$ is invertible in Section D.5, so that the above expression is well-defined. Since $\Sigma \succeq 0$ (it is a covariance matrix), to get a lower bound on $\operatorname{Tr}(\Sigma)$, it suffices to get a lower bound on the largest eigenvalue of Σ. Looking at the factors on the right hand side of eq. (5.12), we note first that Theorem 74 ensures an exponential lower bound on *all* eigenvalues of $\nabla_\theta^2 L(\theta_*)^{-1}$. The bulk of the proof towards lower bounding the largest eigenvalue of Σ consists of lower bounding $\operatorname{Var}\left[v^\top \cdot \sqrt{n}\nabla_\theta L^n(\theta_*)\right]$), the *directional* variance of $\sqrt{n}\nabla_\theta L^n(\theta_*)$ along a suitably chosen direction v in terms of $v^\top \nabla_\theta^2 L(\theta_*) v$. In Section 5.3.2 and Section 5.3.3, we use anti-concentration bounds to prove such variance lower bounds.

5.3.1 Gaussian limit of $\sqrt{n}(\hat{\theta}_n - \theta_*)$

To begin, we will show that $\sqrt{n}(\hat{\theta}_n - \theta_*)$ behaves as a Gaussian random variable as $n \to \infty$. Recall that the empirical NCE loss is given by eq. (5.2):

$$L^n(\theta) = \frac{1}{n}\sum_{i=1}^n -\frac{1}{2}\ln\frac{p_\theta(x_i)}{p_\theta(x_i) + q(x_i)} + \frac{1}{n}\sum_{i=1}^n -\frac{1}{2}\ln\frac{q(y_i)}{p_\theta(y_i) + q(y_i)},$$

where $x_i \sim P_*$ and $y_i \sim Q$ are i.i.d. Let $\hat{\theta}_n$ be the optimizer for L^n. Then, by the Taylor expansion of $\nabla_\theta L^n$ around θ_*, we have

$$\sqrt{n}\left(\hat{\theta}_n - \theta_*\right) = -\nabla_\theta^2 L^n(\theta_*)^{-1} \cdot \sqrt{n}\nabla_\theta L^n(\theta_*) - \sqrt{n} \cdot O\left(\left\|\hat{\theta}_n - \theta_*\right\|^2\right) \tag{5.13}$$

by [GH12], who also show in their Theorem 2 that $\hat{\theta}_n$ is a consistent estimator of θ_*; hence, as $n \to \infty$, $\left\|\hat{\theta}_n - \theta_*\right\|^2 \to 0$. Gutmann and Hyvärinen [GH12, Lemma 12] also assert[1] that the Hessian of the empirical NCE loss (eq. (5.2)) at θ_* converges in probability to the Hessian of the true NCE loss (definition 72) at θ_*, i.e., $\nabla_\theta^2 L^n(\theta_*)^{-1} \xrightarrow{P} \nabla_\theta^2 L(\theta_*)^{-1}$. On the other hand, by the Central Limit Theorem, $\sqrt{n}\nabla_\theta L^n(\theta_*)$ converges to a Gaussian with mean $\mathbb{E}[\sqrt{n}\nabla_\theta L^n(\theta_*)] =$

[1] Translating notation: $T_d = n$, $J_{T_d}(\theta) = -2L^n(\theta)$ and setting $\nu = 1$ gives $\mathcal{I}_\nu = 2\nabla^2 L(\theta_*)$ as in eq. (5.6).

$\sqrt{n}\nabla_\theta L(\theta^*) = 0$, and covariance $\mathrm{Var}[\sqrt{n}\nabla_\theta L^n(\theta_*)]$. With these considerations, we conclude that the random variable $\sqrt{n}(\hat{\theta}_n - \theta_*)$ in eq. (5.13) is asymptotically a Gaussian with mean 0 and covariance $\Sigma = \nabla^2_\theta L(\theta_*)^{-1}\mathrm{Var}[\sqrt{n}\nabla_\theta L^n(\theta_*)]\nabla^2_\theta L(\theta_*)^{-1}$, as defined in eq. (5.12).

Next, we introduce some quantities which will be useful in the subsequent calculations. As we already have a handle on the spectrum of $\nabla^2_\theta L(\theta_*)$ from Theorem 74, the main object of our focus in eq. (5.12) is the term $\mathrm{Var}[\sqrt{n}\nabla_\theta L^n(\theta_*)]$. In particular, since we are concerned with the directional variance of Σ, we will reason about $\mathrm{Var}\big[v^\top \cdot \sqrt{n}\nabla_\theta L^n(\theta_*)\big]$ for a fixed vector of ones, i.e., $v = 1^{d+1}$. This vector has the property that for all $x, v^\top T(x) \geq 1$, as all non-constant coordinates of T are non-negative, and the remaining coordinate is 1. Note that

$$\nabla_\theta L^n(\theta_*) = -\frac{1}{2n}\sum_{i=1}^n \frac{q(x_i)T(x_i)}{p_*(x_i)+q(x_i)} + \frac{1}{2n}\sum_{i=1}^n \frac{p_*(y_i)T(y_i)}{p_*(y_i)+q(y_i)}$$

where $x_i \sim P_*$ and $y_i \sim Q$. Writing out the variance term explicitly, we have

$$\mathrm{Var}\big[v^\top \cdot \sqrt{n}\nabla_\theta L^n(\theta_*)\big] = n\cdot\frac{1}{4n}\mathrm{Var}_{x\sim p_*}\left[\frac{q(x)\cdot v^\top T(x)}{p_*(x)+q(x)}\right] + n\cdot\frac{1}{4n}\mathrm{Var}_{y\sim q}\left[\frac{p_*(y)\cdot v^\top T(y)}{p_*(y)+q(y)}\right]$$

(using linearity and independence)

$$= \frac{1}{4}\mathrm{Var}_{x\sim p_*}\underbrace{\left[\frac{q(x)\cdot v^\top T(x)}{p_*(x)+q(x)}\right]}_{A(x)} + \frac{1}{4}\mathrm{Var}_{y\sim q}\underbrace{\left[\frac{p_*(y)\cdot v^\top T(y)}{p_*(y)+q(y)}\right]}_{B(y)}. \tag{5.14}$$

Define $A(x) = \frac{q(x)\cdot v^\top T(x)}{p_*(x)+q(x)} = \frac{R_1(x)}{1+R_1(x)}v^\top T(x)$ where $R_1(x) = \frac{q(x)}{p_*(x)}$ and $B(y) = \frac{p_*(y)\cdot v^\top T(y)}{p_*(y)+q(y)} = \frac{R_2(y)}{1+R_2(y)}v^\top T(y)$ where $R_2(y) = \frac{p_*(y)}{q(y)}$. To show that $\mathrm{Var}_{x\sim p_*}[A(x)]$ and $\mathrm{Var}_{y\sim q}[B(y)]$ are large, we will need anti-concentration bounds on $R_1(x)$ and $R_2(y)$.

5.3.2 Anti-concentration of $R_1(x), R_2(y)$

Next, we show that R_1 and R_2 satisfy (quantitative) anti-concentration. We show this by a relatively straightforward application of the Berry-Esseen Theorem, and the proof is given in Section D.4. Precisely, we show:

Lemma 79. *Let $d > 0$ be sufficiently large. Let $p = \hat{p}^d$ and $q = \hat{q}^d$ be any product distributions, and define $R(x) = \frac{q(x)}{p(x)}$. Suppose we have the following third moment bound: $\mathbb{E}_{x\sim\hat{p}}\left[\left(\log\frac{\hat{q}}{\hat{p}}\right)^3\right] < \infty$. Then, for any ϵ, there exist constants $\alpha = \alpha(\hat{p},\hat{q},\epsilon)$, $\mu = \mu(\hat{p},\hat{q},\epsilon) < 0$ such that*

$$\mathbb{P}_{x\sim p}\Big[R(x) \leq \exp\Big(\mu d - \alpha\sqrt{d}\Big)\Big] \geq \frac{1}{2} - \epsilon \text{ and } \mathbb{P}_{x\sim p}\Big[R(x) \geq \exp\Big(\mu d + \alpha\sqrt{d}\Big)\Big] \geq \frac{1}{2} - \epsilon.$$

Instantiating Lemma 79 for the pair (p_*, q) gives us the anti-concentration result for R_1, while instantiating it for the reversed pair (q, p_*) gives us the anti-concentration result for R_2. We can verify that the third moment condition holds in both instantiations, since in both the cases, $\log(\hat{q}/\hat{p})$ is a polynomial. Crucially, we will also utilize the fact that the constant μ is negative (as it equals $-\mathrm{KL}(\hat{p}||\hat{q})$). We are now ready to bound the variance of $A(x)$ and $B(y)$.

5.3.3 Bounding the variance of $A(x), B(y)$

Recall that $A(x) = \frac{R_1(x) \cdot v^\top T(x)}{1+R_1(x)}$ and $B(y) = \frac{R_2(y) \cdot v^\top T(y)}{1+R_2(y)}$. Let μ, α be the constants given by Lemma 79 for p_*, q, ϵ. Further, let $L_1 = \exp\left(\mu d - \alpha\sqrt{d}\right)$ and $L_2 = \exp\left(\mu d + \alpha\sqrt{d}\right)$. Since the mapping $x \mapsto \frac{x}{1+x}$ is monotonically increasing in x,

$$\mathbb{P}_{x\sim p_*}[R_1(x) \le L_1] = \mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)}{1+R_1(x)} \le \frac{L_1}{1+L_1}\right] \ge \frac{1}{2} - \epsilon \tag{5.15}$$

$$\mathbb{P}_{x\sim p_*}[R_1(x) \ge L_2] = \mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)}{1+R_1(x)} \ge \frac{L_2}{1+L_2}\right] \ge \frac{1}{2} - \epsilon. \tag{5.16}$$

Let T_{up} be such that

$$\mathbb{P}_{x\sim p_*}\Big[\|T(x)\| \le T_{\mathrm{up}}\Big] \ge \frac{7}{8} \qquad \text{and} \qquad \mathbb{P}_{x\sim q}\Big[\|T(x)\| \le T_{\mathrm{up}}\Big] \ge \frac{7}{8}. \tag{5.17}$$

In Section D.6, we show that some $T_{\mathrm{up}} = O(\sigma^2\sqrt{d})$ suffices for this to hold. Then, from eq. (5.15), we have

$$\begin{aligned}
&\mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)}{1+R_1(x)} \le \frac{L_1}{1+L_1}\right] \ge \frac{1}{2} - \epsilon \\
\implies\; &\mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)\cdot v^\top T(x)}{1+R_1(x)} \le \frac{L_1\sqrt{d+1}\|T(x)\|}{1+L_1}\right] \ge \frac{1}{2} - \epsilon && \text{(Cauchy-Schwarz)} \\
\implies\; &\mathbb{P}_{x\sim p_*}\left[\left(\frac{R_1(x)\cdot v^\top T(x)}{1+R_1(x)} \le \frac{L_1\sqrt{d+1}\|T(x)\|}{1+Li_1}\right) \wedge \left(\|T(x)\| \le T_{\mathrm{up}}\right)\right] \ge \frac{3}{8} - \epsilon \\
& && \text{(union bound with eq. (5.17))} \\
\implies\; &\mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)v^\top T(x)}{1+R_1(x)} \le \frac{\sqrt{d+1}L_1T_{\mathrm{up}}}{1+L_1}\right] \ge \frac{3}{8} - \epsilon \\
\implies\; &\mathbb{P}_{x\sim p_*}\left[A(x) \le \frac{\sqrt{d+1}L_1T_{\mathrm{up}}}{1+L_1}\right] \ge \frac{1}{4},
\end{aligned}$$

for $\epsilon \le \frac{1}{8}$. On the other hand, recall also that v satisfies $v^\top T(x) \ge 1$ for all x. Therefore, we have

$$\begin{aligned}
&\mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)}{1+R_1(x)} \ge \frac{L_2}{1+L_2}\right] \ge \frac{1}{2} - \epsilon \\
\implies\; &\mathbb{P}_{x\sim p_*}\left[\frac{R_1(x)\cdot v^\top T(x)}{1+R_1(x)} \ge \frac{L_2}{1+L_2}\right] \ge \frac{1}{2} - \epsilon \implies \quad \mathbb{P}_{x\sim p_*}\left[A(x) \ge \frac{L_2}{1+L_2}\right] \ge \frac{1}{4}.
\end{aligned}$$

Now, consider the event $A_1 = \left\{A(x) \in \left[\frac{1}{2}\mathbb{E}_{x\sim p_*}[A(x)], \frac{3}{2}\mathbb{E}_{x\sim p_*}[A(x)]\right]\right\}$. If this event were to intersect both the events $A_2 = \left\{A(x) \le \frac{\sqrt{d+1}L_1T_{\mathrm{up}}}{1+L_1}\right\}$ and $A_3 = \left\{A(x) \ge \frac{L_2}{1+L_2}\right\}$, then we would have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{x\sim p_*}[A(x)] \le \frac{\sqrt{d+1}L_1T_{\mathrm{up}}}{1+L_1} \qquad \text{and} \qquad \frac{3}{2}\mathbb{E}_{x\sim p_*}[A(x)] \ge \frac{L_2}{1+L_2} \\
\implies\; &\frac{L_2}{L_1} \cdot \frac{1}{T_{\mathrm{up}}\sqrt{d+1}} \cdot \frac{L_1+1}{L_2+1} \le 3.
\end{aligned}$$

We will show that this cannot be the case. Recall that $\mu < 0$, which means that $L_2 = \exp(\mu d + \alpha\sqrt{d}) < 1$ for sufficiently large d. This means that for sufficiently large d we have:

$$\begin{aligned}
& \exp(\mu d + \alpha\sqrt{d}) < 1 \\
\implies \quad & \exp(\mu d + \alpha\sqrt{d}) - 2\exp(\mu d - \alpha\sqrt{d}) < 1 \\
\implies \quad & 1 + \exp(\mu d + \alpha\sqrt{d}) < 2 + 2\exp(\mu d - \alpha\sqrt{d}) \\
\implies \quad & \frac{1 + \exp(\mu d - \alpha\sqrt{d})}{1 + \exp(\mu d + \alpha\sqrt{d})} > \frac{1}{2} \\
\implies \quad & \frac{L_1 + 1}{L_2 + 1} > \frac{1}{2}.
\end{aligned}$$

Further, since $\frac{L_2}{L_1} = \exp(2\alpha\sqrt{d})$ and $T_{\text{up}} = O(\sigma^2\sqrt{d})$, we get that

$$\frac{L_2}{L_1} \cdot \frac{1}{T_{\text{up}}\sqrt{d+1}} \cdot \frac{L_1 + 1}{L_2 + 1} > \frac{\exp(2\alpha\sqrt{d})}{O(\sigma^2 d)} \cdot \frac{1}{2} > 3,$$

where the last inequality follows for large enough d since the numerator grows faster than the denominator. Hence for large enough d, A_1 cannot intersect both A_2 and A_3. If the event A_1 is disjoint from A_2, then

$$\begin{aligned}
& \mathbb{P}_{x\sim p_*}[A_1 \cup A_2] = \mathbb{P}_{x\sim p_*}[A_1] + \mathbb{P}_{x\sim p_*}[A_2] \leq 1 \\
\implies \quad & \mathbb{P}_{x\sim p_*}[A_1] \leq 1 - \mathbb{P}_{x\sim p_*}[A_2] \\
\implies \quad & \mathbb{P}_{x\sim p_*}\left[A(x) \in \left[\frac{1}{2}\mathbb{E}_{x\sim p_*}[A(x)], \frac{3}{2}\mathbb{E}_{x\sim p_*}[A(x)]\right]\right] \leq \frac{3}{4} \\
\implies \quad & \mathbb{P}_{x\sim p_*}\left[|A - \mathbb{E}_{p_*}A| \geq \frac{1}{2}\mathbb{E}_{p_*}A\right] \geq \frac{1}{4}.
\end{aligned}$$

This finally lower-bounds the variance of A as

$$\mathrm{Var}_{p_*}[A] = \mathbb{E}\left[(A - \mathbb{E}_{p_*}A)^2\right] \geq \frac{1}{4}(\mathbb{E}_{p_*}A)^2 \cdot \mathbb{P}\left[(A - \mathbb{E}_{p_*}A)^2 \geq \frac{1}{4}(\mathbb{E}_{p_*}A)^2\right] \geq \frac{1}{16}(\mathbb{E}_{p_*}A)^2.$$

and thus $\mathbb{E}_{p_*}(A^2) - (\mathbb{E}_{p_*}A)^2 = \mathrm{Var}_{p_*}[A] \geq \frac{1}{16}(\mathbb{E}_{p_*}A)^2$, so that $(\mathbb{E}_{p_*}A)^2 \leq \frac{16}{17}\mathbb{E}_{p_*}(A^2)$.

Altogether, we get $\mathrm{Var}_{p_*}[A] \geq \frac{1}{17}\mathbb{E}_{p_*}(A^2)$. An analogous argument in the case when A_1 is disjoint with A_3 yields the same bound on the variance. Using an identical argument for R_2 and B, we get that for large enough d, $\mathrm{Var}_q[B] \geq \frac{1}{17}\mathbb{E}_q(B^2)$.

5.3.4 Putting things together

Putting together the lower bounds $\mathrm{Var}_{p_*}[A] \geq \frac{1}{17}\mathbb{E}_{p_*}(A^2)$ and $\mathrm{Var}_q[B] \geq \frac{1}{17}\mathbb{E}_q(B^2)$ we showed in the previous subsection, and recalling eq. (5.14), we get

$$\begin{aligned}
\mathrm{Var}\left[v^\top \cdot \sqrt{n}\nabla_\theta L^n(\theta_*)\right] &= \frac{1}{4}\mathrm{Var}_{p_*}[A] + \frac{1}{4}\mathrm{Var}_{p_*}[B] \geq \frac{1}{68}\left(\mathbb{E}_{p_*}\left[A^2\right] + \mathbb{E}_q\left[B^2\right]\right) \\
&= \frac{1}{68}\left(\int_x \left(\frac{q(x)^2 p_*(x) + q(x)p_*(x)^2}{(p_*(x)+q(x))^2}\right) v^\top T(x)T(x)^\top v\, dx\right) \\
&= \frac{1}{68} v^\top \cdot \int_x \frac{p_*(x)q(x)}{p_*(x)+q(x)} T(x)T(x)^\top dx \cdot v = \frac{1}{34} v^\top \nabla_\theta^2 L(\theta_*) v \\
&\qquad \text{(from eq. (5.6))}.
\end{aligned}$$

Finally, since $\nabla_\theta^2 L(\theta_*)$ is invertible as claimed earlier (Lemma 200, Section D.5), let w be such that $v = \nabla_\theta^2 L(\theta_*)^{-1} w$. Then, recalling the expression for Σ in eq. (5.12), we can conclude that

$$\begin{aligned}
w^\top \Sigma w = v^\top \mathrm{Var}\left[\sqrt{n}\nabla_\theta L^n(\theta_*)\right] v &= \mathrm{Var}\left[v^\top \cdot \sqrt{n}\nabla_\theta L^n(\theta_*)\right] \\
&\geq \frac{1}{34} v^\top \nabla_\theta^2 L(\theta_*) v = \frac{1}{34} w^\top \nabla_\theta^2 L(\theta_*)^{-1} w, \qquad (5.18)
\end{aligned}$$

which gives us the desired bound on the MSE, namely

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\theta}_n - \theta_*\right\|_2^2\right] &\geq \frac{1}{n}\mathrm{Tr}(\Sigma) \geq \frac{1}{n}\sup_z \frac{z^\top \Sigma z}{\|z\|^2} \\
&\geq \frac{1}{n}\frac{w^\top \Sigma w}{\|w\|^2} \geq \frac{1}{34n}\frac{w^\top \nabla_\theta^2 L(\theta_*)^{-1} w}{\|w\|^2} \geq \frac{1}{34n}\inf_z \frac{z^\top \nabla_\theta^2 L(\theta_*)^{-1} z}{\|z\|^2} \geq \frac{\exp(\Omega(d))}{n},
\end{aligned}$$

where the last inequality follows from Theorem 74 and the fact that $\lambda_{\max}(\nabla_\theta^2 L(\theta_*))^{-1} = \lambda_{\min}(\nabla_\theta^2 L(\theta_*)^{-1})$. This concludes the proof of Theorem 75.

5.4 Simulations

We also verify our results with simulations. Precisely, we study the MSE for the empirical NCE loss as a function of the ambient dimension, and recover the dependence from Theorem 75. For dimension $d \in \{70, 72, \ldots, 120\}$, we generate $n = 500$ samples from the distribution P_* we construct in the theorem. We generate an equal number of samples from the noise distribution $Q = \mathcal{N}(0, I_d)$, and run gradient descent to minimize the empirical NCE loss to obtain $\hat{\theta}_n$. Since we explicitly know what θ_* is, we can compute the squared error $\|\hat{\theta}_n - \theta_*\|^2$. We run 100 trials of this, where we obtain fresh samples each time from P_* and Q, and average the squared errors over the trials to obtain an estimate of the MSE.

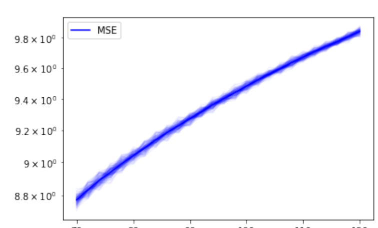


Figure 5.1: Log MSE versus Dimension—Theorem 75 suggests this plot should be linear, as is observed.

Figure 5.1 shows the plot of $\log \text{MSE}$ versus dimension - we can see that the graph is nearly linear. This corroborates the bound in Theorem 75, which tells us that as $n \to \infty$, the MSE scales exponentially with d. This behavior is robust even when the proportion of noise samples to true data samples is changed to 70:30 (though our theory only addresses the 50:50 case). Finally, we note that optimizing the empirical NCE loss becomes numerically unstable with increasing d (due to very large ratios in the loss), which is why we used comparatively moderate values of d.

5.5 Conclusion

Despite significant interest in alternatives to maximum likelihood—for example NCE (considered in this paper), score matching, etc.—there is little understanding of what there is to "sacrifice" with these losses, either algorithmically or statistically. In this paper, we provided formal lower bounds on the asymptotic sample complexity of NCE, when using a common choice for the noise distribution Q, a Gaussian with matching mean and covariance. Thus, it is likely that even for moderately complex distributions in practice, more involved techniques like Gao et al. [Gao+20] and Rhodes, Xu, and Gutmann [RXG20] will have to be used, in which one learns a noise distribution Q simultaneously with the NCE minimization or "anneals" the NCE objective. There is very little theoretical understanding of such techniques, and this seems like a very fruitful direction for future work.

Chapter 6

Provable Benefits of Score Matching

Energy-based models are a flexible class of probabilistic models with wide-ranging applications. They are parameterized by a class of energies $E_\theta(x)$ which in turn determines the distribution

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$$

up to a constant of proportionality Z_θ that is called the partition function. One of the major challenges of working with energy-based models is designing efficient algorithms for fitting them to data. Statistical theory tells us that the maximum likelihood estimator (MLE)—i.e., the parameters θ which maximize the likelihood—enjoys good statistical properties including consistency and asymptotic efficiency.

However, there is a major computational impediment to computing the MLE: Both evaluating the log-likelihood and computing its gradient with respect to θ (i.e., implementing zeroth and first order oracles, respectively) seem to require computing the partition function, which is often computationally intractable. More precisely, the gradient of the negative log-likelihood depends on $\nabla_\theta \log Z_\theta = \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(x)]$. A popular approach is to estimate this quantity by using a Markov chain to approximately sample from p_θ. However in high-dimensional settings, Markov chains often require many, sometimes even exponentially many, steps to mix.

Score matching [Hyv05] is a popular alternative that sidesteps needing to compute the partition function of sample from p_θ. The idea is to fit the score of the distribution, in the sense that we want θ such that $\nabla_x \log p(x)$ matches $\nabla_x \log p_\theta(x)$ for a typical sample from p. This approach turns out to have many nice properties. It is consistent in the sense that minimizing the objective function yields provably good estimates for the unknown parameters. Moreover, while the definition depends on the unknown $\nabla_x \log p(x)$, by applying integration by parts, it is possible to transform the objective into an equivalent one that can be estimated from samples.

The main question is to bound its statistical performance, especially relative to that of the maximum likelihood estimator. Recent work by [KHR22] showed that the cost can be quite steep. They gave explicit examples of distributions that have bad isoperimetric properties (i.e., large Poincaré constant) and showed how such properties can cause poor statistical performance.

Despite wide usage, there is little rigorous understanding of when score matching *helps*. This amounts to finding a general setting where maximizing the likelihood with standard first-order optimization is provably hard, and yet score matching is both computationally and statistically efficient, with only a polynomial loss in sample complexity relative to the MLE. In this work, we show the first such guarantees, and we do so for a natural class of exponential families defined by polynomials. As we discuss in Section 6.0.1, our results parallel recent developments in learning graphical models—where it is known that pseudolikelihood methods allow efficient learning of distributions that are hard to sample from—and can be viewed as a continuous analogue of such results.

In general, an exponential family on $\mathbb{R}^n$ has the form $p_\theta(x) \propto h(x)\exp(\langle \theta, T(x)\rangle)$ where $h(x)$ is the *base measure*, θ is the *parameter vector*, and $T(x)$ is the vector of *sufficient statistics*. Exponential families are one of the most classic parametric families of distributions, dating back to works by [Dar35], [Koo36] and [Pit36]. They have a number of natural properties, including: (1) The parameters θ are uniquely determined by the expectation of the sufficient statistics $\mathbb{E}_{p_\theta}[T]$; (2) The distribution p_θ is the maximum entropy distribution, subject to having given values for $\mathbb{E}_{p_\theta}[T]$; (3) They have conjugate priors [Bro86], which allow characterizations of the family for the posterior of the parameters given data.

For any (odd positive integer) constant d and norm bound $B \geq 1$, we study a natural exponential family $\mathcal{P}_{n,d,B}$ on $\mathbb{R}^n$ where

1. The *sufficient statistics* $T(x) \in \mathbb{R}^{M-1}$ consist of all monomials in $x_1, \ldots, x_n$ of degree at least 1 and at most d $\left(\text{where } M = \binom{n+d}{d}\right)$.

2. The *base measure* is defined as $h(x) = \exp(-\sum_{i=1}^n x_i^{d+1})$.[1]

3. The *parameters* θ lie in an l_∞-ball: $\theta \in \Theta_B = \{\theta \in \mathbb{R}^{M-1} : \|\theta\|_\infty \leq B\}$.

Towards stating our main results, we formally define the maximum likelihood and score matching objectives, denoting by $\hat{\mathbb{E}}$ the empirical average over the training samples drawn from some $p \in \mathcal{P}_{n,d,B}$:

$$\begin{aligned} L_{\mathrm{MLE}}(\theta) &= \hat{\mathbb{E}}_{x\sim p}[\log p_\theta(x)] \\ L_{\mathrm{SM}}(\theta) &= \frac{1}{2}\hat{\mathbb{E}}_{x\sim p}[\|\nabla \log p(x) - \nabla \log p_\theta(X)\|^2] + K_p \\ &= \hat{\mathbb{E}}_{x\sim p}\left[\mathrm{Tr}\,\nabla^2 \log p_\theta(x) + \frac{1}{2}\|\nabla \log p_\theta(x)\|^2\right] \end{aligned} \tag{6.1}$$

where K_p is a constant depending only on p and (6.1) follows by integration by parts [Hyv05]. In the special case of exponential families, (6.1) is a quadratic, and in fact the optimum can be written in closed form:

$$\operatorname*{argmin}_\theta L_{\mathrm{SM}}(\theta) = -\hat{\mathbb{E}}_{x\sim p}[(JT)_x(JT)_x^T]^{-1}\hat{\mathbb{E}}_{x\sim p}\Delta T(x) \tag{6.2}$$

where $(JT)_x : (M-1)\times n$ is the Jacobian of T at the point x, $\Delta f = \sum_i \partial_i^2 f$ is the Laplacian, applied coordinate wise to the vector-valued function f.

With this setting in place, we show the following intractability result.

[1]We note that the choice of base measure is for convenience in ensuring tail bounds necessary in our proof.

Theorem 80 (Informal, computational lower bound). *Unless $\mathsf{RP} = \mathsf{NP}$, there is no* $\mathrm{poly}(n, N)$*-time algorithm that evaluates $L_{\mathrm{MLE}}(\theta)$ and $\nabla L_{\mathrm{MLE}}(\theta)$ given $\theta \in \Theta_B$ and arbitrary samples $x_1, \ldots, x_N \in \mathbb{R}^n$, for $d = 7, B = \mathrm{poly}(n)$. Thus, optimizing the MLE loss using a zeroth-order or first-order method is computationally intractable.*

The main idea of the proof is to construct a polynomial $F_{\mathcal{C}}(x)$ which has roots exactly at the satisfying assignments of a given 3-SAT formula $\mathcal{C}$. We then argue that $\exp(-\gamma F_{\mathcal{C}}(x))$, for sufficiently large $\gamma > 0$, concentrates near the satisfying assignments. Finally, we show sampling from this distribution or approximating $\log Z_\theta$ or $\nabla_\theta \log Z_\theta$ (where $\theta \in \mathbb{R}^{M-1}$ is the parameter vector corresponding to the polynomial $-\gamma F_{\mathcal{C}}(x)$) would enable efficiently finding a satisfying assignment.

Our next result shows that MLE, though computationally intractable to compute via implementing zeroth or first order oracles, has (asymptotic) sample complexity $\mathrm{poly}(n, B)$ (for constant d).

Theorem 81 (Informal, efficiency of MLE). *The MLE estimator $\hat{\theta}_{\mathrm{MLE}} = \mathrm{argmax}_\theta L_{\mathrm{MLE}}(\theta)$ has asymptotic sample complexity polynomial in n. That is, for all sufficiently large N it holds with probability at least* 0.99 *(over N samples drawn from p_{θ^*}) that:*

$$\|\hat{\theta}_{\mathrm{MLE}} - \theta^*\|^2 \leq O\left(\frac{(nB)^{\mathrm{poly}(d)}}{N}\right).$$

The main proof technique for this is an anticoncentration bound of low-degree polynomials, for distributions in our exponential family.

Lastly, we prove that score matching *also* has polynomial (asymptotic) statistical complexity.

Theorem 82 (Informal, efficiency of SM). *The score matching estimator $\hat{\theta}_{\mathrm{SM}} = \mathrm{argmax}_\theta L_{\mathrm{SM}}(\theta)$ also has asymptotic sample complexity at most polynomial in n. That is, for all sufficiently large N it holds with probability at least* 0.99 *(over N samples drawn from p_{θ^*}) that:*

$$\|\hat{\theta}_{\mathrm{SM}} - \theta^*\|^2 \leq O\left(\frac{(nB)^{\mathrm{poly}(d)}}{N}\right). \tag{6.3}$$

The main ingredient in this result is a bound on the *restricted Poincaré constant*—namely, the Poincaré constant, when restricted to functions that are linear in the sufficient statistics T. We bound this quantity for the exponential family we consider in terms of the condition number of the Fisher matrix of the distribution, which we believe is a result of independent interest. With this tool in hand, we can use the framework of [KHR22], which relates the asymptotic sample complexity of score matching to the asymptotic sample complexity of maximum likelihood, in terms of the restricted Poincaré constant of the distribution.

6.0.1 Discussion and related work

Score matching: Score matching was proposed by [Hyv05], who also gave conditions under which it is consistent and asymptotically normal. Asymptotic normality is also proven for various kernelized variants of score matching in [Bar+19]. [KHR22] prove that the statistical sample complexity of score matching is not much worse than the sample complexity of maximum likelihood when the distribution satisfies a (restricted) Poincaré inequality. While we leverage machinery from [KHR22], their work only bounds the sample complexity of score matching by a quantity polynomial in the ambient dimension for a specific distribution in a specific bimodal exponential family. By contrast, we can handle an entire class of exponential families with low-degree sufficient statistics.

Poincaré vs Restricted Poincaré: We note that while Poincaré inequalities are directly related to isoperimetry and mixing of Markov chains, sample efficiency of score matching only depends on the Poincaré inequality holding for a *restricted* class of functions, namely, functions linear in the sufficient statistics. Hence, hardness of sampling only implies sample complexity lower bounds in cases where the family is expressive enough—indeed, the key to exponential lower bounds for score matching in [KHR22] is augmenting the sufficient statistics with a function defined by a bad cut. This gap means that we can hope to have good sample complexity for score matching even in cases where sampling is hard—which we take advantage of in this work.

Learning exponential families: Despite the fact that exponential families are both classical and ubiquitous, both in statistics and machine learning, there is relatively little understanding about the computational-statistical tradeoffs to learn them from data, that is, what sample complexity can be achieved with a computationally efficient algorithm. [Ren+21] consider a version of the "interaction screening" estimator, a close relative of pseudolikelihood, but do not prove anything about the statistical complexity of this estimator. [SSW21] consider a related estimator, and analyze it under various low-rank and sparsity assumptions of reshapings of the sufficient statistics into a tensor. Unfortunately, these assumptions are somewhat involved, and it's unclear if they are needed for designing computationally and statistically efficient algorithms.

Discrete exponential families (Ising models): Ising models have the form $p_J(x) \propto \exp\big(\sum_{i\sim j} J_{ij} x_i x_j + \sum_i J_i x_i\big)$ where $\sim$ denotes adjacency in some (unknown) graph, and J_{ij}, J_i denote the corresponding pairwise and singleton potentials. [Bre15] gave an efficient algorithm for learning any Ising model over a graph with constant degree (and l_∞-bounds on the coefficients); see also the more recent work [Dag+21]. In contrast, it is a classic result [AB09] that approximating the partition function of members in this family is NP-hard.

Similarly, the exponential family we consider is such that it contains members for which sampling and approximating their partition function is intractable (the main ingredient in the proof of Theorem 80). Nevertheless, by Theorem 6.3, we can learn the parameters for members in this family computationally efficiently, and with sample complexity comparable to the optimal one (achieved by maximum likelihood). This also parallels other developments in Ising models [BGS14; Mon15], where it is known that restricting the type of learning algorithm (e.g., requiring it to work with sufficient statistics only) can make a tractable problem become intractable.

The parallels can be drawn even on an algorithmic level: a follow up work to [Bre15] by [Vuf+16] showed that similar results can be shown in the Ising model setting by using the "screening estimator", a close relative of the classical pseudolikelihood estimator [Bes77] which tries to learn a distribution by matching the conditional probability of singletons, and thereby avoids having to evaluate a partition function. Since conditional probabilities of singletons capture changes in a single coordinate, they can be viewed as a kind of "discrete gradient"—a further analogy to score matching in the continuous setting.[2]

6.1 Preliminaries

We consider the following exponential family. Fix positive integers $n, d, B \in \mathbb{N}$ where d is odd. Let $h(x) = \exp(-\sum_{i=1}^n x_i^{d+1})$, and let $T(x) \in \mathbb{R}^{M-1}$ be the vector of monomials in $x_1, \ldots, x_n$ of degree at least 1 and at most d (so that $M = \binom{n+d}{d}$). Define $\Theta \subseteq \mathbb{R}^{M-1}$ by $\Theta = \{\theta \in \mathbb{R}^{M-1} : \|\theta\|_\infty \leq B\}$. For any $\theta \in \Theta$ define $p_\theta : \mathbb{R}^n \to [0, \infty)$ by

$$p_\theta(x) := \frac{h(x) \exp(\langle \theta, T(x) \rangle)}{Z_\theta}$$

where $Z_\theta = \int_{\mathbb{R}^n} h(x) \exp(\langle \theta, T(x) \rangle)\, dx$ is the normalizing constant. Then we consider the family $\mathcal{P}_{n,d,B} := (p_\theta)_{\theta \in \Theta_B}$. Throughout, we will assume that $B \geq 1$.

Polynomial notation: Let $\mathbb{R}[x_1, \ldots, x_n]_{\leq d}$ denote the space of polynomials in $x_1, \ldots, x_n$ of degree at most d. We can write any such polynomial f as $f(x) = \sum_{|\mathbf{d}| \leq d} a_{\mathbf{d}} x_{\mathbf{d}}$ where $\mathbf{d}$ denotes a degree function $\mathbf{d} : [n] \to \mathbb{N}$, and $|\mathbf{d}| = \sum_{i=1}^n \mathbf{d}(i)$, and we write $x_{\mathbf{d}}$ to denote $\prod_{i=1}^n x_i^{\mathbf{d}(i)}$. Note that every $\mathbf{d}$ with $1 \leq |\mathbf{d}| \leq d$ corresponds to an index of T, i.e. $T(x)_{\mathbf{d}} = x_{\mathbf{d}}$.

Let $\|\cdot\|_{\mathsf{mon}}$ denote the ℓ^2 norm of a polynomial in the monomial basis; that is, $\|\sum_{\mathbf{d}} a_{\mathbf{d}} x_{\mathbf{d}}\|_{\mathsf{mon}} = \left(\sum_{\mathbf{d}} a_{\mathbf{d}}^2\right)^{1/2}$. For any function $f : \mathbb{R}^n \to \mathbb{R}$, let $\|f\|_{L^2([-1,1]^n)}^2 = \mathbb{E}_{x \sim \mathrm{Unif}([-1,1]^n)} f(x)^2$.

Statistical efficiency of MLE: For any $\theta \in \mathbb{R}^{M-1}$, the Fisher information matrix of p_θ with respect to the sufficient statistics $T(x)$ is defined as

$$\mathcal{I}(\theta) := \mathbb{E}_{x \sim p_\theta}[T(x)T(x)^\top] - \mathbb{E}_{x \sim p_\theta}[T(x)]\mathbb{E}_{x \sim p_\theta}[T(x)]^\top.$$

It is well-known that for any exponential family with no affine dependencies among the sufficient statistics (see e.g., Theorem 4.6 in [Van00]), it holds that for any $\theta^* \in \mathbb{R}^{M-1}$, given N independent samples $x^{(1)}, \ldots, x^{(N)} \sim p_{\theta^*}$, the estimator $\hat{\theta}_{\mathrm{MLE}} = \hat{\theta}_{\mathrm{MLE}}(x^{(1)}, \ldots, x^{(N)})$ satisfies

$$\sqrt{N}\left(\hat{\theta}_{\mathrm{MLE}} - \theta^*\right) \to \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}).$$

Statistical efficiency of score matching: Our analysis of the statistical efficiency of score matching is based on a result due to [KHR22]. We state a requisite definition followed by the result.

[2] In fact, ratio matching, proposed in [Hyv07] as a discrete analogue of score matching, relies on exactly this intuition.

Definition 83 (Restricted Poincaré for exponential families)**.** The restricted Poincaré constant of $p \in \mathcal{P}_{n,d,B}$ is the smallest $C_P > 0$ such that for all $w \in \mathbb{R}^{M-1}$, it holds that

$$\mathrm{Var}_p(\langle w, T(x)\rangle) \le C_P \mathbb{E}_{x\sim p}\|\nabla_x \langle w, T(x)\rangle\|_2^2.$$

Theorem 84 ([KHR22])**.** *Under certain regularity conditions (see Lemma 180), for any p_{θ^*} with restricted Poincaré constant C_P and with $\lambda_{min}(\mathcal{I}(\theta^*)) > 0$, given N independent samples $x^{(1)}, \dots, x^{(N)} \sim p_{\theta^*}$, the estimator $\hat{\theta}_{\mathrm{SM}} = \hat{\theta}_{\mathrm{SM}}(x^{(1)}, \dots, x^{(N)})$ satisfies*

$$\sqrt{N}(\hat{\theta}_{\mathrm{SM}} - \theta^*) \to \mathcal{N}(0, \Gamma)$$

where Γ satisfies

$$\|\Gamma\|_{op} \le \frac{2C_P^2(\|\theta\|_2^2 \mathbb{E}_{x\sim p_{\theta^*}}\|(JT)(x)\|_{op}^4 + \mathbb{E}_{x\sim p_{\theta^*}}\|\Delta T(x)\|_2^2)}{\lambda_{min}(\mathcal{I}(\theta^*))^2}$$

where $(JT)(x)_i = \nabla_x T_i(x)$ and $\Delta T(x) = \mathrm{Tr}\,\nabla_x^2 T(x)$.

6.2 Hardness of Implementing Optimization Oracles for $\mathcal{P}_{n,7,\mathrm{poly}(n)}$

In this section we prove NP-hardness of implementing approximate zeroth-order and first-order optimization oracles for maximum likelihood in the exponential family $\mathcal{P}_{n,7,Cn^2\log(n)}$ (for a sufficiently large constant C) as defined in Section 6.1; we also show that approximate sampling from this family is NP-hard. See Theorems 89, 91, and 94 respectively. All of the hardness results proceed by reduction from 3-SAT and use the same construction.

The idea is that for any formula $\mathcal{C}$ on n variables, we can construct a non-negative polynomial $F_{\mathcal{C}}$ of degree at most 6 in variables $x_1, \dots, x_n$, which has roots exactly at the points of the hypercube $\mathcal{H} := \{-1, 1\}^n \subseteq \mathbb{R}^n$ that correspond to satisfying assignments (under the bijection that $x_i = 1$ corresponds to True and $x_i = -1$ corresponds to False). Intuitively, the distribution with density proportional to $\exp(-\gamma F_{\mathcal{C}}(x))$ will, for sufficiently large $\gamma > 0$, concentrate on the satisfying assignments. It is then straightforward to see that sampling from this distribution or efficiently computing either $\log Z_\theta$ or $\nabla_\theta \log Z_\theta$ (where $\theta \in \mathbb{R}^{M-1}$ is the parameter vector corresponding to the polynomial $-\gamma F_{\mathcal{C}}(x)$) would enable efficiently finding a satisfying assignment.

The remainder of this section makes the above intuition precise; important details include (1) incorporating the base measure $h(x) = \exp(-\sum_{i=1}^n x_i^8)$ into the density function, and (2) showing that a polynomially-large temperature γ suffices.

Definition 85 (Clause/formula polynomials)**.** Given a 3-clause formula of the form $C = \tilde{x}_i \vee \tilde{x}_j \vee \tilde{x}_k$ where $\tilde{x}_i = x_i$ or $\tilde{x}_i = \neg x_i$, we construct a polynomial $H_C \in \mathbb{R}[x_1, \dots, x_n]_{\le 6}$ defined by

$$H_C(x) = f_i(x_i)^2 f_j(x_j)^2 f_k(x_k)^2$$

where

$$f_i(t) = \begin{cases} (t+1) & \text{if } x_i \text{ is negated in } C \\ (t-1) & \text{otherwise} \end{cases}.$$

For example, if $C = x_1 \vee x_2 \vee \neg x_3$, then $H_C = (x_1-1)^2(x_2-1)^2(x_3+1)^2$. Further, given a 3-$\mathsf{SAT}$ formula $\mathcal{C} = C_1 \wedge \cdots \wedge C_m$ on m clauses[3], we define the polynomial

$$H_{\mathcal{C}}(x) = H_{C_1}(x) + \cdots + H_{C_m}(x).$$

It can be seen that any $x \in \mathcal{H}$ corresponds to a satisfying assignment for $\mathcal{C}$ if and only if $H_{\mathcal{C}}(x) = 0$. Note that there are possibly points outside $\mathcal{H}$ which satisfy $H_{\mathcal{C}}(x) = 0$. To avoid these solutions, we introduce another polynomial:

Definition 86 (Hypercube polynomial)**.** We define $G : \mathbb{R}^n \to \mathbb{R}$ by $G(x) = \sum_{i=1}^n (1-x_i^2)^2$.

Note that $G(x) \geq 0$ for all x, and the roots of $G(x)$ are precisely the vertices of $\mathcal{H}$. Therefore for any $\alpha, \beta > 0$, the roots (in $\mathbb{R}^n$) of the polynomial $F_{\mathcal{C}}(x) = \alpha H_{\mathcal{C}}(x) + \beta G(x)$ are precisely the vertices of $\mathcal{H}$ that correspond to satisfying assignments for $\mathcal{C}$.

Definition 87. Let $\mathcal{C}$ be a 3-CNF formula with n variables and m clauses. Let $\alpha, \beta > 0$. Then we define a distribution $P_{\mathcal{C},\alpha,\beta}$ with density function

$$p_{\mathcal{C},\alpha,\beta}(x) := \frac{h(x)\exp(-\alpha H_{\mathcal{C}}(x) - \beta G(x))}{Z_{\mathcal{C},\alpha,\beta}}$$

where $Z_{\mathcal{C},\alpha,\beta} = \int_{\mathbb{R}^n} h(x)\exp(-\alpha H_{\mathcal{C}}(x) - \beta G(x))\,dx$.

This distribution lies in the exponential family $\mathcal{P}_{n,d,B}$, for $d = 7$ and $B = \Omega(\beta + m\alpha)$ (Lemma 174). Thus, if $\theta(\mathcal{C}, \alpha, \beta)$ is the parameter vector that induces $P_{\mathcal{C},\alpha,\beta}$, then it suffices to show that (a) approximating $\log Z_{\theta(\mathcal{C},\alpha,\beta)}$, (b) approximating $\nabla_\theta \log Z_{\theta(\mathcal{C},\alpha,\beta)}$, and (c) sampling from $P_{\mathcal{C},\alpha,\beta}$ are NP-hard (under randomized reductions).

Additional notation. Given a point $v \in \mathcal{H}$, let $\mathcal{O}(v) := \{x \in \mathbb{R}^n : x_i v_i \geq 0; \forall i \in [n]\}$ denote the octant containing v, and let $\mathcal{B}_r(v) := \{x \in \mathbb{R}^n : \|x - v\|_\infty \leq r\}$ denote the ball of radius r with respect to ℓ_∞ norm.

6.2.1 Hardness of approximating $\log Z_{\mathcal{C},\alpha,\beta}$

In order to prove (a), we bound the mass of $P_{\mathcal{C},\alpha,\beta}$ in each orthant of $\mathbb{R}^n$. In particular, we show that for $\alpha = \Omega(n)$ and $\beta = \Omega(m \log m)$, any orthant corresponding to a satisfying assignment has exponentially larger contribution to $Z_{\mathcal{C},\alpha,\beta}$ than any orthant corresponding to an unsatisfying assignment (Lemma 175). A consequence is that the partition function $Z_{\mathcal{C},\alpha,\beta}$ is exponentially larger when the formula $\mathcal{C}$ is satisfiable than when it isn't:

[3] It suffices to work with $m = O(n)$, see Theorem 173.

Lemma 88. *Fix $n, m \in \mathbb{N}$ and let $\alpha \geq 2(n+1)$ and $\beta \geq 6480m \log(13n\sqrt{m})$. There is a constant $A = A(n, m, \alpha, \beta)$ so that the following hold for every 3-CNF formula $\mathcal{C}$ with n variables and m clauses:*

- *If $\mathcal{C}$ is unsatisfiable, then $Z_{\mathcal{C},\alpha,\beta} \leq A$*
- *If $\mathcal{C}$ is satisfiable, then $Z_{\mathcal{C},\alpha,\beta} \geq (2/e)^n A$.*

Proof. If $\mathcal{C}$ is unsatisfiable, then by the second part of Lemma 175, we have

$$Z = Z \sum_{w \in \mathcal{H}} \Pr_{x \sim p}(x \in \mathcal{O}(w)) \leq 2^n e^{-\alpha} \left(\int_0^\infty \exp(-x^{d+1} - \beta(1-x^2)^2)\, dx \right)^n =: A_{\text{unsat}}.$$

On the other hand, if $\mathcal{C}$ is satisfiable, then by the first part of Lemma 175 with $r = 1/\sqrt{162m}$,

$$Z \geq Z \Pr_{x \sim p}(x \in \mathcal{B}_r(v)) \geq e^{-1-\alpha/2} \left(\int_0^\infty \exp(-x^{d+1} - \beta(1-x^2)^2)\, dx \right)^n =: A_{\text{sat}}.$$

Since $\alpha \geq 2(n+1)$, we get

$$A_{\text{unsat}} \leq (2/e)^n A_{\text{sat}}$$

as claimed. □

But then approximating $Z_{\mathcal{C},\alpha,\beta}$ allows distinguishing a satisfiable formula from an unsatisfiable formula, which is NP-hard. This implies the following theorem:

Theorem 89. *Fix $n \in \mathbb{N}$ and let $B \geq Cn^2$ for a sufficiently large constant C. Unless RP = NP, there is no* $\text{poly}(n)$*-time algorithm which takes as input an arbitrary $\theta \in \Theta_B$ and outputs an approximation of* $\log Z_\theta$ *with additive error less than* $n \log 1.16$.

Proof. First, observe that the following problem is NP-hard (under randomized reductions): given two 3-CNF formulas $\mathcal{C}, \mathcal{C}'$ each with n variables and at most $10n$ clauses, where it is promised that exactly one of the formulas is satisfiable, determine which of the formulas is satisfiable. Indeed, this follows from Theorem 173: given a 3-CNF formula $\mathcal{C}$ with n variables, at most $5n$ clauses, and at most one satisfying assignment, consider adjoining either the clause x_i or the clause $\neg x_i$ to $\mathcal{C}$. If $\mathcal{C}$ has a satisfying assignment v^*, then exactly one of the resulting formulas is satisfiable, and determining which one is satisfiable identifies v_i^*. Repeating this procedure for all $i \in [n]$ yields an assignment v, which satisfies $\mathcal{C}$ if and only if $\mathcal{C}$ is satisfiable.

