

CONVERGENCE OF LANGEVIN AIS FOR MULTIMODAL DISTRIBUTIONS

AKSHAT AGARWAL, GAUTAM IYER, AIDAN JAMESON, SEUNGJAE SON,
AND WYATT WIMMER

ABSTRACT. We study convergence rates of the *annealed importance sampling algorithm* (Neal '01) combined with *Langevin Monte Carlo* when the target is a multimodal Gibbs measure. The main result shows that for a fixed error threshold, the time complexity is *quadratic in the inverse temperature*. We identify a simple and useful quantity that controls the sampling error for AIS in a general setting, and then bound this quantity in our setting using spectral estimates. We also study an autonormalized version and obtain bounds for the time complexity in terms of the inverse temperature.

1. Introduction

1.1. Main Results. We begin by stating our main results, following which we survey the literature and place our results in the context of the current literature. Let $\mathcal{X}$ be a configuration space and $U: \mathcal{X} \rightarrow \mathbb{R}$ be an energy function. Given $\varepsilon > 0$, the Gibbs measure with temperature ε is defined by

$$(1.1) \quad \pi_\varepsilon \stackrel{\text{def}}{=} \frac{\tilde{\pi}_\varepsilon}{Z_\varepsilon} \quad \text{where} \quad \tilde{\pi}_\varepsilon = e^{-U/\varepsilon}, \quad \text{and} \quad Z_\varepsilon = \int_{\mathcal{X}} \tilde{\pi}_\varepsilon dx,$$

where the last integral is carried out with respect to some reference measure on $\mathcal{X}$. We typically only consider the case where $\mathcal{X}$ is the d -dimensional Euclidean space $\mathbb{R}^d$, or the d -dimensional torus $\mathbb{T}^d$. In these cases the reference measure in (1.1) is the Lebesgue measure.

Consider the overdamped Langevin equation

$$(1.2) \quad dX_t^\varepsilon = -\nabla U(X_t^\varepsilon) + \sqrt{2\varepsilon} dB_t,$$

where B is a standard d -dimensional Brownian motion. It is well known (see for instance [Pav14]) that the stationary distribution of the Langevin equation is π_ε , and the *Langevin Monte Carlo (LMC)* algorithm obtains samples from π_ε by simulating (1.2) for long time.

The disadvantage of LMC is that when U is not convex, the rate at which solutions to (1.2) converge to the stationary distribution is exponentially small in $1/\varepsilon$. This is the Arrhenius law [Arr89], and is described in more detail below. We show that this slow rate of convergence can be overcome if one uses *annealed importance sampling (AIS)* [Nea01], or an autonormalized version of AIS.

2020 *Mathematics Subject Classification.* Primary: 60J22, Secondary: 65C05, 65C40.

Key words and phrases. Annealed importance sampling; Langevin Monte Carlo; Multimodal distributions; Sampling.

This work has been partially supported by the National Science Foundation under grants DMS-2406853, DMS-2342349 and the Center for Nonlinear Analysis.

1.1.1. *Langevin Annealed Importance Sampling.* For convenience, we first state the Langevin AIS algorithm.

Algorithm 1 Langevin Annealed Importance Sampling (Langevin AIS)

Tunable parameters:

- (1) Target temperature $\varepsilon > 0$.
 - (2) Annealing schedule $K \in \mathbb{N}$ and $\varepsilon_1 > \dots > \varepsilon_{K+1} = \varepsilon$
 - (3) LMC simulation time T .
 - 1: Start with $X_0 \in \mathbb{T}^d$ arbitrary, and weight $\tilde{w}_0 = 1$.
 - 2: **for** $k = 1, \dots, K$ **do**
 - 3: Simulate (1.2) with temperature ε_k , starting from X_{k-1} , and let X_k be the state after time T .
 - 4: $\tilde{w}_k \leftarrow \tilde{w}_{k-1} \frac{\pi_{\varepsilon_{k+1}}(X_k)}{\pi_{\varepsilon_k}(X_k)}$.
 - 5: **end for**
 - 6: **return** X_K and the *unnormalized* weight $\tilde{w}_K$.
-

Our main result shows that if we obtain N independent samples using Algorithm 1, and empirically normalize the weights, then we obtain good samples from π_ε with time complexity that is quadratic in $1/\varepsilon$.

Theorem 1.1. *Suppose the configuration space $\mathcal{X} = \mathbb{T}^d$, and U is a regular double-well potential with non-degenerate wells of equal depth. Given $\nu > 0$, and $\varepsilon_1 > 0$ there exist constants $C_T = C_T(\nu, U, \varepsilon_1)$, $\bar{C}_w(\nu, U, \varepsilon_1)$ such that the following holds.*

For every $\delta > 0$, $\varepsilon \in (0, \varepsilon_1)$, choose

$$(1.3) \quad K = \left\lceil \frac{1}{\varepsilon \nu} \right\rceil,$$

$$(1.4) \quad T > \max \left\{ C_T \left(\frac{1}{\varepsilon} + \log K \right), (4 + |\log_2 \delta|) t_{\text{mix}, \varepsilon_1}^\infty \right\},$$

$$(1.5) \quad N = \left\lceil \frac{64 \bar{C}_w}{\delta^2} \right\rceil.$$

Choose $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ so that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. Let $(X_K^1, \tilde{w}_K^1), \dots, (X_K^N, \tilde{w}_K^N)$ be the points and unnormalized weights obtained from N independent runs of Algorithm 1 with these parameters. Define the empirical measure μ_N by

$$(1.6) \quad \mu_N \stackrel{\text{def}}{=} \sum_{i=1}^N w_i \delta_{X_K^i} \quad \text{where} \quad w_i = \frac{\tilde{w}_K^i}{\tilde{W}_{K,N}} \quad \text{and} \quad \tilde{W}_{K,N} \stackrel{\text{def}}{=} \sum_{i=1}^N \tilde{w}_K^i.$$

Then for every bounded test function f

$$(1.7) \quad \mathbf{E} |\langle f, \mu_N \rangle - \langle f, \pi_\varepsilon \rangle|^2 < \|f\|_{\text{osc}}^2 \delta^2.$$

Moreover, for every $s > d/2$ there exists an explicit dimensional constant C_s (independent of $\varepsilon_1, \varepsilon, \nu, \delta$) such that

$$(1.8) \quad \mathbf{E} \|\mu_N - \pi_\varepsilon\|_{\dot{H}^{-s}}^2 \leq C_s \delta^2.$$

Here $t_{\text{mix}, \varepsilon_1}^\infty$ is the uniform mixing time of (1.2) on $\mathbb{T}^d$ with $\varepsilon = \varepsilon_1$