For each $n \in \mathbb{N}$ define $\alpha = 2(n+1)$ and $\beta = 64800n \log(13n\sqrt{10n})$. Let $B > 0$ be chosen later. Suppose that there is a $\text{poly}(n)$-time algorithm which, given $\theta \in \Theta_B$, computes an approximation of $\log Z_\theta$ with additive error less than $n \log 1.16$. Then given two formulas $\mathcal{C}$ and $\mathcal{C}'$ with n variables and at most $10n$ clauses each, we can compute $\theta = \theta(\mathcal{C}, \alpha, \beta)$ and $\theta' = \theta(\mathcal{C}', \alpha, \beta)$. By Lemma 174, we have $\theta, \theta' \in \Theta_B$ so long as $B \geq Cn^2$ for a sufficiently large constant C. Hence by assumption we can compute approximations $\tilde{Z}_\theta$ and $\tilde{Z}_{\theta'}$ of Z_θ and $Z_{\theta'}$ respectively, with multiplicative error less than 1.16^n. However, by Lemma 88 and the assumption that exactly one of $\mathcal{C}$ and $\mathcal{C}'$ is satisfiable, we know that $\tilde{Z}_\theta > \tilde{Z}_{\theta'}$ if and only if $\mathcal{C}$ is satisfiable. Thus, NP = RP. □

6.2.2 Hardness of approximating $\nabla_\theta \log Z_{\theta(\mathcal{C},\alpha,\beta)}$

Note that $\nabla_\theta \log Z_\theta = \mathbb{E}_{x\sim p_\theta}[T(x)]$, so in particular approximating the gradient yields an approximation to the mean $\mathbb{E}_{x\sim p_\theta}[x]$. Since $P_{\mathcal{C},\alpha,\beta}$ is concentrated in orthants corresponding to satisfying assignments of $\mathcal{C}$, we would intuitively expect that if $\mathcal{C}$ has exactly one satisfying assignment v^*, then $\operatorname{sign}(\mathbb{E}_{p_\theta}[x])$ corresponds to this assignment. Formally, we show that if $\alpha = \Theta(n)$ and $\beta = \Omega(mn\log m)$, then $\mathbb{E}_{x\sim p_{\mathcal{C},\alpha,\beta}}[v_i^* x_i] \geq 1/20$ for all $i \in [n]$:

Lemma 90. *Let $\mathcal{C}$ be a 3-CNF formula with m clauses and n variables, and exactly one satisfying assignment $v^* \in \mathcal{H}$. Let $\alpha = 4n$ and $\beta \geq 25920mn\log(102n\sqrt{mn})$, and define $p := p_{\mathcal{C},\alpha,\beta}$ and $Z := Z_{\mathcal{C},\alpha,\beta}$. Then $\mathbb{E}_{x\sim p}[v_i^* x_i] \geq 1/20$ for all $i \in [n]$.*

Proof. Without loss of generality take $i = 1$ and $v_i^* = 1$. Set $r = 1/(\sqrt{648mn})$, $\alpha = 4n$, and $\beta \geq 40r^{-2}\log(4n/r)$. We want to show that $\mathbb{E}_{x\sim p}[x_1] \geq 1/20$. We can write

$$\begin{aligned}\mathbb{E}[x_1] &= \mathbb{E}[x_1 \mathbb{1}[x \in B_r(v^*)]] + \mathbb{E}[x_1 \mathbb{1}[x \in \mathcal{O}(v^*) \setminus B_r(v^*)]] + \sum_{v \in \mathcal{H}\setminus\{v^*\}} \mathbb{E}[x_1 \mathbb{1}[x \in \mathcal{O}(v)]] \\ &\geq (1-r)\Pr[x \in B_r(v^*)] - 2^n \max_{v\in\mathcal{H}\setminus\{v^*\}} \mathbb{E}[|x_1| \mathbb{1}[x \in \mathcal{O}(v)]] \end{aligned} \tag{6.4}$$

since $x_1 \geq 1 - r$ for $x \in B_r(v^*)$ and $x_1 \geq 0$ for $x \in \mathcal{O}(v^*)$. Now observe that on the one hand,

$$\Pr(x \in B_r(v^*)) \geq \frac{e^{-1-81m\alpha r^2}}{Z}\left(\int_0^\infty \exp(-x^* - \beta g(x))\,dx\right)^n \tag{6.5}$$

by Lemma 175. On the other hand, for any $v \in \mathcal{H} \setminus \{v^*\}$,

$$\begin{aligned}\mathbb{E}[|x_1|\mathbb{1}[x \in \mathcal{O}(v)]] &= \frac{1}{Z}\int_{\mathcal{O}(v)} |x_1| \exp\left(-\sum_{i=1}^n x_i^8 - \alpha H(x) - \beta G(x)\right) dx \\ &\leq \frac{e^{-\alpha}}{Z}\int_{\mathcal{O}(v)} |x_1| \exp\left(-\sum_{i=1}^n x_i^8 - \beta G(x)\right) dx \\ &= \frac{e^{-\alpha}}{Z}\left(\int_0^\infty x\exp(-x^8 - \beta g(x))\,dx\right)\left(\int_0^\infty \exp(-x^8 - \beta g(x))\,dx\right)^{n-1} \\ &\leq \frac{2e^{-\alpha}}{Z}\left(\int_0^\infty \exp(-x^8 - \beta g(x))\,dx\right)^n \end{aligned} \tag{6.6}$$

where the second inequality is by Lemma 177 with $k = 1$. Combining (6.5) and (6.6) with (6.4), we have

$$\begin{aligned}
\mathbb{E}[x_1] &\geq \frac{(1-r)e^{-1-81m\alpha r^2} - 2^{n+1}e^{-\alpha}}{Z}\left(\int_0^\infty \exp(-x^8 - \beta g(x))\,dx\right)^n \\
&\geq \frac{1}{10Z}\left(\int_0^\infty \exp(-x^8 - \beta g(x))\,dx\right)^n \\
&\geq \frac{1}{10Z}\int_{\mathcal{O}(v^*)} \exp\left(-\sum_{i=1}^n x_i^8 - \alpha H(x) - \beta G(x)\right)\,dx \\
&= \frac{1}{10}\Pr[x \in \mathcal{O}(v^*)] \\
&\geq \frac{1}{20}
\end{aligned}$$

where the second inequality is by choice of α and r; the third inequality is by nonnegativity of $H(x)$; and the fourth inequality is by Lemma 93 and uniqueness of the satisfying assignment v^*. □

Since solving a formula with a unique satisfying assignment is still NP-hard, we get the following theorem:

Theorem 91. *Fix $n \in \mathbb{N}$ and let $B \geq Cn^2 \log(n)$ for a sufficiently large constant C. Unless RP = NP, there is no $\mathrm{poly}(n)$-time algorithm which takes as input an arbitrary $\theta \in \Theta_B$ and outputs an approximation of $\nabla_\theta \log Z_\theta$ with additive error (in an l_∞ sense) less than $1/20$.*

Proof. Suppose that such an algorithm exists. Set $\alpha = 4n$ and $\beta = 129600n^2 \log(102n^2\sqrt{5})$. Given a 3-CNF formula $\mathcal{C}$ with n variables, at most $5n$ clauses, and exactly one satisfying assignment $v^* \in \mathcal{H}$, we can compute $\theta = \theta(\mathcal{C}, \alpha, \beta)$. Let $E \in \mathbb{R}^n$ be the algorithm's estimate of $\nabla_\theta \log Z_\theta = \mathbb{E}_{x\sim p_{\mathcal{C},\alpha,\beta}} T(x)$. Then $\left\|E - \mathbb{E}_{x\sim p_{\mathcal{C},\alpha,\beta}} T(x)\right\|_\infty < 1/20$. But by Lemma 90, for each $i \in [n]$, the i-th entry of $\mathbb{E}_{x\sim p_{\mathcal{C},\alpha,\beta}} T(x)$, which corresponds to the monomial x_i, has sign v_i^* and magnitude at least $1/20$. Thus, $\mathrm{sign}(E_i) = v_i^*$. So we can compute v^* in polynomial time. By Theorem 173, it follows that NP = RP. □

With the above two theorems in hand, we are ready to present the formal version of Theorem 80; the proof is immediate from the definition of $L_{\mathrm{MLE}}(\theta)$.

Corollary 92. *Fix $n, N \in \mathbb{N}$ and let $B \geq Cn^2 \log n$ for a sufficiently large constant C. Unless RP = NP, there is no $\mathrm{poly}(n, N)$-time algorithm which takes as input an arbitrary $\theta \in \Theta_B$, and an arbitrary sample $x_1, \ldots, x_N \in \mathbb{R}^n$, and outputs an approximation of $L_{\mathrm{MLE}}(\theta)$ up to additive error of $n \log 1.16$, or $\nabla_\theta L_{\mathrm{MLE}}(\theta)$ up to an additive error of $1/20$.*

Proof. Recall that $\log p_\theta(x) = \log h(x) + \langle \theta, T(x) \rangle - \log Z_\theta$. Therefore $L_{\mathrm{MLE}}(\theta) = \hat{\mathbb{E}} \log h(x) + \langle \theta, \hat{\mathbb{E}} T(x) \rangle - \log Z_\theta$ and $\nabla_\theta L_{\mathrm{MLE}}(\theta) = \hat{\mathbb{E}} T(x) - \nabla_\theta \log Z_\theta$. Note that we can compute $\hat{\mathbb{E}} \log h(x)$ and $\hat{\mathbb{E}} T(x)$ exactly. It follows that if we can approximate $L_{\mathrm{MLE}}(\theta)$ up to an additive error of $n \log 1.16$, then we can compute $\log Z_\theta$ up to an additive error of $n \log 1.16$. Similarly, if we can compute $\nabla_\theta L_{\mathrm{MLE}}(\theta)$ up to an additive error of $1/20$, then we can compute $\nabla_\theta \log Z_\theta$ up to an additive error of $1/20$. This contradicts Theorems 89 and 91 respectively, completing the proof. □

6.2.3 Hardness of approximate sampling

We show that for $\alpha = \Omega(n)$ and $\beta = \Omega(m \log m)$, the likelihood that $x \sim P_{\mathcal{C},\alpha,\beta}$ lies in an orthant corresponding to a satisfying assignment for $\mathcal{C}$ is at least $1/2$ (Lemma 93). Hardness of approximate sampling follows immediately (Theorem 94). Hence, although we show that score matching can efficiently estimate θ^* from samples produced by nature, knowing θ^* isn't enough to efficiently *generate* samples from the distribution.

Lemma 93. *Let $\mathcal{C}$ be a satisfiable instance of 3-SAT with m clauses and n variables. Let $\alpha, \beta > 0$ satisfy $\alpha \geq 2(n+1)$ and $\beta \geq 6480m \log(13n\sqrt{m})$. Set $p := p_{\mathcal{C},\alpha,\beta}$ and $Z := Z_{\mathcal{C},\alpha,\beta}$. If $\mathcal{V} \subseteq \mathcal{H}$ is the set of satisfiable assignments for $\mathcal{C}$, then*

$$\sum_{v \in \mathcal{V}} \Pr_{x \sim p}(x \in \mathcal{O}(v)) \geq \frac{1}{2}.$$

Proof. Let $v \in \mathcal{H}$ be any assignment that satisfies $\mathcal{C}$, and let $w \in \mathcal{H}$ be any assignment that does not satisfy $\mathcal{C}$. By Lemma 175 with $r = 1/\sqrt{162m}$, we have

$$\begin{aligned}
\Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(v)) &\geq \Pr_{x \sim p_{\mathcal{C}}}(x \in B_r(v)) \\
&\geq \frac{e^{-1-\alpha/2}}{Z} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx \right)^n \\
&\geq e^{-1+\alpha/2} \Pr(x \in \mathcal{O}(w)).
\end{aligned}$$

Since we chose α sufficiently large that $e^{-1+\alpha/2} \geq 2^n$, we get that

$$\Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(v)) \geq \sum_{w \in \mathcal{H} \setminus \mathcal{V}} \Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(w)).$$

Hence,

$$\sum_{v \in \mathcal{V}} \Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(v)) \geq \sum_{w \in \mathcal{H} \setminus \mathcal{V}} \Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(w)) = 1 - \sum_{v \in \mathcal{V}} \Pr_{x \sim p_{\mathcal{C}}}(x \in \mathcal{O}(v)).$$

The lemma statement follows. $\square$

Theorem 94. *Let $B \geq Cn^2$ for a sufficiently large constant C. Unless RP = NP, there is no algorithm which takes as input an arbitrary $\theta \in \Theta_B$ and outputs a sample from a distribution Q with $\mathrm{TV}(P_\theta, Q) \leq 1/3$ in $\mathrm{poly}(n)$ time.*

Proof. Suppose that such an algorithm exists. For each $n \in \mathbb{N}$ define $\alpha = 2(n+1)$ and $\beta = 32400n \log(13n\sqrt{5n})$. Given a 3-CNF formula $\mathcal{C}$ with n variables and at most $5n$ clauses, we can compute $\theta = \theta(\mathcal{C}, \alpha, \beta)$. By Lemma 174 we have $\theta \in \Theta_B$ so long as $B \geq Cn^2$ for a sufficiently large constant C. Thus, by assumption we can generate a a sample from a distribution Q with $\mathrm{TV}(P_{\mathcal{C},\alpha,\beta}, Q) \leq 1/3$. But by Lemma 93, we have $\Pr_{x \sim P_{\mathcal{C},\alpha,\beta}}[\mathrm{sign}(x) \text{ satisfies } \mathcal{C}] \geq 1/2$. Thus, $\Pr_{x \sim Q}[\mathrm{sign}(x) \text{ satisfies } \mathcal{C}] \geq 1/6$. It follows that we can find a satisfying assignment with $O(1)$ invocations of the sampling algorithm in expectation. By Theorem 173 we get NP = RP. $\square$

6.3 Statistical Efficiency of Maximum Likelihood

In this section we prove Theorem 81 by showing that for any $\theta \in \Theta_B$, we can lower bound the smallest eigenvalue of the Fisher information matrix $\mathcal{I}(\theta)$. Concretely, we show:

Theorem 95. *For any $\theta \in \Theta_B$, it holds that*

$$\lambda_{\min}(\mathcal{I}(\theta)) \geq (nB)^{-O(d^3)}.$$

As a corollary, given N samples from p_θ, it holds as $N \to \infty$ that $\sqrt{N}(\hat{\theta}_{\mathrm{MLE}} - \theta) \to N(0, \Gamma_{\mathrm{MLE}})$ where $\|\Gamma_{\mathrm{MLE}}\|_{op} \leq (nB)^{O(d^3)}$. Moreover, for sufficiently large N, with probability at least 0.99 it holds that $\left\|\hat{\theta}_{\mathrm{MLE}} - \theta\right\|_2^2 \leq (nB)^{O(d^3)}/N$.

Once we have the bound on $\lambda_{\min}(\mathcal{I}(\theta))$, the first corollary follows from standard bounds for MLE (Section 6.1), and the second corollary follows from Markov's inequality (see e.g., Remark 4 in [KHR22]). Lower-bounding $\lambda_{\min}(\mathcal{I}(\theta))$ itself requires lower-bounding the variance of any polynomial (with respect to p_θ) in terms of its coefficients. The proof consists of three parts. First, we show that the norm of a polynomial in the monomial basis is upper-bounded in terms of its L^2 norm on $[-1,1]^n$:

Lemma 96. *For $f \in \mathbb{R}[x_1, \dots, x_n]_{\leq d}$, we have $\|f\|_{mon}^2 \leq \binom{n+d}{d}(4e)^d \|f\|_{L^2([-1,1]^n)}^2$.*

The key idea behind this proof is to work with the basis of (tensorized) Legendre polynomials, which is orthonormal with respect to the L^2 norm. Once we write the polynomial with respect to this basis, the L^2 norm equals the Euclidean norm of the coefficients. Given this observation, all that remains is to bound the coefficients after the change-of-basis. The formal proof is given below.

Proof of Lemma 96. We use the fact that the Legendre polynomials

$$L_k(x) = \frac{1}{2^k} \sum_{j=0}^{k} \binom{k}{j}^2 (x-1)^{k-j}(x+1)^j,$$

for integers $0 \leq k \leq d$, form an orthogonal basis for the vector space $\mathbb{R}[x]_{\leq d}$ with respect to $L^2[-1,1]$ (see e.g. [Koe98]). We consider the normalized versions $\hat{L}_k = \sqrt{\frac{2k+1}{2}} L_k$, so that $\left\|\hat{L}_k\right\|_{L^2[-1,1]} = 1$. By tensorization, the set of products of Legendre polynomials

$$\hat{L}_{\mathbf{d}}(x) = \prod_{i=1}^{n} \hat{L}_{\mathbf{d}(i)}(x_i),$$

as $\mathbf{d}$ ranges over degree functions with $|\mathbf{d}| \leq d$, form an orthonormal basis for $\mathbb{R}[x_1, \dots, x_n]_{\leq d}$ with respect to $L^2([-1,1]^n)$.

Using the formula for L_k, we obtain that the sum of absolute values of coefficients of L_k (in the monomial basis) is at most $\frac{1}{2^k}\sum_{j=0}^{k} \binom{k}{j}^2 2^k = 2^k$. By the bound $\|\cdot\|_2 \leq \|\cdot\|_1$ and the definition of $\hat{L}_k$,

$$\left\|\hat{L}_k\right\|_{\mathsf{mon}}^2 \leq \frac{2k+1}{2} \|L_k\|_{\mathsf{mon}}^2 \leq \frac{2k+1}{2} 2^{2k}$$

and hence for any degree function $\mathbf{d}$ with $|\mathbf{d}| \leq d$,

$$\left\|\hat{L}_{\mathbf{d}}\right\|_{\mathsf{mon}}^2 = \prod_{i=1}^n \left\|\hat{L}_{\mathbf{d}(i)}\right\|_{\mathsf{mon}}^2 \leq \prod_{i=1}^n \frac{2\mathbf{d}(i)+1}{2} 2^{2\mathbf{d}(i)}$$
$$\leq \prod_{i=1}^n e^{\mathbf{d}(i)} 2^{2\mathbf{d}(i)} \leq (4e)^d.$$

Consider any polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]_{\leq d}$, and write $f = \sum_{|\mathbf{d}| \leq d} a_{\mathbf{d}} \hat{L}_{\mathbf{d}}$. By orthonormality, it holds that $\sum_{|\mathbf{d}| \leq d} a_{\mathbf{d}}^2 = \|f\|_{L^2([-1,1]^n)}^2$. Thus, by the triangle inequality and Cauchy-Schwarz,

$$\|p\|_{\mathsf{mon}}^2 = \left\| \sum_{|\mathbf{d}| \leq d} a_{\mathbf{d}} \hat{L}_{\mathbf{d}} \right\|_{\mathsf{mon}}^2 \leq \sum_{|\mathbf{d}| \leq d} a_{\mathbf{d}}^2 \cdot \sum_{|\mathbf{d}| \leq d} \left\|\hat{L}_{\mathbf{d}}\right\|_{\mathsf{mon}}^2$$
$$\leq \|p\|_{L^2([-1,1]^n)}^2 \binom{n+d}{d} (4e)^d$$

as claimed. $\square$

Next, we show that if a polynomial $f : \mathbb{R}^n \to \mathbb{R}$ has small variance with respect to p, then there is some box on which f has small variance with respect to the uniform distribution. This provides a way of comparing the variance of f with its L^2 norm (after an appropriate rescaling).

Lemma 97. *Fix any $\theta \in \Theta_B$ and define $p := p_\theta$. Define $R := 2^{d+3} nBM$. Then for any $f \in \mathbb{R}[x_1, \ldots, x_n]_{\leq d}$, there is some $z \in \mathbb{R}^n$ with $\|z\|_\infty \leq R$ and some $\epsilon \geq 1/(2(d+1)MR^d(n+B))$ such that*

$$\operatorname{Var}_p(f) \geq \frac{1}{2e} \operatorname{Var}_{\tilde{\mathcal{U}}}(f),$$

where $\tilde{\mathcal{U}}$ is the uniform distribution on $\{x \in \mathbb{R}^n : \|x - z\|_\infty \leq \epsilon\}$.

In order to prove this result, we pick a random box of radius ϵ (within a large bounding box of radius R). In expectation, the variance on this box (with respect to p) is not much less than $\operatorname{Var}_p(f)$. Moreover, for sufficiently small ϵ, the density function of p on this box has bounded fluctuations, allowing comparison of $\operatorname{Var}_p(f)$ and $\operatorname{Var}_{\tilde{\mathcal{U}}}(f)$. This argument is formalized below. First, we require the following fact that monomials of bounded degree are Lipschitz within a bounding box:

Lemma 98. *Fix $R > 0$. For any degree function $\mathbf{d} : [n] \to \mathbb{N}$ with $|\mathbf{d}| \leq d$, and for any $u, v \in \mathbb{R}^n$ with $\|u\|_\infty, \|v\|_\infty \leq R$, it holds that*

$$|u_{\mathbf{d}} - v_{\mathbf{d}}| \leq dR^{d-1} \|u - v\|_\infty.$$

Proof. Define $m(x) = x_{\mathbf{d}} = \prod_{i=1}^n x_i^{\mathbf{d}(i)}$. Then

$$|m(u) - m(v)| \leq \|u - v\|_\infty \sup_{x \in \mathcal{B}_R(0)} \|\nabla_x m(x)\|_1$$
$$= \|u - v\|_\infty \sup_{x \in \mathcal{B}_R(0)} \sum_{i \in [n] : \mathbf{d}(i) > 0} \alpha_i \prod_{j=1}^n x_i^{\mathbf{d}(i) - \mathbb{1}[i=j]}$$
$$\leq \|u - v\|_\infty \cdot dR^{d-1}$$

as claimed. $\square$

Proof of Lemma 97. Let $f \in \mathbb{R}[x_1, \dots, x_n]_{\le d}$ be a polynomial of degree at most d in $x_1, \dots, x_n$. Define $g(x) = f(x) - \mathbb{E}_{x\sim p} f(x)$. Set $\epsilon = 1/(2(d+1)MR^d(n+B))$ and let $(W_i)_{i\in I}$ be ℓ_∞-balls of radius ϵ partitioning $\{x \in \mathbb{R}^n : \|x\|_\infty \le R\}$. Define random variable $X \sim p|\{\|X\|_\infty \le R\}$ and let $\iota \in I$ be the random index so that $X \in B_\iota$. Then

$$\begin{aligned}\operatorname{Var}_p(f) &= \mathbb{E}_{x\sim p}[g(x)^2] \\ &\ge \frac{1}{2}\mathbb{E}[g(X)^2] \\ &= \frac{1}{2}\mathbb{E}_\iota \mathbb{E}_X[g(X)^2 | X \in W_\iota]\end{aligned}$$

where the inequality uses guarantee (c) of Lemma 104 that $\Pr_{x\sim p}[\|x\|_\infty > R] \le 1/2$.

Thus, there exists some $\iota^* \in I$ such that $\mathbb{E}_X[g(X)^2 | X \in W_{\iota^*}] \le 2\operatorname{Var}_p(f)$. Let $q : \mathbb{R}^n \to \mathbb{R}_+$ be the density function of $X | X \in W_{\iota^*}$. Since $q(x) \propto p(x)\mathbb{1}[x \in W_{\iota^*}]$, for any $u, v \in W_{\iota^*}$ we have that

$$\begin{aligned}\frac{q(u)}{q(v)} = \frac{p(u)}{p(v)} &= \frac{h(u)\exp(\langle\theta, T(u)\rangle)}{h(v)\exp(\langle\theta, T(v)\rangle)} \\ &= \exp\left(\sum_{i=1}^n v_i^{d+1} - u_i^{d+1} + \langle\theta, T(u) - T(v)\rangle\right).\end{aligned}$$

Applying Lemma 98, we get that

$$\begin{aligned}\frac{q(u)}{q(v)} &\le \exp\left(n(d+1)R^d\|u-v\|_\infty + MB\|T(u) - T(v)\|_\infty\right) \\ &\le \exp\left((n+B)\cdot M(d+1)R^d\|u-v\|_\infty\right) \\ &\le \exp(2\epsilon(n+B)\cdot M(d+1)R^d) \\ &\le \exp(1)\end{aligned}$$

by choice of ϵ. It follows that if $\tilde{\mathcal{U}}$ is the uniform distribution on W_{ι^*}, then $q(x) \ge e^{-1}\tilde{\mathcal{U}}(x)$ for all $x \in \mathbb{R}^n$. Thus,

$$\operatorname{Var}_p(f) \ge \frac{1}{2}\mathbb{E}_X[g(X)^2 | X \in W_{\iota^*}] \ge \frac{1}{2e}\mathbb{E}_{x\sim\tilde{\mathcal{U}}}[g(x)^2] \ge \frac{1}{2e}\operatorname{Var}_{\tilde{\mathcal{U}}}(g) = \frac{1}{2e}\operatorname{Var}_{\tilde{\mathcal{U}}}(f)$$

as desired. $\square$

Together, Lemma 96 and 97 allow us to lower bound the variance $\operatorname{Var}_p(f)$ in terms of $\|f\|_{\mathsf{mon}}$.

Lemma 99. *Fix any $\theta \in \Theta_B$ and define $p := p_\theta$. Define $R := 2^{d+3}nBM$. Then for any $f \in \mathbb{R}[x_1, \dots, x_n]_{\le d}$ with $f(0) = 0$, it holds that*

$$\operatorname{Var}_p(f) \ge \frac{1}{2^{2d}(d+1)^{2d}(4e)^{d+1}M^{2d+3}R^{2d^2+2d}(n+B)^{2d}}\|f\|_{\mathsf{mon}}^2.$$

Proof. By Lemma 97, there is some $z \in \mathbb{R}^n$ with $\|z\|_\infty \le R$ and some $\epsilon \ge 1/(2(d+1)MR^d(n+B))$ so that if $\tilde{\mathcal{U}}$ is the uniform distribution on $\{x \in \mathbb{R}^n : \|x - z\|_\infty \le \epsilon\}$, then

$$\mathrm{Var}_p(f) \ge \frac{1}{2e} \mathrm{Var}_{\tilde{\mathcal{U}}}(f).$$

Define $g : \mathbb{R}^n \to \mathbb{R}$ by $g(x) = f(\epsilon x + z) - \mathbb{E}_{\tilde{\mathcal{U}}} f$. Then by Lemma 96,

$$\begin{aligned} \|g\|_{\mathsf{mon}}^2 &\le (4e)^d M \mathbb{E}_{x \sim \mathrm{Unif}([-1,1]^n)} g(x)^2. \\ &= (4e)^d M \,\mathrm{Var}_{\tilde{\mathcal{U}}}(f) \\ &\le (4e)^{d+1} M \,\mathrm{Var}_p(f). \end{aligned}$$

Write $f(x) = \sum_{1 \le |\mathbf{d}| \le d} \alpha_{\mathbf{d}} x_{\mathbf{d}}$ and $g(x) = \sum_{1 \le |\mathbf{d}| \le d} \beta_{\mathbf{d}} x_{\mathbf{d}}$. We know that $f(x) = g(\epsilon^{-1}(x - z)) + \mathbb{E}_{\tilde{\mathcal{U}}} f$. Thus, for any nonzero degree function $\mathbf{d}$, we have

$$\alpha_{\mathbf{d}} = \sum_{\substack{\mathbf{d}' \ge \mathbf{d} \\ |\mathbf{d}'| \le d}} \epsilon^{-|\mathbf{d}'|} (-z)^{\mathbf{d}' - \mathbf{d}} \beta_{\mathbf{d}'}.$$

Thus $|\alpha_{\mathbf{d}}| \le \epsilon^{-d} R^d \|\beta\|_1 \le \epsilon^{-d} R^d \sqrt{M} \, \|g\|_{\mathsf{mon}}$, and so summing over monomials gives

$$\|f\|_{\mathsf{mon}}^2 \le M^2 \epsilon^{-2d} R^{2d} \, \|g\|_{\mathsf{mon}}^2 \le (4e)^{d+1} M^3 \epsilon^{-2d} R^{2d} \,\mathrm{Var}_p(f).$$

Substituting in the choice of ϵ from Lemma 97 completes the proof. □

We are now ready to finish the proof of Theorem 95.

Proof of Theorem 95. Fix $\theta \in \Theta_B$. Pick any $w \in \mathbb{R}^M$ and define $f(x) = \langle w, T(x) \rangle$. By definition of $\mathcal{I}(\theta)$, we have $\mathrm{Var}_{p_\theta}(f) = w^\top \mathcal{I}(\theta) w$. Moreover, $\|f\|_{\mathsf{mon}}^2 = \|w\|_2^2$. Thus, Lemma 99 gives us that $w^\top \mathcal{I}(\theta) w \ge (nB)^{-O(d^3)} \|w\|_2^2$, using that $R = 2^{d+3} nBM$ and $M = \binom{n+d}{d}$. The bound $\lambda_{\min}(\mathcal{I}(\theta)) \ge (nB)^{-O(d^3)}$ follows. □

6.4 Statistical Efficiency of Score Matching

In this section we prove Theorem 82. The main technical ingredient is a bound on the restricted Poincaré constants of distributions in $\mathcal{P}_{n,d,B}$. For any fixed $\theta \in \Theta_B$, we show (Lemma 102) that C_P can be bounded in terms of the *condition number* of the Fisher information matrix $\mathcal{I}(\theta)$.

Fix $\theta, w \in \mathbb{R}^{M-1}$ and define $f(x) := \langle w, T(x) \rangle$. First, we need to upper bound $\mathrm{Var}_{p_\theta}(f)$. This is where (the first half of) the condition number appears. Using the crucial fact that the restricted Poincaré constant only considers functions f that are linear in the sufficient statistics, and the definition of $\mathcal{I}(\theta)$, we get the following bound on $\mathrm{Var}_{p_\theta}(f)$ in terms of the coefficient vector w.

Lemma 100. *Fix $\theta, w \in \mathbb{R}^{M-1}$ and define $f(x) := \langle w, T(x) \rangle$. Then*

$$\|w\|_2^2 \lambda_{\min}(\mathcal{I}(\theta)) \le \mathrm{Var}_{p_\theta}(f) \le \|w\|_2^2 \lambda_{\max}(\mathcal{I}(\theta)).$$

Proof. We have

$$\begin{aligned}\mathrm{Var}_{p_\theta}(f) &= \mathbb{E}_{x\sim p_\theta}[f(x)^2] - \mathbb{E}_{x\sim p_\theta}[f(x)]^2 \\ &= w^\top \mathbb{E}_{x\sim p_\theta}[T(x)T(x)^\top]w - w^\top \mathbb{E}_{x\sim p_\theta}[T(x)]\mathbb{E}_{x\sim p_\theta}[T(x)]^\top w \\ &= w^\top \mathcal{I}(\theta) w,\end{aligned}$$

and since

$$\|w\|_2^2 \lambda_{\min}(\mathcal{I}(\theta)) \le w^\top \mathcal{I}(\theta) w \le \|w\|_2^2 \lambda_{\max}(\mathcal{I}(\theta),$$

the lemma statement follows. □

Next, we lower bound $\mathbb{E}_{x\sim p_\theta}\|\nabla_x f(x)\|_2^2$. To do so, we could pick an orthonormal basis and bound $\mathbb{E}\langle u, \nabla_x f(x)\rangle^2$ over all directions u in the basis; however, it is unclear how to choose this basis. Instead, we pick $u \sim \mathcal{N}(0, I_n)$ randomly, and use the following identity:

$$\mathbb{E}_{x\sim p_\theta}[\|\nabla_x f(x)\|_2^2] = \mathbb{E}_{x\sim p_\theta}\mathbb{E}_{u\sim N(0,I_n)}\langle u, \nabla_x f(x)\rangle^2$$

For any fixed u, the function $g(x) = \langle u, \nabla_x f(x)\rangle$ is also a polynomial. If this polynomial had no constant coefficient, we could immediately lower bound $\mathbb{E}\langle u, \nabla_x f(x)\rangle^2$ in terms of the remaining coefficients, as above. Of course, it may have a nonzero constant coefficient, but with some case-work over the value of the constant, we can still prove the following bound:

Lemma 101. *Fix $\theta, \tilde{w} \in \mathbb{R}^{M-1}$ and $c \in \mathbb{R}$, and define $g(x) := \langle \tilde{w}, T(x)\rangle + c$. Then*

$$\mathbb{E}_{x\sim p_\theta}[g(x)^2] \ge \frac{c^2 + \|\tilde{w}\|_2^2}{4 + 4\|\mathbb{E}[T(x)]\|_2^2} \min(1, \lambda_{min}(\mathcal{I}(\theta))).$$

Proof. We have

$$\begin{aligned}\mathbb{E}_{x\sim p_\theta}[g(x)^2] &= \mathrm{Var}_{p_\theta}(g) + \mathbb{E}_{x\sim p_\theta}[g(x)]^2 \\ &= \mathrm{Var}_{p_\theta}(g - c) + (c + \tilde{w}^\top \mathbb{E}_{x\sim p_\theta}[T(x)])^2 \\ &\ge \|\tilde{w}\|_2^2 \lambda_{\min}(\mathcal{I}(\theta)) + (c + \tilde{w}^\top \mathbb{E}_{x\sim p_\theta}[T(x)])^2\end{aligned}$$

where the inequality is by Lemma 100. We now distinguish two cases.

Case I. Suppose that $|c + \tilde{w}^\top \mathbb{E}_{x\sim p_\theta}[T(x)]| \ge c/2$. Then

$$\mathbb{E}_{x\sim p_\theta}[g(x)^2] \ge \|\tilde{w}\|_2^2 \lambda_{\min}(\mathcal{I}(\theta)) + \frac{c^2}{4} \ge \frac{c^2 + \|\tilde{w}\|_2^2}{4} \min(1, \lambda_{\min}(\mathcal{I}(\theta))).$$

Case II. Otherwise, we have $|c + \tilde{w}^\top \mathbb{E}_{x\sim p_\theta}[T(x)]| < c/2$. By the triangle inequality, it follows that $|\tilde{w}^\top \mathbb{E}_{x\sim p_\theta}[T(x)]| \geq c/2$, so $\|\tilde{w}\|_2 \geq c/(2\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2)$. Therefore

$$c^2 + \|\tilde{w}\|_2^2 \leq (1 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2)\|\tilde{w}\|_2^2,$$

from which we get that

$$\mathbb{E}_{x\sim p_\theta}[g(x)^2] \geq \|\tilde{w}\|_2^2 \lambda_{\min}(\mathcal{I}(\theta)) \geq \frac{c^2 + \|\tilde{w}\|_2^2}{1 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2} \lambda_{\min}(\mathcal{I}(\theta))$$

as claimed. □

With Lemma 100 and Lemma 101 in hand (taking $g(x) = \langle u, \nabla_x f(x)\rangle$ in the latter), all that remains is to relate the squared monomial norm of $\langle u, \nabla_x f(x)\rangle$ (in expectation over u) to the squared monomial norm of f. This crucially uses the choice $u \sim N(0, I_n)$. We put together the pieces in the following lemma.

Lemma 102. *Fix $\theta, w \in \mathbb{R}^{M-1}$. Define $f(x) := \langle w, T(x)\rangle$. Then*

$$\mathrm{Var}_{p_\theta}(f) \leq (4 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2) \frac{\lambda_{max}(\mathcal{I}(\theta))}{\min(1, \lambda_{min}(\mathcal{I}(\theta)))} \mathbb{E}_{x\sim p_\theta}[\|\nabla_x f(x)\|_2^2].$$

Proof. Since $f(x) = \sum_{1\leq|\mathbf{d}|\leq d} w_\mathbf{d} x_\mathbf{d}$, we have for any $u \in \mathbb{R}^n$ that

$$\langle u, \nabla_x f(x)\rangle = \sum_{i=1}^n u_i \sum_{0\leq|\mathbf{d}|<d} (1 + \mathbf{d}(i)) w_{\mathbf{d}+\{i\}} x_\mathbf{d} = c(u) + \sum_{1\leq|\mathbf{d}|<d} \tilde{w}(u)_\mathbf{d} x_\mathbf{d}$$

where $c(u) := \sum_{i=1}^n u_i w_{\{i\}}$ and $\tilde{w}(u)_\mathbf{d} := \sum_{i=1}^n u_i(1 + \mathbf{d}(i)) w_{\mathbf{d}+\{i\}}$. But now

$$\begin{aligned}
\mathbb{E}_{x\sim p_\theta}[\|\nabla_x f(x)\|_2^2] &= \mathbb{E}_{x\sim p_\theta} \mathbb{E}_{u\sim N(0,I_n)} \langle u, \nabla_x f(x)\rangle^2 \\
&= \mathbb{E}_{u\sim N(0,I_n)} \mathbb{E}_{x\sim p_\theta} (c(u) + \langle \tilde{w}(u), T(x)\rangle)^2 \\
&\geq \mathbb{E}_{u\sim N(0,I_n)} \frac{c(u)^2 + \|\tilde{w}(u)\|_2^2}{4 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2} \min(1, \lambda_{\min}(\mathcal{I}(\theta))).
\end{aligned}$$

where the last inequality is by Lemma 101. Finally,

$$\begin{aligned}
\mathbb{E}_{u\sim N(0,I_n)} \left[c(u)^2 + \|\tilde{w}(u)\|_2^2\right] &= \sum_{0\leq|\mathbf{d}|<d} \mathbb{E}_{u\sim N(0,I_n)} \left[\left(\sum_{i=1}^n u_i(1 + \mathbf{d}(i)) w_{\mathbf{d}+\{i\}}\right)^2\right] \\
&= \sum_{0\leq|\mathbf{d}|<d} \sum_{i=1}^n (1 + \mathbf{d}(i))^2 w_{\mathbf{d}+\{i\}}^2 \qquad \geq \|w\|_2^2
\end{aligned}$$

where the second equality is because $\mathbb{E}[u_i u_j] = \mathbb{1}[i = j]$ for all $i, j \in [n]$, and the last inequality is because every term $w_{\mathbf{d}}^2$ in $\|w\|_2^2$ appears in at least one of the terms of the previous summation (and has coefficient at least one). Putting everything together gives

$$\begin{aligned}
\mathbb{E}_{x\sim p_\theta}[\|\nabla_x f(x)\|_2^2] &\geq \frac{\|w\|_2^2}{4 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2} \min(1, \lambda_{\min}(\mathcal{I}(\theta))) \\
&\geq \frac{1}{4 + 4\|\mathbb{E}[T(x)]\|_2^2} \frac{\min(1, \lambda_{\min}(\mathcal{I}(\theta)))}{\lambda_{\max}(\mathcal{I}(\theta))} \operatorname{Var}_{p_\theta}(f)
\end{aligned}$$

where the last inequality is by Lemma 100. □

Finally, putting together Lemma 102, Theorem 95 (that lower bounds $\lambda_{\min}(\mathcal{I}(\theta))$), and Lemma 104 (that upper bounds $\lambda_{\max}(\mathcal{I}(\theta))$ – a straightforward consequence of the distributions in $\mathcal{P}_{n,d,B}$ having bounded moments), we can prove the following formal version of Theorem 82:

Theorem 103. *Fix $n, d, B, N \in \mathbb{N}$. Pick any $\theta^* \in \Theta_B$ and let $x^{(1)}, \ldots, x^{(N)} \sim p_{\theta^*}$ be independent samples. Then as $N \to \infty$, the score matching estimator $\hat{\theta}_{\mathrm{SM}} = \hat{\theta}_{\mathrm{SM}}(x^{(1)}, \ldots, x^{(N)})$ satisfies*

$$\sqrt{N}(\hat{\theta}_{\mathrm{SM}} - \theta^*) \to N(0, \Gamma)$$

where $\|\Gamma\|_{\mathsf{op}} \leq (nB)^{O(d^3)}$. As a corollary, for all sufficiently large N it holds with probability at least 0.99 that $\left\|\hat{\theta}_{\mathrm{SM}} - \theta^*\right\|_2^2 \leq (nB)^{O(d^3)}/N$.

Proof. We apply Theorem 84. By Lemma 180 and the fact that $\lambda_{\min}(I(\theta^*)) > 0$ (Theorem 95), the necessary regularity conditions are satisfied so that the score matching estimator is consistent and asymptotically normal, with asymptotic covariance Γ satisfying

$$\|\Gamma\|_{\mathsf{op}} \leq \frac{2C_P^2(\|\theta\|_2^2 \mathbb{E}_{x\sim p_{\theta^*}} \|(JT)(x)\|_{\mathsf{op}}^4 + \mathbb{E}_{x\sim p_{\theta^*}} \|\Delta T(x)\|_2^2)}{\lambda_{\min}(\mathcal{I}(\theta^*))^2} \tag{6.7}$$

where C_P is the restricted Poincaré constant for p_{θ^*} with respect to linear functions in $T(x)$ (see Definition 83). By Lemma 102, we have

$$\begin{aligned}
C_P &\leq (4 + 4\|\mathbb{E}_{x\sim p_\theta}[T(x)]\|_2^2) \frac{\lambda_{\max}(\mathcal{I}(\theta^*))}{\min(1, \lambda_{\min}(\mathcal{I}(\theta^*))} \\
&\leq (4 + 4B^{2d} M^{2d+2} 2^{2d(d+1)+1}) \frac{B^{2d} M^{2d+1} 2^{2d(d+1)+1}}{(nB)^{-O(d^3)}} \quad \leq (nB)^{O(d^3)}
\end{aligned}$$

using parts (a) and (b) of Lemma 104; Theorem 95; and the fact that $M = \binom{n+d}{d}$. Substituting into (6.7) and bounding the remaining terms using Lemma 179 and a second application of Theorem 95, we conclude that $\|\Gamma\|_{\mathsf{op}} \leq (nB)^{O(d^3)}$ as claimed. The high-probability bound now follows from Markov's inequality; see Remark 4 in [KHR22] for details. □

All that remains is to prove the upper bounds on $\lambda_{\max}(\mathcal{I}(\theta))$ and $\|\mathbb{E}_{x\sim p_\theta} T(x)\|_2^2$, which are encapsulated in parts (a) and (b) of Lemma 104 below (part (c) is used in the proof of Lemma 97). These bounds follow from the fact that distributions in $\mathcal{P}_{n,d,B}$ have bounded moments (Lemma 178).

Lemma 104 (Largest eigenvalue bound). *For any $\theta \in \Theta_B$, it holds that*

$$\mathbb{E}_{x\sim p_\theta} T(x)T(x)^\top \preceq B^{2d} M^{2d+1} 2^{2d(d+1)+1}.$$

We also have the following consequences:

(a) $\|\mathbb{E}_{x\sim p_\theta} T(x)\|_2^2 \le B^{2d} M^{2d+2} 2^{2d(d+1)+1}$,

(b) $\lambda_{max}(\mathcal{I}(\theta)) \le B^{2d} M^{2d+1} 2^{2d(d+1)+1}$,

(c) $\Pr_{x\sim p_\theta}[\|x\|_\infty > 2^{d+3} nBM] \le 1/2$.

Proof. Fix any $u, v \in [M]$. Then $T(x)_u T(x)_v = \prod_{i=1}^n x_i^{\gamma_i}$ for some nonnegative integers $\gamma_1, \dots, \gamma_n$ where $d' := \sum_{i=1}^n \gamma_i \le 2d$. Therefore

$$\mathbb{E}_{x\sim p_\theta} T(x)_u T(x)_v = \mathbb{E}_{x\sim p_\theta} \prod_{i=1}^n x_i^{\gamma_i} \le \prod_{i=1}^n \left(\mathbb{E}_{x\sim p_\theta} x_i^{d'}\right)^{\gamma_i/d'} \le B^{2d} M^{2d} 2^{2d(d+1)+1}$$

by Holder's inequality and Lemma 178 (with $\ell = 2d$). The claimed spectral bound follows. To prove (a), observe that

$$\|\mathbb{E}_{x\sim p_\theta} T(x)\|_2^2 \le \mathbb{E}_{x\sim p_\theta} \|T(x)\|_2^2 = \operatorname{Tr} \mathbb{E}_{x\sim p_\theta} T(x)T(x)^\top \le M\lambda_{\max}(\mathbb{E}_{x\sim p_\theta} T(x)T(x)^\top)$$

To prove (b), observe that $\mathcal{I}(\theta) \preceq \mathbb{E}_{x\sim p_\theta} T(x)T(x)^\top$. To prove (c), observe that for any $i \in [n]$,

$$\Pr_{x\sim p_\theta}[|x_i| > 2^{d+3} nBM] \le \frac{\mathbb{E}_{x\sim p_\theta} x_i^{2d}}{(2^{d+3} nBM)^{2d}} \le \frac{1}{2n}.$$

A union bound over $i \in [n]$ completes the proof. $\square$

6.5 Conclusion

We have provided a concrete example of an exponential family—namely, exponentials of bounded degree polynomials—where score matching is significantly more computationally efficient than maximum likelihood estimation (through optimization with a zero- or first-order oracle), while still achieving the same sample efficiency up to polynomial factors. While score matching was designed to be more computationally efficient for exponential families, the determination of statistical complexity is more challenging, and we give the first separation between these two methods for a general class of functions.

As we have restricted our attention to the asymptotic behavior of both of the methods, an interesting future direction is to see how the finite sample complexities differ. One could also give a more fine-grained comparison between the polynomial dependencies of score matching and MLE, which we have not attempted to optimize. Finally, it would be interesting to relate our results with similar results and algorithms for learning Ising and higher-order spin glass models in the discrete setting, and give a more unified treatment of psueudo-likelihood or score/ratio matching algorithms in these different settings.

Chapter 7

An Universal Approximation result for Normalizing Flows

Normalizing flows [DKB14; RM15] are a class of generative models parametrizing a distribution in $\mathbb{R}^d$ as the pushfoward of a simple distribution (e.g. Gaussian) through an invertible map $g_\theta : \mathbb{R}^d \to \mathbb{R}^d$ with trainable parameter θ. The fact that g_θ is invertible allows us to write down an explicit expression for the density of a point x through the change-of-variables formula, namely $p_\theta(x) = \phi(g_\theta^{-1}(x))\det(Dg_\theta^{-1}(x))$, where ϕ denotes the density of the standard Gaussian. For different choices of parametric families for g_θ, one gets different families of normalizing flows, e.g. affine coupling flows [DKB14; DSB16; KD18], Gaussianization flows [Men+20], sum-of-squares polynomial flows [JSY19].

In this paper we focus on affine coupling flows – arguably the family that has been most successfully scaled up to high resolution datasets [KD18]. The parametrization of g_θ is chosen to be a composition of so-called *affine coupling blocks*, which are maps $f : \mathbb{R}^d \to \mathbb{R}^d$, s.t. $f(x_S, x_{[d]\setminus S}) = (x_S, x_{[d]\setminus S} \odot s(x_S) + t(x_S))$, where $\odot$ denotes entrywise multiplication and s, t are (typically simple) neural networks. The choice of parametrization is motivated by the fact that the Jacobian of each affine block is triangular, so that the determinant can be calculated in linear time.

Despite the empirical success of this architecture, theoretical understanding remains elusive. The most basic questions revolve around the representational power of such models. Even the question of universal approximation was only recently answered by three concurrent papers [HDC20; Zha+20; KMR20]—though in a less-than-satisfactory manner, in light of how normalizing flows are trained. Namely, [HDC20; Zha+20] show that any (reasonably well-behaved) distribution p, once padded with zeros and treated as a distribution in $\mathbb{R}^{d+d'}$, can be arbitrarily closely approximated by an affine coupling flow. While such padding can be operationalized as an algorithm by padding the training image with zeros, it is never done in practice, as it results in an ill-conditioned Jacobian. This is expected, as the map that always sends the last d' coordinates to 0 is not injective. [KMR20] prove universal approximation without padding; however their construction *also* gives rise to a poorly conditioned Jacobian: namely, to approximate a distribution p to within accuracy ϵ in the Wasserstein-1 distance, the Jacobian of the network they construct will have smallest singular value on the order of ϵ.

Importantly, for all these constructions, the condition number of the resulting affine coupling map is poor *no matter how nice the underlying distribution it's trying to approximate is*. In other words, the source of this phenomenon isn't that the underlying distribution is low-dimensional or otherwise degenerate. Thus the question arises:

Question: *Can well-behaved distributions be approximated by an affine coupling flow with a well-conditioned Jacobian?*

In this paper, we answer the above question in the affirmative for a broad class of distributions – log-concave distributions – if we pad the input distribution not with zeroes, but with independent Gaussians. This gives theoretical grounding of an empirical observation in [KMR20] that Gaussian padding works better than zero-padding, as well as no padding.

The practical relevance of this question is in providing guidance on the type of distributions we can hope to fit via training using an affine coupling flow. Theoretically, our techniques uncover some deep connections between affine coupling flows and two other (seeming unrelated) areas of mathematics: *stochastic differential equations* (more precisely *underdamped Langevin dynamics*, a "momentum" variant of the standard overdamped Langevin dynamics) and *dynamical systems* (more precisely, a family of dynamical systems called *Hénon-like maps*).

7.1 Overview of results

In order to state our main result, we introduce some notation and definitions.

7.1.1 Notation

Definition 105. An *affine coupling block* is a map $f : \mathbb{R}^d \to \mathbb{R}^d$, s.t. $f(x_S, x_{[d]\setminus S}) = (x_S, x_{[d]\setminus S} \odot s(x_S) + t(x_S))$ for some set of coordinates S, where $\odot$ denotes entrywise multiplication and s, t are trainable (generally non-linear) functions. An *affine coupling network* is a finite sequence of affine coupling blocks. Note that the partition $(S, [d] \setminus S)$, as well as s, t may be different between blocks. We say that the non-linearities are in a class $\mathcal{F}$ (e.g., neural networks, polynomials, etc.) if $s, t \in \mathcal{F}$.

The appeal of affine coupling networks comes from the fact that the Jacobian of each affine block is triangular, so calculating the determinant is a linear-time operation.

We will be interested in the *conditioning* of f—that is, an upper bound on the largest singular value $\sigma_{\max}(Df)$ and lower bound on the smallest singular value $\sigma_{\min}(Df)$ of the Jacobian Df of f. Note that this is a slight abuse of nomenclature – most of the time, "condition number" refers to the ratio of the largest and smallest singular value. As training a normalizing flow involves evaluating $\det(Df)$, we in fact want to ensure that neither the smallest nor largest singular values are extreme.

The class of distributions we will focus on approximating via affine coupling flows is *log-concave* distributions:

Definition 106. A distribution $p : \mathbb{R}^d \to \mathbb{R}^+, p(x) \propto e^{-U(x)}$ is *log-concave* if $\nabla^2 U(x) = -\nabla^2 \ln p(x) \succeq 0$.

Log-concave distributions are typically used to model distributions with Gaussian-like tail behavior. What we will leverage about this class of distributions is that a special stochastic differential equation (SDE), called *underdamped Langevin dynamics*, is well-behaved in an analytic sense. Finally, we recall the definitions of positive definite matrices and Wasserstein distance, and introduce a notation for truncated distributions.

Definition 107. We say that a symmetric matrix is *positive semidefinite (PSD)* if all of its eigenvalues are non-negative. For symmetric matrices A, B, we write $A \succeq B$ if and only if $A - B$ is PSD.

Definition 108. Given two probability measures μ, ν over a metric space (M, d), the *Wasserstein*-1 distance between them, denoted $W_1(\mu, \nu)$, is defined as

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)\, d\gamma(x, y)$$

where $\Gamma(\mu, \nu)$ is the set of couplings, i.e. measures on $M \times M$ with marginals μ, ν respectively. For two probability *distributions* p, q, we denote by $W_1(p, q)$ the Wasserstein-1 distance between their associated measures. In this paper, we set $M = \mathbb{R}^d$ and $d(x, y) = \|x - y\|_2$.

Definition 109. Given a distribution q and a compact set $\mathcal{C}$, we denote by $q|_{\mathcal{C}}$ the distribution q truncated to the set $\mathcal{C}$. The truncated measure is defined as $q|_{\mathcal{C}}(A) = \frac{1}{q(\mathcal{C})} q(A \cap \mathcal{C})$.

7.1.2 Main result

Our main result states that we can approximate any log-concave distribution in Wasserstein-1 distance by a *well-conditioned* affine-coupling flow network. Precisely, we show:

Theorem 110. *Let $p(x) : \mathbb{R}^d \to \mathbb{R}^+$ be of the form $p(x) \propto e^{-U(x)}$, such that:*

1. *$U \in C^2$, i.e., $\nabla^2 U(x)$ exists and is continuous.*
2. *$\ln p$ satisfies $\mathrm{I}_d \preceq -\nabla^2 \ln p(x) \preceq \kappa \mathrm{I}_d$.*

Furthermore, let $p_0 := p \times \mathcal{N}(0, \mathrm{I}_d)$. Then, for every $\epsilon > 0$, there exists a compact set $\mathcal{C} \subset \mathbb{R}^{2d}$ and an invertible affine-coupling network $f : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ with polynomial non-linearities, such that

$$W_1(f_\#(\mathcal{N}(0, \mathrm{I}_{2d})|_{\mathcal{C}}), p_0) \leq \epsilon.$$

Furthermore, the map defined by this affine-coupling network f is well conditioned over $\mathcal{C}$, that is, there are positive constants $A(\kappa), B(\kappa) = \kappa^{O(1)}$ such that for any unit vector w,

$$A(\kappa) \leq \|D_w f(x, v)\| \leq B(\kappa)$$

for all $(x, v) \in \mathcal{C}$, where D_w is the directional derivative in the direction w. In particular, the condition number of $Df(x, v)$ is bounded by $\frac{B(\kappa)}{A(\kappa)} = \kappa^{O(1)}$ for all $(x, v) \in \mathcal{C}$.

We make several remarks regarding the statement of the theorem:

Remark 111. The Gaussian padding (i.e. setting $p_0 = p \times \mathcal{N}(0, \mathrm{I}_d)$) is essential for our proofs. All the other prior works on the universal approximation properties of normalizing flows (with or without padding) result in ill-conditioned affine coupling networks. This gives theoretical backing of empirical observations on the benefits of Gaussian padding in [KMR20].

Remark 112. The choice of non-linearities s, t being polynomials is for the sake of convenience in our proofs. Using standard universal approximation results, they can also be chosen to be neural networks with a smooth activation function.

Remark 113. The Jacobian Df has both upper-bounded largest singular value, and lower-bounded smallest singular value—which of course bounds the determinant $\det(Df)$. As remarked in Section 7.1.1, merely bounding the ratio of the two quantities would not suffice for this. Moreover, the bound we prove *only* depends on properties of the distribution (i.e., κ), and does not worsen as $\epsilon \to 0$, in contrast to [KMR20].

Remark 114. The region $\mathcal{C}$ where the pushforward of the Gaussian through f and p_0 are close is introduced solely for technical reasons—essentially, standard results in analysis for approximating smooth functions by polynomials can only be used if the approximation needs to hold on a compact set. Note that $\mathcal{C}$ can be made arbitrarily large by making ϵ arbitrarily small.

Remark 115. We do not provide an explicit computation of the number of affine coupling blocks in the constructed network, although a bound of $\mathrm{polylog}(\epsilon)/\epsilon^{O(k)}$ can be extracted from our proofs.

Remark 116. Our proof also implies a well-conditioned universal approximation result for other related normalizing flow models. Lemma 124 proves that the flow map of underdamped Langevin dynamics is well conditioned for all $t \in [0, T]$. However, as indicated in [Che+18], underdamped Langevin dynamics is a continuous normalizing flow, thus the claim applies to such flows as well. Similarly, the particular affine coupling layers we construct in eq. (7.13) also form a residual block, so the claim also holds for residual flows [Beh+18].

7.2 Preliminaries

Our techniques leverage tools from stochastic differential equations and dynamical systems. We briefly survey the relevant results.

7.2.1 Langevin Dynamics

Broadly, Langevin diffusions are families of stochastic differential equations (SDEs) which are most frequently used as algorithmic tools for sampling from distributions specified up to a constant of proportionality. They have also recently received a lot of attention as tools for designing generative models [SE19; Son+20].

In this paper, we will only make use of *underdamped Langevin dynamics*, a momentum-like analogue of the more familiar *overdamped Langevin dynamics*, defined below. Our construction will involve simulating underdamped Langevin dynamics using affine coupling blocks.

Definition 117 (Underdamped Langevin Dynamics)**.** *Underdamped Langevin dynamics* with potential U and parameters ζ, γ is the pair of SDEs

$$\begin{cases} dx_t & = -\zeta v_t dt \\ dv_t & = -\gamma\zeta v_t dt - \nabla U(x_t)dt + \sqrt{2\gamma}\, dB_t. \end{cases} \tag{7.1}$$

The stationary distribution of the SDEs (limiting distribution as $t \to \infty$) is given by $p^*(x, v) \propto e^{-U(x)-\frac{\zeta}{2}\|v\|^2}$.

The variable v_t can be viewed as a "velocity" variable and x_t as a "position" variable – in that sense, the above SDE is an analogue to momentum methods in optimization.

The convergence of (7.1) can be bounded when the distribution $p(x) \propto \exp(-U(x))$ satisfies an analytic condition, namely has a bounded *log-Sobolev* constant. Though we don't use the log-Sobolev constant in any substantive manner in this paper, we include the definition for completeness.

Definition 118. A distribution $p : \mathbb{R}^d \to \mathbb{R}^+$ satisfies a *log-Sobolev* inequality with constant $C > 0$ if $\forall g : \mathbb{R}^d \to \mathbb{R}$, s.t. $g^2, g^2|\log g^2| \in L^1(p)$, we have

$$\mathbb{E}_p[g^2 \log g^2] - \mathbb{E}_p[g^2] \log \mathbb{E}_p[g^2] \leq 2C\mathbb{E}_p\|\nabla g\|^2. \tag{7.2}$$

In the context of Markov diffusions (and in particular, designing sampling algorithms using diffusions), the interest in this quantity comes as it governs the convergence rate of *overdamped* Langevin diffusion in the KL divergence sense. Namely, if p_t is the distribution of overdamped Langevin after time t, one can show

$$\mathrm{KL}(p_t||p) \leq e^{-Ct}\mathrm{KL}(p_0||p).$$

We will only need the following fact about the log-Sobolev constant:

Fact 119 ([BÉ85; BGL13])**.** Let the distributions $p(x) \propto \exp(-U(x))$ be such that $U(x) \succeq \lambda I$. Then, p has log-Sobolev constant bounded by λ.

We will also need the following result characterizing the convergence time of *underdamped* Langevin dynamics in terms of the log-Sobolev constant, as shown in [Ma+19]:

Theorem 120 ([Ma+19])**.** *Let $p^*(x) \propto \exp(-U(x))$ have a log-Sobolev constant bounded by ρ. Furthermore, for a distribution $p : \mathbb{R}^d \to \mathbb{R}^+$, let*

$$\mathcal{L}[p] := \mathrm{KL}(p\|p^*) + \mathbb{E}_p\left[\left\langle \nabla \frac{\delta\, \mathrm{KL}(p\|p^*)}{\delta p}, S\nabla \frac{\delta\, \mathrm{KL}(p\|p^*)}{\delta p} \right\rangle\right],$$

where S is a positive definite matrix given by $S = \frac{1}{\kappa}\begin{bmatrix} \frac{1}{4}I_{d\times d} & \frac{1}{2}I_{d\times d} \\ \frac{1}{2}I_{d\times d} & 2I_{d\times d}\end{bmatrix}$. If p_t is the distribution of (x_t, v_t) which evolve according to (7.1), we have

$$\frac{d}{dt}\mathcal{L}[p_t] \leq -\frac{\rho}{10}\mathcal{L}[p_t] \tag{7.3}$$

whenever p^* satisfies a log-Sobolev inequality with constant ρ.

We note that the above theorem uses a non-standard Lyapunov function $\mathcal{L}$, which combines KL divergence with an extra term, since the generator of underdamped Langevin is not self-adjoint—this makes analyzing the drop in KL divergence difficult. As $\mathcal{L}$ is clearly an upper bound on $KL(p||p^*)$, so it suffices to show $\mathcal{L}$ decreases rapidly.

We will also need a less-well-known *deterministic* form of the updates which is equivalent to (7.1). Precisely, we convert (7.1) an equivalent ODE (with time-dependent coefficients). The proof of this fact (via a straightforward comparison of the Fokker-Planck equation) can be found in [Ma+19].

Theorem 121. *Let $p_t(x_t, v_t)$ be the probability distribution of running (7.1) for time t. If started from $(x_0, v_0) \sim p_0$, the probability distribution of the solution (x_t, v_t) to the ODEs*

$$\frac{d}{dt}\begin{bmatrix} x_t \\ v_t \end{bmatrix} = \begin{bmatrix} O & I_d \\ -I_d & -\gamma I_d \end{bmatrix} (\nabla \ln p_t - \nabla \ln p^*) \tag{7.4}$$

is also $p_t(x_t, v_t)$.

7.2.2 Dynamical systems and Henon maps

We also build on work from dynamical systems, more precisely, a family of maps called *Hénon-like maps* [Hén76].

Definition 122 ([Tur02])**.** A pair of ODEs forms a *Hénon-like map* if it has the form

$$\begin{cases} \frac{dx}{dt} = v \\ \frac{dv}{dt} = -x + \nabla J(x) \end{cases} \tag{7.5}$$

for a smooth function $J : \mathbb{R}^d \to \mathbb{R}$.

This special family of ODEs is a continuous-time generalization of a classical discrete dynamical system of the same name [Hén76]. The property that is useful for us is that the Euler discretization of this map can be written as a sequence of affine coupling blocks.

In [Tur02], it was proven that these ODEs are *universal approximators* in some sense. Namely, the iterations of this ODE can approximate any *symplectic diffeomorphism*: a continuous map which preserves volumes (i.e. the Jacobian of the map is 1). These kinds of diffeomorphisms have their genesis in Hamiltonian formulations of classical mechanics [AM08].

At first blush, symplectic diffeomorphisms and underdamped Langevin seem to have nothing to do with each other. The connection comes through the so-called Hamiltonian representation theorem [Pol12], which states that any symplectic diffeomorphism from $\mathcal{C} \subseteq \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ can be written as the iteration of the following *Hamiltonian* system of ODEs for some (time-dependent) Hamiltonian $H(x, v, t)$:

$$\begin{cases} \frac{dx}{dt} = \frac{d}{dv} H(x, v, t) \\ \frac{dv}{dt} = -\frac{d}{dx} H(x, v, t) \end{cases} \tag{7.6}$$

In fact, in our theorem, we will use techniques inspired by those in [Tur02], who shows:

Theorem 123 ([Tur02]). *For any function $H(x, v, t) : \mathbb{R}^{2d} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ which is polynomial in (x, v), there exists a polynomial $V(x, v, t)$, s.t. the time-τ map of the system*

$$\begin{cases} \frac{dx}{dt} = \frac{\partial}{\partial v} H(x, v, t) \\ \frac{dv}{dt} = -\frac{\partial}{\partial x} H(x, v, t) \end{cases} \tag{7.7}$$

is uniformly $O(\tau^2)$-close to the time-2π map of the system

$$\begin{cases} \frac{dx}{dt} = v \\ \frac{dv_j}{dt} = -\Omega_j^2 x_j - \tau \frac{\partial}{\partial x_j} V(x, t) \end{cases} \tag{7.8}$$

for some integers $\{\Omega_i\}_{i=1}^d$.

We will prove a generalization of this theorem that applies to underdamped Langevin dynamics.

7.3 Proof Sketch of Theorem 110

7.3.1 Overview of strategy

We wish to construct an affine coupling network that (approximately) pushes forward a Gaussian $p^* = \mathcal{N}(0, \mathrm{I}_{2d})$ to the distribution we wish to model with Gaussian padding, i.e. $p_0 = p \times \mathcal{N}(0, \mathrm{I}_d)$. Because the inverse of an affine coupling network is an affine coupling network, we can invert the problem, and instead attempt to map p_0 to $N(0, \mathrm{I}_{2d})$. [1]

There is a natural map that takes p_0 to $p^* = N(0, \mathrm{I}_{2d})$, namely, underdamped Langevin dynamics (7.1). Hence, our proof strategy involves understanding and simulating underdamped Langevin dynamics with the initial distribution $p_0 = p \times \mathcal{N}(0, \mathrm{I}_d)$, and the target distribution $p^* = \mathcal{N}(0, \mathrm{I}_{2d})$, and comprises of two important steps.

First, we show that the flow-map for Langevin is well-conditioned (Lemma 124 below). Here, by flow-map, we mean the map which assigns each x to its evolution over a certain amount of time t according to the equations specified by (7.1).

Second, we break the simulation of underdamped Langevin dynamics for a certain time t into intervals of size τ, and show that the *inverse* flow-map over each τ-sized interval of time can be approximated well by a composition of affine-coupling maps (Lemma 129 below). To show this, we consider a more general system of ODEs than the one in [Tur02] (in particular, a non-Hamiltonian system), which can be applied to *underdamped* Langevin dynamics. We then show that the *inverse* flow-map of this system of ODEs can be approximated by a sequence of affine-coupling blocks. We note that for this argument, it is critical that we use underdamped rather than overdamped Langevin dynamics, as overdamped Langevin dynamics do not have the required form for affine-coupling blocks.

[1] As an aside, a similar strategy is taken in practice by recent SDE-based generative models ([Son+20]).

7.3.2 Underdamped Langevin is well-conditioned

Consider running underdamped Langevin dynamics with stationary distribution p^* equal to the standard Gaussian, started at a log-concave distribution with bounded condition number κ. The following lemma says that the flow map is well-conditioned, with condition number depending polynomially on κ.

Lemma 124. *Consider underdamped Langevin dynamics (7.1) with $\zeta = 1$, friction coefficient $\gamma < 2$ and starting distribution p which satisfies all the assumptions in Theorem 110. Let T_t denote the flow-map from time 0 to time t induced by (7.4). Then for any $x_0, v_0 \in \mathbb{R}^d$ and unit vector w, the directional derivative of T_t at x_0, v_0 in direction w satisfies*

$$\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa - 1)\right)^{-2/\gamma} \le \|D_w T_t(x_0)\| \le \left(1 + \frac{2+\gamma}{2-\gamma}(\kappa - 1)\right)^{2/\gamma}.$$

Therefore, the condition number of T_t is bounded by $\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa - 1)\right)^{4/\gamma}$.

We sketch the proof below and include a complete proof in Section B.3.

First, using (7.4) and the chain rule shows that the Jacobian of the flow map at x_0, $D_t = DT_t(x_0)$, satisfies

$$\frac{d}{dt} D_t = \begin{bmatrix} O & I_d \\ -I_d & -\gamma I_d \end{bmatrix} \nabla^2 (\ln p_t - \ln p^*) D_t, \tag{7.9}$$

i.e., it is bounded by the difference of the Hessians of the log-pdfs of the current distribution and the stationary distribution. We will show that $\nabla^2 \ln p_t$ decays exponentially towards $\nabla^2 \ln p^* = \mathrm{I}_{2d}$.

To accomplish this, consider how $\nabla^2 \ln p_t$ evolves if we replace (7.1) by its discretization,

$$\begin{aligned} \widetilde{x}_{t+\eta} &= \widetilde{x}_t + \eta \widetilde{v}_t \\ \widetilde{v}_{t+\eta} &= (1 - \eta\gamma)\widetilde{v}_t - \eta \widetilde{x}_t + \xi_t, \quad \xi_t \sim N(0, 2\gamma\eta \mathrm{I}_d). \end{aligned}$$

Note that because the stationary distribution is a Gaussian, $\nabla U(x_t) = x_t$ in (7.1), and the above equations take a particularly simple form: we apply a linear transformation to $\begin{bmatrix} \widetilde{x}_t \\ \widetilde{v}_t \end{bmatrix}$, and then add Gaussian noise, which corresponds to convolving the current distribution by a Gaussian. We keep track of upper and lower bounds for $\nabla^2 \ln p_t$, and compute how they evolve under this linear transformation and convolution by a Gaussian. Taking $\eta \to 0$, we obtain differential equations for the upper and lower bounds for $\nabla^2 \ln p_t$, which we can solve. A Grönwall argument shows that these bounds decay exponentially towards $\nabla^2 \ln p^* = \mathrm{I}_{2d}$. The decay rate can be bounded as a power of $\frac{1}{\kappa}$.

From (7.9), we then obtain that the condition number of D_t is bounded by the integral of a exponentially decaying function, and hence is bounded independent of t. In particular, we may take t large enough so that p_t is ε-close to the stationary distribution. Because the decay rate of the exponential is $\frac{1}{\kappa^{O(1)}}$, the bound is $\kappa^{O(1)}$.

Note that we vitally used the fact that the stationary distribution p is a standard Gaussian, as our argument requires that $\nabla^2 \ln p^*$ be constant everywhere.

7.3.3 ODE approximation by affine-coupling blocks

Next, we analyze a more general version of the Hamiltonian system of ODEs considered in [Tur02], which we recalled in (7.7). In particular, the system of ODEs we will be considering is:

$$\begin{cases} \frac{dx}{dt} = \frac{\partial}{\partial v} H(x, v, t) \\ \frac{dv}{dt} = -\frac{\partial}{\partial x} H(x, v, t) - \gamma \frac{\partial}{\partial v} H(x, v, t) \end{cases} \tag{7.10}$$

Note that substituting $H(x, v, t) = \ln p_t(x, v) - \ln p^*(x, v)$ above gives us the underdamped Langevin dynamics.

The first step is to restrict our considerations to H being a polynomial in x, v, rather than a general smooth function. Towards this, we recall the notion of closeness in the C^1 topology:

Definition 125. Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact set. Let $f, g : \mathcal{C} \to \mathbb{R}$ be two continuously differentiable functions. Then we say that f, g are uniformly ϵ-close over $\mathcal{C}$ in C^1 topology if

$$\sup_{x \in \mathcal{C}} (\|f(x) - g(x)\| + \|Df(x) - Dg(x)\|) \leq \epsilon$$

The following lemma (a generalization of the Stone-Weierstrass Theorem) then establishes that it suffices to focus on H being polynomial in x, v:

Lemma 126 (Theorem 5, [Pee07])**.** *Let $\mathcal{C} \subset \mathbb{R}^d$ be a compact set. For any C^2 function $H : \mathbb{R}^d \to \mathbb{R}$, and any $\epsilon > 0$, there is a multivariate polynomial $P : \mathbb{R}^d \to \mathbb{R}$ such that P, H are uniformly ϵ-close over $\mathcal{C}$ in C^1 topology.*

Focusing on the case of polynomials, Lemma 127 below shows that instead of flowing the pair of ODEs given by (7.10) over an interval of time τ, we can instead run a different ODE for time 2π, such that the flow-maps corresponding to both these ODEs are $O(\tau^2)$-close.

Lemma 127. *Let $\mathcal{C} \subset \mathbb{R}^{2d}$ be a compact set. For any function $H(x, v, t) : \mathbb{R}^{2d} \to \mathbb{R}$ which is polynomial in (x, v), there exist polynomial functions J, F, G, s.t. the time-$(t_0 + \tau, t_0)$ flow map of the system*

$$\begin{cases} \frac{dx}{dt} = \frac{\partial}{\partial v} H(x, v, t) \\ \frac{dv}{dt} = -\frac{\partial}{\partial x} H(x, v, t) - \gamma \frac{\partial}{\partial v} H(x, v, t) \end{cases} \tag{7.11}$$

is uniformly $O(\tau^2)$-close over $\mathcal{C}$ in C^1 topology to the time-2π map of the system

$$\begin{cases} \frac{dx}{dt} = v - \tau F(v, t) \odot x \\ \frac{dv_j}{dt} = -\Omega_j^2 x_j - \tau J_j(x, t) - \tau v_j G_j(x, t) \end{cases} \tag{7.12}$$

Here, $\odot$ denotes component-wise product, and the constants inside the $O(\cdot)$ depend on $\mathcal{C}$ and the coefficients of H.

The complete proof of this lemma is included in Appendix B.4; we provide a brief sketch here. First, we consider the first order ($O(\tau^2)$) approximation of the flow map of a standard ODE of the form $\dot{y} = Dy$ (where D is diagonal), and observe that for small τ, we can think of (7.12) as a perturbed version of such an ODE with an appropriate choice of D. Using standard ODE perturbation techniques, we can approximately express the time-t evolution of (7.12) up to first-order in τ, in terms of polynomials F, G, J and trigonometric functions.

Then, we compare this map to the first-order approximation of flowing the pair of ODEs (7.11) for time τ via Taylor's theorem. Furthermore, this approximation is a polynomial in (x, v) since H is a polynomial in (x, v).

The crucial step involves choosing the functional form of $F(z,t), J(z,t), G(z,t)$ suitably, so that they are polynomials in z with coefficients in terms of $\sin(\Omega t)$, $\cos(\Omega t)$. After simplification, both expressions can be expressed in terms of polynomials in x, v where coefficients can be expressed in terms of $\int_0^{2\pi} \sin^p(\Omega s)\cos^q(\Omega s)\, ds$, which either integrate to 0 or a constant. Thus, to ensure that the two approximations match, we are left with a problem of making two multivariate polynomials in (x, v) equal.

This final step can of course be written as a linear system of equations. We identify a special structure in this system, which helps us show that the system is full-rank, and hence has a solution. □

Finally, consider discretizing the newly constructed ODE (7.12) into small steps of size η by a simple Euler schema i.e.,

$$\begin{cases} x_{n+1} = x_n + \eta(v_n - \tau F(v_n, \eta n) \odot x_n) \\ v_{n+1,j} = v_{n,j} - \eta(\Omega_j^2 x_{n,j} - \tau J_j(x_n, \eta n) - \tau v_{n,j} G_j(x_n, \eta n)) \end{cases} \tag{7.13}$$

We note that each step above can be written as a composition of two affine coupling blocks given by $(x_n, v_n) \mapsto (x_n, v_{n+1}) \mapsto (x_{n+1}, v_{n+1})$. Namely, the map $(x_n, v_n) \mapsto (x_n, v_{n+1})$ can be written as

$$\begin{cases} x_n = x_n \\ v_{n+1} = v_n \odot (1-\tau) G(x_n, \eta n) - \eta(\Omega^2 \odot x_n - \tau J(x_n, \eta n)) \end{cases}$$

This map is an affine coupling block with $s(x_n) = (1-\tau) \odot G(x_n, \eta n)$ and $t(x_n) = -\eta(\Omega^2 \odot x_n - \tau J(x_n, \eta n))$. The map $(x_n, v_{n+1}) \mapsto (x_{n+1}, v_{n+1})$ can be written as

$$\begin{cases} v_{n+1} = v_{n+1} \\ x_{n+1} = x_n + \eta(v_{n+1} - \tau F(v_{n+1}, \eta n) \odot x_n) \end{cases}$$

which is an affine coupling block with $s(v_{n+1}) = 1 - \eta\tau F(v_{n+1}, \eta n)$ and $t(v_{n+1}) = \eta v_{n+1}$.

The composition of the two maps above yields an affine coupling network $(x_n, v_n) \mapsto (x_{n+1}, v_{n+1})$ precisely as given by Equation (7.13) with non-linearities s, t in each of the blocks given by polynomials. The following lemma bounds the error resulting from this discretization:

Lemma 128 (Euler's discretization method)**.** [2] *Let $\mathcal{C} \subset \mathbb{R}^{2d}$ be a compact set. Consider discretizing the time from 0 to t into $\frac{t}{\eta}$ steps and performing the update given by (7.13) at each of these steps.*

[2] This result is well known in the C^0 topology, we provide an analysis for the C^1 bound in Section B.5.1.

Let the map obtained as a result of discretizing thus be denoted by T'_t and let the original flow map be denoted by T_t. Then T_t and T'_t are uniformly $O(\eta)$ close over $\mathcal{C}$ in C^1 topology, and the constants inside the $O(\cdot)$ depend on $\mathcal{C}$, and bounds on the derivatives of T_t over $\mathcal{C}$.

7.3.4 Simulating by breaking into τ-sized intervals

Let $T_{s,t}$ denote the time-s,t flow-map of (7.10) from time s to time t. Since the flow maps are invertible, $T_{s,t}$ and $T_{t,s}$ are inverses. We are now ready to state the following lemma which says that the underdamped Langevin flow-map $T_{\phi,0}$ can be written as a composition of affine-couplings maps:

Lemma 129. *Let $\mathcal{C} \subset \mathbb{R}^{2d}$ be a compact set. Suppose that $T_{\phi,0}(x, v)$ is the time-$(\phi, 0)$ flow-map of the ODE's*

$$\begin{cases} \frac{dx}{dt} = \frac{\partial}{\partial v} H(x, v, t) \\ \frac{dv}{dt} = -\frac{\partial}{\partial x} H(x, v, t) - \gamma \frac{\partial}{\partial v} H(x, v, t) \end{cases} \tag{7.14}$$

where H is C^∞. Then for any $\epsilon_1, \phi \in \mathbb{R}_+$, there exists an integer $N = N(\epsilon_1, \phi, \mathcal{C})$ and affine-coupling blocks $f_1, \dots, f_N$ such that the composition $f = f_N \circ \cdots \circ f_1$ is ϵ_1-close to $T_{\phi,0}$ in the C^1 topology over $\mathcal{C}$.

The proof of Lemma 129 is in Section B.5. We provide a brief sketch here: from Lemma 126, we know that it suffices to show the result for a polynomial H. Thereafter, we break the time for which we want to flow the ODE given by (7.14) into small chunks of length τ. Lemmas 127 and 128 then show that the flow map over this chunk can be written as an affine coupling network. Composing the affine coupling networks over all the chunks of time gives us the result.

7.3.5 Putting components together

The previous sections established that for any t and any compact set $\mathcal{C}$, there is a affine-coupling network f with polynomial non-linearities such that $T_{t,0}$ and f are uniformly close over $\mathcal{C}$. We will now pick an appropriate value of t and set $\mathcal{C}$ such that $W_1(f_\#(p^*|_\mathcal{C}), p_0) \leq \epsilon$ where $p^* = \mathcal{N}(0, \mathrm{I}_{2d})$, which is the required result of Theorem 110. First, using Theorem 120, for

$$\phi > -10 \log \epsilon_1 + \log 2 + \log \mathcal{L}[p_0]$$

we have that $\mathrm{KL}(T_{0,\phi\#}(p_0), p^*) \leq \frac{\epsilon_1^2}{2}$. We use the following *transportion cost inequality* to convert this to a Wasserstein bound.

Theorem 130 (Talagrand [Tal96]). *The standard Gaussian p on $\mathbb{R}^d$ satisfies a* transportation cost inequality*: For every distribution q on $\mathbb{R}^d$ with finite second moment, $W_1(p, q)^2 \leq 2KL(q\|p)$.*

This gives us that $W_1(T_{0,\phi\,\#}(p_0), p^*) \leq \epsilon_1$. A simple argument in Lemma 186 (Section B.5.2) then gives

$$W_1(p_0, T_{\phi,0\,\#}(p^*)) = W_1(T_{\phi,0\,\#}(T_{0,\phi\,\#}(p_0)), T_{\phi,0\,\#}(p^*)) \leq \mathrm{Lip}(T_{\phi,0})\epsilon_1 \tag{7.15}$$

A subsequent argument stated as Lemma 187 in Section B.5.2, shows that if f and $T_{\phi,0}$ are uniformly ϵ_1-close in C^0 topology on some $\mathcal{C}$, then their pushforwards through $p^*|_{\mathcal{C}}$ are indeed close, i.e.,

$$W_1(T_{\phi,0\,\#}(p^*|_{\mathcal{C}}), f_\#(p^*|_{\mathcal{C}})) \leq \epsilon_1. \tag{7.16}$$

Next, we establish a bound on the Wasserstein distance between the standard Gaussian and its truncation on a compact set, proved in Section B.5.3.

Lemma 131. *Let $p^* = \mathcal{N}(0, \mathrm{I}_{2d})$. Then for every $\delta \in \mathbb{R}_+$, there exists a compact set $\mathcal{C} = B(0, R)$ such that $W_1(p^*, p^*|_{\mathcal{C}}) \leq \delta$, where $B(0, R)$ denotes the ball of radius R centered at the origin.*

We now choose a compact set $\mathcal{C}$ such that Lemma 131 holds for $\delta = \epsilon_1$. Then Lemma 186 again implies that

$$W_1(T_{\phi,0\,\#}(p^*), T_{\phi,0\,\#}\ (p^*|_{\mathcal{C}})) \leq \mathrm{Lip}(T_{\phi,0})\epsilon_1 \tag{7.17}$$

Equations (7.15), (7.16), (7.17) and the triangle inequality together imply

$$W_1(f_\#(p^*|_{\mathcal{C}}), p_0) \leq (2\mathrm{Lip}(T_{\phi,0}) + 1)\epsilon_1 \leq \epsilon$$

for small enough ε_1. We can indeed set ϵ_1 small enough so as to satisfy the last inequality above, because of the global bound $\mathrm{Lip}(T_{\phi,0}) \leq \left(1 + \frac{2+\gamma}{2-\gamma}(\kappa - 1)\right)^{2/\gamma}$ established in Lemma 124. This gives us the statement of Theorem 110. Note that the final value of ϕ depends on ϵ, κ, γ and $\mathcal{L}[p_0]$.

7.4 Related Work

The landscape of normalizing flow models is rather rich. The inception of the ideas was in [RM15] and [DKB14], and in recent years, an immense amount of research has been dedicated to developing different architectures of normalizing flows. The focus of this paper are affine coupling flows, which were introduced in [DKB14], introduced the idea of using pushforward maps with triangular Jacobians for computational efficiency. This was further developed in [DSB16] and culminated in [KD18], who introduced 1x1 convolutions in the affine coupling framework to allow for "trainable" choices of partitions. We note, there have been variants of normalizing flows in which the Jacobian is non-triangular, e.g. [Gra+18; DDT19; Beh+18], but these models still don't scale beyond datasets the size of CIFAR-10.

In terms of theoretical results, the most closely related works are [HDC20; Zha+20; KMR20]. The former two show universal approximation of affine couplings—albeit if the input is padded with zeros. This of course results in maps with singular Jacobians, which is why this strategy isn't used in practice. [KMR20] show universal approximation without padding—though their constructions results in a flow model with condition number $1/\epsilon$ to get approximation ϵ in the Wasserstein sense, regardless of how well-behaved the distribution to be approximated is. Furthemore, [KMR20] provide some empirical evidence that padding with iid Gaussians (as in our paper) is better than both zero padding (as in [HDC20; Zha+20]) and no padding on small-scale data.

7.5 Conclusion

In this paper, we provide the first guarantees on universal approximation with *well-conditioned* affine coupling networks. The conditioning of the network is crucial when the networks are trained using gradient-based optimization of the likelihood. Mathematically, we uncover connections between stochastic differential equations, dynamical systems and affine coupling flows. Our construction uses Gaussian padding, which lends support to the empirical observation that this strategy tends to result in better-conditioned flows [KMR20]. We leave it as an open problem to generalize beyond log-concave distributions.

Chapter 8

Robust subspace approximation in stream

A fundamental problem in large-scale machine learning is that of subspace approximation. Given a set of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$ and an integer k, we wish to find a linear subspace S of dimension k for which $\sum_i M(\operatorname{dist}(S, a_i))$ is minimized, where $\operatorname{dist}(S, x) := \min_{y \in S} \|x - y\|_2$, and $M(\cdot)$ is some loss function. When $M(\cdot) = (\cdot)^2$, this is the well-studied least squares subspace approximation problem. The minimizer in this case can be computed exactly by computing the truncated SVD of the data matrix.

Otherwise M is often chosen from $(\cdot)^p$ for some $p \geq 0$, or from a class of functions called M-estimators, with the goal of providing a more robust estimate than least squares in the face of outliers. Indeed, for $p < 2$, since one is not squaring the distances to the subspace, one is placing less emphasis on outliers and therefore capturing more of the remaining data points. For example, when M is the identity function, we are finding a subspace so as to minimize the sum of distances to it, which could arguably be more natural than finding a subspace so as to minimize the sum of squared distances. We can write this problem in the following form:

$$\min_{S \dim k} \sum_i \operatorname{dist}(S, a_i) = \min_{X \operatorname{rank} k} \sum_i \|(A - AX)_{i*}\|_2$$

where A is the matrix in which the i-th row is the vector a_i. This is the form of robust subspace approximation that we study in this work. We will be interested in the approximate version of the problem for which the goal is to output a k-dimensional subspace S' for which with high probability,

$$\sum_i \operatorname{dist}(S', a_i) \leq (1 + \epsilon) \sum_i \operatorname{dist}(S, a_i) \tag{8.1}$$

The particular form with M equal to the identity was introduced to the machine learning community by Ding et al. [Din+06], though these authors employed heuristic solutions. The series of work in [DTV11],[Gur+10] and [DV07a; Fel+10a; SV12; CW15a] shows that if $M(\cdot) = |\cdot|^p$ for $p \neq 2$, there is no algorithm that outputs a $(1 + 1/\operatorname{poly}(d))$ approximation to this problem unless P = NP. However, [CW15a] also show that for any p there is an algorithm that runs in $O(\mathrm{nnz}(A) + (n + d)\,\mathrm{poly}(k/\epsilon) + \exp(\mathrm{poly}(k/\epsilon))$ time and outputs a k-dimensional subspace whose cost is within a $(1 + \epsilon)$ factor of the optimal solution cost. This provides a considerable computational savings since in most applications $k \ll d \ll n$. Their work builds upon techniques developed in [Fel+10b] and [FL11] which give $O\left(nd \cdot \mathrm{poly}(k/\epsilon) + \exp\left((k/\epsilon)^{O(p)}\right)\right)$ time algorithms for the $p \geq 1$ case. These in turn build on the weak coreset construction of [DV07b]. In other related work [CW15b] give algorithms for performing regression with a variety of M-estimator loss functions.

Our Contributions. We give the first sketching-based solution to this problem. Namely, we show it suffices to compute $Z \cdot A$, where Z is a $d \log n\,\mathrm{poly}(k\epsilon^{-1}) \times n$ random matrix with entries chosen obliviously to the entries of A. The matrix Z is a block matrix with blocks consisting of independent Gaussian entries, while other blocks consist of independent Cauchy random variables, and yet other blocks are sparse matrices with non-zero entries in $\{-1, 1\}$. Previously such sketching-based solutions were known only for $M(\cdot) = (\cdot)^2$. Prior algorithms [DV07a; Fel+10a; SV12; CW15a] also could not be implemented as single-shot sketching algorithms since they require first making a pass over the data to obtain a crude approximation, and then using (often adaptive) sampling methods in future passes to refine to a $(1 + \epsilon)$-approximation. Our sketching-based algorithm, achieving $O(\mathrm{nnz}(A)+(n+d)\,\mathrm{poly}(k/\epsilon)+\exp(\mathrm{poly}(k/\epsilon))$ time, matches the running time of previous algorithms, but has considerable benefits as described below.

Streaming Model. Since Z is linear and oblivious, one can maintain $Z \cdot A$ in the presence of insertions and deletions to the entries of A. Indeed, given the update $A_{i,j} \leftarrow A_{i,j} + \Delta$ for some $\Delta \in \mathbb{R}$, we simply update the j-th column ZA_j in our sketch to $ZA_j + \Delta \cdot Z \cdot e_i$, where e_i is the i-th standard unit vector. Also, the entries of Z can be represented with limited independence, and so Z can be stored with a short random seed. Consequently, we obtain the first algorithm with $d \log n\,\mathrm{poly}(k\epsilon^{-1})$ memory for this problem in the standard turnstile data stream model [Mut05]. In this model, $A \in \mathbb{R}^{n \times d}$ is initially the zero matrix, and we receive a stream of updates to A where the i-th update is of the form (x_i, y_i, c_i), which means that A_{x_i,y_i} should be incremented by c_i. We are allowed one pass over the stream, and should output a rank-k matrix X' which is a $(1 + \epsilon)$ approximation to the robust subsace estimation problem, namely $\sum_i \|(A - AX')_{i*}\|_2 \leq (1 + \epsilon) \min_{X \text{ rank } k} \sum_i \|(A - AX)_{i*}\|_2$. The space complexity of the algorithm is the total number of words required to store this information during the stream. Here, each word is $O(\log(nd))$ bits. Our algorithm achieves $d \log n\,\mathrm{poly}(k\epsilon^{-1})$ memory, and so only logarithmically depends on n. This is comparable to the memory of streaming algorithms when $M(\cdot) = (\cdot)^2$ [CW09; Gha+16], which is the only prior case for which streaming algorithms were known.

Distributed Model. Since our algorithm maintains $Z \cdot A$ for an oblivious linear sketch Z, it is parallelizable, and can be used to solve the problem in the distributed setting in which there are s machines holding $A^1, A^2, \ldots, A^s$, respectively, and $A = \sum_{i=1}^{s} A^i$. This is called the *arbitrary partition model* [KVW14]. In this model, we can solve the problem in one round with $s \cdot d \log n\,\mathrm{poly}(k\epsilon^{-1})$ communication by having each machine agree upon (a short seed describing) Z, and sending ZA^i to a central coordinator who computes and runs our algorithm on $Z \cdot A = \sum_i ZA^i$. The arbitrary partition model is stronger than the so-called row partition model, in which the points (rows of A) are partitioned across machines. For example, if each machine corresponds to a shop, the rows of A correspond to customers, the columns of A correspond to items, and $A^i_{c,d}$ indicates how many times customer c purchased item d at shop i, then the row partition model requires customers to make purchases at a single shop. In contrast, in the arbitrary partition model, customers can purchase items at multiple shops.

8.1 Notation and Terminology

For a matrix A, let A_{i*} denote the i-th row of A, and A_{*j} denote the j-th column of A.

Definition 132. For a matrix $A \in \mathbb{R}^{n\times m}$, let:

$$\|A\|_{2,1} \equiv \sum_i \|A_{i*}\|_2 \qquad \|A\|_{1,2} \equiv \left\|A^T\right\|_{2,1} = \sum_j \|A_{*j}\|_2$$

$$\|A\|_F \equiv \sqrt{\sum_i \|A_{i*}\|_2^2} \qquad \|A\|_{1,1} \equiv \sum_i \|A_{i*}\|_1 \qquad \|A\|_{\mathrm{med},1} \equiv \sum_j \|A_{*j}\|_{\mathrm{med}}$$

where $\|\cdot\|_{\mathrm{med}}$ denotes the function that takes the median of absolute values.

Definition 133 (X^*, Δ^*)**.** Let:

$$\Delta^* \equiv \min_{X \text{ rank } k} \|A - AX\|_{2,1} \qquad X^* \equiv \operatorname*{argmin}_{X \text{ rank } k} \|A - AX\|_{2,1}$$

Definition 134. For a matrix $A \in \mathbb{R}^{n\times d}$ and a target rank k, W is an (α, β)-coreset if its row space is an α-dimensional subspace of $\mathbb{R}^d$ that contains a β-approximation to X^*. Formally:

$$\operatorname*{argmin}_{X \text{ rank } k} \|A - AXW\|_{2,1} \le \beta\Delta^*$$

We also use the following notation: $[n]$ denotes the set $\{1, 2, 3, \cdots n\}$. $[\![E]\!]$ denotes the indicator function for event E. $\mathrm{nnz}(A)$ denotes the number of non-zero entries of A. A^- denotes the pseudoinverse of A.

8.2 Algorithm Overview

At a high level we follow the framework put forth in [CW15a] which gives the first input sparsity time algorithm for the robust subspace approximation problem. In their work Clarkson and Woodruff first find a crude $(\mathrm{poly}(k), K)$-coreset for the problem. They then use a non-adaptive implementation of a residual sampling technique from [DV07b] to improve the approximation quality but increase the dimension, yielding a $(K\,\mathrm{poly}(k), 1 + \epsilon)$-coreset. From here they further use dimension reducing sketches to reduce to an instance with parameters that depend only polynomially on k/ϵ. Finally they pay a cost exponential only in $\mathrm{poly}(k/\epsilon)$ to solve the small problem via a black box algorithm of [BPR94].

There are several major obstacles to directly porting this technique to the streaming setting. For one, the construction of the crude approximation subspace uses leverage score sampling matrices which are non-oblivious and thus not usable in 1-pass turnstile model algorithms. We circumvent this difficulty in Section 8.3.1 by showing that if T is a sparse $\text{poly}(k) \times n$ matrix of Cauchy random variables, the row span of TA contains a rank-k matrix which is a $\log(d)\,\text{poly}(k)$ approximation to the best rank-k matrix under the $\|\cdot\|_{2,1}$ norm.

Second, the residual sampling step requires sampling rows of A with respect to probabilities proportional to their distance to the crude approximation (in our case TA). This is challenging because one does not know TA until the end of the stream, much less the distances of rows of A to TA. We handle this in Section 8.3.2 using a row-sampling data structure of [MW10] developed for regression, which for a matrix B maintains a sketch HB in a stream from which one can extract samples of rows of B according to probabilities given by their norms. By linearity, it suffices to maintain HA and TA in parallel in the stream, and apply the sample extraction procedure to $HA \cdot (\text{Id} - P_{TA})$, where $P_{TA} = TA((TA)^T TA)^-(TA)^T$ is the projection onto the rowspace of TA. Unfortunately, the extraction procedure only returns noisy perturbations of the original rows which majorly invalidates the analysis in [CW15a] of the residual sampling. In Section 8.3.2 we give a novel analysis of non-adaptive noisy residual sampling which we name BootstrapCoreset. This is one of our key contributions and may be of independent interest. This gives a procedure for transforming our $\text{poly}(k)$-dimensional space containing a $\log(d)\,\text{poly}(k)$ approximation into a $\text{poly}(k)\log(d)$-dimensional space containing a $3/2$ factor approximation.

Third, requiring the initial crude approximation to be oblivious yields a coarser $\log(d)\,\text{poly}(k)$ initial approximation than the constant factor approximation of [CW15a]. Thus the dimension of the subspace after residual sampling is $\text{poly}(k)\log(d)$. Applying dimension reduction techniques reduces the problem to instance with $\text{poly}(k)$ rows by $\log(d)\,\text{poly}(k)$ columns. Here the black box algorithm of [BPR94] would take time $d^{\text{poly}(k)}$ which is no longer fixed parameter tractable as desired. Our key insight is that finding the best rank-k matrix under the Frobenius norm, which can be done efficiently, is a $\sqrt{\log d}\log\log d\,\text{poly}(k)$ approximation to the $\|\cdot\|_{2,1}$ norm minimizer. From here we can repeat the residual sampling argument which this time yields a small instance with $\text{poly}(k)$ rows by $\sqrt{\log d}\log\log d\,\text{poly}(k/\epsilon)$ columns. Sublogarithmic in d makes all the difference and now enumerating can be done in time $(n + d)\,\text{poly}(k/\epsilon) + \exp(\text{poly}(k/\epsilon)$. All this is done in parallel in a single pass of the stream.

Lastly, the sketching techniques applied after the residual sampling are not oblivious in [CW15a]. We instead use an obvlious median based embedding in Section 8.4.1, and show that we can still use the black box algorithm of [BPR94] to find the minimizer under the $\|\cdot\|_{\text{med},1}$ norm in Section 8.4.2.

We present our results as two algorithms for the robust subspace approximation problem. The first runs in fully polynomial time but gives a coarse approximation guarantee, which corresponds to stopping before repeating the residual sampling a second time. The second algorithm captures the entire procedure, and uses the first as a subroutine.

Algorithm 5 CoarseApprox

Input: $A \in \mathbb{R}^{n\times d}$ as a stream
Output: $X \in \mathbb{R}^{d\times d}$ such that $\|A - AX\|_{2,1} \le \sqrt{\log d}\log\log d\,\mathrm{poly}(k)\Delta^*$

$T \in \mathbb{R}^{\mathrm{poly}(k)\times n} \leftarrow$ Sparse Cauchy matrix // as in Thm. 137
$C_1 \in \mathbb{R}^{\mathrm{poly}(k)\times n} \leftarrow$ Sparse Cauchy matrix // as in Thm. 149
$S_1 \in \mathbb{R}^{\log d\cdot\mathrm{poly}(k/\epsilon)\times d} \leftarrow$ Count Sketch composed with Gaussian // as in Thm. 146
$R_1 \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times d} \leftarrow$ Count Sketch composed with Gaussian // as in Thm. 146
$G_1 \in \mathbb{R}^{\log d\cdot\mathrm{poly}(k/\epsilon)\times\log d\cdot\mathrm{poly}(k/\epsilon)} \leftarrow$ Gaussian matrices // as in Thm. 149
Compute TA online
Compute C_1A online
$U^T \in \mathbb{R}^{\log d\,\mathrm{poly}(k)\times d} \leftarrow$ BootstrapCoreset$(A, TA, 1/2)$ // as in Alg. 7
$\hat{X} \in \mathbb{R}^{\mathrm{poly}(k)\times\log d\,\mathrm{poly}(k)} \leftarrow \mathrm{argmin}_{X\,\mathrm{rank}\,k}\left\|C_1(A - AR_1^TXU^T)S_1^TG_1\right\|_F$ // as in Fact 150
return $R_1^T\hat{X}U^T$

Theorem 135 (Coarse Approximation in Polynomial Time)**.** *Given a matrix $A \in R^{n\times d}$, Algorithm 5 is a one-pass streaming algorithm that with constant probability computes a rank k matrix $X \in \mathbb{R}^{d\times d}$ such that:*

$$\|A - AX\|_{2,1} \le \sqrt{\log d}\log\log d\cdot\mathrm{poly}(k)\cdot\|A - AX^*\|_{2,1}$$

that runs in space $O(d\log n\,\mathrm{poly}(k))$ and runs in time $O(\mathrm{nnz}(A) + (n+d)\,\mathrm{poly}(k))$.

Proof Sketch We show the following are true in subsequent sections:

1. The row span of TA is a $(\mathrm{poly}(k), \log d\cdot\mathrm{poly}(k))$-coreset for A (Section 8.3.1) with probability $24/25$.

2. BootstrapCoreset$(A, TA, 1/2)$ is a $(\log d\cdot\mathrm{poly}(k), 3/2)$-coreset with probability $99/100$ (Section 8.3.2).

3. If:

$$\hat{X} = \underset{X\,\mathrm{rank}\,k}{\mathrm{argmin}}\left\|C_1AS_1^TG_1 - C_1AR_1^TXU^TS_1^TG_1\right\|_F$$

then with probability $47/50$:

$$\left\|A - AR_1^T\hat{X}U^T\right\|_{2,1} \le \mathrm{poly}(k/\epsilon)\sqrt{\log d}\log\log d\cdot\Delta^*$$

(Sections 8.3.3 and 8.3.4).

By a union bound, with probability $89/100$ all the statements above hold, and the theorem is proved. $\square$

Algorithm 6 $(1+\epsilon)$-Approx

Input: $A \in \mathbb{R}^{n \times d}$ as a stream
Output: $X \in \mathbb{R}^{d \times d}$ such that $\|A - AX\|_{2,1} \leq (1+\epsilon)\Delta^*$

$\hat{X} \in \mathbb{R}^{\mathrm{poly}(k) \times \log d \, \mathrm{poly}(k)} \leftarrow$ CoarseApprox(A) // as in Thm. 135
$C_2 \in \mathbb{R}^{\mathrm{poly}(k/\epsilon) \times n} \leftarrow$ Sparse Cauchy matrix // as in Thm. 151
$S_2 \in \mathbb{R}^{\log d \cdot \mathrm{poly}(k/\epsilon) \times d} \leftarrow$ Count Sketch composed with Gaussian // as in Thm. 146
$R_2 \in \mathbb{R}^{\mathrm{poly}(k/\epsilon) \times d} \leftarrow$ Count Sketch composed with Gaussian // as in Thm. 146
$G_2 \in \mathbb{R}^{\log d \cdot \mathrm{poly}(k/\epsilon) \times \log d \cdot \mathrm{poly}(k/\epsilon)} \leftarrow$ Gaussian matrices // as in Thm. 151
Compute $C_2 A$ online
Let $V \in \mathbb{R}^{\log d \, \mathrm{poly}(k) \times k}$ be such that $\hat{X} = WV^T$ is the rank-k decomposition of $\hat{X}$
$U'^T \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\sqrt{\log d} \log\log d \times d} \leftarrow$ BootstrapCoreset$(A, V^T U^T, \epsilon)$ // as in Alg. 7
$\hat{X}' \in \mathbb{R}^{poly(k/\epsilon) \times \mathrm{poly}(k/\epsilon)\sqrt{\log d} \log\log d} \leftarrow \mathrm{argmin}_{X \text{ rank } k} \left\| C_2(A - AR_2^T X U'^T) S_2^T G_2 \right\|_{\mathrm{med},1}$ // as in Thm. 154
return $R_2^T \hat{X}' U'^T$

Theorem 136 ($(1+\epsilon)$-Approximation)**.** *Given a matrix $A \in R^{n \times d}$, Algorithm 6 is a one-pass streaming algorithm that with constant probability computes a rank k matrix $X \in \mathbb{R}^{d \times d}$ such that:*

$$\|A - AX\|_{2,1} \leq (1+\epsilon)\|A - AX^*\|_{2,1}$$

that runs in space $O\left(d \log(n) \,\mathrm{poly}(k/\epsilon)\right)$ and runs in time $O(\mathrm{nnz}(A) + (n+d)\,\mathrm{poly}(k/\epsilon) + \exp(\mathrm{poly}(k/\epsilon)))$.

Proof Sketch We show the following are true in subsequent sections:

1. If V is such that $\hat{X} = WV^T$, then V^T is a $(\mathrm{poly}(k), \mathrm{poly}(k)\sqrt{\log d} \log\log d)$-coreset with probability 89/100 (Theorem 135).

2. BootstrapCoreset$(A, V^T U^T, \epsilon')$ is a $(\mathrm{poly}(k/\epsilon')\sqrt{\log d} \log\log d, (1+\epsilon'))$-coreset with probability 99/100 (Reusing Section 8.3.2).

3. If:
$$\hat{X}' \leftarrow \underset{X}{\mathrm{argmin}} \left\| C_2(A - AR_2^T X U'^T) S_2^T G_2 \right\|_{\mathrm{med},1}$$
then with probability 19/20:
$$\left\| A - AR_2^T \hat{X}' U^T \right\|_{2,1} \leq (1 + O(\epsilon'))\Delta^*$$
(Reusing Section 8.3.3 and Section 8.4.1).

4. A black box algorithm of [BPR94] computes $\hat{X}'$ to within $(1 + O(\epsilon'))$ (Section 8.4.2).

By a union bound, with probability 83/100 all the statements above hold. Setting ϵ' appropriately small as a function of ϵ, the theorem is proved. □

We give further proofs and details of these theorems in subsequent sections. Refer to the supplementary materials for all the details, and for details regarding the streaming implementation.

8.3 Coarse Approximation

8.3.1 Initial Coreset Construction

We construct a $(\text{poly}(k), O(\log d))$-coreset which will serve as our starting point.

Theorem 137. *If $T \in \mathbb{R}^{\text{poly}(k)\times n}$ is a matrix of i.i.d. Cauchy random variables, then the row space of TA contains a k dimensional subspace with corresponding projection matrix X' such that with probability 24/25:*

$$\|A - AX'\|_{2,1} \leq O(\log d) \min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

Proof. In order to deal with the awkward $\|\cdot\|_{2,1}$ norm, we make use of a well known theorem due to Dvoretzky to convert it into an entrywise 1-norm.

Fact 138 (Dvoretzky's Theorem (Special Case), Section 3.3 of [Ind01])**.** There exists an appropriately scaled Gaussian Matrix $G \in \mathbb{R}^{d\times \frac{d\log(1/\epsilon)}{\epsilon^2}}$ such that w.h.p. the following holds for all $y \in \mathbb{R}^d$ simultaneously

$$\left\|y^T G\right\|_1 \in (1 \pm \epsilon)\left\|y^T\right\|_2$$

Applying this to all rows at once: $\|AX - A\|_{2,1} \in (1 \pm \epsilon)\|AXG - AG\|_{1,1}$.
We also use some existing machinery for input sparsity time ℓ_1 subspace embeddings.

Fact 139 (Theorem 4 from [MM12])**.** For any given $D \in \mathbb{R}^{s\times t}$, let $\Pi \in \mathbb{R}^{r\times s}$ be a random Sparse Cauchy matrix with $r = O(t^5 \log^5 t)$ defined as follows: $\Pi = SC$ where $S \in \mathbb{R}^{r\times s}$ has each column uniformly and independently chosen from the r standard basis vectors in $\mathbb{R}^r$, and where $C \in \mathbb{R}^{s\times s}$ is a diagonal matrix with diagonal entries chosen independently from the standard Cauchy distribution. Then with probability 99/100 simultaneously for all $x \in \mathbb{R}^t$:

$$\frac{1}{O(t^2 \log^2 t)} \cdot \|Dx\|_1 \leq \|\Pi Dx\|_1 \leq O(t \log t) \cdot \|Dx\|_1$$

Fact 140 (Lemma D.25 from [SWZ16])**.** If $\Pi \in \mathbb{R}^{r\times s}$ is a Sparse Cauchy matrix as defined above, and $B \in \mathbb{R}^{s\times t}$ is a fixed matrix, then with probability at least 99/100:

$$\|\Pi B\|_1 \leq O(\log(rt))\|B\|_1$$

Finally, we also need a couple of structural lemmas which we state here without proof:

Lemma 141 (Lemma 29 from [CW15a])**.** *For a fixed (B, D) pair such that $B \in \mathbb{R}^{r\times s}, D \in R^{r\times t}$, if $S \in \mathbb{R}^{s/\text{poly}(\epsilon)\times r}$ is a CountSketch Matrix composed with a matrix of i.i.d. Gaussians (for background on such sketching matrices, we refer the reader to the monograph [Woo14]), then with probability 99/100 both of the properties below hold:*

1. $\|S(BX - D)\|_{1,2} \geq (1 - \epsilon)\|BX - D\|_{1,2}$ *for any X.*
2. *If $X^* = \text{argmin}_{X \text{ rank } k}\|BX - D\|_{1,2}$, then $\|S(BX^* - D)\|_{1,2} \leq (1 + \epsilon)\|BX^* - D\|_{1,2}$.*

Clarkson and Woodruff [CW15a] call such an S a lopsided embedding for (B, D) with respect to the $(1,2)$-norm.

Lemma 142 (Lemma 31 from [CW15a])**.** *If R is a lopsided embedding for (A_k^T, A^T), then:*

$$\min_{X \text{ rank } k} \left\|AR^T X - A\right\|_{2,1} \leq (1+3\epsilon)\Delta^*$$

Let $X' = \operatorname{argmin}_X \left\|TAR^T X - TA\right\|_{2,1}$, $R \in \mathbb{R}^{d\times O(k)}$ as in the lemma above and $\epsilon = O(1)$.

Define E_1 to be the event that the condition in Dvoretzky's theorem is satisfied, E_2 to be the event that Fact 139 holds for $D = AR$, E_3 to be the event that Fact 140 holds for $B = AR^T X^* G - AG$, and E_4 to be the event that R satisfies Lemma 142.

E_1 holds w.h.p., E_2, E_3, E_4 each separately hold with probability 99/100 (for a suitable choice of K). By a union bound, they all hold simultaneously with probability at least 24/25. Conditioned on this happening:

$$\begin{aligned}
&\left\|AR^T X' - A\right\|_{2,1} \leq \left\|AR^T X^* - A\right\|_{2,1} + \left\|AR^T(X^* - X')\right\|_{2,1} && (1)\\
&\leq \left\|AR^T X^* - A\right\|_{2,1} + \operatorname{poly}(k)\cdot\left\|TAR^T(X^* - X')G\right\|_{1,1} && (2)\\
&\leq \operatorname{poly}(k)\left(\left\|AR^T X^* - A\right\|_{2,1} + \left\|T(AR^T X^* - A)G\right\|_{1,1} + \left\|T(AR^T X' - A)G\right\|_{1,1}\right) && (3)\\
&\leq \operatorname{poly}(k)\left(\left\|AR^T X^* - A\right\|_{2,1} + 2\left\|T(AR^T X^* - A)G\right\|_{1,1}\right) && (4)\\
&\leq \operatorname{poly}(k)\left(\left\|AR^T X^* - A\right\|_{2,1} + O(\log d)\left\|(AR^T X^* - A)G\right\|_{1,1}\right) && (5)\\
&\leq \log d\cdot\operatorname{poly}(k)\left\|AR^T X^* - A\right\|_{2,1} && (6)
\end{aligned}$$

(1) and (3) hold by the triangle inequality, (2) since E_1 and E_2 hold, (4) by E_1 again and since X' is the minimizer of the expression $\left\|TAR^T X - TA\right\|_{2,1}$, (5) since E_3 holds, and (6) by E_1 again.

X' lies in the rowspace of TA, since otherwise there is a rank-k projection Z onto the rows of TA with $\|TAX'Z - TAZ\|_{2,1} = \|TAX'Z - TA\|_{2,1}$ smaller than $\|TAX' - TA\|_{2,1}$. Since E_4 holds, $\left\|AR^T X^* - A\right\|_{2,1} \leq O(1)\Delta^*$ and thus the rowspace of TA contains a $\log d\cdot\operatorname{poly}(k)$ approximation. □

Thus if P is the rowspace of TA as in Theorem 137 then P is a $(\operatorname{poly}(k), \log d\cdot\operatorname{poly}(k))$-coreset for A.

8.3.2 Bootstrapping a Coreset

Given a poor coreset for A, we now show how to leverage known results about residual sampling from [DV07b] and [CW15a] to obtain a better coreset of slightly larger dimension.

Theorem 143. *Given P, a (t, K)-coreset for A, Algorithm 7 returns a $(t + K\operatorname{poly}(k/\epsilon), (1+\epsilon))$ coreset for A.*

Algorithm 7 BootstrapCoreset

Input: $A \in \mathbb{R}^{n\times d}$, $P \in \mathbb{R}^{t\times d}$ (t, K)-coreset, $\epsilon \in (0, 1)$

Output: $U \in \mathbb{R}^{(t+K\,\mathrm{poly}(k/\epsilon))\times d}$ $(t + K\,\mathrm{poly}(k/\epsilon), (1 + \epsilon))$-coresets

1: In parallel compute $\{H_i A\}_{i=1}^{O(K)\,\mathrm{poly}(k/\epsilon)}$ online // each H_i as in Lem. 145

2: $Q \leftarrow \mathrm{poly}(k/\epsilon)O(K)$ samples from $\mathcal{P}(A(\mathrm{Id} - P))$ // as in Lem. 144

3: $U \leftarrow$ Orthonormal basis for $\mathrm{RowSpan}\left(\left[\frac{P}{Q}\right]\right)$

4: **return** U

Proof. Consider the following idealized noisy sampling process that samples rows of a matrix B. Sample a row B_i of B with probability at least $\frac{\|B_i\|_1}{\|B\|_1}$ and add a noise vector E with $\|E\|_1 \le \nu\|B\|_1$. Supposing we had such a process $\mathcal{P}^*(B)$, we can prove the following lemma.

Lemma 144. *If P is a (t, K)-coreset for A, and A' is a noisy subset of rows of the residual $A(Id - P)$ sampled according to $\mathcal{P}^*(A(Id - P)G)$, with G an appropriately scaled Gaussian matrix as in Fact 138, then with probability 99/100, $P + Span(A')$ is an $O(t + K\,\mathrm{poly}(k/\epsilon))$ dimensional subspace containing a k-dimensional subspace with corresponding projection matrix X' such that:*

$$\|A - AX'\|_{2,1} \le (1 + \epsilon)\Delta^*$$

Proof. Our theorem is identical to Theorem 45 from [CW15a], which is in turn an adaptation of Theorem 9 from [DV07b], except that our sampling procedure produces noisy samples instead of actual rows of $A(\mathrm{Id} - P)$. We highlight the difference between our proof and the originals, and refer the reader to the sources for a full description.

Let H_ℓ denote the span of the rows of P adjoined with ℓ samples from $\mathcal{P}^*(A(\mathrm{Id} - P))$. The analysis considers $k + 1$ phases during the construction of H_ℓ, where phase j is defined such that there exists a subspace X_j with:

(i) the dimension of $\mathrm{RowSpan}(X_j) \cap H_\ell \ge j$.

(ii) and letting $\delta = \epsilon/2k$ we have: $\|A(\mathrm{Id} - X_j)\|_{2,1} \le (1 + \delta)^j \min_{X \text{ rank } k}\|A - AX\|_{2,1}$

In other words, the cost of the solution X_j slowly gets worse with j, but H_ℓ recovers more of it. Note that in phase k, $\|A(\mathrm{Id} - X_k)\|_{2,1} \le (1 + \epsilon) \min_{X \text{ rank } k}\|A - AX\|_{2,1}$, and furthermore $X_k \subseteq H_\ell$.

Let Y_ℓ denote the rank-k projection whose row space is that of X_j, but rotated about the intersection $\mathrm{RowSpan}(X_j) \cap H_\ell$ such that it also contains the vector in H_ℓ realizing the smallest nonzero principle angle with X_j. Note that Y_ℓ satisfies condition (i) for some $j' > j$, so it remains to show that with high probability, with a small number of new samples, condition (ii) is also satisfied. In particular, we show that if condition (ii) is violated, and thus if:

$$\|A(\mathrm{Id} - Y_\ell)\|_{2,1} > (1 + \delta)\|A(\mathrm{Id} - X_j)\|_{2,1}$$

then with probability greater than $\delta/5K$ we sample a witness row A_{i*} with the property:

$$\left\|\hat{A}_{i*}(\mathrm{Id} - Y_\ell)\right\|_2 \ge (1 + \delta/2)\left\|\hat{A}_{i*}(\mathrm{Id} - X_j)\right\|_2, \tag{8.2}$$

where $\hat{A}_{\ell'*}$ is defined below.

By the Angle Drop Lemma (Lemma 13 of [DV07b]), this witness implies that the smallest nonzero principle angle between X_j and H_ℓ (and so Y_ℓ) decreases. By the analysis on page 16 of their paper, once the angle is small enough, Y_ℓ will satisfy (ii).

$\mathcal{P}^*$ produces a row of $A(\mathrm{Id}-P)$ plus some noise. Call this noisy sample $\hat{A}_{\ell'*}$ and call the noise $E_{\ell'}$. After sampling $\hat{A}_{\ell'*}$, our subspace contains the point $A_{\ell*}P+\hat{A}_{\ell'*} = (\mathrm{Id}-P)A_{\ell'*}+PA_{\ell'*}+E_{\ell'} = A_{\ell'*}+E_{\ell'}$.

We condition on $\mathcal{P}^*$ producing errors that satisfy $\|E_{\ell'}\|_2 \le \nu\|A_{\ell'*}(\mathrm{Id}-P)\|_2$, where $\nu = \delta/(40K)$.

Let W denote the set of *witness* rows, that is, set of all i that satisfy (8.2). We want to show that

$$\sum_{i\in W}\|A_{i*}(\mathrm{Id}-P)\|_2 \ge \frac{\delta}{5K}\|A(\mathrm{Id}-P)\|_{2,1} \tag{8.3}$$

Suppose that (8.3) is false. The definitions of X_j, Y_ℓ and H_ℓ imply that all elements of H_ℓ are closer to Y_ℓ than to X_j. Let $\tilde{X}_\ell$ be a matrix projecting onto H_ℓ.

$$\begin{aligned}
&\left\|\hat{A}_{i*}(\mathrm{Id}-Y_\ell)\right\|_2 \le \left\|\hat{A}_{i*}(\mathrm{Id}-\tilde{X}_\ell)\right\|_2 + \left\|\hat{A}_{i*}\tilde{X}_\ell(\mathrm{Id}-Y_\ell)\right\|_2\\
&\le \left\|\hat{A}_{i*}(\mathrm{Id}-\tilde{X}_\ell)\right\|_2 + \left\|\hat{A}_{i*}\tilde{X}_\ell(\mathrm{Id}-X_j)\right\|_2 \le 2\left\|\hat{A}_{i*}(\mathrm{Id}-\tilde{X}_\ell)\right\|_2 + \left\|\hat{A}_{i*}\tilde{X}_\ell\right\|_2\\
&\le 2\left\|\hat{A}_{i*}(\mathrm{Id}-P)\right\|_2 + \left\|\hat{A}_{i*}(\mathrm{Id}-X_j)\right\|_2
\end{aligned}$$

The first and third inequalities are the triangle inequality, the second is from distance property above, and the last since $P \in H_\ell$. We bound $i \in W$ using the bound above. For $i \notin W$, by definition $\left\|\hat{A}_{i*}(\mathrm{Id}-Y_\ell)\right\|_2 \le (1+\delta/2)\left\|\hat{A}_{i*}(\mathrm{Id}-X_j)\right\|_2$. Combining both the bounds we have for all i;

$$\left\|\hat{A}_{i*}(\mathrm{Id}-Y_\ell)\right\|_2 \le (1+\delta/2)\left\|\hat{A}_{i*}(\mathrm{Id}-X_j)\right\|_2 + [\![i\in W]\!]\left\|\hat{A}_{i*}(\mathrm{Id}-P)\right\|_2$$

Summing over all i,

$$\left\|\hat{A}(\mathrm{Id}-Y_\ell)\right\|_{2,1} \le (1+\delta/2)\left\|\hat{A}(\mathrm{Id}-X_j)\right\|_{2,1} + 2\left\|\hat{A}_{W*}(\mathrm{Id}-P)\right\|_2$$

$$\left[\begin{array}{rc} & \|A(\mathrm{Id}-Y_\ell)\|_{2,1}\\ - & \|E(\mathrm{Id}-Y_\ell)\|_{2,1}\end{array}\right] \le \left[\begin{array}{rc} & (1+\delta/2)\|A(\mathrm{Id}-X_j)\|_{2,1} + 2\|A_{W*}(\mathrm{Id}-P)\|_{2,1}\\ + & (1+\delta/2)\|E(\mathrm{Id}-X_j)\|_{2,1} + 2\|E(\mathrm{Id}-P)\|_{2,1}\end{array}\right]$$

$$\|A(\mathrm{Id}-Y\ell)\|_{2,1} \le (1+\frac{\delta}{2})\|A(\mathrm{Id}-X_j)\|_{2,1} + \frac{2\delta}{5K}K\|A(\mathrm{Id}-X_j)\|_{2,1} + 4\|E\|_{2,1} \tag{4}$$

$$\|A(\mathrm{Id}-Y_\ell)\|_{2,1} \le (1+9\delta/2+4\nu)\|A(\mathrm{Id}-X_j)\|_{2,1} \le (1+\delta)\|A(\mathrm{Id}-X_j)\|_{2,1}$$

Which is a contradiction. (4) follows from the assumption that (8.3) is false. Note that this proof goes through for any error matrix E satisfying $\|E_i\| \le \nu\|A_i\|$ for all i. Also, as written in [CW15a], the proof guarantees success with constant probability. We can repeat the sampling a constant number of times, keep all samples, and guarantee success with probability 99/100. $\square$

It remains to show such a process $\mathcal{P}^*$ exists, which is nearly Lemma 17 from [SW11].

Lemma 145 (Lemma 17 from [SW11])**.** *There exists an oblivious sketching matrix $H \in \mathbb{R}^{d \log n \operatorname{poly}\left(\frac{k}{\nu}\right) \times n}$ and a row sampling process $\mathcal{P}$ such that for a given matrix $B \in \mathbb{R}^{n \times d}$, $\mathcal{P}(B)$ samples the rows of HB according to a distribution that has total variation distance at most $1/100$ from the idealized noisy sampling process $\mathcal{P}^*(B)$ above.*

Proof. Consider the algorithms Sampler and Extract from Appendix C of [SW11]. First fix an appropriate $\ell = O(\log n)$, and sample η uniformly from the interval $[1, 2]$.

Algorithm 8 Sampler

Input: $B \in \mathbb{R}^{n \times d}$
Output: $HB \in \mathbb{R}^{d \log n \operatorname{poly}\left(\frac{k}{\nu}\right) \times d}$

for level $j \in [\ell]$ **do**
  Create hash tables $H^{(j)}$ with $w = \operatorname{poly}(\frac{k\ell}{\nu})$ buckets and assign them independent hash functions $h_j : [n] \to [w]$ (each bucket stores a d dimensional vector)
  for hashtable $H^{(j)}$ **do**
    Subsample a set $J_j \subset [n]$ where each $i \in [n]$ is included with probability $p_j = \min\left(1, \frac{C}{2^j}\right)$ where $C = \operatorname{poly}(\frac{k}{\nu})$
    for $v \in [w]$ **do**
      for $k \in [d]$ **do**
        $H_v^{(j)} = \sum_{i \in J_j} \chi(h_j(i) = v) \cdot B_{i*}$
      end for
    end for
  end for
end forreturn $\{H^{(j)}\}_j$ as a matrix in $\mathbb{R}^{d \log n \operatorname{poly}\left(\frac{k}{\nu}\right) \times d}$

Algorithm 9 Extract

Input: $HB \in \mathbb{R}^{d \log n \operatorname{poly}\left(\frac{k}{\nu}\right) \times d}$
Output: $HB \in \mathbb{R}^{d \log n \operatorname{poly}\left(\frac{k}{\nu}\right) \times d}$

$F \leftarrow \varnothing$
for level $j \in [\ell]$ **do**
  for bucket $v \in [w]$ **do**
    if $(1-\nu) \cdot \frac{\eta \|B\|_{1,1}}{2^j} \leq \left\| H_v^{(j)} \right\|_1 \leq (1+\nu) \cdot 2 \cdot \frac{\eta \|B\|_{1,1}}{2^j}$ **then**
      return H_v^j with weight $\frac{1}{p_j}$
    end if
  end for
end for

Let $L_j = \left\{ B_{i*} : \|B_{i*}\|_1 \in \left[\frac{\eta \|B\|_{1,1}}{2^j}, 2 \cdot \frac{\eta \|B\|_{1,1}}{2^j}\right] \right\}$ be the j-th level set of row norms of B.

By the proof of Lemma 17 in Appendix D of [SW11] (more precisely Claims 18-21), there is a choice of constant $C' = \mathrm{poly}(\frac{k\ell}{\nu})$ such that with probability $99/100$ over the choice of η, all of the following events hold simultaneously for all levels j.

(i) No row i subsampled in the set J_j has the property that $\|B_{i*}\|_1 \in \left[(1-2\nu)\frac{\eta\|B\|_{1,1}}{2^j}, \frac{\eta\|B\|_{1,1}}{2^j}\right]$ or $\|B_{i*}\|_1 \in \left[(1-\nu)\cdot 2\cdot \frac{\eta\|B\|_{1,1}}{2^j}, 2\cdot \frac{\eta\|B\|_{1,1}}{2^j}\right]$

(ii) Every row in $\bigcup_{j'\leq j+\log C'} L_{j'}$ is hashed to a different bucket in $H^{(j)}$.

(iii) The noise $N_{v,j}$ in every bucket v of $H^{(j)}$ is small, formally:

$$\|N_{v,j}\|_1 = \left\|\sum_{i\in[n]} \chi(i \in \bigcup_{j'>j+\log C'} J_{j'})\cdot \chi(h_j(i)=v)\cdot B_{i*}\right\|_1 \leq \nu\cdot \frac{\eta\|B\|_{1,1}}{2^j}$$

(iv) No row in $\bigcup_{j'>\ell} B_{j'}$ is sampled.

If all the events above hold, the combination of Sampler and Extract exactly perform the sampling process $\mathcal{P}^*$, since every hash table $H^{(j)}$ samples the level set L_j uniformly with probability proportional to the 1-norm of the heaviest element in L_j, sends these to distinct buckets, and then adds small noise. □

Combining the two lemmas in this section, it follows that $\mathrm{RowSpan}(P) + \mathrm{RowSpan}(\mathcal{P}(A(\mathrm{Id}-P)))$ is a $(t + K\,\mathrm{poly}(k/\epsilon))$-dimensional subspace containing a $(1+\epsilon)$ approximation to the original problem. Note that each sketch HA generates one sample, and thus we need $K\,\mathrm{poly}(k/\epsilon)$ copies to generate enough samples for the residual sampling. □

8.3.3 Right Dimension Reduction

We show how to reduce the right dimension of our problem. This result is used in both Algorithm 5 and Algorithm 6.

Theorem 146. *If U is a (t, K)-coreset, $S \in \mathbb{R}^{\log d\cdot \mathrm{poly}(k/\epsilon)\times d}$ is a CountSketch matrix composed with a matrix of i.i.d. Gaussians, and $R \in \mathbb{R}^{d\times \mathrm{poly}(k/\epsilon)}$ is a CountSketch matrix composed with a Gaussian, then with probability $49/50$, if $X' = \mathrm{argmin}_X \left\|AS^T - AR^TXU^TS^T\right\|_{2,1}$ then:*

$$\left\|A - AR^TX'U^T\right\|_{2,1} \leq (1+O(\epsilon)) \min_{X \text{ rank } k} \left\|A - AXU^T\right\|_{2,1}$$

Proof. We need a couple lemmas from [CW15a].

Lemma 147 (Lemma 30 from [CW15a])**.** *If S is a lopsided embedding for (B, D), then if X'' has the property that $\|SBX'' - SD\|_{1,2} \leq \kappa \min_{X\in\mathcal{C}} \|SBX - SD\|_{1,2}$ for some κ, then: $\|BX'' - D\|_{1,2} \leq \kappa(1+3\epsilon)\min_{X\in\mathcal{C}} \|BX - D\|_{1,2}$.*

Lemma 148. *If $U \in \mathbb{R}^{d\times t}$ and $R \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times d}$ is a CountSketch matrix composed with a matrix of i.i.d. Gaussians, then with probability* $99/100$*:* $\min_{X \text{ rank } k} \left\|A - AR^T X U^T\right\|_{2,1} \le (1+3\epsilon)\Delta^*$.

Proof. Let $V^* = \mathrm{argmin}_{V \text{ rank } k} \left\|UV - A^T\right\|_{1,2}$ and let $V = V_1 V_2$ be its rank factorization. Applying Lemmas 141 and 147, R is a lopsided embedding for (UV_1, A^T) with probability $99/100$. If $Y = \mathrm{argmin}_{Y \text{ rank } k} \left\|R(UV_1 Y - A^T)\right\|_{1,2}$ then:

$$\left\|UV_1 Y - A^T\right\|_{2,1} \le (1+3\epsilon)\left\|UV^* - A^T\right\|_{1,2} \le (1+3\epsilon)\Delta^*$$

But $Y = (RUV_1)^- RA^T$, and taking transposes this means that:

$$\min_{X \text{ rank } k} \left\|A - AR^T X U^T\right\|_{2,1} \le \left\|A - AR^T((RUV_1)^-)^T V_1^T U^T\right\|_{2,1} \le (1+3\epsilon)\Delta^*$$

□

From the last lemma, a solution to $\min_{X \text{ rank } k} \left\|A - AR^T X U^T\right\|_{2,1}$ will yield a $(1+\epsilon)\cdot O(K)$-approximate solution to the original problem. Lemma 148 holds with probability $99/100$. Applying Lemma 141, with probability $99/100$, an $S \in \mathbb{R}^{d\times \log d\, \mathrm{poly}(k)}$ CountSketch composed with a Gaussian is a lopsided embedding for (U, A^T). Union bounding over these events, and applying Lemma 147 with $\mathcal{C}$ as the set of matrices in $\mathrm{RowSpan}(RA^T)$ proves the claim with probability $49/50$. □

8.3.4 Left Dimension Reduction

We show how to reduce the left dimension of our problem. Together with results from Section 8.3.3, this preserves the solution to X^* to within a coarse $\sqrt{\log d}\log\log d \cdot \mathrm{poly}(k/\epsilon)$ factor.

Theorem 149. *If $C \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times n}$ is a Sparse Cauchy matrix, and $G \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times \mathrm{poly}(k/\epsilon)}$ is a matrix of appropriately scaled i.i.d. Gaussians (as in Fact 138), and*

$$\hat{X} = \underset{X \text{ rank } k}{\mathrm{argmin}} \left\|CAS^T G - CAR^T X U^T S^T G\right\|_F$$

then with probability $24/25$*:* $\left\|AS^T - AR^T \hat{X} U^T S^T\right\|_{2,1} \le \sqrt{\log d}\log\log d \cdot \mathrm{poly}(k/\epsilon)\cdot \Delta^*$

Proof. Define E_1 to be the event that the condition in Dvoretzky's theorem is satisfied, E_2 to be the event that Fact 139 holds for $D = AR$, and E_3 to be the event that Fact 140 holds for $B = (AS^T - AR^T X^* U^T S^T)G$.

E_1 holds w.h.p., E_2, E_3 each separately hold with probability $99/100$ (for a suitable choice of K). By a union bound, they all hold simultaneously with probability at least $24/25$. Conditioned on this happening:

$$
\begin{aligned}
\left\|AS^T - AR^T\hat{X}U^TS^T\right\|_{2,1} &\leq \left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} + \left\|AR(X^* - \hat{X})U^TS^T\right\|_{2,1} && (1)\\
&\leq \left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} + \mathrm{poly}(k/\epsilon)\left\|CAR(X^* - \hat{X})U^TS^TG\right\|_{1,1} && (2)\\
&\leq \mathrm{poly}(k/\epsilon)\begin{bmatrix} & \left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} + \left\|C(A - AR^TX^*U^T)S^TG\right\|_{1,1} \\ + & \left\|C(A - AR^T\hat{X}U^T)S^TG\right\|_{1,1} \end{bmatrix} && (3)\\
&\leq \mathrm{poly}(k/\epsilon)\begin{bmatrix} & \left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} + \left\|C(AS^T - AR^TX^*U^TS^T)G\right\|_{1,1} \\ + & \sqrt{\log d}\left\|C(A - AR^T\hat{X}U^T)S^TG\right\|_{F} \end{bmatrix} && (4)\\
&\leq \mathrm{poly}(k/\epsilon)\left[\left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} + \sqrt{\log d}\left\|C(AS^T - AR^TX^*U^TS^T)G\right\|_{1,1}\right] && (5)\\
&\leq \mathrm{poly}(k/\epsilon)\begin{bmatrix} & \left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} \\ + & \sqrt{\log d}\log\log d\left\|(AS^T - AR^TX^*U^TS^T)G\right\|_{1,1} \end{bmatrix} && (6)\\
&\leq \sqrt{\log d}\log\log d\,\mathrm{poly}(k/\epsilon)\left\|AS^T - AR^TX^*U^TS^T\right\|_{2,1} && (7)
\end{aligned}
$$

(1) and (3) hold by triangle inequality, (2) since E_1 and E_2 hold, (4) comes from the relationship between the 1-norm and 2-norm, (5) since $\hat{X}$ is the minimizer of the expression $\left\|C(A - CAR^TXU^T)S^TG\right\|_F$ and p-norms decrease with p, (6) since E_3 holds, and (7) by E_1 again. $\square$

The rank constrained Frobenius norm minimization problem above has a closed form solution.

Fact 150. For a matrix M, let $U_M\Sigma_MV_M^T$ be the SVD of M. Then:

$$\operatorname*{argmin}_{X\text{ rank }k}\|Y - ZXW\|_F = Z^-[U_ZU_Z^TYV_WV_W^T]_kW^-$$

8.4 $(1+\epsilon)$-Approximation

8.4.1 Left Dimension Reduction

The following median based embedding allows us to reduce the left dimension of our problem. Together with results from Section 8.3.3, this preserves the solution to X^* to within a $(1 + O(\epsilon))$ factor.

Theorem 151. *If $C \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times n}$ is a Sparse Cauchy matrix, and $G \in \mathbb{R}^{\mathrm{poly}(k/\epsilon)\times\mathrm{poly}(k/\epsilon)}$ is a matrix of appropriately scaled i.i.d. Gaussians (as in Fact 138), and:*

$$\hat{X} = \operatorname*{argmin}_{X\text{ rank }k}\left\|CAS^TG - CAR^TXU^TS^TG\right\|_{\mathrm{med},1}$$

then with probability $99/100$:

$$\left\|AS^TG - AR^TX'U^TS^TG\right\|_{1,1} \leq (1+\epsilon)\min_{X\text{ rank }k}\left\|AS^TG - AUXR^TS^TG\right\|_{1,1}$$

Proof. The following fact is known:

Fact 152 (Lemma F.1 from [Bac+16])**.** Let L be a t dimensional subspace of $\mathbb{R}^s$. Let $C \in \mathbb{R}^{m\times s}$ be a matrix with $m = O\left(\frac{1}{\epsilon^2} t \log \frac{t}{\epsilon}\right)$ and i.i.d. standard Cauchy entries. With probability 99/100, for all $x \in L$ we have

$$(1-\epsilon)\|x\|_1 \le \|Cx\|_{\text{med}} \le (1+\epsilon)\|x\|_1$$

The theorem statement is simply the lemma applied to $L = \text{ColSpan}\left([AS^T \mid AR^T]\right)$. □

8.4.2 Solving Small Instances

Given problems of the form $\hat{X} = \text{argmin}_{X \text{ rank } k} \|Y - ZXW\|_{\text{med},1}$, we leverage an algorithm for checking the feasibility of a system of polynomial inequalities as a black box.

Lemma 153. *[BPR94] Given a set $K = \{\beta_1, \cdots, \beta_s\}$ of polynomials of degree d in k variables with coefficients in $\mathbb{R}$, the problem of deciding whether there exist $X_1, \cdots X_k \in \mathbb{R}$ for which $\beta_i(X_1, \cdots, X_k) \ge 0$ for all $i \in [s]$ can be solved deterministically with $(sd)^{O(k)}$ arithmetic operations over $\mathbb{R}$.*

Theorem 154. *Fix any $\epsilon \in (0,1)$ and $k \in [0, \min(m_1, m_2)]$. Let $Y \in \mathbb{R}^{m'\times m''}$, $Z \in \mathbb{R}^{m'\times m_1}$, and $W \in \mathbb{R}^{m_2 \times m''}$ be any matrices. Let $C \in \mathbb{R}^{poly(m'/\epsilon)\times m'}$ be a matrix of i.i.d. Cauchy random variables, and $G \in \mathbb{R}^{m''\times poly(m''/\epsilon)}$ be a matrix of scaled i.i.d. Gaussian random variables. Then conditioned on C satisfying Theorem 152 for $[Y \mid Z]$ and G satisfying the condition of Fact 138, a rank-k projection matrix X can be found that minimizes $\|C(Y - ZXW)G\|_{\text{med},1}$ up to a $(1+\epsilon)$-factor in time $\text{poly}(m'm''/\epsilon)^{O(mk+m')}$, where $m = \max(m_1, m_2)$.*

Proof. We write $X = PQ$, where P is $m_1 \times k$ and Q is $k \times m_2$, to ensure that X is rank $\le k$.

Guess a permutation π_j for each column j of $C(ZXW - Y)G$ and define constraints enforcing the permutation. Since the (i,j)-th entry of the matrix is $\sum_{k,\ell}(CZ)_{ik}X_{k\ell}(WG)_{\ell j} - (CYG)_{ij}$ these constraints are of the form $((C(ZXW - Y)G)_{\pi_j(i)j})^2 \le ((C(ZXW - Y)G)_{\pi_j(i+1)j})^2$. Then define the median of the j-th column to be:

$$M_j = \left(|(C(ZXW - Y)G)_{\pi_j(\lfloor m''/2 \rfloor)j}| + |(C(ZXW - Y)G)_{\pi_j(\lceil m''/2 \rceil)j}|\right)/2$$

Thus we have $mk + poly(m''/\epsilon)$ variables in our polynomial inequality system, $O(mk)$ variables to describe P and Q, and $poly(m''/\epsilon)$ variables to describe the column medians M_j. We have $\text{poly}(m'm''/\epsilon)$ constraints, each involving polynomials of $O(1)$ degree. By Lemma 153, checking the feasibility of this system takes time $\text{poly}(m'm''/\epsilon)^{O(mk)+poly(m''/\epsilon)}$.

We can minimize the objective $\sum_j M_j$ using binary search. This requires a lower bound on the objective value, which we can get by noting from Theorem 152 that:

$$\min_X \|CZXWG - CYG\|_{\text{med},1} \ge (1-\epsilon)\min_X \|ZXW - Y\|_{1,1} \ge (1-\epsilon)\min_X \|ZXW - Y\|_{2,1}$$

By the proof of Theorem 51 in [CW15a], the right hand side is lower bounded by $\frac{1}{\text{poly}(d)}(\sigma_{k+1}(Y))^{1/2}$ (where $\sigma_{k+1}(Y)$ is the $k+1$st singular value of Y), which itself is lower bounded by $\left(\frac{1}{\exp(\text{poly}(m'm''))}\right)^k$. Thus we can do binary search in $\text{poly}(m'm''/\epsilon)$ steps.

Finally, since there are $m'' \cdot m'!$ possible permutation guesses, the entire procedure takes time $\mathrm{poly}(m'm''/\epsilon)^{O(mk)+\mathrm{poly}(m'm''/\epsilon)}$. $\square$

We remark that if we set $m = \log\log d\sqrt{\log d}$ and $m', m'' = \mathrm{poly}(k/\epsilon)$, as we do in our algorithm, we can write our overall runtime as $O(\mathrm{nnz}(A) + (n+d)\,\mathrm{poly}(k/\epsilon) + \exp(\mathrm{poly}(k/\epsilon)))$. If $\mathrm{poly}(k/\epsilon) \leq \sqrt{\log d}/\log\log d$, then this final step is captured in the $(n+d)\,\mathrm{poly}(k/\epsilon)$ term. Otherwise this step is captured in the $\exp(\mathrm{poly}(k/\epsilon))$ term.

8.5 Experiments

In this section we empirically demonstrate the effectiveness of Algorithm 5 compared to the truncated SVD. We experiment on both real and synthetic data sets. Since the algorithm is randomized, we run it 20 times and take the best performing run.

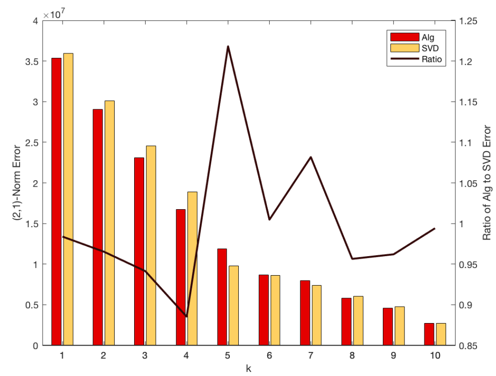
For the real data, we use two data sets. In Figure 8.1a we run on the FIDAP data set[1], which is a 27×27 matrix with 279 real asymmetric non-zero entries. In Figure 8.1b we use the KOS blog entries matrix[2], which represents word frequencies in blogs, and is 3430×6906 with 353160 non-zero entries.

For the synthetic data, we use four example matrices all of dimension 100×10. In Figure 8.1c, we use a random ± 1 matrix. In Figure 8.1*d* we use a random sparse matrix generated as follows: set each entry to 0 with probability 0.95, and otherwise assign it a uniformly random entry from $[0, 1]$. In Figure 8.1e we use a Rank-3 matrix with additional large outlier noise. First we sample U random 100×3 matrix and V random 3×10 matrix. Then we create a random sparse matrix W as before but with probability 0.99 and scaled by a factor of 100. We use $UV + W$. Finally in Figure 8.1f we create a simple Rank-2 matrix with a large outlier. The first row is 100 followed by all zeros. All subsequent rows are 0 followed by all ones.

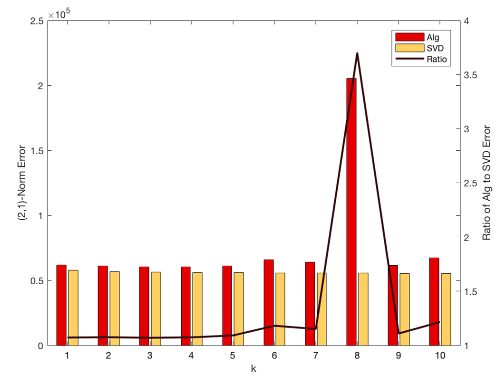
While the approximation guarantee of Algorithm 5 is weak, we find that it performs well against the SVD baseline in practice on several of our examples, namely when the data has large outliers rows. The final example in particular serves as a good demonstration of the robustness of the (2,1)-norm to outliers in comparison to the Frobenius norm. When $k = 1$, the truncated SVD which is the Frobenius norm minimizer recovers the first row of large magnitude, whereas our algorithm recovers the subsequent rows. Note that both our algorithm and the SVD recover the matrix exactly when k is greater than or equal to rank. For example this means that the matrix in Figure 8.1e has rank 8.

[1] https://math.nist.gov/MatrixMarket/data/SPARSKIT/fidap/fidap005.html

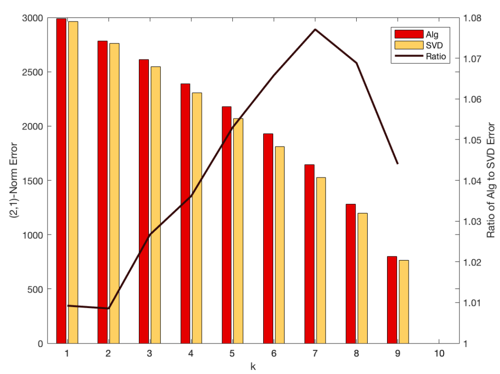
[2] https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/



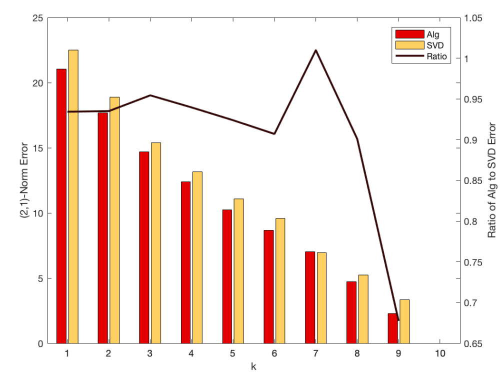
(a) Fidap Matrix



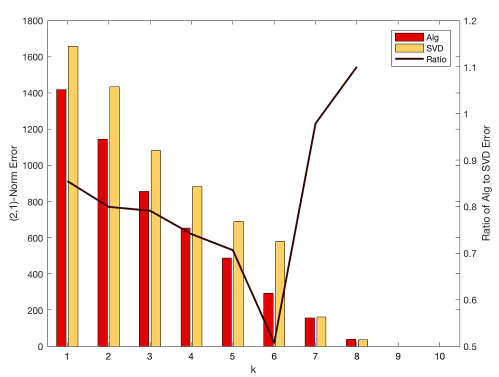
(b) Kos Matrix



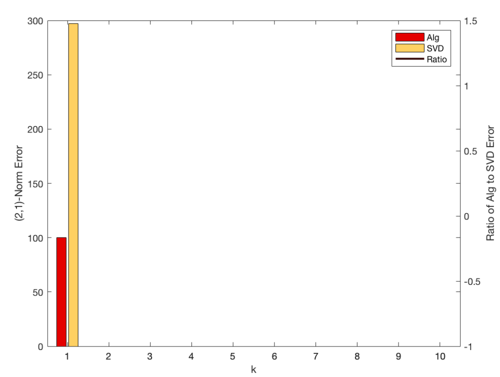
(c) Random ±1 Matrix



(d) Random Sparse Matrix



(e) Random Rank-3 Matrix Plus Large Outliers



(f) Large Outlier Rank-2 Matrix

Figure 8.1: Comparison of Algorithm 5 on real and synthetic examples.

Appendix A

Properties of gadgets

A.1 Quantitative Bounds for Properties of Gadgets

This section will provide quantitative bounds to some properties of (ε, D)-copies of a gadget T. We will give bounds on ε and D in order to satisfy Observation 9, and a slightly stronger version of it. First, we set up some notation. Given a sets $S \subseteq V \subseteq \mathbb{R}^2$ and an Hamiltonian path P on V, we say that S is connected to $V \setminus S$ through a pair of edges e_1, e_2 in P if $e_1, e_2 \in \delta(S, V \setminus S)$, and e_1 and e_2 are connected in P through a path completely contained in S.

Lemma 155. *Let S be a gadget with diameter d, and let P be an optimal Hamiltonian path through V. Given $\varepsilon > 0$ and $\theta > 0$, there is $D \geq D(\varepsilon, \theta, d)$ such that if S_1 is any (ε, D)-copy of the gadget S such that there are two or more pairs of edges joining S_1 to $V \setminus S_{\varepsilon,D}$ in P then the angle between any connecting pair of edges is at least $\pi - \theta$. In particular,*

$$D(\varepsilon, \theta, d) = \frac{6d + 12\varepsilon}{1 - \cos\theta} \tag{A.1}$$

suffices.

Proof. Suppose e_1, e_2 is a pair of edges connecting S_1 to $V \setminus S_1$. Let $e_i = \{p_i, x_i\}$ where $x_i \in S_1$, $p_i \notin S_1$ for $i = 1, 2$. First, we make a precise definition of the angle between these two edges using the cosine formula.

Definition 156. The angle between $\overrightarrow{x_1p_1}$ and $\overrightarrow{x_2p_2}$ denoted by $\measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{x_2p_2})$ is the angle $\phi \in [0, \pi]$ such that

$$\cos(\phi) = \frac{\langle x_1p_1, x_2p_2 \rangle}{\|x_1p_1\| \cdot \|x_2p_2\|}$$

Let $\phi = \measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{x_2p_2})$. Let f_1, f_2 be any other pair of edges connecting S_1 to $V \setminus S_1$. Let $f_i = \{q_i, y_i\}$ where $y_i \in S_1, q_i \notin S_1$ for $i = 1, 2$. Since P is optimal Hamiltonian path, *short-cutting* p_1, p_2 must give a longer path. To be precise, the path Q obtained by deleting edges e_1, e_2, y_1z where $z \neq q_1$, and adding edges p_1p_2, y_1x_1, x_2z, is longer than the path P. In particular, we must have

$$\ell(p_1p_2) + 2d + 4\varepsilon \geq \ell(p_1x_1) + \ell(p_2x_2) \tag{A.2}$$

Let p_2' be a point such that $x_1y_1p_2p_2'$ is a parallelogram. Therefore, $\ell(p_1p_2) \leq \ell(p_1p_2') + d + 2\varepsilon$. Hence, it must hold that

$$\ell(p_1p_2') + 3d + 6\varepsilon \geq \ell(p_1x_1) + \ell(p_2x_2) \tag{A.3}$$

Let $a = \ell(p_1x_1), b = \ell(p_2x_2), c = \ell(p_1p_2')$. Then by definition of ϕ,

$$c^2 = a^2 + b^2 - 2ab\cos\phi$$

Using this, we get

$$\begin{aligned}
\ell(p_1x_1) + \ell(p_2x_2) - \ell(p_1p_2') &= \frac{(a+b)^2 - c^2}{a+b+c} \\
&\geq \frac{(a+b)^2 - c^2}{2(a+b)} && \text{Since } a+b \leq c \\
&= \frac{2ab(1+\cos\phi)}{2(a+b)} = \frac{ab(1-\cos\phi)}{a+b}
\end{aligned}$$

Since S_1 is an (ε, D) copy of S, $a, b \geq D$. Under this condition $\frac{ab}{a+b}$ is minimized at $a = b = D$, implying that $\ell(p_1x_1) + \ell(p_2x_2) - \ell(p_1p_2') \geq \frac{D(1+\cos\phi)}{2}$. Hence, for Equation (A.3) to hold, we must have

$$3d + 6\varepsilon \geq \frac{D(1+\cos\phi)}{2}$$

In particular, if

$$D \geq \frac{6d + 12\varepsilon}{1 - \cos\theta}$$

then $1 + \cos\phi \leq 1 - \cos\theta \implies \phi \geq \pi - \theta$, which completes the proof giving us the bound in Equation (A.1). □

Lemma 157. *Let S be a gadget with diameter d, and let P be an optimal Hamiltonian path through V. Given $\varepsilon > 0$ and $\frac{\pi}{4} \geq \theta > 0$, there is $D \geq D(\varepsilon, \theta, d)$ such that if S_1 is any (ε, D)-copy of the gadget S then there are at most 2 pairs of edges joining S_1 to $V \setminus S_1$. Further, all the four edges joining S_1 to $V \setminus S_1$ make an acute angle of at most 2θ with each other. In particular,*

$$D(\varepsilon, \theta, d) = \frac{6d + 12\varepsilon}{1 - \cos\theta} \tag{A.4}$$

suffices.

Proof. Let e_1, e_2 be a pair of edges joining S_1 to $V \setminus S_1$ such that $e_i = \{x_i, p_i\}$ where $x_i \in S_1, p_i \notin S_1$ for $i = 1, 2$. Let f_1, f_2 be a pair of edges joining S_1 to $V \setminus S_1$ such that $f_i = \{y_i, q_i\}$ where $y_i \in S_1, q_i \notin S_1$ for $i = 1, 2$. Further, let p_2 and q_1 be through portion of P that does not contain x_2.

Since P is an optimal Hamiltonian path, the Hamiltonian path Q obtained by deleting edges p_1x_1, q_1y_1 and adding edges x_1y_1, p_1q_1, must by as long. Therefore, we must have

$$d \geq \ell(p_1x_1) + \ell(q_1y_1) - \ell(x_1y_1)$$

By the computations in Lemma 155, for $D \geq \frac{6d+12\varepsilon}{1-\cos\theta}$, this hold only if $\measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{y_1q_1}) \geq \pi - \theta$. This observation combined with Lemma 155 implies that all four edges e_1, e_2, f_1, f_2 make an acute angle of at most 2θ with each other (This holds even if they are not coplanar!).

Now, assume that there is another pair of edges g_1, g_2 joining S_1 to $V \setminus S_1$, such that $g_i = \{z_i, r_i\}$ where $z_i \in S_1, r_i \notin S_1$ for $i = 1, 2$ and q_2 and r_1 are connected through portion of P that does not contain y_2. Then we have

$$\measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{y_1q_1}) \geq \pi - \theta$$
$$\measuredangle(\overrightarrow{y_1q_1}, \overrightarrow{r_1z_1}) \geq \pi - \theta$$
$$\measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{r_1z_1}) \geq \pi - \theta$$

This leads to contradiction, since first two equations imply $\overrightarrow{x_1, p_1}$ and $\overrightarrow{r_1z_1}$ are on the same side of hyperplane $\langle v, q_1 - y_1 \rangle = 0$. But, the third equation implies otherwise! □

A.2 Properties of Hamiltonian Paths in the Gadgets

In this section, we will provide proofs of various geometrical lemma regarding properties of the gadgets in this section. These include proofs of Lemmas 29, 31 and 33.

A.2.1 Proof of Lemma 29

Let us begin by recall definition of $\Pi(t, h, w)$ and $\Pi_S = \Pi(S, t, h, w)$ (Definitions 28 and 30):

Definition 28. We define the gadget $\Pi(t, h, w)$ for $t \in \mathbb{Z}_{\geq 0}$ and $h, w \in \mathbb{R}_{\geq 0}$, given by points $\pi_1 = \left(-\frac{w}{2}, 0\right)$, $\pi_2 = \left(\frac{w}{2}, 0\right)$, $\pi_3 = \left(-\frac{w}{2}, h\right)$, $\pi_4 = \left(\frac{w}{2}, h\right)$ and points $v_1, \ldots, v_t$ which are evenly spaced along $(0,0), (0,h)$, with $v_1 = (0,0)$ and $v_t = (0,h)$. We will refer to sets $\{\pi_1\pi_2\}$ and $\{\pi_3\pi_4\}$ as *shorter sides* of the gadget, and sets $\{\pi_1\pi_3\}$ and $\{\pi_2\pi_4\}$ as *longer sides* of the gadget.

Definition 30. We construct the gadget $\Pi(S(k), t, h, w)$ by replacing points in C by copies of $S(k)$ centered at each point $\pi_i \in C$. We let S_i denote the copy centered at π_i.

Now we are ready to provide proofs of lemmas in Section 2.2.3.

Lemma 29. *Let p, q be two points on the opposite sides of the horizontal line $y = \frac{h}{2}$ such that*

$$\mathrm{dist}(\{x, y\}, \Pi(t, h, w)) \geq D.$$

Let P be a shortest Hamiltonian path from p to q in $\Pi(t, h, w) \cup \{p, q\}$. Suppose all of the following inequalities hold:

$$D \geq \frac{h^2+w^2}{4w} \qquad h \geq 2w \qquad t \geq \frac{16h}{w}$$

Then for at least two $i \in 1, 2, 3, 4$ we have that neither neighbor v_i^1, v_i^2 of π_i on P is not in $\{p, q\}$ and moreover, v_i^1, v_i^2 are two points in $\{v_1, \ldots, v_t\}$ closest to π_i.

Proof. We begin with a few observations:

Observation 158. If $P' = av_{i_1} \dots v_{i_k}\pi_i$ is a contiguous segment in P, then either $i_1 < \dots < i_k$ or $i_k < \dots < i_1$.

Suppose not. Let $j_1, \dots, j_k$ be a sorting of $i_1, \dots, i_k$ in increasing order. Then j_1 and j_k appear somewhere in P'. Suppose j_1 appears before j_k. For notational convenience, let $\ell(a_1 \dots a_j)$ denote the length of the path $a_1, \dots a_j$.

$$\begin{aligned}\ell(av_{i_1} \dots v_{i_k}\pi_i) &\geq \ell(av_{i_1}) + \ell(v_{i_1}v_{j_1}) + \ell(v_{j_1}v_{j_k}) + \ell(v_{j_k}v_{i_k}) + \ell(v_{i_k}\pi_i) \\ &\geq \ell(av_{j_1}) + \ell(v_{j_1}v_{j_k}) + \ell(v_{j_k}\pi_i) && \text{Triangle Inequality} \\ &\geq \ell(av_{j_1} \dots v_{j_k}\pi_i)\end{aligned}$$

Similarly, in the case when j_k appears before j_1, we get

$$\ell(av_{i_1} \dots v_{i_k}\pi_i) \geq \ell(av_{j_k} \dots v_{j_1}\pi_i)$$

Observation 159. If $P' = av_{i_1} \dots v_{i_k}b$, then we can assume that $i_1 \dots i_k$ is a continuous subset of $[t]$.

First, we can by Observation 158 assume $i_1, \dots, i_k$ are sorted either in increasing order or decreasing order. Without loss of generality, let $i_1 < i_k$. Further, let p be an index such that $i_1 < p < i_k$ that is not contained in the set $\{i_1, \dots, i_k\}$. Then we can insert v_p into $v_{i_1} \dots v_{i_k}$ without changing the total length of the portion P'. On the other hand, shortcut through v_p in P whenever v_p was present may decrease the total length. Thus, this replacement can only get us a shorter path.
Using the two observations, we can assume that the shortest Hamiltonian path P looks like this: $p\overline{v_{i_1}v_{j_1}}c_1\overline{v_{i_2}v_{j_2}}c_2 \dots c_4\overline{v_{i_5}v_{j_5}}q$ Where by $\overline{v_{i_1}v_{j_1}}$ we mean the path containing all the vertices between v_{i_1} and v_{j_1}. Let $\mathcal{C} = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ denote the set of four corners.

Observation 160. Let p such that $\operatorname{dist}(p, \Pi(t, h, w)) \geq D$, and let v_i, v_j be any points in $\{v_1, \dots, v_t\}$. Then if $D \geq \frac{h^2+w^2}{4w}$ and $h \geq 2w$ then

$$\ell(pv_i) + \ell(v_jc) \geq \operatorname{dist}(p, \mathcal{C}) + \frac{w}{4} \tag{A.5}$$

for $c \in \{\pi_1, \pi_2, \pi_3, \pi_4\}$.

Suppose $p = (x_1, y_1) \in \mathbb{R}^2$. We will prove the result by working on different cases based on (x_1, y_1).

Case 161. $y_1 \geq h$: Without loss of generality, assume that $x_1 \geq 0$. Then $\ell(pv_i) \geq \ell(pv_t)$ and $\ell(v_jc) \geq \frac{w}{2} = \ell(v_t\pi_3)$. Therefore, by triangle inequality,

$$\ell(pv_i) + \ell(v_jc) \geq \ell(pv_t) + \ell(v_t\pi_3) = \frac{w}{2} + \ell(pv_t)$$

If $\ell(pv_t) \geq \ell(p\pi_3)$, we get the result in this case. Therefore, we can assume that $x \leq \frac{w}{4}$. Since $\ell(pv_t) \geq (y_1 - h)$, we it suffices to show that

$$\left(\bar{y}_1 + \frac{w}{2} - \frac{w}{4}\right)^2 \geq \operatorname{dist}(p, \mathcal{C})^2 = \bar{y}_1^2 + \left(x - \frac{w}{2}\right)^2$$

where $\bar{y}_1 = y_1 - h$. Since $0 \le x_1 \le \frac{w}{4}$, it suffices to show that

$$\left(\bar{y}_1 + \frac{w}{4}\right)^2 \ge \bar{y}_1^2 + \frac{w^2}{4}$$

This is satisfied when $y' \ge \frac{3w}{8}$. Since $\mathrm{dist}(p, \Pi(t,h,w)) \ge \bar{y}_1$, this holds when $D \ge \frac{h^2+w^2}{4w}$ and $h \ge 2w$.

Case 162. $y_1 \le 0$: This case holds due to computations similar to Case 161.

Case 163. $0 \le y_1 \le h$ and $x_1 > 0$: In this case $\ell(xv_i) \ge x_1$ and $\ell(v_j c) \ge \frac{w}{2}$, therefore, Equation (A.5) holds if and only if

$$\left(x_1 + \frac{w}{4}\right)^2 \ge \left(x_1 - \frac{w}{2}\right)^2 + y_1^2 \qquad \text{or} \qquad \left(x_1 + \frac{w}{4}\right)^2 \ge \left(x_1 - \frac{w}{2}\right)^2 + (y_1 - h)^2$$

We will look at the region where a stronger condition holds, namely

$$x_1^2 \ge \left(x_1 - \frac{w}{2}\right)^2 + y_1^2 \qquad \text{or} \qquad x_1^2 \ge \left(x_1 - \frac{w}{2}\right)^2 + (y_1 - h)^2$$

These constraints define region bounded by parabolas, and point of intersection of these two parabolas is the point furthest away from $\Pi(t,h,w)$ where both the conditions fail. The point of intersection of the parabolas is given by $p = \left(\frac{h^2+w^2}{4w}, \frac{h}{2}\right)$. Therefore, Equation (A.5) holds for all points p satisfying $x_1 \ge \frac{h^2+w^2}{4w}$. Since all points outside both the parabolas satisfy $x_1 \ge \mathrm{dist}(p, \mathcal{C})$, result holds for $D = \frac{h^2+w^2}{4w}$, since

Case 164. $0 \le y_1 \le h$ and $x_1 < 0$: Following the same computations as in Case 163, we get the exact same condition on D.

Now we are ready to prove structure of P, but first we need one definition.

Definition 165. Consider any Hamiltonian path P that looks like $p\overline{v_{i_1}v_{j_1}}c_1\overline{v_{i_2}v_{j_2}}c_2 \dots c_4\overline{v_{i_5}v_{j_5}}q$. For a subpath $p'\overline{v_i v_j}q'$ of P, where $p', q' \in \{p, q, c_1, c_2, c_3, c_4\}$, we define $d(p'q')$ as follows:

- $d(p'q') = \ell(p'v_i) + \ell(q'v_j)$ if $\overline{v_i v_j} \ne \varnothing$
- $d(p'q') = \ell(p'q')$ if $\overline{v_i v_j} = \varnothing$

Observation 160 implies that $d(p, c_1) \ge \mathrm{dist}(p, \mathcal{C})$. Further, for $1 \le a \le 3$, we have $d(c_\alpha, c_{\alpha+1}) \ge \min(h, w) = w$, since if $\overline{v_{i_{\alpha+1}}v_{j_{\alpha+1}}} \ne \varnothing$, $\ell(c_\alpha v_{i_{\alpha+1}}) + \ell(v_{j_{\alpha+1}c_{\alpha+1}}) \ge \frac{w}{2} + \frac{w}{2} = w$. There for we have the lower bound on length of any optimal Hamiltonian path P from p to q:

$$d(p,c_1)+d(c_1,c_2)+d(c_2,c_3)+d(c_3,c_4)+d(c_4,q)+\sum_{i=1}^{5} l(v_{i_1}v_{j_1}) \ge \mathrm{dist}(p,\mathcal{C})+\mathrm{dist}(q,\mathcal{C})+3w+h\left(1-\frac{4}{t}\right)$$

Note that since p, q are on different sides of line $y = \frac{h}{2}$, the nearest corners from p, q respectively are different and are not on the same short side of the gadget. Therefore, we can construct a Hamiltonian path Q such that

$$\ell(Q) \le \mathrm{dist}(p,\mathcal{C}) + \mathrm{dist}(q,\mathcal{C}) + 3w + h$$

In the path P, if the path $c_1c_2c_3c_4$ contains two longer sides of the gadget, then we have

$$\ell(P) \geq \text{dist}(p, \mathcal{C}) + \text{dist}(q, \mathcal{C}) + w + 2h$$

which is longer that Q if $h \geq 2w$. Observation 160 further implies that if $\overline{v_{i_1}v_{j_1}} \neq \varnothing$, then

$$\ell(P) \geq \text{dist}(p, \mathcal{C}) + \text{dist}(q, \mathcal{C}) + 3w + \frac{w}{4} + h - \frac{4h}{t}$$

Therefore, when $\frac{w}{4} \geq \frac{4h}{t}$ or equivalently $t \geq \frac{16h}{w}$, $\overline{v_{i_1}v_{j_1}} = \varnothing$ and $\overline{v_{i_5}v_{j_5}} = \varnothing$. Thus, the shortest Hamiltonian path P, is determined by choice of $\overline{v_{i_\alpha}v_{j_\alpha}}$ for $\alpha = 2, 3, 4$. Suppose without loss of generality that c_1c_2 is the shorter side of the gadget given by $y = 0$. Then the values of i_α, j_α that minimize $d(c_1c_2) + d(c_2c_3) + d(c_3c_4)$ are given by $i_2 = j_2 = 0, i_3 = 1, j_3 = t - 1, i_4 = j_4 = t$. This completely describes the shortest Hamiltonian path P, and both points c_2, c_3 satisfy the condition in the lemma, completing the proof. $\square$

A.2.2 Proof of Lemma 31

Lemma 31. *Let p, q be two points on the opposite sides of the line $y = \frac{h}{2}$ such that*

$$dist(\{p, q\}, \Pi(t, h, w)) \geq D.$$

Let P be a shortest Hamiltonian path from p to q in $\Pi(S(k), t, h, w) \cup \{p, q\}$. Suppose all of the following inequalities hold:

$$D \geq \tfrac{h^2+w^2}{4w} \qquad h \geq 2w \qquad w \geq 100 \qquad t \geq 2h \qquad \frac{h}{t} \leq \frac{4\pi}{k}$$

Then there is a Hamiltonian path Q from p to q in $\Pi(S(k), t, h, w) \cup \{p, q\}$ such that Q visits each S_i at most once, $\ell(Q) \leq \ell(P) + O(1/k)$ and for at least two $i \in 1, 2, 3, 4$ we have that neither neighbor v_i^1, v_i^2 of S_i on Q is not in $\{p, q\}$ and moreover, v_i^1, v_i^2 are two points in $\{v_1, \ldots, v_t\}$ closest to S_i.

Proof. Let π_i denote the center of S_i. Let $\mathcal{S} = \bigcup_{i=1}^4 S_i$ and $\mathcal{C} = \{\pi_1, \ldots, \pi_4\}$. Note that Observation 160 holds with when $D \geq \frac{h^2+w^2}{4w}$ with

$$\ell(pv_i) + \text{dist}(v_jS) \geq \text{dist}(pS) + \frac{w}{4} - 8$$

Since $\ell(pv_i) + \ell(v_jc) \geq \ell(pc) + \frac{w}{4}$ for center c of the gadget S and $\text{dist}\, v_jS \geq \ell(v_jc) - 4$ and $\ell(pc) \geq \text{dist}(pS) - 4$. Further, we can extend the path that we obtain in the proof of the previous lemma by including an Hamiltonian path through S_i when the path is supposed to visit π_i to get a Hamiltonian path P_1 from p to q with length at most

$$\ell(P_1) \leq \text{dist}(p, \mathcal{S}) + \text{dist}(q, \mathcal{S}) + 3(w - 8) + h + 4(10\pi + 8) + 16 \tag{A.6}$$

since length of tour in each gadget is $10\pi + 8$, actual distance between two closest gadgets is $w - 8$, and since we must enter and exit next in adjacent vertices to extend the tours as defined in Section 2.2.2, we pay an additional factor of 8. Now, we extend Definition 165 to sets:

Definition 166. Given a Hamiltonian path P in $\{p,q\} \cup \Pi(S(k),t,h,w)$ from p to q, which can be represented as $pv_{i_1}v_{j_1}T_1 \dots v_{i_u}v_{j_u}T_u v_{i_{u+1}}v_{j_{u+1}}q$ where for each i, T_i is a path such that $T_i \subseteq S_j$ for some $j \in \{1,\dots,4\}$. For any two sets $R_1, R_2 \in \{\{p\},\{q\},T_1,\dots,T_u\}$, such that there is a subpath $p'\overline{v_iv_j}q'$ in P, we define $d(R_1,R_2)$ as follows:

- $d(R_1R_2) = \mathrm{dist}(R_1v_i) + \mathrm{dist}(R_2v_j)$ if $\overline{v_iv_j} \neq \varnothing$
- $d(R_1R_2) = \mathrm{dist}(R_1R_2)$ if $\overline{v_iv_j} = \varnothing$

Observation 167. There is an absolute constant C such that when $k \geq C$, then P visits each S_i exactly once.

We can write P as $p\overline{v_{i_1}v_{j_1}}T_1 \dots \overline{v_{i_u}v_{j_u}}T_u\overline{v_{i_{u+1}}v_{j_{u+1}}}q$, where T_i is a path such that $T_i \subseteq S_j$ for some $j \in \{1,\dots,4\}$. Then we have $d(T_\alpha, T_{\alpha+1}) \geq w-8$ and $d(\{p\},T_1) \geq \mathrm{dist}(p,\mathcal{S})$ and $(\{q\},T_u) \geq \mathrm{dist}(q,\mathcal{S})$. Note that each point in S_i must still be connected to some vertex, and sum of the distances between each vertex and it's nearest neighbor is 40π. This gives the lower bound:

$$\ell(P) \geq \mathrm{dist}(x,\mathcal{C}) + \mathrm{dist}(y,\mathcal{C}) + (u-1)(w-8) + h + 4\left(10\pi + 8 - O\left(\frac{1}{k}\right)\right) - O\left(\frac{u}{k}\right) \tag{A.7}$$

The additive correction $O\left(\frac{u}{k}\right)$ is to account for double counting. All the vertices in $\mathcal{S}$ that are connected to something outside are counted twice, once in 40π and once in $(u-1)(w-8)$. We must subtract their contribution in the 40π term, which is at most $\frac{4\pi}{k}$ for each vertex. Since number of these connecting vertices is at most $2u$, we get the additive correction factor, with 8π being the constant hidden in O-notation. Observe that Since P is a shortest Hamiltonian path, it is shorter than P_1, and hence we must have

$$(u-4)\left(w-8-\frac{8\pi}{k}\right) - \frac{32\pi}{k} - 16 \leq 0$$

It follows that $u \leq 4$ if $w \geq 100$ for $k \geq 16\pi$. This finishes the proof of Observation 167.

Therefore, P looks like $p\overline{v_{i_1}v_{j_1}}T_1 \dots \overline{v_{i_4}v_{j_4}}T_4\overline{v_{i_5}v_{j_5}}q$. If $\overline{v_{i_1}v_{j_1}} \neq \varnothing$ then Observation 160 gives a better lower bound on $\ell(P)$. In particular, it increases the lower bound in Equation (A.7) by $\frac{w-8}{4}$. Comparing this lower bound on $\ell(P)$ with upper bound on $\ell(P_1)$ given in Equation (A.6), following must hold

$$\frac{w-8}{4} - 16 - O\left(\frac{1}{k}\right) - \frac{4h}{t} \leq 0$$

This fails to hold when $w \geq 100$ and $t \geq 2h$ for large enough k. Hence, we can conclude that $\overline{v_{i_1}v_{j_1}} = \overline{v_{i_5}v_{j_5}} = \varnothing$. Hence, if $\frac{h}{t} \approx \frac{4\pi}{k}$, then we can change P to Q by replacing tour inside T_2 and T_3 by the Hamiltonian path described in Section 2.2.2, and connecting it to it's nearest neighbors among $v_1,\dots,v_t$, which are either $\{v_1,v_2\}$ or $\{v_{t-1},v_t\}$ by choice of t. Note that this replacement strictly reduces the total cost outside the gadget, and is optimal inside the gadget up to an additive factor of $O(1/k)$. Therefore, we get the path Q such that

$$\ell(Q) \leq \ell(P) + O\left(\frac{1}{k}\right)$$

□

A.2.3 Proof of Lemma 33

Lemma 33. *Let $\varepsilon > 0$ be positive real. Then there exists constants $D_1, D_2 \geq 0$ such that if P is an optimal Hamiltonian tour over V, and if Δ_1 is any (ε, D_2) copy of $\Delta(D_1, \Pi_S(k))$, then there exists an $i \in \{1, 2, 3\}$ such that P visits Π_i exactly once, where Π_1, Π_2, Π_3 are (ε, D_1)-copies of $\Pi_S(k)$ contained in Δ_1, with centers C_1, C_2, C_3 respectively. Further if p, q are neighbors of T_i in P, then p, q lie on the opposite side of $\overleftrightarrow{OC_i}$, where O is the center of Δ_1. In particular, the values*

$$D_1 = \frac{2000}{1 - \cos\frac{\pi}{10}} \qquad \textit{and} \qquad D_2 = \frac{30000}{\left(1 - \cos\frac{\pi}{10}\right)^2} \tag{2.13}$$

suffice.

Proof. Since we choose $\Pi_S(k) = \Pi(S(k), \frac{200k}{4\pi}, 200, 100)$, the diameter of $\Pi_S(k)$ is at most 300. Let

$$D_1 = \frac{2000}{1 - \cos\frac{\pi}{10}}$$

be chosen to satisfy conditions of Lemmas 155 and 157 for $\Pi_S(k)$ and $\theta = \frac{\pi}{10}$. Then $\Delta(D_1, \Pi_S(k))$ has diameter at most $\frac{5000}{1-\cos(\pi/10)}$. Let

$$D_2 = \frac{30000}{\left(1 - \cos\frac{\pi}{10}\right)^2}$$

be chosen to satisfy conditions of Lemmas 155 and 157 for $\Delta(D_1, \Pi_S(k))$ and $\theta = \frac{\pi}{10}$. It follows that Δ_1 and Π_i for $i = 1, 2, 3$ can be visited by P at most twice, and if they are visited by P exactly twice, then all the four edges exiting the corresponding set are nearly parallel. We will say that P connects two sets $X, Y \subseteq V$ if and only if P contains an edge going from X to Y. Now, we do cases based on how many times these sets are visited.

Case 168. Suppose that there is Π_i such that P visits Π_i twice. Without loss of generality, we will assume that P visits Π_1 twice. Let e_1, e_2 and f_1, f_2 be two pairs of edges connecting Π_1 to $V \setminus \Pi_1$. If g_1 connects Π_1 to Π_2, and g_2 connects Π_1 to Π_3, where $g_1, g_2 \in \{e_1, e_2, f_1, f_2\}$, then g_1 and g_2 have an acute angle of at most $\frac{\pi}{3}$ between them. Since $\frac{\pi}{3} \geq \frac{\pi}{5}$, this contradicts Lemma 157. Hence, P connects Π_1 to exactly one of Π_2, Π_3.

Case 168.1. If Π_1 is connected to neither Π_2, Π_3, then P visits Δ_1 at least 3 times, twice in Π_1, and once in $\Pi_2 \cup \Pi_3$, which is a contradiction to Lemma 157. Without loss of generality, let Π_1 be connected to Π_2. Note that if Π_2 is not connected to Π_3, then P visits Δ_1 at least thrice, twice in $\Pi_1 \cup \Pi_2$, and at least once in Π_3.

Case 168.2. If P visits Π_2 twice, then Π_2 cannot be connected to Π_3, since it is already connected to Π_1, which is a contradiction.

Case 168.3. If P visits Π_2 exactly once, then P must connect Π_2 to both Π_1, Π_3, and since Π_1 and Π_3 are on opposite sides of $\overleftrightarrow{OC_2}$, $i = 2$ satisfies all the conditions of the lemma.

Case 169. Suppose that each of Π_1, Π_2, Π_3 is visited exactly once. Now, we have two cases based on how many times Δ_1 is visited.

Case 169.1. If Δ_1 is visited exactly once, then P must visit Π_1, Π_2, Π_3 in some order, covering the whole set. Suppose this order is $\Pi_{j_1}\Pi_{j_2}\Pi_{j_3}$. Then $i = j_2$ satisfies all the conditions of lemma, since Π_{j_1} and Π_{j_3} are on opposite side of $\overleftrightarrow{OC_{j_2}}$.

Case 169.2. If Δ_1 is visited twice, then let P intersect Δ_1 in two contiguous subpaths, say Q_1, Q_2. Without loss of generality, suppose that $\Pi_1 \subseteq Q_1$ and $\Pi_2, \Pi_3 \subseteq Q_2$. Let e_1, e_2 be pair of edges that connects Π_1 to $V \setminus \Pi_1$. Let $e_i = \{x_i, p_i\}$ where $p_i \notin \Pi_1$, and $x_i \in \Pi_1$. By Lemma 155, $\measuredangle(\overrightarrow{x_1p_1}, \overrightarrow{x_2p_2}) \in \pi \pm \frac{\pi}{10}$. If possible, let p_1, p_2 be on the same side of $\overleftrightarrow{OC_1}$. Further, without loss of generality, let $\measuredangle(\overrightarrow{C_1O}, \overrightarrow{C_1p_1}), \measuredangle(\overrightarrow{C_1O}, \overrightarrow{C_1p_2}) \in [0, \pi]$. Let $\theta_1 = \measuredangle(\overrightarrow{C_1O}, \overrightarrow{x_1p_1})$ and $\theta_2 = \measuredangle(\overrightarrow{C_1O}, \overrightarrow{x_2p_2})$. Since p_1, p_2 are on the same side of $\overleftrightarrow{OC_1}$, we must have

$$d + D_2 \sin\theta_1 \geq 0 \qquad d + D_2 \sin\theta_2 \geq 0$$

This implies that $\sin\theta_i \geq -\frac{d}{D_2}$. Since $|\theta_1 - \theta_2| \in \pi \pm \frac{\pi}{10}$, it implies that

$$\theta_i \in \left[-\frac{\pi}{9}, \frac{\pi}{9}\right] \cup \left[\pi - \frac{\pi}{9}, \pi + \frac{\pi}{9}\right]$$

In fact, each of the two intervals contains exactly one θ_i. Suppose $\theta_1 \in \left[-\frac{\pi}{9}, \frac{\pi}{9}\right]$. Observe that for any $y_2 \in \Pi_2$, $\measuredangle(\overrightarrow{C_10}, \overrightarrow{x_1y_2}) \leq -\frac{\pi}{7}$ and for any $y_3 \in \Pi_3$, $\measuredangle(\overrightarrow{C_10}, \overrightarrow{x_1y_3}) \geq \frac{\pi}{7}$. It follows that for any $y_2 \in \Pi_2$ and $y_3 \in \Pi_3$, p_1 is contained in $\angle y_2x_1y_3$. Since $\ell(x_1y_2), \ell(x_1y_3) \leq D_1 + 4d \leq D_2 \leq \ell(x_1p_2)$, the edge e_1 must intersect edge y_2y_3. Since Q_2 connects Π_2, Π_3, this implies that e_1 intersects and edge in Q_2, implying that P is not planar! But since P is the optimal Hamiltonian path, it must by planar, contradiction!

This covers all the cases, completing the proof of lemma. □

A.2.4 Proof of Lemma 22

Here we provide some more details for the proof of Lemma 22 for sake of completeness.

Lemma 170. *Consider the gadget $S = S(k)$ defined in Definition 7 for large enough k. Let $p, q \in S$ be two points on the outer circle. Then the shortest Hamiltonian path from p to q completely covering S has length at least $10\pi + 8 - \frac{12\pi}{k}$.*

Proof. For this proof, we will approximate smaller segments along the circles by the arcs, the difference between them is $O(k^{-3})$, and since there are $O(k)$ of them, all the computations holds up to $O(k^2)$ error.

Let O_1 denote the set of point on the inner circle of S and let O_2 denote the set of points on the inner circle. Let $G = \{g_1 = (-2, 0), g_2 = (2, 0)\}$ be the set of gap vertices. Let P be the shortest Hamiltonian path from p to q in S. To each vertex in S, we associate the length of the edge leaving that vertex in P as the cost. Cost of each vertex in O_1 is at least $\frac{2\pi}{k}$ and cost of each vertex in O_2 is at least $\frac{4\pi}{k}$. Consider the path P_1 obtained by deleting G from P. Then the path P must leave and enter O_2 at least once, and the number of edges in P that contain exactly one vertex in O_2 is even. Let $2t$ denote number of such edges. Thus, ever such edge costs at least $3 - \frac{4\pi}{k}$ additional length to the path P_1. This gives us the lower bound:

$$\ell(P) \geq \ell(P_1) \geq 10\pi + 2t\left(3 - \frac{4\pi}{k}\right)$$

For $k \geq \frac{4\pi}{3}$, this is an increasing function in t. Further, for $k \geq 4\pi$, value of this function at $t = 2$ is at least $10\pi + 8$. Therefore all the paths with $t \geq 2$ satisfy the required length condition.

Suppose that $t = 1$, but the original path P leaves O_2 more than once. Then, there must be a gap vertex that has both of it's neighbors in O_2. This implies $\ell(P) \geq \ell(P_1) + 4 - \frac{4\pi}{k}$. Since $t = 1$, we have $\ell(P_1) \geq 10\pi + 6 - \frac{8\pi}{k}$, we get the bound

$$\ell(P) \geq 10\pi + 8 - \frac{12\pi}{k}$$

which satisfies the requirement of the theorem. Similarly, if there is a vertex $g \in G$ such that both neighbors of G lie in O_1, then this implies $\ell(P) \geq \ell(P_1) + 2 - \frac{4\pi}{k}$. This leads to exactly the same length bound as above.

Hence, we are left with the case with path P leaves and enters O_2 exactly once and both g_1 and g_2 have exactly one neighbor in O_1 and one in O_2. Suppose p_1 and q_1 are neighbors of g_1 and g_2 respectively in O_1. We claim that any Hamiltonian path Q from p_1 to q_1 in O_1 must have length at least $\mathrm{dist}(p_1, q_1) + 2\pi - \frac{4\pi}{k}$.

Note that line $\overleftrightarrow{p_1q_1}$ divides O_1 in two parts, say H_1 and H_2. For sake of notational convenience, we will include p_1, q_1 in both H_1 and H_2. Let Q be denoted by $p_1 = v_0, \ldots, v_t = q_1$. For each i, define α_i to be the point in H_1 that is furthest away from p_1 and β_i to be the point in H_2 that is furthest away from p_1. We claim that following holds for each i:

1. v_i either equals α_i or β_i.
2. v_{i+1} is neighbor of either α_i or β_i.

We will prove this by induction. First observe that (1) holds for $i = 0$, since $v_0 = p_1$. Assume the strong induction hypothesis that both (1), (2) holds for all $j < i$, and (1) holds for i. We will show that this implies (2) holds for i and (1) holds for $i + 1$, completing the induction. Because of the induction hypothesis, P must have visited all the vertices between p_1 and α_1, β_1 in $\{v_0, \ldots, v_i\}$, since the set of visited vertices forms a contiguous segment on the circle. Suppose that v_{i+1} is not a neighbor of either α_i or β_i. Then there is a vertex v such that v and q_1 are on the opposite sides of line $\overleftrightarrow{v_iv_{i+1}}$. Since Q must visit v before visiting q_1, it must intersect the line $\overleftrightarrow{v_iv_{i+1}}$. Since the segment v_iv_{i+1} completely partitions the convex hull of O_2 into two parts, any path from v to q_1 through the convex hull of O_1 must intersect v_iv_{i+1}, contradicting the planarity of the shortest path. This implies (2). Further, since all the points between α_i and β_i are already visited, v_{i+1} is outside this segment, which implies that v_{i+1} is either α_{i+1} or β_{i+1}.

This proves the claim. The path P must connect $H_1 \setminus \{p_1, q_1\}$ and $H_2 \setminus \{p_1, q_1\}$, and hence it crosses $\overleftrightarrow{p_1q_1}$ at least once. Suppose it crosses the segment more than once. Let v_av_{a+1} and v_bv_{b+1} be the two segments with least indices a, b which cross p_1q_1. Then v_{a+1} is a neighbor of p_1 and v_{b+1} is a neighbor of v_a. Let p_2 be neighbor of p_1 other than v_{a+1}. Note that p_2 is between

p_1 and v_a. Consider the path $Q_1 = \overline{p_1 v_b}\,\overline{p_2 v_{b+1}}$. We claim that this is shorter than the path $Q_2 = \overline{p_1 v_a}\,\overline{v_{a+1} v_b} v_{b+1}$, where $\overline{xy}$ denotes the path covering all the points between x and y which are on the same side of $\overleftrightarrow{p_1 q_1}$ as x, y. Note that $\ell(\overline{p_1 v_a}) = \overline{p_2 v_{b+1}}$ and $\ell(\overline{v_{a+1} v_b}) + \ell(p_1 v_{a+1}) = \ell(\overline{p_1 v_b})$. Therefore, it suffices to show that

$$\ell(v_a v_{a+1}) + \ell(v_b + v_{b+1}) - \ell(p_1 v_{a+1}) - \ell(v_b p_2) \geq 0$$

Let $\angle p_1 O v_{b+1} = \alpha \frac{2\pi}{k}$ and $\angle p_1 O v_b = \beta \frac{2\pi}{k}$ where O is center of O_1. Then we can express all the lengths in terms of sines to get

$$\begin{aligned}
&\ell(v_a v_{a+1}) + \ell(v_b + v_{b+1}) - \ell(p_1 v_{a+1}) - \ell(v_b p_2) \\
=&2\sin\left(\frac{\alpha}{2} \cdot \frac{2\pi}{k}\right) + 2\sin\left(\frac{\alpha+\beta}{2} \cdot \frac{2\pi}{k}\right) - 2\sin\left(\frac{\beta+1}{2} \cdot \frac{2\pi}{k}\right) - 2\sin\left(\frac{1}{2} \cdot \frac{2\pi}{k}\right) \\
=&4\sin\left(\frac{2\alpha+\beta}{2k}\pi\right)\cos\left(\frac{\beta}{2k}\pi\right) - 4\sin\left(\frac{\beta+2}{2k}\pi\right)\cos\left(\frac{\beta}{2k}\pi\right) \\
=&4\cos\left(\frac{\beta}{2k}\pi\right)\left(\sin\left(\frac{2\alpha+\beta}{2k}\pi\right) - \sin\left(\frac{\beta+2}{2k}\pi\right)\right) \\
=&8\cos\left(\frac{\beta}{2k}\pi\right)\cos\left(\frac{\alpha+\beta+1}{2k}\pi\right)\sin\left(\frac{\alpha-1}{2k}\pi\right)
\end{aligned}$$

Since $0 \leq \beta \leq \alpha + \beta + 1 \leq k$, and $\alpha \geq 1$, all the angles in the expression above are between 0 and $\frac{\pi}{2}$, which proves that this expression is always positive implying that Q_1 is shorter than Q_2. We can now replace portion of Q corresponding to Q_2 by Q_1 to get a shorter path if Q crossed $\overleftrightarrow{p_1 q_1}$ more than once. Hence, any optimal Hamiltonian path Q must cross $\overleftrightarrow{p_1 q_1}$ exactly once. Therefore, Q must look like $Q = \overline{p_1 q_2}\,\overline{p_2 q_1}$, where q_2 is a neighbor of q_1 that is on the opposite side of p_2. The points $p_1 p_2 q_1 q_2$ form a cyclic trapezoid, with $p_1 q_1$ and $p_2 q_2$ as diagonals. Therefore,

$$\ell(Q) \geq \text{dist}(p_1 q_1) + 2\pi - \frac{4\pi}{k}$$

Note that we are missing the trivial case when p_1 and q_1 are adjacent, which can be verified to give the exact same bound.

Hence, portion of path P in between two gap vertices has length $\ell(g_1 p_1) + \ell(p_1 q_1) + 2\pi - \frac{4\pi}{k} + \ell(q_1 g_2)$, which is at least $4 + 2\pi - \frac{4\pi}{k}$. Combined with the cost of the path outside two gap vertices, which is at least $8\pi + 4 - \frac{8\pi}{k}$, we get the lower bound

$$\ell(P) \geq 10\pi + 8 - \frac{12\pi}{k} - O(k^{-2})$$

as required. Error term of $O(k^{-2})$ appears from approximating small chords of circle by the arc-lengths. □

In particular, the lemma above gives the lower bound

$$\ell(P) \geq 10\pi + 8 - O\left(\frac{1}{k}\right)$$

as required in Lemma 22.

A.3 Probability bounds for Observation 10

In this section we will provide precise bounds for constant $C^S_{\varepsilon,D}$ defined in Observation 10. More precisely, we will prove the following version of Observation 10:

Lemma 171. *Let d be a fixed integer. Let $\{Y_1, Y_2, \ldots\}$ be a sequence of points drawn uniformly at random from $[0,t]^d$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$, where $t = n^{1/d}$. Given any finite point set $S \subseteq \mathbb{R}^d$ with k points, any $\varepsilon > 0$ and any constant $D > 0$ such that*

1. *ε is smaller than distance between any two points in S; and*
2. *D is larger than the diameter of S*
3. *$\exp(O(k \log(1/\varepsilon))) = o(n)$*

$\mathcal{Y}_n$ contains at least $C^k_{\varepsilon,D} n$ many (ε, D)-copies of S with probability $1 - o(1)$, where

$$C^S_{\varepsilon,D} = \exp(-O(k \log(1/\varepsilon)))$$

where O-notation hides constants dependent on d and D.

Further, if $\exp(O(k \log(1/\epsilon))) \leq \frac{n}{\delta \log n}$ then the result holds with probability $1 - n^{-\delta}$.

Proof. Divide $[0,t]^d$ into boxes of of side length $3D$. Let B denote one such box. Consider a copy of S centered at center of the box B. Let $s_1, \ldots, s_k$ be points in S. For any j, the probability that Y_j at most ε distance from s_i is $\frac{V_d(\varepsilon)}{n}$ where $V_d(R)$ denotes volume of a sphere of radius R in $\mathbb{R}^d$. Given a sequence of points $j_1, \ldots, j_k$, the probability that Y_{j_i} is ε-close to s_i for all i, and there are no other points inside B is given by

$$\left(\frac{V_d(\varepsilon)}{n}\right)^k \left(1 - \frac{(3D)^d}{n}\right)^{n-k}$$

The number of choices for the sequence j_i is exactly

$$\frac{n!}{(n-k)!} \geq \left(1 - \frac{k}{n}\right)^k n^k$$

Since the events corresponding to all the sequences are disjoint, we can simply add these probabilities! Recall that $\log V_d(\varepsilon) = O(-d \log d + d \log \varepsilon)$, and that $1 - x \geq e^{-x/(1-x)} \geq e^{-2x}$ for $x \leq \frac{1}{2}$. Using these two identities, and the two probability bounds above, we get a lower bound on probability that the box B contains an (ε, D)-copy of S:

$$\exp\left(-O\left(dk \log d - dk \log \varepsilon + (n-k)\frac{(3D)^d}{n} + \frac{k^2}{n}\right)\right) = \exp(-O(k \log(1/\varepsilon)))$$

where O hides constants dependent on the gap distance D and dimension d.

There are $\frac{n}{(3D)^d}$ such boxes B. Let these be denoted by $B_1, \ldots, B_s$. Let χ_i be the indicator random variable for box B_i containing an (ε, D)-copy of S. Let $\chi = \sum_i \chi_i$. Then we have

$$\mathbb{E}[\chi] = \sum_i \mathbb{E}[\chi_i] = n \exp(-O(k \log(1/\varepsilon) + d \log D)) = n \exp(-O(k \log(1/\varepsilon)))$$

We will use McDiarmid's Inequality to get a concentration bound on χ.

Lemma 172 (McDiarmid's Inequality)**.** *Suppose a function* $f : \mathcal{Z}_1 \times \cdots \mathcal{Z}_n \to \mathbb{R}$ *satisfies that for all* i

$$\sup_{z_i' \in \mathcal{Z}_i} \left| f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_{i'}, z_{i+1}, \ldots, z_n) \right| \le c_i$$

then for independent random variables $Z_i \sim \mathcal{Z}_i$,

$$\mathbb{P}\Big[f(Z_1, \ldots, Z_n) - \mathbb{E}[f(Z_1, \ldots, Z_n] \le -t \Big] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right)$$

Note that changing a single point $y \in \mathcal{Y}_n$ changes χ by at most 2. Therefore, we can use McDiarmid's Inequality with $c_i = 2$ and for all i and let $t = \frac{1}{2}\mathbb{E}[\chi]$ to get that

$$\mathbb{P}\left[\chi < \frac{1}{2}\mathbb{E}[\chi] \right] \le \exp\left(-\frac{2n^2 \exp(-O(k \log(1/\varepsilon)}{4n} \right) = \exp(-n \exp(-O(k \log(1/\varepsilon)))$$

Note that when $\exp(O(k \log(1/\varepsilon))) = o(n)$, this probability is $o(1)$, which proves that

$$\chi \ge n \exp(-O(k \log(1/\varepsilon)))$$

with probability $1 - o(1)$. Further, if $\exp(k \log(1/\varepsilon)) \le \frac{n}{\delta \log n}$, then the result above holds with probability $1 - \frac{1}{n^\delta}$. ☐

In particular, for $\varepsilon = O(1/k)$, and $k \le \frac{\log n}{\log \log n}$, we have

$$\begin{aligned} \exp(k \log(1/\varepsilon) = \exp(k \log k) &= \exp\left(\frac{\log n}{\log \log n} (\log \log n - \log \log \log n) \right) \\ &= \frac{n}{\exp\left(\frac{\log n \log \log \log n}{\log \log n} \right)} = o\left(\frac{n}{\log n} \right) \end{aligned}$$

since

$$\log \log n = o\left(\frac{\log n \log \log \log n}{\log \log n} \right)$$

Note that this falls under the setting in which we use this result in proof of theorem 6.

Appendix B

Details of

B.1 Technical Details for Section 6.2

Theorem 173 ([VV85; Coo71])**.** *Suppose that there is a randomized* $\mathrm{poly}(n)$*-time algorithm for the following problem: given a 3-CNF formula $\mathcal{C}$ with n variables and at most $5n$ clauses, under the promise that $\mathcal{C}$ has at most one satisfying assignment, determine whether $\mathcal{C}$ is satisfiable. Then, NP = RP.*

Lemma 174. *In the setting of Definition 87, set $d := 7$ and $B := 64m\alpha + 2\beta$. Then $p_{\mathcal{C},\alpha,\beta} \in \mathcal{P}_{n,d,B}$.*

Proof. Since $\alpha H_{\mathcal{C}}(x) + \beta G(x)$ is a polynomial in $x_1, \dots, x_n$ of degree at most 7, there is some $\theta = \theta(\mathcal{C}, \alpha, \beta) \in \mathbb{R}^{M-1}$ such that $\langle \theta, T(x) \rangle + \alpha H_{\mathcal{C}}(x) + \beta G(x)$ is a constant independent of x. Then $h(x) \exp(-\alpha H_{\mathcal{C}}(x) - \beta G(x))$ is proportional to $h(x) \exp(\langle \theta, T(x) \rangle)$, so $p_{\mathcal{C},\alpha,\beta} = p_\theta$. Moreover, for any clause C_j, every monomial of H_{C_j} has coefficient at most 64 in absolute value, so every monomial of $H_{\mathcal{C}}$ has coefficient at most $64m$. Similarly, every monomial of G has coefficient at most 2 in absolute value. Thus, $\|\theta\|_\infty \leq 64m\alpha + 2\beta =: B$, so $p_{\mathcal{C},\alpha,\beta} \in \mathcal{P}_{n,d,B}$. $\square$

Given a point $v \in \mathcal{H}$, let $\mathcal{O}(v) := \{x \in \mathbb{R}^n : x_i v_i \geq 0; \forall i \in [n]\}$ denote the octant containing v, and let $\mathcal{B}_r(v) := \{x \in \mathbb{R}^n : \|x - v\|_\infty \leq r\}$ denote the ball of radius r with respect to ℓ_∞ norm.

Lemma 175. *Let $p := p_{\mathcal{C},\alpha,\beta}$ and $Z := Z_{\mathcal{C},\alpha,\beta}$ for some 3-CNF $\mathcal{C}$ with m clauses and n variables, and some parameters $\alpha, \beta > 0$. Let $r \in (0,1)$. If $\beta \geq 40r^{-2}\log(4n/r)$, then for any $v \in \mathcal{H}$ that is a satisfying assignment for $\mathcal{C}$,*

$$\Pr_{x \sim p}(x \in \mathcal{B}_r(v)) \geq \frac{e^{-1-81m\alpha r^2}}{Z} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx \right)^n.$$

For any $w \in \mathcal{H}$ that is not a satisfying assignment for $\mathcal{C}$,

$$\Pr_{x \sim p}(x \in \mathcal{O}(w)) \leq \frac{e^{-\alpha}}{Z} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx \right)^n.$$

Proof. We begin by lower bounding the probability over $\mathcal{B}_r(v)$. Pick any clause C_ℓ included in $\mathcal{C}$. We claim that $H_{C_\ell}(v') \leq 81r^2$ for all $v' \in \mathcal{B}_r(v)$. Indeed, say that $C_\ell = \tilde{x}_i \vee \tilde{x}_j \vee \tilde{x}_k$. Since v satisfies C_ℓ, at least one of $\{f_i(v_i), f_j(v_j), f_k(v_k)\}$ must be zero. Without loss of generality, say that $f_i(v_i) = 0$; also observe that $|f_j(v_j)|, |f_k(v_k)| \leq 2$. It follows that for any $v' \in \mathcal{B}_r(v)$, $|f_i(v_i')| \leq r$ and $|f_j(v_j')|, |f_j(v_k')| \leq 2 + r \leq 3$ (since $r \leq 1$). Therefore, we have

$$H_{C_\ell}(v') \leq r^2 \cdot (3)^2 \cdot (3)^2 = 81r^2.$$

Summing over all m possible clauses, we have $H_{\mathcal{C}}(v') \leq 81mr^2$ for all $v' \in \mathcal{B}_r(v)$. Hence,

$$\begin{aligned}
\Pr_{x \sim p}(x \in \mathcal{B}_r(v)) &= \frac{1}{Z} \int_{\mathcal{B}_r(v)} \exp\left(-\sum_{i=1}^n x_i^8 - \alpha H_{\mathcal{C}}(x) - \beta G(x)\right) dx \\
&\geq \frac{e^{-81m\alpha r^2}}{Z} \int_{\mathcal{B}_r(v)} \exp\left(-\sum_{i=1}^n x_i^8 - \beta G(x)\right) dx \\
&= \frac{e^{-81m\alpha r^2}}{Z} \left(\int_{1-r}^{1+r} \exp(-x^8 - \beta(1-x^2)^2)\, dx\right)^n \\
&\geq \frac{e^{-81m\alpha r^2}}{Z} \left(1 + \frac{1}{n}\right)^{-n} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx\right)^n \qquad \text{(B.1)} \\
&\geq \frac{e^{-1-81m\alpha r^2}}{Z} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx\right)^n
\end{aligned}$$

where the second inequality (B.1) is by Lemma 176. Next, we upper bound the probability over $\mathcal{O}(w)$. Let C_ℓ be any clause in $\mathcal{C}$ that is not satisfied by w. Say that $C_\ell = \tilde{x}_i \vee \tilde{x}_j \vee \tilde{x}_k$. Then $|f_i(w_i)| = |f_j(w_j)| = |f_k(w_k)| = 2$. Furthermore, for any $w' \in \mathcal{O}^d(w)$, we have $|f_i(w_i')|, |f_j(w_j')|, |f_k(w_k')| \geq 1$, and hence $H_{C_\ell}(w') \geq 1$. Since $H_{C'}(x) \geq 0$ for all x, C', we conclude that $H_{\mathcal{C}}(w') \geq H_{C_\ell}(w') \geq 1$ for all $w' \in \mathcal{O}(w)$. In particular, this gives us

$$\begin{aligned}
\Pr_{x \sim p}(x \in \mathcal{O}(w)) &= \frac{1}{Z} \int_{\mathcal{O}(w)} \exp\left(-\sum_{i=1}^n x_i^8 - \alpha H_{\mathcal{C}}(x) - \beta G(x)\right) dx \\
&\leq \frac{e^{-\alpha}}{Z} \int_{\mathcal{O}(w)} \exp\left(-\sum_{i=1}^n x_i^8 - \beta G(x)\right) dx \\
&= \frac{e^{-\alpha}}{Z} \left(\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\, dx\right)^n
\end{aligned}$$

as claimed. ∎

B.1.1 Integral bounds

Lemma 176. *Fix $\beta > 150$ and $\gamma \in [0,1]$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \gamma x^8 + \beta(1-x^2)^2$. Pick any $r \in (6/\beta, 0.04)$. Then*

$$\int_0^\infty \exp(-f(x))\, dx \leq \left(\frac{1}{1 - \exp(-\beta r^2/8)} + \frac{2\exp(-\beta r/40)}{r}\right) \int_{1-r}^{1+r} \exp(-f(x))\, dx.$$

In particular, for any $m \in \mathbb{N}$, if $\beta \geq 40r^{-2}\log(4m/r)$, then

$$\int_0^\infty \exp(-f(x))\,dx \leq \left(1+\frac{1}{m}\right)\int_{1-r}^{1+r}\exp(-f(x))\,dx.$$

Proof. Set $a = 1/\sqrt{2}$. For any $x \in [a,\infty)$ we have $f''(x) = 56\gamma x^6 - 2\beta + 6\beta x^2 \geq \beta > 0$ for $\beta > 150$. Thus, f has at most one critical point in $[a,\infty)$; call this point t_0. Since $f'(x) = 8\gamma x^7 - 4\beta x(1-x^2)$, we have $f'(1) = 8\gamma \geq 0$ and $f'(1-3/\beta) \leq 8-4\beta(1-3/\beta)(3/\beta)(2-3/\beta) < 0$. Thus, $t_0 \in (1-3/\beta, 1]$. Set $r' = r - 3/\beta \geq r/2$. Then

$$\int_{1-r}^{1+r}\exp(-f(x))\,dx \geq \int_{t_0-r'}^{t_0+r'}\exp(-f(x))\,dx.$$

For every $t \in \mathbb{R}$ define $I(t) = \int_t^{t+r'}\exp(-f(x))\,dx$. Since f is β-strongly convex on $[a,\infty)$, we have for any $t \geq t_0$ that

$$f(t+r') - f(t) \geq r'f'(t) + \frac{r'^2}{2}\beta \geq \frac{r'^2}{2}\beta$$

where the final inequality is because $f'(t) \geq 0$ for $t \in [t_0,\infty)$. Thus, for any $t \geq t_0$,

$$I(t+r') = \int_{t+r'}^{t+2r'}\exp(-f(x))\,dx = \int_t^{t+r}\exp(-f(x+r'))\,dx \leq \exp(-\beta r'^2/2)I(t).$$

By induction, for any $k \in \mathbb{N}$ it holds that $I(t_0+kr') \leq \exp(-\beta k r'^2/2)I(t_0)$, so

$$\int_{t_0}^\infty \exp(-f(x))\,dx = \sum_{k=0}^\infty I(t_0+kr') \leq I(t_0)\sum_{k=0}^\infty \exp(-\beta k r'^2/2) = \frac{I(t_0)}{1-\exp(-\beta r'^2/2)}. \tag{B.2}$$

Similarly, for any $t \in [a+r', t_0]$, we have

$$f(t-r') - f(t) \geq -r'f'(t) + \frac{r'^2}{2}\beta \geq \frac{r'^2}{2}\beta$$

using β-strong convexity on $[a,\infty)$ and the bound $f'(t) \leq 0$ on $[a,t_0]$. Thus, for any $t \in [a, t_0 - r']$,

$$I(t-r') = \int_{t-r'}^{t}\exp(-f(x))\,dx = \int_t^{t+r'}\exp(-f(x-r'))\,dx \leq \exp(-\beta r'^2/2)I(t),$$

so by induction, $I(t_0-kr') \leq \exp(-\beta(k-1)r'^2/2)I(t_0-r')$ for any $1 \leq k \leq K := \lfloor (t_0-a)/r' \rfloor$. It follows that

$$\int_{t_0-Kr'}^{t_0}\exp(-f(x))\,dx = \sum_{k=1}^K I(t_0-kr') \leq I(t_0-r')\sum_{k=1}^K \exp(-\beta(k-1)r'^2/2) \leq \frac{I(t_0-r')}{1-\exp(-\beta r'^2/2)}. \tag{B.3}$$

Finally, note that $t_0-(K-1)r' \leq a+2r' \leq 0.8$. For any $x \in [0, 0.8]$, we have $f'(x) \leq 8x^7 - 0.72\beta x = x(8x^6 - 1.44\beta) \leq 0$, since $\beta > 150$. That is, f is non-increasing on $[0, t_0-(K-1)r']$. It follows that

$$\begin{aligned}
\int_0^{t_0-Kr'} \exp(-f(x))\,dx &\leq \frac{t_0-Kr'}{r'} \int_{t_0-Kr'}^{t_0-(K-1)r'} \exp(-f(x))\,dx \\
&\leq \frac{1}{r'} I(t_0-Kr') \\
&\leq \frac{\exp(-\beta(K-1)r'^2/2)}{r'} I(t_0-r').
\end{aligned}$$

Since $(K-1)r' \geq t_0 - 0.8 \geq 1 - \frac{3}{\beta} - 0.8 \geq 0.1$, we conclude that

$$\int_0^{t_0-Kr'} \exp(-f(x))\,dx \leq \frac{\exp(-\beta r'/20)}{r'} I(t_0-r'). \tag{B.4}$$

Combining (B.2), (B.3), and (B.4), we get

$$\int_0^{\infty} \exp(-f(x))\,dx \leq \left(\frac{1}{1-\exp(-\beta r'^2/2)} + \frac{\exp(-\beta r'/20)}{r'}\right) \int_{t_0-r'}^{t_0+r'} \exp(-f(x))\,dx.$$

Substituting in $r' \geq r/2$ gives the claimed result. □

Lemma 177. *Fix $\beta \geq 160\log(8)$. Then for any $1 \leq k \leq 8$,*

$$\int_0^{\infty} x^k \exp(-x^8 - \beta(1-x^2)^2)\,dx \leq 2^k \int_0^{\infty} \exp(-x^8 - \beta(1-x^2)^2)\,dx.$$

Proof. Define a distribution $q(x) \propto \exp(-x^8 - \beta(1-x^2)^2)$ for $x \in [0, \infty)$. We want to show that $\mathbb{E}_q[x^k] \leq 2^k$. Indeed,

$$\begin{aligned}
\mathbb{E}_q[\exp(x^8)] &= \frac{\int_0^\infty \exp(-\beta(1-x^2)^2)\,dx}{\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\,dx} \\
&\leq \frac{2\int_{1/2}^{3/2} \exp(-\beta(1-x^2)^2)\,dx}{\int_0^\infty \exp(-x^8 - \beta(1-x^2)^2)\,dx} \\
&= 2\mathbb{E}_q[\exp(x^8)\mathbb{1}[1/2 \leq x \leq 3/2]] \\
&\leq 2\exp((3/2)^8)
\end{aligned}$$

where the first inequality is by an application of Lemma 176 with $r = 1/2$ and $m = 1$. Now by Jensen's inequality we get

$$\mathbb{E}_q[x^8] \leq \log \mathbb{E}_q[\exp(x^8)] = \log(2) + (3/2)^8 \leq 2^8$$

and consequently, an application of Hölder inequality gives us $\mathbb{E}_q[x^k] \leq 2^k$, for any $1 \leq k \leq 8$. □

B.2 Moment bounds

Lemma 178 (Moment bound)**.** *For any $\theta \in \Theta_B$, $i \in [n]$, and $\ell \in \mathbb{N}$ it holds that*

$$\mathbb{E}_{x\sim p_\theta} x_i^\ell \leq \max(2\ell^\ell, B^\ell M^\ell 2^{\ell(d+1)+1}).$$

Proof. Without loss of generality assume $i = 1$. Let $L_0 := \max(\ell, BM2^{d+1})$. Then

$$\begin{aligned}
\mathbb{E}_{x\sim p_\theta} x_1^\ell &\leq L_0^\ell + \mathbb{E}_{x\sim p_\theta} x_1^\ell \mathbb{1}[\|x\|_\infty > L_0] \\
&= L_0^\ell + \sum_{k=0}^{\infty} \mathbb{E}_{x\sim p_\theta}\left[x_1^\ell \mathbb{1}[2^k L_0 < \|x\|_\infty \leq 2^{k+1} L_0]\right]
\end{aligned}$$

Now for any $L \geq L_0$,

$$\begin{aligned}
&\mathbb{E}\left[x_1^\ell \mathbb{1}[L < \|x\|_\infty \leq 2L]\right] \\
&= \frac{1}{Z_\theta}\int_{B_{2L}(0)\setminus B_L(0)} x_1^\ell \exp\left(-\sum_{i=1}^n x_i^{d+1} + \langle \theta, T(x)\rangle\right) dx \\
&\leq \frac{(2L)^n}{Z_\theta}(2L)^\ell \exp\left(-L^{d+1} + BM(2L)^d\right) \\
&\leq \frac{(2L)^{n+\ell}\exp(-L^{d+1}/2)}{Z_\theta}.
\end{aligned}$$

We can lower bound Z_θ as

$$\begin{aligned}
Z_\theta &\geq \int_{\mathcal{B}_{1/(BM)}(0)} \exp\left(-\sum_{i=1}^n x_i^{d+1} + \langle \theta, T(x)\rangle\right) dx \\
&\geq (BM)^{-n}\exp(-n(BM)^{-d-1} - BM(BM)^{-d}) \\
&\geq e^{-2}(BM)^{-n}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[x_1^\ell \mathbb{1}[L < \|x\|_\infty \leq 2L]\right] &\leq \exp\left((n+\ell)\log(2L) - \frac{1}{2}L^{d+1} + 2 + n\log(BM)\right) \\
&\leq \exp\left(-\frac{1}{4}L^{d+1}\right)
\end{aligned}$$

since L was assumed to be sufficiently large (recall that we assume $B \geq 1$). We conclude that

$$\begin{aligned}
\mathbb{E}_{x\sim p_\theta} x_1^\ell &\leq L_0^\ell + \sum_{k=0}^{\infty}\exp\left(-\frac{1}{4}2^{k(d+1)}L_0^{d+1}\right) \\
&\leq L_0^\ell + 1 \qquad \leq 2L_0^\ell
\end{aligned}$$

which completes the proof. □

Lemma 179 (Smoothness bounds)**.** *For every $\theta \in \Theta_B$, it holds that*

$$\mathbb{E}_{x\sim p_\theta}\|\Delta T(x)\|_2^2 := \sum_{j=1}^{M} \mathbb{E}_{x\sim p_\theta}(\Delta T_j(x))^2 \le d^4 B^{2d} M^{2d+1} 2^{2d(d+1)+1}$$

and

$$\mathbb{E}_{x\sim p_\theta}\|(JT)(x)\|_{\mathsf{op}}^2 \le n d^2 B^{2d} M^{2d+1} 2^{2d(d+1)+1}.$$

Proof. Fix any $j \in [M]$; then there is a degree function $\mathbf{d}$ with $1 \le |\mathbf{d}| \le d$ so that $T_j(x) = x_{\mathbf{d}} = \prod_{i=1}^n x_i^{\mathbf{d}(i)}$. Therefore

$$\Delta T_j(x) = \sum_{k\in[n]:\mathbf{d}(k)\ge 2} \mathbf{d}(k)(\mathbf{d}(k)-1)x_{\mathbf{d}-2\{k\}} =: \langle w, T(x)\rangle$$

for some $w \in \mathbb{R}^M$ with $\|w\|_2^2 = \sum_{k\in[n]:\mathbf{d}(k)\ge 2} \mathbf{d}(k)^2(\mathbf{d}(k)-1)^2 \le d^4$. By Corollary 104, we conclude that

$$\mathbb{E}_{x\sim p_\theta}(\Delta T_j(x))^2 = \mathbb{E}_{x\sim p_\theta}\langle w, T(x)\rangle^2 \le n^2 d^4 B^{4d} M^{4d+2} 2^{4d(d+2)+1}.$$

Summing over $j \in [M]$ gives the first claimed bound. For the second bound, observe that

$$\mathbb{E}_{x\sim p_\theta}\|(JT)(x)\|_{\mathsf{op}}^4 \le \mathbb{E}_{x\sim p_\theta}\|(JT)(x)\|_F^4 = \mathbb{E}_{x\sim p_\theta}\left(\sum_{j=1}^M\sum_{i=1}^n\left(\frac{\partial}{\partial x_i}T_j(x)\right)^2\right)^2.$$

For any $j \in [M]$ and $i \in [n]$, there is some degree function $\mathbf{d}$ with $|\mathbf{d}| \le d$ and $\frac{\partial}{\partial x_i}T_j(x) = |\mathbf{d}| \cdot x_{\mathbf{d}-\{i\}}$. Thus, by Holder's inequality and Lemma 178 (with $\ell = 4d$), we get

$$\begin{aligned}\mathbb{E}_{x\sim p_\theta}\left(\sum_{j=1}^M\sum_{i=1}^n\left(\frac{\partial}{\partial x_i}T_j(x)\right)^2\right)^2 &= \sum_{j,j'\in[M]}\sum_{i,i'\in[n]}\mathbb{E}_{x\sim p_\theta}\left(\frac{\partial}{\partial x_i}T_j(x)\right)^2\left(\frac{\partial}{\partial x_{i'}}T_{j'}(x)\right)^2 \\ &\le M^2 n^2 d^4 B^{4d} M^{4d} 2^{4d(d+2)+1}\end{aligned}$$

which proves the second bound. ☐

The following regularity conditions are sufficient for consistency and asymptotic normality of score matching, assuming that the restricted Poincaré constant is finite and $\lambda_{\min}(\mathcal{I}(\theta^*)) > 0$ (see Proposition 2 in [FL15] together with Lemma 1 in [KHR22]). We show that these conditions hold for our chosen exponential family.

Lemma 180 (Regularity conditions)**.** *For any $\theta \in \mathbb{R}^M$, the quantities $\mathbb{E}_{x\sim p_\theta}\|\nabla h(x)\|_2^4$, $\mathbb{E}_{x\sim p_\theta}\|\Delta T(x)\|_2^2$, and $\mathbb{E}_{x\sim p_\theta}\|(JT)(x)\|_{\mathsf{op}}^4$ are all finite. Moreover, $p_\theta(x) \to 0$ and $\|\nabla_x p_\theta(x)\|_2 \to 0$ as $\|x\|_2 \to \infty$.*

Proof. Both of the quantities $\|\nabla h(x)\|_2^4$ and $\|\Delta T(x)\|_2^2$ can be written as a polynomial in x. Finiteness of the expectation under p_θ follows from Holder's inequality and Lemma 178 (with parameter B set to $\|\theta\|_\infty$). Finiteness of $\mathbb{E}_{x\sim p_\theta}\|(JT)(x)\|_{\mathsf{op}}^4$ is shown in Lemma 179 (again, with $B := \|\theta\|_\infty$). The decay condition $p_\theta(x) \to 0$ holds because $\log p_\theta(x) + \log Z_\theta = -\sum_{i=1}^n x_i^{d+1} + \langle\theta, T(x)\rangle$. For $x \in \mathbb{R}^n$ with $L \le \|x\|_\infty \le 2L$, the RHS is at most $-L^{d+1} + M\|\theta\|_\infty(2L)^d$, which goes to $-\infty$ as $L \to \infty$. A similar bound shows that $\|\nabla_x p_\theta(x)\|_2 \to 0$. ☐

B.3 Conditioning

We analyze the condition number of underdamped Langevin dynamics with potential $f(x) = \frac{1}{2}\|x\|^2$ and stationary distribution $p(x,v) = e^{-f(x)-\frac{1}{2}\|v\|^2} = e^{-\frac{1}{2}(\|x\|^2+\|v\|^2)}$. Underdamped Langevin dynamics is given by the following SDE's,

$$\begin{aligned} dx_t &= -v_t && \text{(B.5)} \\ dv_t &= -\gamma v_t - \nabla f(x_t) + \sqrt{2}dB_t \\ &= -\gamma v_t - x_t + \sqrt{2}dB_t. && \text{(B.6)} \end{aligned}$$

Given the distribution p_0 at time 0, the distribution p_t at time t is the same as that given by,

$$\begin{bmatrix} \frac{dx}{dt} \\ \frac{dv}{dt} \end{bmatrix} = -\begin{bmatrix} 0 & -\mathrm{I}_d \\ \mathrm{I}_d & \gamma \mathrm{I}_d \end{bmatrix} \begin{bmatrix} \nabla_x \frac{\delta \operatorname{KL}(\mathbf{p}_t \| \mathbf{p}^*)}{\delta \mathbf{p}_t} \\ \nabla_v \frac{\delta \operatorname{KL}(\mathbf{p}_t \| \mathbf{p}^*)}{\delta \mathbf{p}_t} \end{bmatrix} \tag{B.7}$$

which simplifies to

$$d\begin{bmatrix} x_t \\ v_t \end{bmatrix} = \begin{bmatrix} O & \mathrm{I}_d \\ -\mathrm{I}_d & -\gamma \mathrm{I}_d \end{bmatrix} (\nabla \ln p_t - \nabla \ln p). \tag{B.8}$$

Our goal is to prove the following theorem.

Theorem 181. *Consider underdamped Langevin dynamics (B.5)–(B.6) with friction coefficient $\gamma < 2$ and starting distribution p_0 that is C^2. Let T_t denote the transport map from time 0 to time t induced by (B.8). Suppose that the initial distribution $p_0(x,v)$ is such that*

$$\mathrm{I}_{2d} \preceq -\nabla^2 \ln p_0(x,v) \preceq \kappa \mathrm{I}_{2d}.$$

Then for any x_0, v_0 and unit vector w, the directional derivative of T_t at x_0, v_0 in direction w satisfies

$$\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right)^{-2/\gamma} \le \|D_w T_t(x_0)\| \le \left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right)^{2/\gamma}$$

Thus the condition number of T_t is bounded by $\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right)^{4/\gamma}$.

We remark that the exponent is likely loose by a factor of 2, and that taking $\gamma \to 2$ gives the best exponent; however, the case $\gamma = 2$ would require a separate calculation as the matrix appearing in the exponential is not diagonalizable. Note $\gamma = 2$ is the transition between when the dynamics exhibit underdamped and overdamped behavior.

To prove the theorem, we first relate the Jacobian with the Hessian of the log-pdf. By Lemma 188, the Jacobian $D_t = DT_t(x_0)$ satisfies

$$\frac{d}{dt}D_t = \begin{bmatrix} O & \mathrm{I}_d \\ -\mathrm{I}_d & -\gamma \mathrm{I}_d \end{bmatrix} \nabla^2(\ln p_t - \ln p) D_t. \tag{B.9}$$

We will show that $\nabla^2(\ln p_t - \ln p)$ decays exponentially (Lemma 184). First, we need the following bound for convolutions.

B.3.1 Bounding the Hessian of the logarithm of a convolution

Lemma 182. *Suppose that p is a probability density function on $\mathbb{R}^d$ such that $\Sigma_1^{-1} \preceq -\nabla^2 \ln p \preceq \Sigma_2^{-1}$. Let q be the distribution of $N(0,\Sigma)$ (where Σ is not necessarily full-rank). Then*

$$(\Sigma_1+\Sigma)^{-1} \preceq -\nabla^2 \ln(p*q) \preceq (\Sigma_2+\Sigma)^{-1}.$$

Proof. The lower bound is a bound on the strong log-concavity parameter; see Theorem 3.7b in [SW14].

For the upper bound, we first prove the lemma in the case that Σ is full rank. We have $(p*q)(x) = \int_{\mathbb{R}^d} p(u)q(x-u)\,dt$, so

$$\begin{aligned}
\nabla^2[\ln((p*q)(x))] &= \frac{\int_{\mathbb{R}^d} p(u)\nabla^2 q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du} - \left(\frac{\int_{\mathbb{R}^d} p(u)\nabla q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du}\right)\left(\frac{\int_{\mathbb{R}^d} p(u)\nabla q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du}\right)^\top \\
&= \left(\frac{\int_{\mathbb{R}^d} \Sigma^{-1}(x-u)p(u)q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du}\right)\left(\frac{\int_{\mathbb{R}^d} (\Sigma^{-1}(x-u))^\top p(u)q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du}\right) \\
&\quad - \frac{\int_{\mathbb{R}^d} (\Sigma^{-1}(x-u)(x-u)^\top\Sigma^{-1} - \Sigma^{-1})p(u)q(x-u)\,du}{\int_{\mathbb{R}^d} p(u)q(x-u)\,du}
\end{aligned}$$

Let μ_x denote the distribution with density function $\rho(u) \propto p(u)q(x-u)$. Then

$$\begin{aligned}
-\nabla^2[\ln((p*q)(x))] &= [\mathbb{E}_{\mu_x}\Sigma^{-1}(u-x)][\mathbb{E}_{\mu_x}(\Sigma^{-1}(u-x))^\top] - [\mathbb{E}_{\mu_x}\Sigma^{-1}(u-x)(u-x)^\top\Sigma^{-1}] + \Sigma^{-1} \\
&= -\mathbb{E}_{\mu_x}[\Sigma^{-1}(u-\mathbb{E}u)(u-\mathbb{E}u)^\top\Sigma^{-1}] + \Sigma^{-1}.
\end{aligned}$$

It suffices to show for any unit vector v, that

$$-v^\top\nabla^2[\ln((p*q)(x))]v = -\mathbb{E}_{\mu_x}[\langle\Sigma^{-1}v,(u-\mathbb{E}u)\rangle^2] + v^\top\Sigma^{-1}v \le v^\top(\Sigma_2+\Sigma)^{-1}v$$

Note that μ_x satisfies

$$-\nabla^2\ln\mu_x \preceq \Sigma_2^{-1} + \Sigma^{-1},$$

so μ_x can be written as the density of a Gaussian with variance $(\Sigma_2^{-1}+\Sigma^{-1})^{-1}$ multiplied by a log-convex function. By the Brascamp-Lieb moment inequality (Theorem 5.1 in [BL02])[1],

$$\mathbb{E}_{\mu_x}[\langle\Sigma^{-1}v,(u-\mathbb{E}u)\rangle^2] \ge \mathbb{E}_{u\sim N(0,(\Sigma_2^{-1}+\Sigma^{-1})^{-1})}[\langle\Sigma^{-1}v,u\rangle^2] = v^\top\Sigma^{-1}(\Sigma_2^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}v.$$

Hence

$$-v^\top\nabla^2[\ln((p*q)(x))]v \le v^\top\left[-\Sigma^{-1}(\Sigma_2^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}+\Sigma^{-1}\right]v$$

The conclusion then follows from

$$\begin{aligned}
-\Sigma^{-1}(\Sigma_2^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}+\Sigma^{-1} &= -(\Sigma\Sigma_2^{-1}\Sigma+\Sigma)^{-1}+\Sigma^{-1} \\
&= (\Sigma\Sigma_2^{-1}\Sigma+\Sigma)^{-1}(\cancel{-I_d}+\Sigma\Sigma_2^{-1}+\cancel{I_d}) \\
&= (\Sigma+\Sigma_2)^{-1}.
\end{aligned}$$

[1]Note that the sign is flipped in the theorem statement in the log-convex case.

Now for the general case, take the limit as $\Sigma' \to \Sigma$ where Σ' is full-rank. More precisely, let $\Sigma_t = \Sigma + tP$, where P is projection onto $\mathrm{Im}(\Sigma)^\perp$, and let q_t be the density function for $N(0, \Sigma_t)$. Then we have

$$\nabla^2[\ln((p * q_t)(x))] = \frac{\int_{\mathbb{R}^d} \nabla^2 p(x-u) q_t(u)\, du}{\int_{\mathbb{R}^d} p(x-u) q_t(u)\, du} - \left(\frac{\int_{\mathbb{R}^d} \nabla p(x-u) q_t(u)\, du}{\int_{\mathbb{R}^d} p(x-u) q_t(u)\, du}\right)\left(\frac{\int_{\mathbb{R}^d} \nabla p(x-u) q_t(u)\, du}{\int_{\mathbb{R}^d} p(x-u) q_t(u)\, du}\right)^\top$$

Examining the first term, we have

$$\begin{aligned}\int_{\mathbb{R}^d} \nabla^2 p(x-u) q_t(u)\, du &= \int_{\mathrm{Im}(\Sigma)} \int_{\mathrm{Im}(P)} \nabla^2 p(x-u-v) q_t(u+v)\, dv\, du \\ &\to \int_{\mathrm{Im}(\Sigma)} \nabla^2 p(x-u) q_t(u)\, du \text{ as } t \to 0^+\end{aligned}$$

by the dominated convergence theorem. Similarly, the other integrals converge to their counterparts with $q(u)$. Therefore, $\nabla^2[\ln((p * q_t)(x))] \to \nabla^2[\ln((p * q)(x))]$ as $t \to 0^+$. Apply the lemma to the full-rank case; the RHS bound converges to the desired bound: $(\Sigma_2 + \Sigma_t)^{-1} \to (\Sigma_2 + \Sigma)^{-1}$.

□

B.3.2 Bounding the variance proxy for underdamped Langevin

As it is useful to work with the matrices Σ_1 and Σ_2, we make the following definition.

Definition 183. Let p be a probability density on $\mathbb{R}^d$. For a positive definite matrix Σ_1, if $\Sigma_1^{-1} \preceq -\nabla^2 \ln p$, we say that Σ_1 is an **upper variance proxy** for p. For a positive definite matrix Σ_2, if $-\nabla^2 \ln p \preceq \Sigma_2^{-1}$, we say Σ_2 is a **lower variance proxy** for p.

Lemma 184. *Consider underdamped Langevin dynamics (B.5)–(B.6) with with starting distribution $p_0(x, v)$ that is C^2. Suppose p_0 has lower (upper) variance proxy Σ_0. Then p_t has lower (upper) variance proxy*

$$\Sigma_t = \exp\left[\left(\begin{bmatrix} & 1 \\ -1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] (\Sigma_0 - \mathrm{I}_{2d}) \exp\left[\left(\begin{bmatrix} & -1 \\ 1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] + \mathrm{I}_{2d}.$$

Proof. We first consider discretized Lanegevin, given by

$$\begin{aligned}\widetilde{x}_{t+\eta} &= \widetilde{x}_t + \eta \widetilde{v}_t \\ \widetilde{v}_{t+\eta} &= (1 - \eta\gamma)\widetilde{v}_t - \eta \widetilde{x}_t + \xi_t, \quad \xi_t \sim N(0, 2\eta \mathrm{I}_d)\end{aligned}$$

or in matrix form,

$$\begin{bmatrix} \widetilde{x}_{t+\eta} \\ \widetilde{v}_{t+\eta} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_d & \eta \mathrm{I}_d \\ -\eta \mathrm{I}_d & (1-\eta\gamma)\mathrm{I}_d \end{bmatrix} \begin{bmatrix} \widetilde{x}_t \\ \widetilde{v}_t \end{bmatrix} + \xi_t, \quad \xi_t \sim N\left(0, \begin{bmatrix} O & O \\ O & 2\eta \mathrm{I}_d \end{bmatrix}\right).$$

Fix t. Let $\widetilde{p}_t^{(\eta)}$ be the distribution at time t for discretized Langevin with step size η (dividing t). By standard arguments, $\widetilde{p}_t^{(\eta)} \to p_t$ as $\eta \to 0$, in the C^2 topology on any compact set. In particular, for any x, v, $\nabla^2 \ln \widetilde{p}_t^{(\eta)}(x, v) \to \nabla^2 \ln p_t(x, v)$. Hence it suffices to bound $\nabla^2 \ln p_t(x, v)$.

We write the proof for the upper variance proxy; the proof for the lower variance proxy differs only in the direction of the inequality. Suppose $-\ln \widetilde{p}_t(x,v) \succeq \widetilde{\Sigma}_t^{-1}$. Consider breaking the update into two steps,

$$\begin{aligned}
\begin{bmatrix} \widetilde{x}'_{t+\eta} \\ \widetilde{v}'_{t+\eta} \end{bmatrix} &= \begin{bmatrix} \mathrm{I}_d & \eta \mathrm{I}_d \\ -\eta \mathrm{I}_d & (1-\eta\gamma)\mathrm{I}_d \end{bmatrix} \begin{bmatrix} \widetilde{x}_t \\ \widetilde{v}_t \end{bmatrix} \\
\begin{bmatrix} \widetilde{x}_{t+\eta} \\ \widetilde{v}_{t+\eta} \end{bmatrix} &= \begin{bmatrix} \widetilde{x}'_{t+\eta} \\ \widetilde{v}'_{t+\eta} \end{bmatrix} + \xi_t, \quad \xi_t \sim N\left(0, \begin{bmatrix} O & O \\ O & 2\eta \mathrm{I}_d \end{bmatrix}\right).
\end{aligned}$$

Let $\widetilde{p}'_{t+\eta}(x,v)$ denote the distribution of $\begin{bmatrix} \widetilde{x}'_{t+\eta} \\ \widetilde{v}'_{t+\eta} \end{bmatrix}$. Then

$$\widetilde{p}'_{t+\eta}(x,v) = \widetilde{p}_t\left(\begin{bmatrix} \mathrm{I}_d & \eta \mathrm{I}_d \\ -\eta \mathrm{I}_d & (1-\eta\gamma)\mathrm{I}_d \end{bmatrix}^{-1} \begin{bmatrix} x \\ v \end{bmatrix} \right)$$

so

$$\widetilde{\Sigma}'_{t+\eta} := \begin{bmatrix} \mathrm{I}_d & \eta \mathrm{I}_d \\ -\eta \mathrm{I}_d & (1-\eta\gamma)\mathrm{I}_d \end{bmatrix} \widetilde{\Sigma}_t \begin{bmatrix} \mathrm{I}_d & -\eta \mathrm{I}_d \\ \eta \mathrm{I}_d & (1-\eta\gamma)\mathrm{I}_d \end{bmatrix}$$

is an upper variance proxy for $\widetilde{p}'_{t+\eta}$ and by Lemma 182,

$$\widetilde{\Sigma}_{t+\eta} := \widetilde{\Sigma}'_{t+\eta} + \begin{bmatrix} O & O \\ O & 2\eta \mathrm{I}_d \end{bmatrix}$$

is an upper variance proxy for $\widetilde{p}_{t+\eta}$. Note that

$$\widetilde{\Sigma}_{t+\eta} := \widetilde{\Sigma}_t + \left[\begin{bmatrix} & 1 \\ -1 & -\gamma\eta \end{bmatrix} \otimes \mathrm{I}_d\right] \widetilde{S}_t + \widetilde{S}_t \left[\begin{bmatrix} & -1 \\ 1 & -\gamma\eta \end{bmatrix} \otimes \mathrm{I}_d\right] + \begin{bmatrix} 0 & 0 \\ 0 & 2\gamma\eta \end{bmatrix} + O(\eta^2).$$

By the standard analysis of Euler's method, as $\eta \to 0$, the distribution, $\widetilde{\Sigma}_t$ approaches Σ_t defined by

$$\frac{d}{dt}\Sigma_t = \left[\begin{bmatrix} & 1 \\ -1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right] \Sigma_t + \Sigma_t \left[\begin{bmatrix} & -1 \\ 1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right] + \begin{bmatrix} 0 & 0 \\ 0 & 2\gamma \end{bmatrix}.$$

This Σ_t is an upper variance proxy for p_t. The solution to this equation is

$$\Sigma_t = \exp\left[\left(\begin{bmatrix} & 1 \\ -1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] (\Sigma_0 - \mathrm{I}_{2d}) \exp\left[\left(\begin{bmatrix} & -1 \\ 1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] + \mathrm{I}_{2d},$$

as desired. $\square$

B.3.3 Proof that underdamped Langevin is well-conditioned

We are now ready to prove the main theorem.

Proof of Theorem 181. Let $H_t = \nabla^2(-\ln p_t + \ln p)$ and $C = \begin{bmatrix} O & \mathrm{I}_d \\ -\mathrm{I}_d & -\gamma \mathrm{I}_d \end{bmatrix}$. By (B.9) and the chain rule,

$$\frac{d}{dt} D_t D_t^\top = -(CH_t D_t D_t^\top + D_t D_t^\top H_t C^\top). \tag{B.10}$$

Fix w and consider $y_t = D_t w = D_w T_t(x_0)$. Multiplying the above by W on both sides gives[2]

$$\left| \frac{d}{dt} \|y_t\|^2 \right| \le 2 \|CH_t\| \|y_t\|^2$$

so by Grönwall's inequality (Lemma 192),

$$\exp\left[-2\int_0^t \|CH_s\|\, ds\right] \le \|y_t\|^2 \le \exp\left[2\int_0^t \|CH_s\|\, ds\right]. \tag{B.11}$$

By Lemma 184,

$$\mathrm{I}_{2d} \preceq -\nabla^2 \ln p_t \preceq (\kappa - 1)\exp\left[\left(\begin{bmatrix} & 1 \\ -1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] \exp\left[\left(\begin{bmatrix} & -1 \\ 1 & -\gamma \end{bmatrix} \otimes \mathrm{I}_d\right) t\right] + \mathrm{I}_{2d}.$$

The eigenvalues of $A := \begin{bmatrix} & -1 \\ 1 & -\gamma \end{bmatrix}$ are $\frac{-\gamma \pm \sqrt{\gamma^2-4}}{2}$, which have absolute value 1. The absolute value of the inner product of the eigenvectors of A is $\gamma/2$, so the condition number squared of the two exponential factors is bounded by $\frac{1+\frac{\gamma}{2}}{1-\frac{\gamma}{2}} = \frac{2+\gamma}{2-\gamma}$. In full detail, we calculate

$$\exp\left(\begin{bmatrix} & -1 \\ 1 & \gamma \end{bmatrix} t\right) = \underbrace{\begin{bmatrix} 1 & 1 \\ \frac{\gamma - \sqrt{\gamma^2-4}}{2} & \frac{\gamma + \sqrt{\gamma^2-4}}{2} \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} \exp\left(\frac{-\gamma+\sqrt{\gamma^2-4}}{2} t\right) & \\ & \exp\left(\frac{-\gamma-\sqrt{\gamma^2-4}}{2} t\right) \end{bmatrix}}_{D}$$

$$\cdot \underbrace{\frac{1}{\sqrt{\gamma^2-4}} \begin{bmatrix} \frac{\gamma+\sqrt{\gamma^2-4}}{2} & -1 \\ \frac{-\gamma+\sqrt{\gamma^2-4}}{2} & 1 \end{bmatrix}}_{S^{-1}}$$

$$\|S^\dagger S\| = \left\| \begin{bmatrix} 2 & \frac{\gamma^2+\gamma\sqrt{\gamma^2-4}}{2} \\ \frac{\gamma^2-\gamma\sqrt{\gamma^2-4}}{2} & 2 \end{bmatrix} \right\| = 2 + \gamma$$

$$\left\| \exp\left(\begin{bmatrix} & -1 \\ 1 & \gamma \end{bmatrix} t\right) \right\| \le \frac{2+\gamma}{\sqrt{4-\gamma^2}} \exp\left(\frac{-\gamma t}{2}\right) = \sqrt{\frac{2+\gamma}{2-\gamma}} \exp\left(\frac{-\gamma t}{2}\right).$$

[2] The condition number bound in Theorem 181 is the square of what one might expect because we are only able to get obtain a bound on the absolute value here. If this is always increasing or decreasing, then we would save a factor of 2 in the exponent.

Hence $H_t = -\nabla^2 \ln p_t + \mathrm{I}_{2d}$ satisfies

$$
\begin{aligned}
\|CH_s\| &\le 1 - \frac{1}{1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)e^{-\gamma t/2}} \\
\int_0^\infty \|CH_s\|\, ds &\le \int_0^\infty \frac{\frac{2+\gamma}{2-\gamma}(\kappa-1)e^{-\gamma t/2}}{1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)e^{-\gamma t/2}}\, ds \\
&\le \left[\frac{2}{\gamma}\ln\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)e^{-\gamma t/2}\right)\right]_\infty^0 \le \frac{2}{\gamma}\ln\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right).
\end{aligned}
$$

Hence by (B.11),

$$
\left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right)^{-2/\gamma} \le \|y_t\| \le \left(1 + \frac{2+\gamma}{2-\gamma}(\kappa-1)\right)^{2/\gamma},
$$

giving the theorem. To obtain the bound on condition number, note that the condition number of $DT_t(x_0)$ is $\frac{\max_{\|w\|=1}\|D_w T_t(x_0)\|}{\min_{\|w\|=1}\|D_w T_t(x_0)\|}$. □

B.4 Proof of Lemma 127

For the sake of convenience, we restate Lemma 127 again.

Lemma. *Let $\mathcal{C} \in \mathbb{R}^{2d}$ be a compact set. For any function $H(x, v, t) : \mathbb{R}^{2d} \times \mathbb{R}_{\ge 0} \to \mathbb{R}$ which is polynomial in (x, v), there exist polynomial functions J, F, G, s.t. the time-$(t_0 + \tau, t_0)$ flow map of the system*

$$
\begin{cases}
\frac{dx}{dt} = \frac{\partial}{\partial v} H(x, v, t) \\
\frac{dv}{dt} = -\frac{\partial}{\partial x} H(x, v, t) - \gamma \frac{\partial}{\partial v} H(x, v, t)
\end{cases}
\tag{B.12}
$$

is uniformly $O(\tau^2)$-close over $\mathcal{C}$ in C^1 topology to the time-2π map of the system

$$
\begin{cases}
\frac{dx}{dt} = v - \tau F(v, t) \odot x \\
\frac{dv_j}{dt} = -\Omega_j^2 x_j - \tau J_j(x, t) - \tau v_j G_j(x, t)
\end{cases}
\tag{B.13}
$$

for some integers $\{\Omega_j\}_{j=1}^d$. Here, $\odot$ denotes component-wise product, and the constants inside the $O(\cdot)$ depend on $\mathcal{C}$ and the coefficients of H.

Proof. First, note that the time-$(t_0 + \tau, t_0)$ flow map of (B.12) is equal to the time-$(t_0, t_0 + \tau)$ flow map of the system:

$$
\begin{cases}
\frac{dx}{dt} = -\frac{\partial}{\partial v} H(x, v, t_0 + \tau - t) \\
\frac{dv}{dt} = \frac{\partial}{\partial x} H(x, v, t_0 + \tau - t) + \gamma \frac{\partial}{\partial v} H(x, v, t_0 + \tau - t)
\end{cases}
\tag{B.14}
$$

Proceeding ahead, we broadly follow the proof strategy in [Tur02]. For notational convenience, let's denote the initial vector by $x(0), v(0)$ (each coordinate is specified separately). Let

$$x_j^0(t) = x_j(0)\cos\Omega_j t + \frac{1}{\Omega_j} v_j(0)\sin\Omega_j t \tag{B.15}$$

$$v_j^0(t) = -\Omega_j x_j(0)\sin\Omega_j t + v_j(0)\cos\Omega_j t. \tag{B.16}$$

Using perturbative ODE techniques (see section B.5.5), the solution to (B.13) satisfies

$$\begin{cases} x(t) = x^0(t) - \tau\int_0^t \big(\frac{1}{\Omega}\odot J(x^0(s),s)\odot\sin\Omega(t-s) + F(v^0(s),s)\odot\cos\Omega(t-s)\odot x^0(s) \\ \qquad\qquad + \frac{1}{\Omega}\odot G(x^0(s),s)\odot\sin\Omega(t-s)\odot v^0(s)\big)\,ds + O(\tau^2) \\ v(t) = v^0(t) - \tau\int_0^t \big(J(x^0(s),s)\odot\cos\Omega(t-s) - \Omega\odot F(v^0(s),s)\odot\sin\Omega(t-s)\odot x^0(s) \\ \qquad\qquad + G(x^0(s),s)\odot\cos\Omega(t-s)\odot v^0(s)\big)\,ds + O(\tau^2) \end{cases} \tag{B.17}$$

Substituting $t = 2\pi$, the time-2π map of (B.13) is given by

$$\begin{cases} x(2\pi) = x^0(2\pi) - \tau\int_0^{2\pi} \big(-\frac{1}{\Omega}\odot J(x^0(s),s)\odot\sin\Omega s + F(v^0(s),s)\odot\cos\Omega s\odot x^0(s) \\ \qquad\qquad - \frac{1}{\Omega}\odot G(x^0(s),s)\odot\sin\Omega s\odot v^0(s)\big)\,ds + O(\tau^2) \\ v(2\pi) = v^0(2\pi) - \tau\int_0^{2\pi} \big(J(x^0(s),s)\odot\cos\Omega s + \Omega\odot F(v^0(s),s)\odot\sin\Omega s\odot x^0(s) \\ \qquad\qquad + G(x^0(s),s)\odot\cos\Omega s\odot v^0(s)\big)\,ds + O(\tau^2) \end{cases} \tag{B.18}$$

Note that this holds if Ω is integral, and we will choose it to be so.

On the other hand, using Taylor's theorem, the solution to (B.12) satisfies:

$$\begin{cases} x(\tau) = x(0) - \tau\frac{\partial}{\partial v}H(x(0),v(0),t_0+\tau) + O(\tau^2) \\ v(\tau) = v(0) + \tau\frac{\partial}{\partial x}H(x(0),v(0),t_0+\tau) + \tau\gamma\frac{\partial}{\partial v}H(x(0),v(0),t_0+\tau) + O(\tau^2) \end{cases} \tag{B.19}$$

We will now show that for any two polynomials r_1, r_2 of total degree at most M we can choose functions J, F, G, s.t.:

$$\begin{cases} \int_0^{2\pi} \big(-\frac{1}{\Omega}\odot J(x^0(s),s)\odot\sin\Omega s + F(v^0(s),s)\odot\cos\Omega s\odot x^0(s) \\ \qquad\qquad - \frac{1}{\Omega}\odot G(x^0(s),s)\odot\sin\Omega s\odot v^0(s)\big)\,ds = r_1(x(0),y(0)) \\ \int_0^{2\pi} (J(x^0(s),s)\odot\cos\Omega s + \Omega\odot F(v^0(s),s)\odot\sin\Omega s\odot x^0(s) \\ \qquad\qquad + G(x^0(s),s)\odot\cos\Omega s\odot v^0(s))\,ds = r_2(x(0),y(0)) \end{cases} \tag{B.20}$$

We will choose J, F, G of the form:

$$\begin{cases} \forall j\in[d] : J_j(z,t) = \sum_{\mathbf{i}:|\mathbf{i}|\le M} v_{j,\mathbf{i}}^J(t) z^{\mathbf{i}} \\ \forall j\in[d] : F_j(z,t) = \sum_{\mathbf{i}:|\mathbf{i}|\le M-1} v_{j,\mathbf{i}}^F(t) z^{\mathbf{i}} \\ \forall j\in[d] : G_j(z,t) = \sum_{\mathbf{i}:|\mathbf{i}|\le M-1} v_{j,\mathbf{i}}^G(t) z^{\mathbf{i}} \end{cases} \tag{B.21}$$

where $\mathbf{i} = (i_1,\ldots,i_d)$ denotes multi-index, and $|\mathbf{i}| = \sum_{k=1}^d i_k$ and $z^{\mathbf{i}} = \prod_{k=1}^d z_k^{i_k}$. Let

$$r_{1,j}(x(0),v(0)) = \sum_{\mathbf{k}:|\mathbf{k}|\le M}\sum_{\mathbf{p}+\mathbf{q}=\mathbf{k}} h_{j,\mathbf{p},\mathbf{q}}^1 x(0)^{\mathbf{p}} v(0)^{\mathbf{q}} \tag{B.22}$$

$$r_{2,j}(x(0),v(0)) = \sum_{\mathbf{k}:|\mathbf{k}|\le M}\sum_{\mathbf{p}+\mathbf{q}=\mathbf{k}} h_{j,\mathbf{p},\mathbf{q}}^2 x(0)^{\mathbf{p}} v(0)^{\mathbf{q}} \tag{B.23}$$

The equation (B.20) gives us that for all j,

$$\begin{cases} \int_0^{2\pi} \Big(-\frac{1}{\Omega_j} J_j(x^0(s), s) \sin(\Omega_j s) + F_j(v^0(s), s) \cos(\Omega_j s) x_j^0(s) \\ \qquad\qquad -\frac{1}{\Omega_j} G_j(x^0(s), s) \sin(\Omega_j s) v_j^0(s) \Big)\, ds = r_{1,j}(x(0), y(0)) \\ \int_0^{2\pi} \big(J_j(x^0(s), s) \cos(\Omega_j s) + \Omega_j F_j(v^0(s), s) \sin(\Omega_j s) x_j^0(s) \\ \qquad\qquad + G_j(x^0(s), s) \cos(\Omega_j s) v_j^0(s) \big)\, ds = r_{2,j}(x(0), y(0)) \end{cases} \tag{B.24}$$

Let $\binom{\mathbf{k}}{\mathbf{p}} = \prod_{k=1}^d \binom{k_i}{p_i}$. Let $\mathbf{k_j^t}$ be the multi-index $(k_1, \ldots, k_j + t, \ldots, k_d)$. We substitute (B.15)–(B.16), (B.21), and (B.22)–(B.23) into (B.24) and match the coefficients of $x(0)^{\mathbf{p}} v(0)^{\mathbf{q}}$.

If $k_j = 0$, then

$$\begin{aligned} h^1_{j,\mathbf{p},\mathbf{q}} &= \int_0^{2\pi} -\frac{1}{\Omega_j} v^J_{j,\mathbf{k}} \cos(\Omega s)^{\mathbf{p}} \sin(\Omega s)^{\mathbf{q_j^1}} \binom{\mathbf{k}}{\mathbf{p}} ds \\ h^2_{j,\mathbf{p},\mathbf{q}} &= \int_0^{2\pi} v^J_{j,\mathbf{k}} \cos(\Omega s)^{\mathbf{p_j^1}} \sin(\Omega s)^{\mathbf{q}} \binom{\mathbf{k}}{\mathbf{p}} ds \end{aligned}$$

where $v^J_{j,\mathbf{k}} = a \cos(\Omega s)^{\mathbf{p}} \sin(\Omega s)^{\mathbf{q_j^1}} + b \cos(\Omega s)^{\mathbf{p_j^1}} \sin(\Omega s)^{\mathbf{q}}$. Since the function $\delta(s) = \cos(\Omega s)^{\mathbf{p}+\mathbf{p_j^1}} \sin(\Omega s)^{\mathbf{q}+\mathbf{q_j^1}}$ satisfies $\delta(\pi - s) = -\delta(\pi + s)$, this function integrates to zero, and hence the system above reduces to

$$\begin{aligned} h^1_{j,\mathbf{p},\mathbf{q}} &= a \frac{1}{\Omega_j} C \binom{\mathbf{k}}{\mathbf{p}} \\ h^2_{j,\mathbf{p},\mathbf{q}} &= b C \binom{\mathbf{k}}{\mathbf{p}} \end{aligned}$$

for some non-zero constant

$$C = \int_0^{2\pi} \cos(\Omega s)^{2\mathbf{p}} \sin(\Omega s)^{2\mathbf{q_j^1}} ds = \int_0^{2\pi} \cos(\Omega s)^{2\mathbf{p_j^1}} \sin(\Omega s)^{2\mathbf{q}} ds$$

Note that the integral is non-zero since the function inside is positive as all the powers are even.

If $k_j > 0$, then substituting the forms of $x^0(s), v^0(s)$ from (B.15) in the LHS of (B.24), and expanding using the binomial theorem, we get that

$$\begin{aligned}
h^1_{j,\mathbf{p},\mathbf{q}} = &\frac{1}{\Omega^{\mathbf{q_j^1}}}\int_0^{2\pi} -v^J_{j,\mathbf{k}}\cos(\Omega s)^{\mathbf{p}}\sin(\Omega s)^{\mathbf{q_j^1}}\binom{\mathbf{k}}{\mathbf{p}}ds\\
&+\Omega^{\mathbf{p_j^{-1}}}\int_0^{2\pi} v^F_{j,\mathbf{k_j^{-1}}}(-\mathbf{1})^{\mathbf{p_j^{-1}}}\sin(\Omega s)^{\mathbf{p_j^{-1}}}\cos(\Omega s)^{\mathbf{q_j^2}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}}ds\\
&+\Omega^{\mathbf{p_j^{-1}}}\int_0^{2\pi} v^F_{j,\mathbf{k_j^{-1}}}(-\mathbf{1})^{\mathbf{p}}\sin(\Omega s)^{\mathbf{p_j^1}}\cos(\Omega s)^{\mathbf{q}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p}}ds\\
&+\frac{1}{\Omega^{\mathbf{q}}}\int_0^{2\pi}\left(v^G_{j,\mathbf{k_j^{-1}}}\cos(\Omega s)^{\mathbf{p_j^{-1}}}\sin(\Omega s)^{\mathbf{q_j^2}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} - v^G_{j,\mathbf{k_j^{-1}}}\cos(\Omega s)^{\mathbf{p_j^1}}\sin(\Omega s)^{\mathbf{q}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p}}\right)ds\\
h^2_{j,\mathbf{p},\mathbf{q}} = &\frac{1}{\Omega^{\mathbf{q}}}\int_0^{2\pi} v^J_{j,\mathbf{k}}\cos(\Omega s)^{\mathbf{p_j^1}}\sin(\Omega s)^{\mathbf{q}}\binom{\mathbf{k}}{\mathbf{p}}ds\\
&+\Omega^{\mathbf{p}}\int_0^{2\pi} v^F_{j,\mathbf{k_j^{-1}}}(-\mathbf{1})^{\mathbf{p_j^{-1}}}\sin(\Omega s)^{\mathbf{p}}\cos(\Omega s)^{\mathbf{q_j^1}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}}ds\\
&+\Omega^{\mathbf{p}}\int_0^{2\pi} v^F_{j,\mathbf{k_j^{-1}}}(-\mathbf{1})^{\mathbf{p}}\sin(\Omega s)^{\mathbf{p_j^2}}\cos(\Omega s)^{\mathbf{q_j^{-1}}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p}}ds\\
&+\frac{1}{\Omega^{\mathbf{q_j^{-1}}}}\int_0^{2\pi}\left(-v^G_{j,\mathbf{k_j^{-1}}}\cos(\Omega s)^{\mathbf{p}}\sin(\Omega s)^{\mathbf{q_j^1}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} + v^G_{j,\mathbf{k_j^{-1}}}\cos(\Omega s)^{\mathbf{p_j^2}}\sin(\Omega s)^{\mathbf{q_j^{-1}}}\binom{\mathbf{k_j^{-1}}}{\mathbf{p}}\right)ds
\end{aligned}$$

Let $g_{\mathbf{k},\mathbf{p}}(s) = \cos(\Omega s)^{\mathbf{p}}\sin(\Omega s)^{\mathbf{k}-\mathbf{p}}$ for all $\mathbf{p} \le \mathbf{k}$. Crucially, let us assume that $v^J_{j,\mathbf{k}}, v^F_{j,\mathbf{k}}, v^G_{j,\mathbf{k}}$ are all of the form

$$\begin{cases}
v^F_{j,\mathbf{k}} = \sum_{\mathbf{r}\le\mathbf{k_j^2}} \alpha_{\mathbf{k_j^2},\mathbf{r}} g_{\mathbf{k_j^2},\mathbf{r}}(s)\\
v^G_{j,\mathbf{k}} = \sum_{\mathbf{r}\le\mathbf{k_j^2}} \beta_{\mathbf{k_j^2},\mathbf{r}} g_{\mathbf{k_j^2},\mathbf{r}}(s)\\
v^J_{j,\mathbf{k}} = \sum_{\mathbf{r}\le\mathbf{k_j^1}} \gamma_{\mathbf{k_j^1},\mathbf{r}} g_{\mathbf{k_j^1},\mathbf{r}}(s)
\end{cases} \tag{B.25}$$

Substituting,

$$h^1_{j,\mathbf{p},\mathbf{q}} = \frac{1}{\Omega^{\mathbf{q}^1_j}} \int_0^{2\pi} -\sum_{\mathbf{r}\le\mathbf{k}^1_j} \gamma_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}}(s) \binom{\mathbf{k}}{\mathbf{p}} ds$$

$$+ \Omega^{\mathbf{p}^{-1}_j} \int_0^{2\pi} \left((-\mathbf{1})^{\mathbf{p}^{-1}_j} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{q}^2_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + (-\mathbf{1})^{\mathbf{p}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{q}}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right) ds$$

$$+ \frac{1}{\Omega^{\mathbf{q}}} \int_0^{2\pi} \left(\sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}^{-1}_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} - \sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}^1_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right) ds$$

$$h^2_{j,\mathbf{p},\mathbf{q}} = \frac{1}{\Omega^{\mathbf{q}}} \int_0^{2\pi} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \gamma_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}^1_j}(s) \binom{\mathbf{k}}{\mathbf{p}} ds$$

$$+ \Omega^{\mathbf{p}} \int_0^{2\pi} \left((-\mathbf{1})^{\mathbf{p}^{-1}_j} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{q}^1_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + (-\mathbf{1})^{\mathbf{p}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{q}^{-1}_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right) ds$$

$$+ \frac{1}{\Omega^{\mathbf{q}^{-1}_j}} \int_0^{2\pi} \left(-\sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + \sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} g_{\mathbf{k}^1_j,\mathbf{r}}(s) g_{\mathbf{k}^1_j,\mathbf{p}^2_j}(s) \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right) ds$$

Now, let $\langle f, g\rangle = \int_0^{2\pi} f(s)g(s)ds$ denote the ℓ_2 inner product. Then, we can rewrite the above system as

$$h^1_{j,\mathbf{p},\mathbf{q}} = -\frac{1}{\Omega^{\mathbf{q}^1_j}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \gamma_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}}(s) \right\rangle \binom{\mathbf{k}}{\mathbf{p}}$$

$$+ \Omega^{\mathbf{p}^{-1}_j} \left[(-\mathbf{1})^{\mathbf{p}^{-1}_j} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{q}^2_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + (-\mathbf{1})^{\mathbf{p}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{q}}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right]$$

$$+ \frac{1}{\Omega^{\mathbf{q}}} \left[\sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}^{-1}_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} - \sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}^1_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right]$$

$$h^2_{j,\mathbf{p},\mathbf{q}} = \frac{1}{\Omega^{\mathbf{q}}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \gamma_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}^1_j}(s) \right\rangle \binom{\mathbf{k}}{\mathbf{p}}$$

$$+ \Omega^{\mathbf{p}} \left[(-\mathbf{1})^{\mathbf{p}^{-1}_j} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{q}^1_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + (-\mathbf{1})^{\mathbf{p}} \sum_{\mathbf{r}\le\mathbf{k}^1_j} \alpha_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{q}^{-1}_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right]$$

$$+ \frac{1}{\Omega^{\mathbf{q}^{-1}_j}} \left[-\sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}^{-1}_j} + \sum_{\mathbf{r}\le\mathbf{k}^1_j} \beta_{\mathbf{k}^1_j,\mathbf{r}} \left\langle g_{\mathbf{k}^1_j,\mathbf{r}}(s), g_{\mathbf{k}^1_j,\mathbf{p}^2_j}(s) \right\rangle \binom{\mathbf{k}^{-1}_j}{\mathbf{p}} \right]$$

Now, we will add a few redundant constraints in the system. These are added to ensure that the system has a nice matrix form; they are all of the type $0 = 0$. To do this, we allow $\mathbf{p} \geq \mathbf{0}_{\mathbf{j}}^{-\mathbf{1}}$, instead of $\mathbf{p} \geq \mathbf{0}$. Note that if $p_j = -1$, then $q_j = k_j + 1$ since $\mathbf{p} + \mathbf{q} = \mathbf{k}$. Again, we follow the convention that $\binom{n}{i} = 0$ if $i < 0$ or $i > n$, as well as $g_{\mathbf{k},\mathbf{p}} = 0$ if $\mathbf{p}$ is not between $\mathbf{0}$ and $\mathbf{k}$, both inclusive. Also define $h^1_{\mathbf{p},\mathbf{q}} = h^2_{\mathbf{p},\mathbf{q}} = 0$ if either $\mathbf{p}$ or $\mathbf{q}$ are not between $\mathbf{0}$ and $\mathbf{k}$. Thus, all the new constraints added are indeed of the type $0 = 0$.

After these modifications, the system obtained has one constraint corresponding to $h^t_{\mathbf{p},\mathbf{q}}$ for each $\mathbf{0} \leq \mathbf{q} \leq \mathbf{k}_{\mathbf{j}}^{\mathbf{1}}$ (or equivalently $\mathbf{0}_{\mathbf{j}}^{-\mathbf{1}} \leq \mathbf{p} \leq \mathbf{k}$), $\mathbf{p} + \mathbf{q} = \mathbf{k}$, $t = 1, 2$ with variables $\alpha_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, \beta_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, \gamma_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}$ for $\mathbf{0} \leq \mathbf{r} \leq \mathbf{k}_{\mathbf{j}}^{\mathbf{1}}$. Further, let

$$n_{j,\mathbf{k}} = |D_{\mathbf{k}}| \qquad\qquad D_{\mathbf{k}} = \{\mathbf{r} : \mathbf{0} \leq \mathbf{r} \leq \mathbf{k}\}$$

We will write this system in a matrix form, given by a matrix $A_{j,\mathbf{k}}$ of dimension $2n_{j,\mathbf{k}_{\mathbf{j}}^{\mathbf{1}}} \times 3n_{j,\mathbf{k}_{\mathbf{j}}^{\mathbf{1}}}$ such that

$$A_{j,\mathbf{k}} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} h^1_j \\ h^2_j \end{bmatrix}$$

Here $\xi = (\xi_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}})$ is the vector of dimension $n_{j,\mathbf{k}_{\mathbf{j}}^{\mathbf{1}}}$ for $\xi \in \{\alpha, \beta, \gamma\}$. For notational convenience, we will fix j and $\mathbf{k}$ and denote $A = A_{j,\mathbf{k}}$. We will index rows of A by $(\mathbf{p}, t)$ and columns by $(\mathbf{r}, \xi)$ where $\mathbf{r}, \mathbf{p}_{\mathbf{j}}^{\mathbf{1}} \in D_{\mathbf{k}_j^1}$, $t \in \{1, 2\}$, $\xi \in \{\alpha, \beta, \gamma\}$. Further, we will denote by $A_{t,\xi}$ the submatrix of A corresponding to the rows $(\mathbf{p}, t)$ and columns $(\mathbf{r}, \xi)$, that is, $A_{t,\xi}(\mathbf{p}, \mathbf{r}) = A((\mathbf{p}, t), (\mathbf{r}, \xi))$. Matrix A has only $2n_{j,\mathbf{k}}$ non-trivial rows, namely the rows which correspond to $\mathbf{p}$ such that $\mathbf{p} \geq 0$. Hence to show that the system above has a solution, it suffices to prove that matrix A has rank $2n_{j,\mathbf{k}}$.

Define X, Y to be $n_{j,\mathbf{k}} \times n_{j,\mathbf{k}}$ matrices with rows and columns indexed by elements of $D_{\mathbf{k}}$ such that

$$X(\mathbf{p}, \mathbf{r}) = \left\langle g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{p}_{\mathbf{j}}^{\mathbf{1}}} \right\rangle$$

$$Y(\mathbf{p}, \mathbf{r}) = (-\mathbf{1})^{\mathbf{p}_{\mathbf{j}}^{\mathbf{1}}} \left\langle g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{k}_{\mathbf{j}}^{\mathbf{1}}-\mathbf{p}_{\mathbf{j}}^{\mathbf{1}}} \right\rangle$$

Now, assign $\Omega_1 = 1$, $\Omega_j = \frac{M^j-1}{M-1}$ for $j > 1$. For this choice of Ω_j's, it is shown in [Tur02] that the functions $g_{\mathbf{k},\mathbf{s}}$ for $\mathbf{0} \leq \mathbf{s} \leq \mathbf{k}$ are linearly independent. It follows from this that the matrices X and Y are full rank. Let P be the permutation matrix that takes row $\mathbf{r}$ of this matrix to row $\mathbf{r}_{\mathbf{j}}^{\mathbf{1}}$ unless $r_j = k_j$, in which case it takes row $\mathbf{r}$ to $\mathbf{s}$ where $s_i = r_i$ for all $i \neq j$ and $s_j = -1$. Thus, for any matrix M, $PM(\mathbf{p}, \mathbf{r}) = M(\mathbf{p}_{\mathbf{j}}^{-\mathbf{1}}, \mathbf{r})$ when $p_j \neq -1$, and $PM(\mathbf{p}, \mathbf{r}) = M(\mathbf{p}', \mathbf{r})$ where $p'_i = p_i$ for $i \neq j$ and $p'_i = k_j$ if $p_j = -1$. In particular,

$$PX(\mathbf{p}, \mathbf{r}) = X(\mathbf{p}_{\mathbf{j}}^{-\mathbf{1}}, \mathbf{r}) = \left\langle g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{p}} \right\rangle$$

$$PY(\mathbf{p}, \mathbf{r}) = Y(\mathbf{p}_{\mathbf{j}}^{-\mathbf{1}}, \mathbf{r}) = (-1)^{\mathbf{p}} \left\langle g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{r}}, g_{\mathbf{k}_{\mathbf{j}}^{\mathbf{1}},\mathbf{k}_{\mathbf{j}}^{\mathbf{1}}-\mathbf{p}} \right\rangle$$

when $\mathbf{p} \geq \mathbf{0}$. Define $n_{j,\mathbf{k}} \times n_{j,\mathbf{k}}$ diagonal matrices D_1, D_2, D_3 such that

$$D_1(\mathbf{p}, \mathbf{p}) = \binom{\mathbf{k}_{\mathbf{j}}^{-\mathbf{1}}}{\mathbf{p}} \qquad D_2(\mathbf{p}, \mathbf{p}) = \binom{\mathbf{k}_{\mathbf{j}}^{-\mathbf{1}}}{\mathbf{p}_{\mathbf{j}}^{-\mathbf{1}}} \qquad D_3(\mathbf{p}, \mathbf{p}) = \binom{\mathbf{k}}{\mathbf{p}}$$

for $\mathbf{0_j^{-1}} \leq \mathbf{p} \leq \mathbf{k}$. Recalling that $\mathbf{q} = \mathbf{k} - \mathbf{p}$, we see that

$$
\begin{aligned}
A_{1,\alpha}(\mathbf{p},\mathbf{r}) &= \Omega^{\mathbf{p_j^{-1}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} (-\mathbf{1})^{\mathbf{p_j^{-1}}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{k_j^1}-\mathbf{p_j^{-1}}} \right\rangle + \Omega^{\mathbf{p_j^{-1}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p}} (-\mathbf{1})^{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{k_j^1}-\mathbf{p_j^1}} \right\rangle \\
&= \Omega^{\mathbf{p_j^{-1}}} D_2(\mathbf{p},\mathbf{p}) P^2 Y(\mathbf{p},\mathbf{r}) - \Omega^{\mathbf{p_j^{-1}}} D_1(\mathbf{p},\mathbf{p}) Y(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{1,\alpha} &= \Omega^{\mathbf{p_j^{-1}}} (D_2 P^2 - D_1) Y \\
A_{1,\beta}(\mathbf{p},\mathbf{r}) &= \frac{1}{\Omega^{\mathbf{q}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p_j^{-1}}} \right\rangle - \frac{1}{\Omega^{\mathbf{q}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p_j^1}} \right\rangle \\
&= \frac{1}{\Omega^{\mathbf{q}}} D_2(\mathbf{p},\mathbf{p}) P^2 X(\mathbf{p},\mathbf{r}) - \frac{1}{\Omega^{\mathbf{q}}} D_1(\mathbf{p},\mathbf{p}) X(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{1,\beta} &= \frac{1}{\Omega^{\mathbf{q}}} (D_2 P^2 - D_1) X \\
A_{1,\gamma}(\mathbf{p},\mathbf{r}) &= -\frac{1}{\Omega^{\mathbf{q_j^1}}} \binom{\mathbf{k}}{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p}} \right\rangle \\
&= -\frac{1}{\Omega^{\mathbf{q_j^1}}} D_3(\mathbf{p},\mathbf{p}) P X(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{1,\gamma} &= -\frac{1}{\Omega^{\mathbf{q_j^1}}} D_3 P X \\
A_{2,\alpha}(\mathbf{p},\mathbf{r}) &= \Omega^{\mathbf{p}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} (-\mathbf{1})^{\mathbf{p_j^{-1}}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{k_j^1}-\mathbf{p}} \right\rangle + \Omega^{\mathbf{p}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p}} (-\mathbf{1})^{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{k_j^1}-\mathbf{p_j^2}} \right\rangle \\
&= -\Omega^{\mathbf{p}} D_2(\mathbf{p},\mathbf{p}) P Y(\mathbf{p},\mathbf{r}) + \Omega^{\mathbf{p}} D_1(\mathbf{p},\mathbf{p}) P^{-1} Y(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{2,\alpha} &= \Omega^{\mathbf{p}} (-D_2 P + D_1 P^{-1}) Y \\
A_{2,\beta}(\mathbf{p},\mathbf{r}) &= -\frac{1}{\Omega^{\mathbf{q_j^{-1}}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p_j^{-1}}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p}} \right\rangle + \frac{1}{\Omega^{\mathbf{q_j^{-1}}}} \binom{\mathbf{k_j^{-1}}}{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p_j^2}} \right\rangle \\
&= -\frac{1}{\Omega^{\mathbf{q_j^{-1}}}} D_2(\mathbf{p},\mathbf{p}) P X(\mathbf{p},\mathbf{r}) + \frac{1}{\Omega^{\mathbf{q_j^{-1}}}} D_1(\mathbf{p},\mathbf{p}) P^{-1} X(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{2,\beta} &= \frac{1}{\Omega^{\mathbf{q_j^{-1}}}} (-D_2 P + D_1 P^{-1}) X \\
A_{2,\gamma}(\mathbf{p},\mathbf{r}) &= \frac{1}{\Omega^{\mathbf{q}}} \binom{\mathbf{k}}{\mathbf{p}} \left\langle g_{\mathbf{k_j^1},\mathbf{r}}, g_{\mathbf{k_j^1},\mathbf{p_j^1}} \right\rangle \\
&= \frac{1}{\Omega^{\mathbf{q}}} D_3(\mathbf{p},\mathbf{p}) X(\mathbf{p},\mathbf{r}) \\
\Rightarrow A_{2,\gamma} &= \frac{1}{\Omega^{\mathbf{q}}} D_3 X
\end{aligned}
$$

For the above equations to go through as is, we need to check the case when $p_j = -1$, since definitions of PX and PY are different for this case. But, in this case, $D_1(\mathbf{p},\mathbf{p}) = D_2(\mathbf{p},\mathbf{p}) = 0$, and hence the equations hold. Similarly, we need to check the case $p_j = 0$ for blocks $A_{1,\alpha}$ and $A_{1,\beta}$, but again, $D_2(\mathbf{p},\mathbf{p}) = 0$ and hence the equations hold. Thus, we can write A as

$$\begin{bmatrix} \mathrm{I} & 0 \\ 0 & \Omega_j \mathrm{I} \end{bmatrix} \begin{bmatrix} D_2P^2 - D_1 & D_2P^2 - D_1 & -D_3P \\ -D_2P + D_1P^{-1} & -D_2P + D_1P^{-1} & D_3 \end{bmatrix} \begin{bmatrix} \Omega^{\mathbf{p}_\mathbf{j}^{-1}}\mathrm{I} & 0 & 0 \\ 0 & \frac{1}{\Omega^{\mathbf{q}}}\mathrm{I} & 0 \\ 0 & 0 & \frac{1}{\Omega^{\mathbf{q}_\mathbf{j}^{1}}}\mathrm{I} \end{bmatrix} \begin{bmatrix} Y & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{bmatrix}$$

To show that A has rank $2n_{j,\mathbf{k}}$, it suffices to show that the matrix

$$B = \begin{bmatrix} D_2P^2 - D_1 & -D_3P \\ -D_2P + D_1P^{-1} & D_3 \end{bmatrix}$$

has rank $2n_{j,\mathbf{k}}$. Let us index rows of B using $(\mathbf{p}, s)$ and columns using $(\mathbf{p}, t)$ for $s, t \in \{1, 2\}$. Since P is a permutation matrix, post multiplying by P takes column $\mathbf{r}$ of this matrix to column $\mathbf{r}_\mathbf{j}^{-1}$, where the indices cycle whenever they are out of bounds. More specifically,

$$MP(\mathbf{p},\mathbf{r}) = P^{-1}M^\intercal(\mathbf{r},\mathbf{p}) = M^\intercal(\mathbf{r}_\mathbf{j}^{\mathbf{1}},\mathbf{p}) = M(\mathbf{p},\mathbf{r}_\mathbf{j}^{\mathbf{1}}).$$

Hence, for a fixed row $(\mathbf{p}, 1)$ the non-zero entries in B are in columns $(\mathbf{p}_\mathbf{j}^{-\mathbf{2}}, 1), (\mathbf{p}, 1), (\mathbf{p}_\mathbf{j}^{-\mathbf{1}}, 2)$. Similarly, non-zero entries in the row $(\mathbf{p}, 2)$ are in columns $(\mathbf{p}_\mathbf{j}^{-\mathbf{1}}, 1), (\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}, 2)$. Observe that rows $(\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1)$ and $(\mathbf{p}, 2)$ have non-zero entries in the same columns. This gives us a procedure to convert this matrix into a lower triangular matrix using row operations, where indices are ordered using any order $<_R$ that respects

1. $(\mathbf{p}, t) <_R (\mathbf{q}, t)$ if $p_j < q_j$
2. $(\mathbf{p}, 1) <_R (\mathbf{q}, 2)$ for all $\mathbf{0}_\mathbf{j}^{-\mathbf{1}} \leq \mathbf{p}, \mathbf{q} \leq \mathbf{k}$

In particular, any lexicographical ordering with highest priority to the j^{th} coordinate works.

Note that only upper triangular non-zero entries using any such ordering are of the type $((\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}, 2))$. Now, we eliminate these using the following row operations:

$$R(\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1) \leftarrow R(\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1) + C_\mathbf{p} R(\mathbf{p}, 2))$$

for all $\mathbf{p}$ such that $0 \leq \mathbf{p} \leq \mathbf{k}_\mathbf{j}^{-\mathbf{1}}$. Here

$$C_\mathbf{p} = -\frac{B((\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}, 2))}{B((\mathbf{p}, 2), (\mathbf{p}, 2))} = -\frac{-\binom{\mathbf{k}}{\mathbf{p}_\mathbf{j}^{\mathbf{1}}}}{\binom{\mathbf{k}}{\mathbf{p}}} = \frac{\binom{k_j}{p_j+1}}{\binom{k_j}{p_j}} = \frac{k_j - p_j}{p_j + 1}$$

Note that after this set of operations, $B((\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}, 2)) \leftarrow 0$. On the other hand,

$$\begin{aligned} B((\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1)) \leftarrow & B((\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1), (\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1)) + \frac{k_j - p_j}{p_j + 1} B((\mathbf{p}, 2), (\mathbf{p}_\mathbf{j}^{\mathbf{1}}, 1)) \\ = & -\binom{\mathbf{k}_\mathbf{j}^{-\mathbf{1}}}{\mathbf{p}_\mathbf{j}^{\mathbf{1}}} + \frac{k_j - p_j}{p_j + 1}\binom{\mathbf{k}_\mathbf{j}^{-\mathbf{1}}}{\mathbf{p}} \\ = & \binom{\mathbf{k}_\mathbf{j}^{-\mathbf{1}}}{\mathbf{p}}\left(-\frac{k_j - p_j - 1}{p_j + 1} + \frac{k_j - p_j}{p_j + 1}\right) \\ = & \frac{1}{p_j + 1}\binom{\mathbf{k}_\mathbf{j}^{-\mathbf{1}}}{\mathbf{p}} \neq 0 \end{aligned}$$

The only non-zero entries in the upper triangle after this operation corresponds to positions $((\mathbf{p_j^1}, 1), (\mathbf{p}, 2))$, for $\mathbf{0_j^{-1}} \leq \mathbf{p} \leq \mathbf{k_j^{-1}}$, such that $p_j = -1$. To eliminate these, we perform the following row operations:

$$R(\mathbf{p_j^1}, 1) \leftrightarrow R(\mathbf{p}, 2)$$

for all $\mathbf{0_j^{-1}} \leq \mathbf{p} \leq \mathbf{k_j^{-1}}$ such that $p_j = -1$. Hence,

$$B((\mathbf{p}, 2), (\mathbf{p}, 2)) \leftarrow B((\mathbf{p_j^1}, 1), (\mathbf{p}, 2)) = \binom{\mathbf{k}}{\mathbf{p_j^1}} \neq 0$$

Note that $R(\mathbf{p}, 2) = 0$ since this row corresponds to a dummy constraint. Also, the other two non-zero entries in $R(\mathbf{p_j^1}, 1)$ are in the first half, and hence this does not create any upper triangular entries. Hence, this matrix is in fact lower triangular, in the given ordering $<_R$ of indices.

After the operations, among the diagonal terms, $B((\mathbf{p}, 2), (\mathbf{p}, 2)) \neq 0$ for $\mathbf{0_j^{-1}} \leq \mathbf{p} \leq \mathbf{k}$. Also, $B((\mathbf{p}, 1), (\mathbf{p}, 1)) \neq 0$ for $\mathbf{0_j^1} \leq \mathbf{p} \leq \mathbf{k}$. Therefore, the total number of non-zero diagonal entries is

$$n_{j,\mathbf{k}} \left(\frac{k_j + 1}{k_j} + \frac{k_j - 1}{k_j} \right) = 2n_{j,\mathbf{k}}$$

This proves that the matrix has rank $2n_{j,\mathbf{k}}$, which is the same as the number of non-trivial rows, and hence the system has a solution for any r_1, r_2. Consequently, we can always find polynomial functions J, F, G as required. □

B.5 Proof of Lemma 129

Proof. From Lemma 126, it suffices to focus on H being a polynomial. We break the time from ϕ to 0 for which we want to flow the ODE given by (7.14) into $(n+1)$ small chunks of length τ, i.e., let $\tau = \phi/(n+1)$. Further, let $A_i = T_{(n-i+1)\tau,(n-i)\tau}$. Then, the time-$\phi$ flow map can be write as the composition of $n+1$ maps, that is

$$T_{\phi,0} = T_{\tau,0} \circ \cdots \circ T_{\phi,\phi-\tau} = A_n \circ \cdots \circ A_0$$

Let $\mathcal{C}_0 = T_{0,\phi}(\mathcal{C})$. Let $\mathcal{C}_1, \ldots, \mathcal{C}_{n+1}$ be a sequence of compact sets such that $A_i(\mathcal{C}_i)$ is in the interior of $\mathcal{C}_{i+1}$; by choosing them small enough, we can make $\mathcal{C}_{n+1}$ an arbitrary compact set containing $\mathcal{C}$ in its interior. Below, we treat $A_0, \ldots, A_n$ (and their approximations) as maps $\mathcal{C}_0 \to \mathcal{C}_1 \to \cdots \to \mathcal{C}_{n+1}$, and when we take the C^1 norm, we do it on the appropriate compact set. For small enough η, the η-discretized maps will stay inside the $\mathcal{C}_i$.

Let S_i denote the time-2π flow map obtained by running the ODE system (7.12) from Lemma 127 above which approximates the map $T_{(n-i+1)\tau,(n-i)\tau} = A_i$. Further, let S_i' denote the map obtained by discretizing the ODE system as in (7.13) with step size η. Then, we have that for each i, as $\eta \to 0$,

$$\begin{aligned}
\|S_i' - A_i\|_{C^1} &\leq \|S_i' - S_i + S_i - A_i\|_{C^1} \\
&\leq \|S_i' - S_i\|_{C^1} + \|S_i - A_i\|_{C^1} \\
&\leq O(\eta) + O(\tau^2) && \text{(by Lemmas 127 and 128)}
\end{aligned}$$

We choose $\eta = \tau^2$. Using the definition of C^1 norm, this implies that

$$\|S'_i - A_i\| = O(\tau^2) \qquad\qquad \|DS'_i - DA_i\| = O(\tau^2),$$

where $\|\cdot\|$ denotes L^∞ norm on $\mathcal{C}_i$; for matrix-valued functions $M(x)$ on $\mathcal{C}_i$, $\|M\| = \sup_{x\in\mathcal{C}_i} \|M(x)\|_2$, where $\|\cdot\|_2$ denotes spectral norm. Again, using the definition of the C^1 norm,

$$\begin{aligned}
&\|A_n \circ \cdots \circ A_0 - S'_n \circ \cdots \circ S'_0\|_{C^1} \\
\leq &\|A_n \circ \cdots \circ A_0 - S'_n \circ \cdots \circ S'_0\| + \|D(A_n \circ \cdots \circ A_0) - D(S'_n \circ \cdots \circ S'_0)\|
\end{aligned}$$

We will bound each term individually. For the first term, note that

$$\begin{aligned}
&\|A_n \circ \cdots \circ A_0 - S'_n \circ \cdots \circ S'_0\| \\
&\leq \|A_n \circ \cdots \circ A_1 \circ A_0 - A_n \circ \cdots \circ A_1 \circ S'_0\| + \|A_n \circ \cdots \circ A_1 \circ S'_0 - S'_n \circ \cdots \circ S'_1 \circ S'_0\| \\
&\hspace{30em}\text{(by triangle inequality)} \\
&= \|T_{\phi-\tau,0} \circ A_0 - T_{\phi-\tau} \circ S'_0\| + \|A_n \circ \cdots \circ A_1 \circ S'_0 - S'_n \circ \cdots \circ S'_1 \circ S'_0\| \\
&\leq \|DT_{\phi-\tau,0}\|\|S'_0 - A_0\| + \|A_n \circ \cdots \circ A_1 \circ S'_0 - S'_n \circ \cdots \circ S'_1 \circ S'_0\| \\
&\leq O(\tau^2) + \|A_n \circ \cdots \circ A_1 \circ S'_0 - S'_n \circ \cdots \circ S'_1 \circ S'_0\|
\end{aligned} \tag{B.26}$$

Observe that

$$\begin{aligned}
&\sup_x \|A_n \circ \cdots \circ A_1 \circ S'_0(x) - S'_n \circ \cdots \circ S'_1 \circ S'_0(x)\| \\
&= \sup_{y=S'_0(x)} \|A_n \circ \cdots \circ A_1(y) - S'_n \circ \cdots \circ S'_1(y)\| \\
&\leq \sup_y \|A_n \circ \cdots \circ A_1(y) - S'_n \circ \cdots \circ S'_1(y)\| \\
&= \|A_n \circ \cdots \circ A_1(y) - S'_n \circ \cdots \circ S'_1(y)\|
\end{aligned} \tag{B.27}$$

Using (B.27), (B.26), and induction, we get that

$$\|A_n \circ \cdots \circ A_0 - S'_n \circ \cdots \circ S'_0\| \leq O(n\tau^2)$$

Now, we bound the derivatives:

$$
\begin{aligned}
&\|D(A_n \circ \cdots \circ A_0) - D(S'_n \circ \cdots \circ S'_0)\| \\
&\le \|D(A_n \circ \cdots \circ A_1 \circ A_0) - D(A_n \circ \cdots \circ A_1 \circ S'_0)\| \\
&\quad + \|D(A_n \circ \cdots \circ A_1 \circ S'_0) - D(S'_n \circ \cdots \circ S'_1 \circ S'_0)\| && \text{(by triangle inequality)} \\
&= \sup_x \big\| DT_{\phi-\tau,0}|_{A_0(x)} DA_0(x) - DT_{\phi-\tau,0}|_{S'_0(x)} DS'_0(x) \big\| \\
&\quad + \sup_x \big\| D(A_n \circ \cdots \circ A_1)|_{S'_0(x)} DS'_0(x) - D(S'_n \circ \cdots \circ S'_1)|_{S'_0(x)} DS'_0)(x) \big\| && \text{(by chain rule)} \\
&\le \sup_x \big\| DT_{\phi-\tau,0}|_{A_0(x)} DA_0(x) - DT_{\phi-\tau,0}|_{S'_0(x)} DA_0(x) \big\| \\
&\quad + \sup_x \big\| DT_{\phi-\tau,0}|_{S'_0(x)} DA_0(x) - DT_{\phi-\tau,0}|_{S'_0(x)} DS'_0(x) \big\| && \text{(by triangle inequality)} \\
&\quad + \|DS'_0\| \|D(A_n \circ \cdots \circ A_1) - D(S'_n \circ \cdots \circ S'_1)\| && \text{(B.28)} \\
&\le \sup_x \big\| DT_{\phi-\tau,0}|_{A_0(x)} - DT_{\phi-\tau,0}|_{S'_0(x)} \big\| \|DA_0\| \\
&\quad + \sup_x \big\| DT_{\phi-\tau,0}|_{S'_0(x)} \big\| \|DA_0 - DS_0\| \\
&\quad + \|DS'_0\| \|D(A_n \circ \cdots \circ A_1) - D(S'_n \circ \cdots \circ S'_1)\| \\
&\le \big\| D^2 T_{\phi-\tau,0} \big\| \|S'_0 - A'_0\| \|DA_0\| + \|DT_{\phi-\tau,0}\| \|DA_0 - DS'_0\| \\
&\quad + \|DS'_0\| \|D(A_n \circ \cdots \circ A_1) - D(S'_n \circ \cdots \circ S'_1)\| \\
&\le O(\tau^2) + \Big(\|DA_0\| + O(\tau^2) \Big) \|D(A_n \circ \cdots \circ A_1) - D(S'_n \circ \cdots \circ S'_1)\| && \text{(B.29)}
\end{aligned}
$$

where, for a 3-tensor $\mathcal{T}$, we define $\|\mathcal{T}\| = \sup_{\|u\|=1} \|\mathcal{T}u\|_2$, where $\|\mathcal{T}u\|_2$ is the spectral norm of the matrix $\mathcal{T}u$, and we define $\|D^2 T_{\phi-\tau,0}\| = \sup_x \|D^2 T_{\phi-\tau,0}(x)\|$. In the last step, we use the fact that $\|DT_{s,t}\|, \|D^2 T_{s,t}\|$ are bounded for all $s, t > 0$; this follows from Lemma 185 below. (Alternatively, note that $\|DT_{s,t}\|$ can also be more directly bounded by Theorem 181.)

In the above, (B.28) follows using an argument similar to (B.27), (B.29) follows since $\|DA_0 - DS'_0\| = O(\tau^2)$. Further, differentiating (B.33), we get

$$
DA_0 = \mathrm{I} + \tau D_{(x,v)} F(x, v, t) + O(\tau^2)
$$

where F denotes the defining equation of the ODE system in (7.14). Therefore, we get

$$
\|DA_0\| \le 1 + \tau L + O(\tau^2)
$$

where L is the upper bound on $\|Df\|$ over all the appropriate compact sets. Using this bound and induction, we get that

$$
\|D(A_n \circ \cdots \circ A_0) - D(S'_n \circ \cdots \circ S'_0)\| \le O(n\tau^2)(1 + \tau L + O(\tau^2))^n = O(n\tau^2 e^{n\tau L})
$$

for small enough τ. Substituting $n\tau = \phi$, we get the overall C^1 bound of

$$
\|A_n \circ \cdots \circ A_0 - S'_n \circ \cdots \circ S'_0\|_{C^1} = O(\phi\tau e^{\phi L}).
$$

Now, we can choose τ small enough so that the two maps are ϵ_1-close, finishing the proof.

Concretely, we can write each S_i' as a composition of affine-coupling maps (which constitute the $f_1, \ldots, f_N$ in the lemma statement). In this manner, we can compose these compositions of affine coupling maps over each τ-sized chunk of time so as to get a map which is overall close to the required flow map. $\square$

Lemma 185. *Consider the ODE $\frac{d}{dt}x(t) = F(x(t), t)$ for $F(x,t)$ that is C^ℓ in $x \in \mathbb{R}^d$ and continuous in t. Let $\mathcal{C}$ be a compact set and suppose solutions exist for any $(x(0), v(0)) \in \mathcal{C}$ up to time T. Let $T_{s,t}$ be the flow map from time s to time t, for any $0 \le s, t \le T$. Then for any $0 \le r \le \ell$, $D^r T_{s,t}$ is bounded on $T_s(\mathcal{C})$.*

Proof. Let $\partial_{i_1 \cdots i_r} = \frac{\partial^r}{\partial x_{i_1} \cdots \partial x_{i_r}}$. Using the chain rule as in Lemma 188, we find by induction that

$$\frac{d}{dt}\partial_{i_1 \cdots i_r}(T_t(x)) = \sum_{i=1}^{d} \partial_i F(x(t), t)\partial_{i_1 \cdots i_r}(T_t(x)_i) + G(DF, \ldots, D^r F, DT_t, \ldots, D^{r-1}T_t). \tag{B.30}$$

for some polynomial G. For $r = 1$, the differential equation is given by Lemma 188. By a Grönwall argument, a bound on DF gives an upper and lower bound on the singular values of DT_t as in (B.10). We use induction on r; for $r > 1$, let $v(t)$ be equal to $(\partial_{i_1 \cdots i_r}(T_t(x)))_{i_1 \cdots i_r}$ written as one large vector. By the chain rule and (B.30),

$$\frac{d}{dt}\|v(t)\|^2 \le \langle |v(t)|, A|v(t)| + b\rangle \le \left(\sigma_{\max}(A) + \frac{1}{2}\right)\|v(t)\|^2 + \frac{1}{2}\|b\|^2$$

for some A, b depending on $DF, \ldots, D^r F, DT_t, \ldots, D^{r-1}T_t$, where $\sigma_{\max}$ denotes the maximum singular value and $|v|$ denotes entrywise absolute value. Grönwall's inequality (Lemma 192) applied to $\|v(t)\|^2$ then gives bounds on $\|v(t)\|^2$ and hence $\left|\frac{d}{dt}\partial_{i_1 \cdots i_r}(T_t(x))\right|$. This shows $D^r T_{s,t}$ is bounded when $s \le t$ (by starting the flow at time s).

When $s > t$, note that the computation of the rth derivative of an inverse map involves up-to-r derivatives of the forward map, and inverses of the first derivative. As we have a lower bound on the singular value of DF, this implies that $D^r T_{s,t}$ is bounded. $\square$

B.5.1 Proof of Lemma 128

We consider a more general ODE than the specific one in (7.12), of the form

$$\begin{cases} \frac{d}{dt}(x(t)) = f(x(t), v(t), t) \\ \frac{d}{dt}(v(t)) = g(x(t), v(t), t) \end{cases} \tag{B.31}$$

where f, g are C^2 functions in x, v, t. Given a compact set $\mathcal{C}$, suppose that the solutions are well-defined for any $(x(0), v(0)) \in \mathcal{C}$ up to time T. Consider discretizing these ODEs into steps of size η, as follows:

$$\begin{cases} \widetilde{T}_i^x(X_i) = X_{i+1} = X_i + \eta f(X_i, V_{i+1}, t_i) \\ \widetilde{T}_i^v(V_i) = V_{i+1} = V_i + \eta g(X_i, V_i, t_i) \end{cases} \tag{B.32}$$

where $t_i = i\eta$. We call this the alternating Euler update. The actual flow maps are given by

$$\begin{cases} T_i^x(x_i) = x_{i+1} = x_i + \eta f(x_i, v_i, t_i) + \int_{i\eta}^{(i+1)\eta} \int_{i\eta}^t x''(s)\,ds\,dt \\ T_i^v(v_i) = v_{i+1} = v_i + \eta g(x_i, v_i, t_i) + \int_{i\eta}^{(i+1)\eta} \int_{i\eta}^t v''(s)\,ds\,dt \end{cases} \tag{B.33}$$

We bound the local truncation error. This consists of two parts. First, we have the integral terms in (B.33):

$$\left\| \begin{bmatrix} \int_{i\eta}^{(i+1)\eta} \int_{i\eta}^t x''(s)\,ds\,dt \\ \int_{i\eta}^{(i+1)\eta} \int_{i\eta}^t v''(s)\,ds\,dt \end{bmatrix} \right\| \le \frac{1}{2}\eta^2 \max_{s\in[0,t_i]} \left\| \begin{bmatrix} x''(s) \\ v''(s) \end{bmatrix} \right\|. \tag{B.34}$$

Second we bound the error from using $\widetilde{v}_{i+1} := v_i + \eta g(x_i, v_i, t_i)$ instead of v_i in the x update,

$$\begin{aligned} \|\eta[f(x_i, v_i + \eta g(x_i, v_i, t_i), t_i) - f(x_i, v_i, t_i)]\| &\le \left\| \eta \int_0^\eta D_v f(x_i, v_i + sg(x_i, v_i, t_i), t_i) g(x_i, v_i, t_i)\,ds \right\| \\ &\le \eta^2 \max_{\mathcal{C}'} \|D_v f\| \max_{\mathcal{C}'} \|g\|. \end{aligned} \tag{B.35}$$

where $D_v f(x, v, t)$ denotes the Jacobian in the v variables (rather than the directional derivative), and where we define

$$\mathcal{C}' := \{(x, v + sg(x, v, t), t) : (x, v) = T_t(x_0, v_0) \text{ for some } (x_0, v_0) \in \mathcal{C}, 0 \le s \le T\},$$

which ensures that it contains $(x_i, v_i + sg(x_i, v_i, t_i), t_i)$ and (x_i, v_i, t_i). The local truncation error is then at most the sum of (B.34) and (B.35).

Supposing that $\begin{bmatrix} f \\ g \end{bmatrix}$ is L-Lipschitz in $(x, v) \in \mathbb{R}^{2d}$ for each t, we obtain by a standard argument (similar to the proof for the usual Euler's method, see e.g., [AG11, §16.2]) that the global error at any step is bounded by

$$\left\| \begin{bmatrix} \widetilde{x}_i \\ \widetilde{v}_i \end{bmatrix} - \begin{bmatrix} x_i \\ v_i \end{bmatrix} \right\| \le \eta \cdot \frac{e^{Lt_i} - 1}{L} \left(\max_{\mathcal{C}'} \|D_v f\| \max_{\mathcal{C}'} \|g\| + \frac{1}{2} \max_{s\in[0,t_i]} \left\| \begin{bmatrix} x''(s) \\ v''(s) \end{bmatrix} \right\| \right). \tag{B.36}$$

In the case when $\begin{bmatrix} f \\ g \end{bmatrix}$ is not globally Lipschitz, we show that we can restrict the argument to a compact set on which it is Lipschitz. Let $\mathcal{C}''$ be a compact set which contains $\{(x, v, t) : (x, v) = T_t(x_0, v_0) \text{ for some } (x_0, v_0) \in \mathcal{C}, 0 \le s \le T\}$ in its interior. Apply the argument to $\hat{f}$ and $\hat{g}$ which are defined to be equal to f, g on $\mathcal{C}''$, and are globally Lipschitz. Then the error bound applies to the system defined by $\hat{f}, \hat{g}$. Hence, for small enough step size, the trajectory of the discretization stays inside $\mathcal{C}''$, and is the same as that for the system defined by f, g. Then (B.36) holds for small enough η and L equal to the Lipschitz constant in (x, v) on $\mathcal{C}''$.

To get a bound in C^1 topology, we need to bound the derivatives of these maps as well. Let $T_{s,t}(x, v)$ denote the flow map of system (B.31). Let $h(x, v, t) = (f(x, v, t), g(x, v, t))$. Now, consider the system of ODEs

$$\begin{cases} \frac{d}{dt}(x(t)) = f(x(t), v(t), t) \\ \frac{d}{dt}(v(t)) = g(x(t), v(t), t) \\ \frac{d}{dt}(\alpha(t)) = D_{(x,v)} f(x(t), v(t), t) \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} \\ \frac{d}{dt}(\beta(t)) = D_{(x,v)} g(x(t), v(t), t) \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} \end{cases} \tag{B.37}$$

where $\alpha(t), \beta(t)$ are $d \times 2d$ matrices. Note that setting $\begin{bmatrix} \alpha(0) \\ \beta(0) \end{bmatrix} = \mathrm{I}_{2d}$ and $\begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} = D_{(x,v)} T_{0,t}(x(0), v(0))$ satisfies (B.37) by Lemma 188.

Now we claim that applying the alternating Euler update to $(x, \alpha), (v, \beta)$, the resulting (α_i, β_i) is exactly the Jacobian of the flow map that arises from alternating Euler applied to x, v. This means that we can bound the errors for α, β using the bound for the alternating Euler method.

The claim follows from noting that the alternating Euler update on α, β is

$$\begin{aligned} \alpha_{i+1} &= (\mathrm{I}_d, O) + D_{(x,v)} f(x_i, v_{i+1}, t_i) \begin{bmatrix} \alpha_i \\ \beta_{i+1} \end{bmatrix} \\ \beta_{i+1} &= (O, \mathrm{I}_d) + D_{(x,v)} f(x_i, v_i, t_i) \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \end{aligned}$$

which is the same recurrence that is obtained from differentiating X_{i+1}, V_{i+1} in (B.32) with respect to X_0, V_0, and using the chain rule.

Thus we can apply (B.36) to get a bound for the Jacobians of the flow map. The constants in the $O(\eta)$ bound depend on up to the second derivatives of the x, v, α, β for the true solution, Lipschitz constants for $\begin{bmatrix} f \\ g \end{bmatrix}$, $D \begin{bmatrix} f \\ g \end{bmatrix}$ (on a suitable compact set), and bounds for $D_v f, g, D_v D_{(x,v)} f, D_{(x,v)} g$ (on a suitable compact set).

B.5.2 Wasserstein bounds

Lemma 186. *Given two distributions p, q and a function g with Lipschitz constant $L = \mathrm{Lip}(g)$,*

$$W_1(g_\# p, g_\# q) \le L W_1(p, q)$$

Proof. Let $\epsilon > 0$. Then there exists a coupling $(x, t) \sim \gamma$ such that

$$\int \|x - y\|_2 d\gamma(x, y) \le W_1(p, q) + \epsilon$$

Consider the coupling (x', y') given by $(x', y') = (g(x), g(y))$ where $(x, y) \sim \gamma$. Then

$$\begin{aligned} W_1(g_{\#}p, g_{\#}q) &\leq \int \|g(x) - g(y)\|_2 \, d\gamma(x, y) \\ &\leq \mathrm{Lip}(g) \int \|x - y\| \, d\gamma(x, y) \\ &\leq LW_1(p, q) + L\epsilon. \end{aligned}$$

Since this holds for all $\epsilon > 0$, we get that

$$W_1(g_{\#}p, g_{\#}q) \leq LW_1(p, q)$$

□

Lemma 187. *Given two functions $f, g : \mathbb{R}^d \to \mathbb{R}^d$ that are uniformly ϵ_1-close over a compact set $\mathcal{C}$ in C^1 topology, and a probability distribution p,*

$$W_1(f_{\#}(p|_{\mathcal{C}}), g_{\#}(p|_{\mathcal{C}})) \leq \epsilon_1$$

Proof. Consider the coupling γ, where a sample $(x, y) \sim \gamma$ is generated as follows: first, we sample $z \sim p|_{\mathcal{C}}$, and then compute $x = f(z)$, $y = g(z)$. By definition of the pushforward, the marginals of x and y are $f_{\#}(p|_{\mathcal{C}})$ and $g_{\#}(p|_{\mathcal{C}})$ respectively. However, we are given that for this γ, $\|x - y\| \leq \epsilon_1$ uniformly. Thus, we can conclude that

$$\begin{aligned} W_1(f_{\#}(p|_{\mathcal{C}}), g_{\#}(p|_{\mathcal{C}})) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2 \, d\gamma(x, y) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \epsilon_1 \, d\gamma(x, y) = \epsilon_1 \end{aligned}$$

□

B.5.3 Proof of Lemma 131

Proof. Fix any $R > 0$, and set $\mathcal{C} = B(0, R)$. Consider the coupling $(X, Y) \sim \gamma$, where a sample (X, Y) is generated as follows: we first sample $X \sim p^* = \mathcal{N}(0, \mathrm{I}_{2d})$. If $X \in B(0, R)$, then we set $Y = X$. Else, we draw Y from $p^*|_{\mathcal{C}}$. Clearly, the marginal of γ on X is p. Furthermore, since p^*

and $p^*|_{\mathcal{C}}$ are proportional within $\mathcal{C}$, the marginal of γ on Y is $p^*|_{\mathcal{C}}$. Then, we have that

$$\begin{aligned}
W_1(p^*, p^*|_{\mathcal{C}}) &\leq \int_{\mathbb{R}^{2d}\times\mathcal{C}} \|x-y\| d\gamma \\
&= \cancel{\int_{\mathcal{C}\times\mathcal{C}} \|x-y\| d\gamma} + \int_{\mathbb{R}^{2d}\backslash\mathcal{C}\times\mathcal{C}} \|x-y\| d\gamma \\
&= \int_{\mathbb{R}^{2d}\backslash\mathcal{C}\times\mathcal{C}} \|x-y\| d\gamma \\
&\leq \int_{\mathbb{R}^{2d}\backslash\mathcal{C}\times\mathcal{C}} (\|x\| + \|y\|) d\gamma \\
&\leq \int_{\mathbb{R}^{2d}\backslash\mathcal{C}\times\mathcal{C}} (\|x\| + R) d\gamma \\
&\leq \int_{\mathbb{R}^{2d}\backslash\mathcal{C}\times\mathcal{C}} (\|x\| + R) d\gamma \\
&= \int_{\mathbb{R}^{2d}\backslash\mathcal{C}} (\|x\| + R) dp^* \\
&\leq \int_{\mathbb{R}^{2d}\backslash\mathcal{C}} 2\|x\| dp^* = \frac{2}{\sqrt{2\pi}} \int_{\mathbb{R}^{2d}\backslash\mathcal{C}} \|x\| e^{-\frac{\|x\|^2}{2}} dx
\end{aligned}$$

Now, note that $\int_{\mathbb{R}^{2d}} \|x\| e^{-\frac{\|x\|^2}{2}} dx < \infty$. Hence, by the Dominated Convergence Theorem,

$$\lim_{R\to\infty} \int_{\mathbb{R}^{2d}\backslash B(0,R)} \|x\| e^{-\frac{\|x\|^2}{2}} dx = 0.$$

Thus, given any $\delta > 0$, we can choose R large enough so that the integral above is smaller than δ, which concludes the proof. □

B.5.4 Derivatives of flow maps

We state and prove a technical lemma about the ODE that the derivative of a flow map satisfies.

Lemma 188. *Suppose $x_t = x(t)$ satisfies the ODE*

$$\dot{x} = F(x, t)$$

with flow map $T(x,t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. Suppose $\alpha(t)$ be the derivative of the map $x \mapsto T(x,t)$ at x_0, then $\alpha(t)$ satisfies

$$\dot{\alpha} = DF(x_t, t)\alpha$$

with $\alpha(0) = \mathrm{I}$.

Proof. Let $T_t(x) = T(x,t)$. Then T_t satisfies

$$T_t(x_0) = \int_0^t F(x_s, s)\, ds.$$

Differentiating, we get

$$\begin{aligned}
\alpha(t) = DT_t(x_0) &= \int_0^t D(F(x_s, s))\, ds \\
&= \int_0^t DF(x_s, s) DT_s(x_0)\, ds && \text{by chain rule} \\
&= \int_0^t DF(x_s, s)\alpha(s)\, ds.
\end{aligned}$$

Now, looking at the derivative with respect to t, we get

$$\dot{\alpha} = DF(x_t, t)\alpha,$$

which is the required result. □

B.5.5 Solving Perturbed ODEs

In this section, we state a result about finding approximate solutions of perturbed differential equations. Consider the ODE having the following general form:

$$\dot{x} = Ax + \epsilon g(x, t)$$

The reason we are concerned with this ODE is that the ODE given by Equation (7.12) has precisely this form, namely with $x \equiv \begin{bmatrix} x \\ v \end{bmatrix}$, $A \equiv \begin{bmatrix} 0 & \mathrm{I}_d \\ -\mathrm{diag}(\Omega^2) & 0 \end{bmatrix}$ and $\epsilon g(x, t) \equiv -\tau \begin{bmatrix} F(v, t) \odot x \\ J(x, t) + G(x, t) \odot v \end{bmatrix}$.

Let $T^x : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be the time t flow map for this ODE. We will find a flow map $T^y : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ such that the maps T^x_t defined by $T^x_t(x) = T^x(t, x)$ and the map T^y_t defined by $T^y_t(y) = T^y(t, y)$ are uniformly ϵ-close over $\mathcal{C}$ in C^r topology for all $0 \le t \le 2\pi$. That is,

$$\sup_x \quad \|T^x_t(x) - T^y_t(x)\| + \|DT^x_t(x) - DT^y_t(x)\| + \cdots + \|D^r T^x_t(x) - D^r T^y_t(x)\|$$

is small, for all $t \in [0, 2\pi]$. Here D^r denotes the r-th derivative, and the norms are defined inductively as follows: for a r-tensor $\mathcal{T}$, we let $\|\mathcal{T}\| = \sup_{\|u\|=1} \|\mathcal{T}u\|$; here $\mathcal{T}u$ is a $(r-1)$-tensor. (The choice of norm is not important; we choose this for convenience.)

Lemma 189. *Consider the ODE*

$$\frac{d}{dt}x(t) = F(x(t), t) + \varepsilon G(x(t), t) \tag{B.38}$$

where $x : [0, t_{\max}] \to \mathbb{R}^n$, $F, G : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, and $F(x, t)$, $G(x, t)$ are C^1, and F is L-Lipschitz. Let $\mathcal{C}$ be a compact set, and suppose that for all $x_0 \in \mathcal{C}$, solutions to (B.38) with $x(0) = x_0$ exist for $0 \le t \le t_{\max}$ and $\varepsilon = 0$. Then there exists ε_0 such that solutions to (B.38) with $x(0) = x_0$ exist for $0 \le t \le t_{\max}$ and $0 \le \varepsilon < \varepsilon_0$.

Moreover, letting $x^{(\varepsilon)}(t)$ be the solution with given ε, we have that as $\varepsilon \to 0$, $\left\|x^{(\varepsilon)}(t) - x^{(0)}(t)\right\| = O(\varepsilon)$, where the constants in the $O(\cdot)$ depend only on L and $\max_{0\le t\le t_{\max}, x_0\in\mathcal{C}} \left\|G(x^{(0)}(t), t)\right\|$ (the maximum of G on the $\varepsilon = 0$ trajectories).

Proof. Let $T^\epsilon(t, x_0)$ be the flow map of (B.38). Let $\mathcal{K} = T^0(\mathcal{C} \times [0, t_{\max}])$ be the image of $\mathcal{C} \times [0, t_{\max}]$ under the flow map T^0. Since F is C^1, T^0 is C^1, which implies that $\mathcal{K}$ is bounded. Fix some $\epsilon_2 > 0$. Let $B(\mathcal{K}, r)$ denote the set

$$B(\mathcal{K}, r) = \{(x, t) \in \mathbb{R}^n \times [0, t_{\max}] : d(\mathcal{K}, x) \leq r\}$$

Let $\mathcal{K}_2 = B(\mathcal{K}, \epsilon_2)$. Note that since $\mathcal{K}$ is compact, so is $\mathcal{K}_2$. Let

$$M = \max\{\sup_{(x,t)\in\mathcal{K}_2\times[0,t_{\max}]} \|F(x, t)\|, \sup_{(x,t)\in\mathcal{K}_2\times[0,t_{\max}]} \|G(x, t)\|\}$$

M is finite since $\mathcal{K}_2$ is compact and F, G are C^1.

Let $h : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz C^1 function such that

$$h(x) = x \text{ if } |x| \leq M$$
$$|h(x)| \leq 2M \text{ for all } x.$$

Let $h_n : \mathbb{R}^n \to \mathbb{R}^n$ be defined as $h_n(x) = \frac{x}{\|x\|} h(\|x\|)$. Then $h_n(x)$ is also C^1 and is the identity function on $B(0, M)$. Let $F_1 = h_n \circ F$ and let $G_1 = h_n \circ F$. Then F_1, G_1 are C^1 functions such that $\|F_1\|, \|G_1\| \leq 2M$. Further, F_1 is L-Lipschitz. Now, we look at the ODE

$$\frac{d}{dt} x(t) = F_1(x(t), t) + \varepsilon G_1(x(t), t) \tag{B.39}$$

Since F_1, G_1 are C^1, note that the function $H_1(x, \epsilon, t) = F_1(x, t) + \epsilon G_1(x, t)$ is C^1 in x, t, ϵ. Therefore, using the existence theorem for parametric ODEs (Theorem 1.2, [Chi06]), there is a $\epsilon_1, t_1 > 0$ such that solutions $x_1^{(\epsilon)}(t)$ to (B.39) exist for all $x_0 \in \mathcal{C}, \epsilon < \epsilon_1$ and $t < t_1$. Further, the extensibility result for the ODEs (Theorem 1.4, [Chi06]) states that if t_1 is largest such value for which such solutions exist, then there exists a $x_0 \in \mathcal{C}$ and $\epsilon < \epsilon_1$ such that $\lim_{t\to t_1} \left\| x_1^{(\epsilon)}(t) \right\| = \infty$.

Now, we will bound $\left\| x_1^{(\epsilon)} - x_1^{(0)} \right\|$ for $t < t_1$. Define $\alpha = x_1^{(0)} - x_1^{(\epsilon)}$. Then $\alpha(t)$ satisfies

$$\frac{d}{dt}\alpha(t) = F_1(x_1^{(0)}(t), t) - F_1(x_1^{(\epsilon)}(t), t) - \epsilon G_1(x_1^{(\epsilon)}(t), t)$$

Therefore,

$$\begin{aligned}
\frac{d}{dt}\|\alpha(t)\|^2 &\leq 2\|\alpha(t)\|\|\frac{d}{dt}\alpha(t)\| \\
&\leq 2\|\alpha(t)\| \left\| F_1(x_1^{(0)}(t), t) - F_1(x_1^{(\epsilon)}(t), t) - \epsilon G_1(x_1^{(\epsilon)}(t), t) \right\| \\
&\leq 2\|\alpha(t)\|(L\|\alpha(t)\| + 2\epsilon M) \\
&\leq 2L\|\alpha(t)\|^2 + 4\epsilon M\|\alpha(t)\| \\
\implies \frac{d}{dt}\|\alpha(t)\| &\leq \frac{1}{2}\|\alpha(t)\|^{-1} \frac{d}{dt}\|\alpha(t)\|^2 \leq L\|\alpha(t)\| + 2\epsilon M
\end{aligned}$$

Now, Grönwall's inequality (Lemma 192) gives us the bound

$$\|\alpha(t)\| \le 2\epsilon t M e^{Lt} \le 2\epsilon t_{\max} M e^{Lt_{\max}} = O(\epsilon) \tag{B.40}$$

Since $t_{\max}, L, M$ are fixed, we can choose ϵ_0 such that $\epsilon_0 < \epsilon_1$ and $2\epsilon_0 t_{\max} M e^{Lt_{\max}} < \epsilon_2$, which ensure that for all $x_0 \in \mathcal{C}, \epsilon < \epsilon_0$ and $t < \min(t_1, t_{\max})$, the point $x_1^{(\epsilon)}(t)$ is in the interior of $\mathcal{K}_2$. Therefore, if $t_1 \le t_{\max}$ then $\lim_{t \to t_1} \left\| x_1^{(\epsilon)}(t) \right\| \in \mathcal{K}_2$, which contradicts the extensibility result. Thus, $t_1 > t_{\max}$, and hence flow maps for (B.39) exists for all $0 \le \epsilon \le \epsilon_0$ and $0 \le t \le t_{\max}$.

Now, we end with the remark that since $F_1 = F$ and $G_1 = G$ in $\mathcal{K}_2$, the flow map of (B.39) is a flow map for (B.38) inside $\mathcal{K}_2$, and therefore, solutions to (B.38) exist for all $x_0 \in \mathcal{C}, 0 \le \epsilon \le \epsilon_0$ and $0 \le t \le t_{\max}$.

Lastly, we will comment on value of M. Let G be L_1-Lipschitz on $\mathcal{K}_2$, and let

$$M' = \max_{0 \le t \le t_{\max}, x_0 \in \mathcal{C}} \left\| G(x^{(0)}(t), t) \right\|$$

Then $M \le M' + \epsilon_0 L_1$. Therefore, we can just choose ϵ_0 small enough so that $M \le 2M' + 1$, which enforces the constants in $O(\cdot)$ notation to depend only on L, M' and $t_{\max}$.

□

Lemma 190. *Consider the ODE's*

$$\begin{aligned} \frac{d}{dt} x(t) &= F(x(t), t) + \epsilon G(x(t), t) \\ \frac{d}{dt} y_0(t) &= F(y_0(t), t) \\ \frac{d}{dt} y(t) &= F(y(t), t) + \varepsilon G(y_0(t), t) \end{aligned} \tag{B.41}$$

such $F, G : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ are in C^{r+1}. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a compact set, and suppose that solutions to (B.41) exist for all $x_0 \in \mathcal{C}$. Let $T^x(x_0)$, $T^{y_0}(x_0)$, and $T^y(x_0)$ be the time $t_{\max}$-flow map corresponding to this ODE for initial values $x(t) = y_0(t) = y(t) = x_0$.

Then as $\varepsilon \to 0$, the maps T_t^x and T_t^y are $O(\epsilon^2)$ uniformly close over $\mathcal{C}$ in C^r topology, for all $t \in [0, t_{\max}]$. The constants in the $O(\cdot)$ depend on $\max_{0 \le k \le r+1, x_0 \in \mathcal{C}, 0 \le t \le t_{\max}} \left\| D^k F(x,t)|_{x = y_0(t)} \right\|$ *(the first $r+1$ derivatives of F on the y_0-trajectories) and* $\max_{0 \le k \le r, x_0 \in \mathcal{C}, 0 \le t \le t_{\max}} \left\| D^k G(x,t)|_{x = y_0(t)} \right\|$, *(the first r derivatives of G on the y_0-trajectories).*

Proof. Let $F_\epsilon(x,t) = F(x,t) + \epsilon G(x,t)$, and let $T_t^\epsilon(x_0)$ denote the flow map of (B.41) starting at x_0. From (B.30), there is a polynomial $P = P_{i_1, \ldots, i_r}$ such that

$$\frac{d}{dt} \partial_{i_1 \cdots i_r} T_t^x(x_0) = \sum_{i=1}^d \partial_i F_\epsilon(x(t), t) \partial_{i_1 \cdots i_r} T_{t,i}^\epsilon + P(DF_\epsilon, \ldots, D^r F_\epsilon, DT_t^x, \ldots, D^{r-1} T_t^x) \tag{B.42}$$

On the other hand, applying (B.30) to y_0 gives

$$\frac{d}{dt} \partial_{i_1 \cdots i_r} T_t^{y_0}(x_0) = \sum_{i=1}^d \partial_i F(y_0(t), t) \partial_{i_1 \cdots i_r} T_{t,i}^{y_0} + P(DF, \ldots, D^r F, DT_t^{y_0}, \ldots, D^{r-1} T_t^{y_0})$$

We will now show that these two trajectories are $O(\epsilon)$ uniformly close by induction on r. Note that the base case ($r = 0$) is proved in Lemma 189. We will first show that

$$\left\|P(DF_\epsilon, \ldots, D^r F_\epsilon, DT_t^x, \ldots, D^{r-1}T_t^x) - P(DF, \ldots, D^r F, DT_t^{y_0}, \ldots, D^{r-1}T_t^{y_0})\right\| = O(\epsilon)$$

Since P is a fixed polynomial that depends on $i_1, \ldots, i_r$, to show the above, we only need to show that the coordinates are $O(\epsilon)$ close, for small enough ϵ.

$$\begin{aligned}
\left\|D^k F_\epsilon(x(t), t) - D^k F(y_0(t), t)\right\| &\leq \left\|D^k F_\epsilon(x(t), t) - D^k F(x(t), t)\right\| + \left\|D^k F(x(t), t) - D^k F(y_0(t), t)\right\| \\
&\leq \epsilon\left\|D^k G(x(t), t)\right\| + \|x(t) - y_0(t)\|(2N_{k+1} + 1) \\
&\leq O(\epsilon(2M_k + 2N_{k+1} + 2))
\end{aligned}$$

where $N_{k+1} = \sup_{x_0\in\mathcal{C}, 0\leq t\leq t_{\max}} \left\|D^{k+1}F(x, t)|_{x=y_0(t)}\right\|$ and $M_k = \sup_{x_0\in\mathcal{C}, 0\leq t\leq t_{\max}} \left\|D^k G(x, t)|_{x=y_0(t)}\right\|$. The second inequality follows since the base case (Lemma 189) implies that $\|x(t) - y_0(t)\| = O(\epsilon)$, and since $D^{k+1}F$ is continuous, it follows that for small enough ϵ, $\left\|D^{k+1}F|_{(x,t)}\right\| \leq 2N_{k+1} + 1$, for all x such that $\|x - y_0(t)\| = O(\epsilon)$. Similarly, note that for small enough ϵ, $\left\|D^k G(x(t), t)\right\| \leq 2M_k + 1$, since G is C^k. Therefore, $\left\|D^k F_\epsilon(x(t), t) - D^k F(y_0(t), t)\right\| = O(\epsilon)$, where constants in $O(\cdot)$ depend M_k and N_{k+1}.

To simplify notation, let $\alpha(t) = \frac{d}{dt}\partial_{i_1\cdots i_r}(T_t^x - T_t^{y_0})$. Then,

$$\begin{aligned}
\frac{d}{dt}\alpha(t) &= \frac{d}{dt}\partial_{i_1\cdots i_r}(T_t^x - T_t^{y_0}) \\
&= \sum_{i=1}^{d} \partial_i F_\epsilon(x(t), t)\partial_{i_1\cdots i_r} T_{t,i}^x - \sum_{i=1}^{d} \partial_i F(y_0(t), t)\partial_{i_1\cdots i_r} T_{t,i}^{y_0} + O(\epsilon) \\
&= \sum_{i=1}^{d} \partial_i F_\epsilon(x(t), t)\partial_{i_1\cdots i_r}(T_{t,i}^x - T_{t,i}^{y_0}) + \sum_{i=1}^{d}(\partial_i F_\epsilon(x(t), t) - \partial_i F(y_0(t), t))\partial_{i_1\cdots i_r} T_{t,i}^{y_0} + O(\epsilon) \\
&= DF_\epsilon(x(t), t)\partial_{i_1\cdots i_r}(T_t^x - T_t^{y_0}) + (DF_\epsilon(x(t), t) - DF(y_0(t), t))\partial_{i_1\cdots i_r} T_t^x + O(\epsilon) \\
&= DF_\epsilon(x(t), t)\alpha(t) + (DF(x(t), t) - DF(y_0(t), t) + \epsilon G(x(t), t))\partial_{i_1\cdots i_r} T_t^{y_0} + O(\epsilon) \\
\Rightarrow \frac{1}{2}\frac{d}{dt}\|\alpha\|^2 &\leq \|DF_\epsilon(x(t), t)\|\|\alpha\|^2 + O(\epsilon(N_2 + M_0))\|\partial_{i_1\cdots i_r} T_t^{y_0}\| + O(\epsilon) \\
\Rightarrow \frac{d}{dt}\|\alpha\| &\leq \|DF(x(t), t)\|\|\alpha\| + O(\epsilon) \\
&\leq (2N_1 + 1)\|\alpha\| + O(\epsilon)
\end{aligned}$$

Now, Grönwall's inequality (Lemma 192) gives us the bound,

$$\|\alpha(t)\| \leq t_{\max} e^{N_1 t_{\max}} O(\epsilon) = O(\epsilon)$$

The constants in the last $O(\cdot)$ notation depend on $t_{\max}$, N_k for $0 \leq k \leq r + 1$ and M_k for $0 \leq k \leq r$.

This tells us that

$$\|T_t^x - T_t^{y_0}\|_{C^r} = O(\epsilon) \tag{B.43}$$

Now, note that T_t^y satisfies

$$\begin{aligned} &\frac{d}{dt}y(t) = F(y(t),t) + \epsilon G(y(t),t) + \epsilon(G(y_0(t),t) - G(y(t),t)) \\ \implies &\frac{d}{dt}y(t) = F(y(t),t) + \epsilon G(y(t),t) + \epsilon^2 H(y(t),t) \end{aligned}$$

where $H(y,t) = \frac{1}{\epsilon}(G(y_0(t),t) - G(y(t),t))$. Consider the system of ODEs

$$\frac{d}{dt}y(t) = F_\epsilon(y(t),t) + \gamma H(y(t),t) \tag{B.44}$$

Note that when $\gamma = 0$, T_t^x is the flow map for this system, and when $\gamma = \epsilon^2$, T_t^y is the flow map for this system. Therefore, applying (B.43) for the system (B.44), we get

$$\|T_t^x - T_t^y\|_{C^r} = O(\gamma) = O(\epsilon^2)$$

where the constants in $O(\cdot)$ notation depend on $\sup_{0\le k\le r, x_0\in\mathcal{C}, 0\le t\le t_{\max}} \left\|D^{k+1}F_\epsilon(x(t),t)\right\|$ which is bounded by $\max_{0\le k\le r}(2N_{k+1}+1)$ for small ϵ, and $M_k' = \sup_{0\le k\le r, x_0\in\mathcal{C}, 0\le t\le t_{\max}} \left\|D^{k+1}H(x(t),t)\right\|$. Using the definition of H,

$$\begin{aligned} \left\|D^k H(x(t),t)\right\| &= \frac{1}{\epsilon}\left\|D^k G(y_0(t),t) - D^k G(x(t),t)\right\| \\ &\le \frac{1}{\epsilon}\|y_0(t) - x(t)\|(2M_{k+1}+1) \\ &= \frac{1}{\epsilon}\cdot O(\varepsilon)\cdot(2M_{k+1}+1) = O(1) \end{aligned}$$

where the constant in the $O(\cdot)$ depends on $M_0,\dots,M_{r+1}$ and $N_1,\dots,N_{r+1}$. This proves the dependence in $O(\cdot)$ notation as stated in the statement, completing the proof. □

Corollary 191. *Consider the ODE*

$$\dot{x} = Ax + \epsilon g(x,t)$$

such that $\|A\| = 1$ and g has bounded $(r+1)^{th}$ derivatives on a compact set $\mathcal{C}$. Let T^x be the flow map corresponding to this ODE. For fixed x_0, let y_0, y_1 be functions satisfying

$$\begin{aligned} \dot{y_0} &= Ay_0 \\ \dot{y_1} &= Ay_1 + g(y_0(t),t) \end{aligned}$$

such that $y_0(0) = x_0$ and $y_1(0) = 0$. Consider the flow map $T^y : \mathbb{R}\times\mathbb{R}^n$ such that $T^y(t,x_0) = y_0(t) + \epsilon y_1(t)$. Then, the maps T_t^x and T_t^y are $O(\epsilon^2)$ uniformly close over $\mathcal{C}$ in C^r topology, for all $t\in[0,2\pi]$. The constants in the $O(\cdot)$ depend on $\|A\|$ and the first r derivatives of g on the trajectories $x(t) = e^{At}x_0, x_0\in\mathcal{C}$.

This follows directly from Lemma 190, after noting $\dot{y} = Ay_0 + \varepsilon Ay_1 + \varepsilon g(y_0(t),t) = Ay + \varepsilon g(y_0(t),t)$. Note that $F(x) = Ax$ is a linear function, so derivatives of F are bounded, and the y_0 trajectories can be computed easily.

B.5.6 Grönwall lemma

The following lemma is very useful for bounding the growth of solutions, or errors from perturbations to ODE's.

Lemma 192 (Grönwall)**.** *If $x(t)$ is differentiable on $t \in [0, t_{\max}]$ and satisfies the differential inequality*

$$\frac{d}{dt}x(t) \le ax(t) + b,$$

then

$$x(t) \le (bt + x(0))e^{at}$$

for all $t \in [0, t_{\max}]$.

Appendix C

Implementation details for Algorithm 5

We justify here that our algorithm solves the robust subspace approximation problem with sublinear space in the general turnstile streaming model within the claimed time bounds.

The only times the input matrix A is involved in computation directly is during left matrix multiplications by Sparse Cauchy matrices (TA, C_1A, C_2A), and in the computation of H_iA from the Sampler Algorithm (Alg 8). All of these are oblivious linear sketches, and thus can be performed online with low space in input sparsity time.

We note that we only make use of limited independence Cauchy variables for the proofs in this paper. Thus we can store each matrix and perform multiplication with each stream update in sublinear space by storing just the random seed for each matrix (see Section J of [SWZ16] for a full description). The Sampler Algorithm was originally a streaming algorithm, and we only keep $\log d \operatorname{poly}(k/\epsilon)$ copies in parallel over the course of the entire algorithm.

The algorithm performs BootstrapCoreset twice: once with TA and once with V^TU^T as input. Note that we cannot compute the projection $A(\mathrm{Id} - TA)$ or $A(\mathrm{Id} - V^TU^T)$ until the after the stream is finished. Fortunately, since H is oblivious, we can right multiply HA by $(\mathrm{Id} - P)$ once P is available, and only then perform the sampling procedure $\mathcal{P}$ from Extract (Alg. 9).

Except for the very last step involving the algorithm of [BPR94], all other steps in the algorithm are standard matrix operations on matrices of small size.

Appendix D

Miscellaneous Technical Tools

D.1 Properties of Poisson Distribution

Let $\mathcal{X}$ be a measure space with measure μ. We consider a poisson point process Φ with parameter λ over $\mathcal{X}$, to be a point process such that for any $B \subseteq \mathcal{X}$ of finite measure,

$$\mathbb{P}[\Phi(B) = n] = \frac{\Lambda(B)^n e^{-\Lambda(B)}}{n!} \tag{D.1}$$

where $\Lambda(B) = \lambda V_\mu(B)$ and $\Phi(B)$ denote the number of points of Φ contained in B. These point processes satisfy the following property:

Proposition 193. *For a poisson process Φ and two fixed disjoint sets B_1, B_2, the random variables $\Phi(B_1)$ and $\Phi(B_2)$ are independent.*

For a complete formal treatment of poisson point processes, see . A simple computation shows that $\mathbb{E}[\Phi(B)] = \lambda V_\mu(B)$ for any set $B \subseteq \mathcal{X}$ of finite measure. Equivalently, we shall also say that $\Phi(B)$ is given by the measure $\lambda\mu$. We define $\Phi(B_1, \dots, B_K)$ to denote the number of tuples of distinct points $(x_1, \dots, x_k) \in \Phi$ such that $x_i \in B_i$. We now claim that $\mathbb{E}[\Phi(B_1, \dots, B_k)] = \lambda^k V_\mu(B_1) \cdots V_\mu(B_k)$.

Lemma 194. *Let $B_1, \dots, B_k \subseteq \mathcal{X}$ be of finite measure such that $V_\mu(B_i \cap B_j) = 0$. Then*

$$\mathbb{E}[\Phi(B_1, \dots, B_k)] = \lambda^k V_\mu(B_1) \cdots V_\mu(B_k).$$

Proof. First, observe that we can construct B_i' which are pairwise disjoint, such that $B_i' = B_i \setminus X_i$ with $V_\mu(X_i) = 0$. Then $\Phi(B_1, \dots, B_k) = \Phi(B_1', \dots, B_k')$ almost surely, and it suffices to show the result on $B_1', \dots, B_k'$. Therefore, we may assume that B_i are pairwise disjoint.

Since $B_1, \dots, B_k$ are pairwise disjoint, $\Phi(B_1, \dots, B_k) = \Phi(B_1) \cdots \Phi(B_k)$. Further, $\Phi(B_i)$ are independent due to Proposition 193. Therefore,

$$\begin{aligned} \mathbb{E}[\Phi(B_1, \dots, B_k)] &= \mathbb{E}[\Phi(B_1) \cdots \Phi(B_k)] \\ &= \mathbb{E}[\Phi(B_1)] \cdots \mathbb{E}[\Phi(B_k)] \\ &= \lambda^k V_\mu(B_1) \cdots V_\mu(B_k) \end{aligned}$$

which completes the proof. $\square$

Lemma 195. *For any $B \subset \mathcal{X}$ of finite measure, let $B_1 = \ldots = B_k = B$. Then*

$$\mathbb{E}[\Phi(B_1, \ldots, B_k)] = \lambda^k V_\mu(B)^k = \lambda^k V_\mu(B_1) \cdots V_\mu(B_k).$$

Proof. Note that if $\Phi(B) = n \geq k$, then $\Phi(B_1, \ldots, B_k) = \frac{n!}{(n-k)!}$, and otherwise $\Phi(B) = 0$. Therefore,

$$\begin{aligned}
\mathbb{E}[\Phi(B_1, \ldots, B_k)] &= \sum_{n=k}^{\infty} \mathbb{P}[\Phi(B) = n] \cdot \frac{n!}{(n-k)!} \\
&= \sum_{n=k}^{\infty} \frac{\Lambda(B)^n e^{-\Lambda(B)}}{n!} \cdot \frac{n!}{(n-k)!} && \text{from eq. (D.1)} \\
&= \Lambda(B)^k e^{-\Lambda(B)} \cdot \sum_{n=k}^{\infty} \frac{\Lambda(B)^{n-k}}{(n-k)!} \\
&= \Lambda(B)^k e^{-\Lambda(B)} \cdot e^{\Lambda(B)} = \Lambda(B)^k
\end{aligned}$$

Since $\Lambda(B) = \lambda V_\mu(B)$, we get the required result. ∎

By using Lemmas 194 and 195, it follows that

Lemma 196. *Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be finite measure subsets such that for all i, j, either $V_\mu(B_i \cap B_j) = 0$ or $B_i = B_j$. Then*

$$\mathbb{E}[\Phi(B_1, \ldots, B_k)] = \lambda^k V_\mu(B_1) \cdots V_\mu(B_k).$$

The proof is essentially the same as Lemma 194, we just group the dependent sets together and apply Lemma 195 to compute the expectation on these sets rather than breaking it up into different parts. Now, we prove the main claim:

Lemma 197. *Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be of finite measure. Then*

$$\mathbb{E}[\Phi(B_1, \ldots, B_k)] = \lambda^k V_\mu(B_1) \cdots V_\mu(B_k).$$

Proof. For each binary string $S \neq 0$ (where 0 indicates all zero binary string) of size k, define $B_S = \bigcap_{i=1}^k C_i$ where $C_i = B_i$ if $S_i = 1$ and $C_i = \bar{B}_i$ otherwise. Let $\mathcal{S} = \left\{S \in 2^k, S \neq 0\right\}$. For each i, define $\mathcal{S}_i = \left\{S \in 2^k : S_i = 1\right\}$. Then we have $B_i = \bigcup_{S \in \mathcal{S}_i} B_S$. Therefore, by linearity of expectation, we know that

$$\mathbb{E}[\Phi(B_1, \ldots, B_k)] = \sum_{S_i \in \mathcal{S}_i} \mathbb{E}[\Phi(B_{S_1}, \ldots, B_{S_k})]$$

Further, for any $S_1, S_2 \in \mathcal{S}$, either $V_\mu(B_{S_1} \cap B_{S_2}) = 0$ or $S_1 = S_2$. Therefore, by Lemma 196,

$$\mathbb{E}[\Phi(B_1, \ldots, B_k)] = \sum_{S_i \in \mathcal{S}_i} \mathbb{E}[\Phi(B_{S_1}, \ldots, B_{S_k})] = \sum_{S_i \in \mathcal{S}_i} \lambda^k V_\mu(B_{S_1}) \cdots V_\mu(B_{S_k})$$

Since we are looking at all possible such sums, we have

$$\lambda^k \sum_{S_i \in \mathcal{S}_i} \prod_{i=1}^k V_\mu(B_{S_i}) = \lambda^k \prod_{i=1}^k \left(\sum_{S_i \in \mathcal{S}_i} V_\mu(B_{S_i}) \right) = \lambda^k \prod_{i=1}^k V_\mu(B_i)$$

Combining the two equations, we have the required result. ∎

Now, consider any set $B \subseteq \mathcal{X}^k$ that is measurable with respect to μ^k. Define $\Phi(B)$ to be the expected number of tuples $(x_1, \ldots, x_k)$, such that x_i are distinct, and the vector $(x_1, \ldots, x_k) \in B$. Then we claim that Φ defines a measure on $\mathcal{X}^k$, given by the measure $\lambda^k \mu^k$. Note that Φ is a measure by linearity of expectation. Hence, it suffices to show that Φ agrees with $\lambda^k \mu^k$ on set of generators of μ^k. Since the family of sets $B_1 \times \cdots \times B_k$ where $B_i \subseteq \mathcal{X}$ is μ measurable forms a basis for the measurable sets of μ^k, we can see that $\lambda^k \mu^k$ and Φ agree due to Lemma 197, which proves the following:

Theorem 198. *Let μ be a measure on $\mathcal{X}$. Let Φ be a poisson process with parameter λ. For any k, and for any $B \subseteq \mathcal{X}^k$ measurable with respect to μ^k, let $\Phi(B)$ denote the number of tuples $(x_1, \ldots, x_n) \in \Phi$ of distinct points such that $(x_1, \ldots, x_k) \in B$. Then $\mathbb{E}[\Phi(B)]$ is given by the measure $\lambda^k \mu^k$. In other words,*

$$\mathbb{E}[\Phi(B)] = \int_B \lambda^k \mu^k = \lambda^k V_\mu(B)$$

D.2 Bounds on Binomial Coefficients

We first recall some exponential bounds on $1 + x$. We have the standard upper bound:

$$e^x \geq 1 + x \qquad \forall x \in \mathbb{R} \tag{D.2}$$

On the other hand, we have the lower bound:

$$e^{\frac{x}{1+x}} \leq 1 + x \leq e^x \qquad \forall\, x > -1 \tag{D.3}$$

This follows since

$$1 - t \leq e^{-t} \implies 1 - \frac{x}{1+x} \leq e^{-\frac{x}{1+x}} \implies \frac{1}{1+x} \leq e^{-\frac{x}{1+x}}$$

We get Equation (D.3) from this by taking reciprocals whenever $\frac{1}{1+x} \geq 0$. Further, Equation (D.3) implies that

$$e^{\frac{x}{2}} \leq 1 + x \leq e^x \qquad \forall\, 0 \leq x \leq 1 \tag{D.4}$$

We also recall the Sterling's Approximation - the non-asymptotic version of Sterling's Approximation is given in Robbins [Rob55] as

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n+1}\right) \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n}\right) \tag{D.5}$$

We can use these exponential bounds on $(1+x)$ to bound the binomial coefficients. In particular, we are interested in bounded the binomial coefficient $\binom{n+x}{k}$ in the case where $x, k \leq \frac{n}{10}$. Recall by the definition of binomial coefficients:

$$\binom{n+x}{k} = \frac{1}{k!} \prod_{i=0}^{k-1} (n + x - i) = \frac{n^k}{k!} \prod_{i=0}^{k-1} \left(1 + \frac{x-i}{n}\right)$$

Using Equation (D.2) we get the following upper bound:

$$\binom{n+x}{k} \leq \frac{n^k}{k!} \exp\left(\sum_{i=0}^{k-1} \frac{x-i}{n}\right)$$
$$\leq \frac{n^k}{k!} \exp\left(\frac{2kx - k^2 + k}{2n}\right)$$

Using Equation (D.4) we get the following lower bound when $n + x - k \geq |x|, k$:

$$\begin{aligned}
\binom{n+x}{k} &\geq \frac{n^k}{k!} \exp\left(\sum_{i=0}^{k-1} \frac{\frac{x-i}{n}}{1 + \frac{x-i}{n}}\right) \\
&= \frac{n^k}{k!} \exp\left(\sum_{i=0}^{k-1} \frac{x-i}{n+x-i}\right) \\
&= \frac{n^k}{k!} \exp\left(\sum_{i=0}^{k-1} \frac{x-i}{n} + \frac{x-i}{n+x-i} - \frac{x-i}{n}\right) \\
&= \frac{n^k}{k!} \exp\left(\sum_{i=0}^{k-1} \frac{x-i}{n} - \frac{(x-i)^2}{n(n+x-i)}\right) \\
&\geq \frac{n^k}{k!} \exp\left(\frac{2kx - k^2 + k}{2n} - \frac{2k(|x| + k)}{n}\right)
\end{aligned}$$

Where the last inequality follows since $(x-i)^2 \leq 2x^2 + 2i^2 \leq 2x^2 + 2k^2 \leq 2(|x| + k)(n + x - i)$ assuming that $n + x - i \geq |x|, k$. Together, we get the following upper and lower bounds on the binomial coefficients:

$$\frac{n^k}{k!} \exp\left(\frac{2kx - k^2 + k}{2n} - \frac{2k|x| + 2k^2}{n}\right) \leq \binom{n+x}{k} \leq \frac{n^k}{k!} \exp\left(\frac{2kx - k^2 + k}{n}\right) \tag{D.6}$$

D.3 Bounding the matrix integral in Equation 5.6

We prove a variant of the Cauchy-Schwarz inequality that gives us a handle on norms of matrix integrals.

Lemma 199. *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $A : \mathbb{R}^d \to \mathbb{R}^n$ be integrable functions, with $M = \int_x f(x)A(x)dx$. Then we have*

$$\|M\|_2^2 = \left\| \int_x f(x)A(x)ds \right\|_2^2 \leq \left(\int_x |f(x)|^2 dx \right) \left(\int_x \|A(x)\|_2^2 dx \right). \tag{D.7}$$

Similarly, if $A : \mathbb{R}^d \to \mathbb{R}^{n \times n}$ is a matrix valued function then

$$\|M\|_F^2 = \left\| \int_x f(x)A(x) \right\|_F^2 \leq \left(\int_x |f(x)|^2 dx \right) \left(\int_x \|A(x)\|_F^2 dx \right). \tag{D.8}$$

Proof. The proof follows from the Cauchy-Schwarz inequality. Since we integrate component-wise, for eq. (D.7) we have that

$$M_i^2 = \left(\int_x f(x)A(x)_i dx\right)^2 \leq \left(\int_x f(x)^2 dx\right)\left(\int_x A(x)_i^2 dx\right).$$

Summing over i, we get the result. The matrix variant eq. (D.8) follows by looking at the matrix M as a vector in $\mathbb{R}^{n^2}$. $\square$

D.4 Proof of Lemma 79

We restate the lemma for convenience:

Lemma 79. *Let $d > 0$ be sufficiently large. Let $p = \hat{p}^d$ and $q = \hat{q}^d$ be any product distributions, and define $R(x) = \frac{q(x)}{p(x)}$. Suppose we have the following third moment bound: $\mathbb{E}_{x\sim\hat{p}}\left[\left(\log\frac{\hat{q}}{\hat{p}}\right)^3\right] < \infty$. Then, for any ϵ, there exist constants $\alpha = \alpha(\hat{p},\hat{q},\epsilon)$, $\mu = \mu(\hat{p},\hat{q},\epsilon) < 0$ such that*

$$\mathbb{P}_{x\sim p}\left[R(x) \leq \exp\left(\mu d - \alpha\sqrt{d}\right)\right] \geq \frac{1}{2} - \epsilon \text{ and } \mathbb{P}_{x\sim p}\left[R(x) \geq \exp\left(\mu d + \alpha\sqrt{d}\right)\right] \geq \frac{1}{2} - \epsilon.$$

Proof. We will analyze the behaviour of $R(x)$ using the Berry-Esseen theorem. Given that $p_* = \hat{p}^d$ and $q = \hat{q}^d$ are product distributions, let $r(x)$ be the random variable defined by $r(x) = \frac{\hat{q}(x)}{\hat{p}(x)}$, $x \sim \hat{p}$. Let $y_i(x) = \log r(x)$ for $1 \leq i \leq d$ be d independent copies of the random variable $r(x)$. Let $\mathbb{E}[y_i] = \mu_r$, $\mathbb{E}\left[\|y_i - \mu_r\|^2\right] = \sigma_r^2$ and $\mathbb{E}\left[\|y_i - \mu_r\|^3\right] = \gamma_r$, all of which are well defined by the hypothesis of the lemma. Let $Y = \sum_{i=1}^d y_i$, and Z be the standard Gaussian in $\mathbb{R}$. Then, by the Berry-Esseen Theorem [Dur19, Theorem 3.4.17],

$$\mathbb{P}\left[\frac{Y - \mu_r d}{\sigma_r\sqrt{d}} \leq -c\right] \geq \mathbb{P}[Z \leq -c] - \frac{C_{\mathrm{BE}}\cdot\gamma_r}{\sigma_r^3\sqrt{d}},$$

where $C_{\mathrm{BE}} < 1$ [Bee72] is an absolute constant. We can now choose $c = c(\epsilon)$ such that $\mathbb{P}[Z \leq c] \geq \frac{1-\epsilon}{2}$. Further, we can choose d large enough so that $\frac{C_{\mathrm{BE}}\cdot\gamma}{\sigma^3\sqrt{d}} \leq \frac{\epsilon}{2}$. Then for $\mu = \mu_r$ and $\alpha = c\sigma_r$, we have

$$\mathbb{P}_{x\sim p}\left[R(x) \leq \exp\left(\mu d - \alpha\sqrt{d}\right)\right] \geq \frac{1}{2} - \epsilon.$$

Since Z is symmetric around 0, Berry-Esseen gives us the other inequality for the same choice of μ and α,

$$\mathbb{P}\left[\frac{Y - \mu_r d}{\sigma_r\sqrt{d}} \geq c\right] \geq \mathbb{P}[Z \geq c] - \frac{C_{\mathrm{BE}}\cdot\gamma_r}{\sigma_r^3\sqrt{d}} \geq \frac{1}{2} - \epsilon.$$

Note that the constants μ and α are independent of d. Further, note that $\mu = \mu_r = -\mathrm{KL}(\hat{p}||\hat{q}) < 0$. $\square$

D.5 Invertibility of the Hessian

We prove that the Hessian of NCE loss for the exponential family given by $T(x) = (x_1^4, \ldots, x_d^4, 1)$ is invertible. In particular, we have the following lemma:

Lemma 200. *Let $Q = \mathcal{N}(0, I_d)$ be the standard Gaussian in $\mathbb{R}^d$. Let $\hat{P}$ be the log concave distribution defined in definition 77. Let $P = \hat{P}^d$. Let q and p denote the density functions of Q and P respectively. Observe that P is in the exponential family given by $T(x) = (x_1^4, \ldots, x_d^4, 1)$, and equals P_{θ_*} for some θ_*. Then the hessian of the NCE loss with respect to distribution P and noise Q given by*

$$H = \nabla_\theta^2 L(\theta_*) = \frac{1}{2}\int_x \frac{p_* q}{p_* + q} T(x)T(x)^\top$$

is invertible.

Proof. For any subset $A \subseteq \mathbb{R}^d$, define

$$H_A = \frac{1}{2}\int_{x \in A} \frac{p_* q}{p_* + q} T(x)T(x)^\top.$$

Observe that the density functions p_* and q of P_* and Q respectively are strictly positive over all of $\mathbb{R}^d$. Therefore, for any subset $A \subseteq \mathbb{R}^d$ and any $v \in \mathbb{R}^{d+1}$, we have

$$v^\top H v \geq \frac{1}{2}\int_{x \in A} \frac{p_* q}{p_* + q} v^\top T(x)T(x)^\top v = v^\top H_A v.$$

Given a vector $v \in \mathbb{R}^{d+1}$, we will pick A such that $\left|T(x)^\top v\right| > 0$ for all $x \in A$. Note that the set $\mathcal{B} = \{e_1 + e_{d+1}, \ldots, e_d + e_{d+1}, e_{d+1}\}$ is a basis. Therefore, if $b^\top v = 0$ for all $b \in \mathcal{B}$, then $v = 0$. Hence, there exists some $x \in \{e_1, \ldots, e_d\}$ such that $\left|T(x)^\top v\right| > 0$. Since $x \mapsto T(x)^\top v$ is a continuous function, we can find an open set A around x such that

$$\left|T(y)^\top v\right| > 0, \qquad \forall y \in A.$$

It follows that

$$v^\top H_A v = \frac{1}{2}\int_{x \in A} \frac{p_* q}{p_* + q} v^\top T(x)T(x)^\top v = \frac{1}{2}\int_{x \in A} \frac{p_* q}{p_* + q} \left|T(x)^\top v\right|^2 > 0.$$

Let $B = \mathbb{R}^d \setminus A$. Since $v^\top H_A v > 0$ and $v^\top H_B v \geq 0$, we have that $v^\top H v > 0$. Since this holds for any arbitrary non-zero vector v, the matrix H must be full rank. Since H is an integral of PSD matrices, it is a full rank PSD matrix and hence invertible. $\square$

D.6 Tail bounds for Equation 5.17

We prove that some $T_{\text{up}} = O(\sigma^2\sqrt{d})$ suffices to obtain the bounds in eq. (5.17). Concretely, we prove tail bounds for $\|T(x)\|$ using tail bounds for P_* and Q. We will use Lemma 1 from [LM00] which proves a bound for χ^2 distributions:

Lemma (Lemma 1, [LM00])**.** *If X is a χ^2 random variable with d degrees of freedom, then for any positive t,*

$$\mathbb{P}\Big[X - d \geq 2\sqrt{td} + 2t\Big] \leq \exp(-t).$$

Then, for $x \sim Q$, $\|x\|^2$ is a χ^2 random variable with d degrees of freedom. Observe that for $t, d \geq 4$, we have $d + 2t + 2\sqrt{td} \leq 2td$. In particular, we have the weaker bound

$$\mathbb{P}_{x\sim Q}\Big[\|x\|^2 \geq 2dt^2\Big] \leq \exp\big(-t^2\big),$$

implying that

$$\mathbb{P}_{x\sim Q}\Big[\|x\| \geq t\Big] \leq \exp\bigg(-\frac{t^2}{2d}\bigg).$$

Further, if $\|x\| \geq \sigma^2\sqrt{d}$, $q(x) \geq p_*(x)$, implying that for $t \geq \sigma^2\sqrt{d}$

$$\mathbb{P}_{x\sim P_*}\Big[\|x\| \geq t\Big] \leq \exp\bigg(-\frac{t^2}{2d}\bigg).$$

In particular, for any δ such that $\log(1/\delta) \geq \sigma^4$, we have

$$\mathbb{P}_{x\sim Q}\Big[\|x\| \geq \sqrt{2d\log(1/\delta)}\Big] \leq \delta \qquad \text{and} \qquad \mathbb{P}_{x\sim P_*}\Big[\|x\| \geq \sqrt{2d\log(1/\delta)}\Big] \leq \delta. \tag{D.9}$$

References

[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[AG11] Uri M Ascher and Chen Greif. *A first course on numerical methods*. SIAM, 2011.

[AM08] Ralph Abraham and Jerrold E Marsden. *Foundations of mechanics*. 364. American Mathematical Soc., 2008.

[App+06] David L Applegate et al. *The traveling salesman problem: a computational study*. Princeton university press, 2006.

[Aro96] Sanjeev Arora. "Polynomial time approximation schemes for Euclidean TSP and other geometric problems". In: *Proceedings of 37th Conference on Foundations of Computer Science*. IEEE. 1996, pp. 2–11.

[Aut+19] Eric Autrey et al. "Metropolized forest recombination for monte carlo sampling of graph partitions". In: *arXiv preprint arXiv:1911.01503* (2019).

[Aut+21] Eric A Autry et al. "Metropolized multiscale forest recombination for redistricting". In: *Multiscale Modeling & Simulation* 19.4 (2021), pp. 1885–1914.

[Bac+16] Arturs Backurs et al. "Nearly-optimal bounds for sparse recovery in generic norms, with applications to k-median sketching". In: *SODA*. 2016.

[Bar+19] Alessandro Barp et al. "Minimum stein discrepancy estimators". In: *Advances in Neural Information Processing Systems* 32 (2019).

[BÉ85] Dominique Bakry and Michel Émery. "Diffusions hypercontractives". In: *Séminaire de Probabilités XIX 1983/84*. Springer, 1985, pp. 177–206.

[Bee72] Paul van Beek. "An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 23.3 (1972), pp. 187–196.

[Beh+18] Jens Behrmann et al. "Invertible Residual Networks". In: (Nov. 2, 2018). arXiv: 1811.00995v3 [cs.LG]. URL: http://arxiv.org/abs/1811.00995v3 (visited on 10/25/2021).

[Bes77] Julian Besag. "Efficiency of pseudolikelihood estimation for simple Gaussian fields". In: *Biometrika* (1977), pp. 616–618.

[BGL13] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Springer Science & Business Media, 2013.

[BGS14] Guy Bresler, David Gamarnik, and Devavrat Shah. "Structure learning of antiferromagnetic Ising models". In: *Advances in Neural Information Processing Systems* 27 (2014).

[BHH59] Jillian Beardwood, John H Halton, and John Michael Hammersley. "The shortest path through many points". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 55 Issue 4. Cambridge University Press. 1959, pp. 299–327.

[BL02] Herm Jan Brascamp and Elliott H Lieb. "On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation". In: *Inequalities*. Springer, 2002, pp. 441–464.

[BPR94] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. "On the Combinatorial and Algebraic Complexity of Quantifier Elimination". In: *J. ACM*. 1994.

[Bre15] Guy Bresler. "Efficiently learning Ising models on arbitrary graphs". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 771–782.

[Bro86] Lawrence D Brown. "Fundamentals of statistical exponential families: with applications in statistical decision theory". In: Ims. 1986.

[BV90] Dimitris J Bertsimas and Garrett Van Ryzin. "An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability". In: *Operations Research Letters* 9.4 (1990), pp. 223–231.

[Car04] Robert Carr. "Separation Algorithms for Classes of STSP Inequalities Arising from a New STSP Relaxation". In: *Mathematics of Operations Research* 29.1 (Feb. 2004), pp. 80–91. DOI: 10.1287/moor.1030.0058. URL: https://doi.org/10.1287%2Fmoor.1030.0058.

[Car97] Robert Carr. "Separating Clique Trees and Bipartition Inequalities Having a Fixed Number of Handles and Teeth in Polynomial Time". In: *Mathematics of Operations Research* 22.2 (1997), pp. 257–265. ISSN: 0364765X, 15265471. URL: http://www.jstor.org/stable/3690263 (visited on 12/21/2022).

[CGH22] Omar Chehab, Alexandre Gramfort, and Aapo Hyvärinen. "The optimal noise in noise-contrastive learning is not what you think". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 307–316.

[Che+18] Ricky T. Q. Chen et al. "Neural Ordinary Differential Equations". In: (June 19, 2018). arXiv: 1806.07366v5 [cs.LG]. URL: http://arxiv.org/abs/1806.07366v5 (visited on 10/25/2021).

[Che+22] Sitan Chen et al. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: (Sept. 22, 2022). arXiv: 2209.11215v3 [cs.LG]. URL: http://arxiv.org/abs/2209.11215v3 (visited on 07/21/2023).

[Chi06] Carmen Chicone. *Ordinary differential equations with applications*. New York Berlin: Springer, 2006. ISBN: 9780387307695.

[Coo71] Stephen A Cook. “The complexity of theorem-proving procedures”. In: *Proceedings of the third annual ACM symposium on Theory of computing*. 1971, pp. 151–158.

[CW09] Kenneth L. Clarkson and David P. Woodruff. “Numerical linear algebra in the streaming model”. In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*. 2009, pp. 205–214.

[CW15a] Kenneth L. Clarkson and David P. Woodruff. “Input Sparsity and Hardness for Robust Subspace Approximation”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (2015), pp. 310–329.

[CW15b] Kenneth L. Clarkson and David P. Woodruff. “Sketching for M-Estimators: A Unified Approach to Robust Regression”. In: *SODA*. 2015.

[Dag+21] Yuval Dagan et al. “Learning Ising models from one or multiple samples”. In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 161–168.

[Dar35] Georges Darmois. “Sur les lois de probabilitéa estimation exhaustive”. In: *CR Acad. Sci. Paris* 260.1265 (1935), p. 85.

[DDS19] Daryl DeFord, Moon Duchin, and Justin Solomon. “Recombination: A family of Markov chains for redistricting”. In: *arXiv preprint arXiv:1911.05725* (2019).

[DDT19] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. “Augmented neural odes”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3134–3144.

[Din+06] Chris H. Q. Ding et al. “R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization”. In: *ICML*. 2006.

[DKB14] Laurent Dinh, David Krueger, and Yoshua Bengio. “Nice: Non-linear independent components estimation”. In: *arXiv preprint arXiv:1410.8516* (2014).

[DM19] Yilun Du and Igor Mordatch. “Implicit Generation and Generalization in Energy-Based Models”. In: *arXiv preprint arXiv:1903.08689* (Mar. 2019). arXiv: `1903.08689v6 [cs.LG]`. URL: http://arxiv.org/abs/1903.08689v6.

[DSB16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (2016).

[DTV11] Amit Deshpande, Madhur Tulsiani, and Nisheeth K. Vishnoi. “Algorithms and Hardness for Subspace Approximation”. In: *SODA*. 2011.

[Dur19] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[DV07a] Amit Deshpande and Kasturi R. Varadarajan. “Sampling-based dimension reduction for subspace approximation”. In: *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*. 2007, pp. 641–650.

[DV07b] Amit Deshpande and Kasturi R. Varadarajan. “Sampling-based dimension reduction for subspace approximation”. In: *STOC*. 2007.

[Fel+10a] Dan Feldman et al. “Coresets and Sketches for High Dimensional Subspace Approximation Problems”. In: *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*. 2010, pp. 630–649.

[Fel+10b] Dan Feldman et al. “Coresets and Sketches for High Dimensional Subspace Approximation Problems”. In: *SODA*. 2010.

[FL11] Dan Feldman and Michael Langberg. “A unified framework for approximating and clustering data”. In: *STOC*. 2011.

[FL15] Peter GM Forbes and Steffen Lauritzen. “Linear estimating equations for exponential families with application to Gaussian linear concentration models”. In: *Linear Algebra and its Applications* 473 (2015), pp. 261–283.

[FP15] Alan Frieze and Wesley Pegden. “Separating subadditive Euclidean functionals”. In: *Random Structures and Algorithms* 51 (Jan. 2015). arXiv: 1501.01944v4 [math.PR]. URL: http://arxiv.org/abs/1501.01944v4.

[FR74] Dominique Foata and John Riordan. *Mappings of acyclic and parking functions*. en. Feb. 1974. DOI: 10.1007/bf01834776. URL: http://dx.doi.org/10.1007/BF01834776.

[Gao+20] Ruiqi Gao et al. “Flow contrastive estimation of energy-based models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7518–7528.

[GB91] Michel X Goemans and Dimitris J Bertsimas. “Probabilistic analysis of the Held and Karp lower bound for the Euclidean traveling salesman problem”. In: *Mathematics of operations research* 16.1 (1991), pp. 72–89.

[GH10] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 297–304.

[GH12] Michael U Gutmann and Aapo Hyvärinen. “Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics.” In: *Journal of machine learning research* 13.2 (2012).

[Gha+16] Mina Ghashami et al. “Frequent Directions: Simple and Deterministic Matrix Sketching”. In: *SIAM J. Comput.* 45.5 (2016), pp. 1762–1792.

[GP86] Martin Grötschel and William R Pulleyblank. “Clique tree inequalities and the symmetric travelling salesman problem”. In: *Mathematics of operations research* 11.4 (1986), pp. 537–569.

[Gra+18] Will Grathwohl et al. “Ffjord: Free-form continuous dynamics for scalable reversible generative models”. In: *arXiv preprint arXiv:1810.01367* (2018).

[Gur+10] Venkatesan Guruswami et al. “Bypassing UGC from some optimal geometric inapproximability results”. In: *ACM Trans. Algorithms* 12 (2010), 6:1–6:25.

[HDC20] Chin-Wei Huang, Laurent Dinh, and Aaron Courville. "Augmented Normalizing Flows: Bridging the Gap Between Generative Flows and Latent Variable Models". In: (Feb. 17, 2020). arXiv: 2002.07101v1 [cs.LG]. URL: http://arxiv.org/abs/2002.07101v1 (visited on 10/25/2021).

[Hén76] Michel Hénon. "A two-dimensional mapping with a strange attractor". In: *The Theory of Chaotic Attractors*. Springer, 1976, pp. 94–102.

[HK71] Michael Held and Richard M Karp. "The traveling-salesman problem and minimum spanning trees: Part II". In: *Mathematical programming* 1.1 (1971), pp. 6–25.

[Hyv05] Aapo Hyvärinen. "Estimation of non-normalized statistical models by score matching." In: *Journal of Machine Learning Research* 6.4 (2005).

[Hyv07] Aapo Hyvärinen. "Some extensions of score matching". In: *Computational statistics & data analysis* 51.5 (2007), pp. 2499–2512.

[Ind01] P. Indyk. "Algorithmic Applications of Low-Distortion Geometric Embeddings". In: *Proceedings of the 42Nd IEEE Symposium on Foundations of Computer Science*. FOCS '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 10–. ISBN: 0-7695-1390-5. URL: http://dl.acm.org/citation.cfm?id=874063.875596.

[JSY19] Priyank Jaini, Kira A. Selby, and Yaoliang Yu. *Sum-of-Squares Polynomial Flow*. 2019. arXiv: 1905.02325 [cs.LG].

[Kar77] Richard M Karp. "Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane". In: *Mathematics of operations research* 2.3 (1977), pp. 209–224.

[KD18] Durk P Kingma and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10215–10224.

[KHR22] Frederic Koehler, Alexander Heckett, and Andrej Risteski. "Statistical Efficiency of Score Matching: The View from Isoperimetry". In: *arXiv preprint arXiv:2210.00726* (2022).

[KMR20] Frederic Koehler, Viraj Mehta, and Andrej Risteski. *Representational aspects of depth and conditioning in normalizing flows*. 2020. arXiv: 2010.01155 [cs.LG].

[Koe98] Wolfram Koepf. "Hypergeometric summation". In: *Vieweg, Braunschweig/Wiesbaden* 5.6 (1998).

[Koo36] Bernard Osgood Koopman. "On distributions admitting a sufficient statistic". In: *Transactions of the American Mathematical society* 39.3 (1936), pp. 399–409.

[KVW14] Ravi Kannan, Santosh Vempala, and David P. Woodruff. "Principal Component Analysis and Higher Correlations for Distributed Data". In: *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*. 2014, pp. 1040–1057.

[KW66] Alan G. Konheim and Benjamin Weiss. "An Occupancy Discipline and Applications". In: *SIAM Journal on Applied Mathematics* 14.6 (1966), pp. 1266–1274. ISSN: 00361399. URL: http://www.jstor.org/stable/2946240.

[Law85] Eugene Lawler. *The Traveling salesman problem : a guided tour of combinatorial optimization*. Chichester West Sussex New York: Wiley, 1985. ISBN: 0471904139.

[Liu+21] Bingbin Liu et al. "Analyzing and Improving the Optimization Landscape of Noise-Contrastive Estimation". In: *arXiv preprint arXiv:2110.11271* (Oct. 2021). arXiv: 2110.11271v1 [cs.LG]. URL: http://arxiv.org/abs/2110.11271v1.

[LM00] Béatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *The Annals of Statistics* 28.5 (Oct. 2000). DOI: 10.1214/aos/1015957395. URL: https://doi.org/10.1214%5C%2Faos%5C%2F1015957395.

[Lov79] László Miklós Lovász. "Combinatorial problems and exercises". In: 1979.

[LPW06] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006. URL: http://scholar.google.com/scholar.bib?q=info:3wf9IU94tyMJ:scholar.google.com/&output=citation&hl=en&as_sdt=2000&ct=citation&cd=0.

[Ma+19] Yi-An Ma et al. "Is There an Analog of Nesterov Acceleration for MCMC?" In: (Feb. 4, 2019). arXiv: 1902.00996v2 [stat.ML]. URL: http://arxiv.org/abs/1902.00996v2 (visited on 10/29/2020).

[Men+20] Chenlin Meng et al. *Gaussianization Flows*. 2020. arXiv: 2003.01941 [cs.LG].

[Mit99] Joseph SB Mitchell. "Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems". In: *SIAM Journal on computing* 28.4 (1999), pp. 1298–1309.

[MM12] Xiangrui Meng and Michael W. Mahoney. "Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression". In: *CoRR* abs/1210.3135 (2012). arXiv: 1210.3135. URL: http://arxiv.org/abs/1210.3135.

[Mon15] Andrea Montanari. "Computational implications of reducing data to sufficient statistics". In: (2015).

[MP15] Sandro Montanari and Paolo Penna. "On sampling simple paths in planar graphs according to their lengths". In: (2015), pp. 493–504.

[Mut05] S. Muthukrishnan. "Data Streams: Algorithms and Applications". In: *Foundations and Trends in Theoretical Computer Science* 1.2 (2005).

[MW10] Morteza Monemizadeh and David P. Woodruff. "1-Pass Relative-Error Lp-Sampling with Applications". In: *SODA*. 2010.

[Pap78] CH Papadimitriou. "The probabilistic analysis of matching heuristics". In: *Proc. 15th Annual Conference Comm. Contr. Computing*. Univ. Illinois Champaign, IL. 1978.

[Pee07] Matthew M. Peet. *Exponentially Stable Nonlinear Systems have Polynomial Lyapunov Functions on Bounded Regions*. 2007. arXiv: 0707.0218 [math.CA].

[Pit36] Edwin James George Pitman. “Sufficient statistics and intrinsic accuracy”. In: *Mathematical Proceedings of the cambridge Philosophical society*. Vol. 32. 4. Cambridge University Press. 1936, pp. 567–579.

[Pol12] Leonid Polterovich. *The geometry of the group of symplectic diffeomorphism*. Birkhäuser, 2012.

[Pyk59] Ronald Pyke. “The Supremum and Infimum of the Poisson Process”. In: *The Annals of Mathematical Statistics* 30.2 (1959), pp. 568–576. DOI: 10.1214/aoms/1177706269. URL: https://doi.org/10.1214/aoms/1177706269.

[Ren+21] Christopher X Ren et al. “Learning Continuous Exponential Families Beyond Gaussian”. In: *arXiv preprint arXiv:2102.09198* (2021).

[RM15] Danilo Rezende and Shakir Mohamed. “Variational inference with normalizing flows”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1530–1538.

[Rob55] Herbert Robbins. “A Remark on Stirling’s Formula”. In: *The American Mathematical Monthly* 62.1 (Jan. 1955), p. 26. DOI: 10.2307/2308012. URL: https://doi.org/10.2307/2308012.

[RXG20] Benjamin Rhodes, Kai Xu, and Michael U Gutmann. “Telescoping density-ratio estimation”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4905–4916.

[Sch03] Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Vol. 24. Springer Science & Business Media, 2003.

[SE19] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems* 32 (2019). URL: https://arxiv.org/abs/1907.05600.

[Son+20] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *arXiv preprint arXiv:2011.13456* (2020).

[SSW21] Abhin Shah, Devavrat Shah, and Gregory Wornell. “A computationally efficient method for learning exponential family distributions”. In: *Advances in neural information processing systems* 34 (2021), pp. 15841–15854.

[ST09] Daniel A. Spielman and Shang-Hua Teng. “Smoothed analysis: an attempt to explain the behaviour of algorithms in practice”. In: *Communications of the ACM* 52.10 (Oct. 2009), pp. 76–84. DOI: 10.1145/1562764.1562785. URL: https://doi.org/10.1145%2F1562764.1562785.

[Staa] Richard Stanley. *Parking Functions*. URL: www-math.mit.edu/~rstan/transparencies/parking.pdf (visited on 04/13/2023).

[Stab] Richard Stanley. *Parking Functions*. URL: www-math.mit.edu/~rstan/transparencies/parking3.pdf (visited on 04/13/2023).

[Ste15] Stefan Steinerberger. “New bounds for the traveling salesman constant”. In: *Advances in Applied Probability* 47.1 (2015), pp. 27–36.

[Ste81] J Michael Steele. "Subadditive Euclidean functionals and nonlinear growth in geometric probability". In: *The Annals of Probability* (1981), pp. 365–376.

[SV12] Nariankadu D. Shyamalkumar and Kasturi R. Varadarajan. "Efficient Subspace Approximation Algorithms". In: *Discrete & Computational Geometry* 47.1 (2012), pp. 44–63.

[SW11] Christian Sohler and David P. Woodruff. "Subspace embeddings for the L1-norm with applications". In: *STOC*. 2011.

[SW14] Adrien Saumard and Jon A Wellner. "Log-concavity and strong log-concavity: a review". In: *Statistics surveys* 8 (2014), p. 45.

[SWZ16] Zhao Song, David P. Woodruff, and Peilin Zhong. "Low Rank Approximation with Entrywise L1-Norm Error". In: *CoRR* abs/1611.00898 (2016). arXiv: 1611.00898. URL: http://arxiv.org/abs/1611.00898.

[Tal96] Michel Talagrand. "Transportation cost for Gaussian and other product measures". In: *Geometric & Functional Analysis GAFA* 6.3 (1996), pp. 587–600.

[Tur02] Dmitry Turaev. "Polynomial approximations of symplectic dynamics and richness of chaos in non-hyperbolic area-preserving maps". In: *Nonlinearity* 16.1 (Nov. 2002), pp. 123–135. DOI: 10.1088/0951-7715/16/1/308. URL: https://doi.org/10.1088%5C%2F0951-7715%5C%2F16%5C%2F1%5C%2F308.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Oct. 1998. DOI: 10.1017/cbo9780511802256. URL: https://doi.org/10.1017%2Fcbo9780511802256.

[Van00] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[Vuf+16] Marc Vuffray et al. "Interaction screening: Efficient and sample-optimal learning of Ising models". In: *Advances in neural information processing systems* 29 (2016).

[VV85] Leslie G Valiant and Vijay V Vazirani. "NP is as easy as detecting unique solutions". In: *Proceedings of the seventeenth annual ACM symposium on Theory of computing*. 1985, pp. 458–463.

[Was20] Larry Wasserman. *Lecture Notes 27*. https://www.stat.cmu.edu/~larry/=stat705/Lecture27.pdf. [Online; accessed 5 May 2022]. 2020.

[Woo14] David P. Woodruff. "Sketching as a Tool for Numerical Linear Algebra". In: *Foundations and Trends in Theoretical Computer Science* 10.1-2 (2014), pp. 1–157. DOI: 10.1561/0400000060. URL: https://doi.org/10.1561/0400000060.

[Zha+20] Han Zhang et al. "Approximation capabilities of neural ODEs and invertible residual networks". In: (2020).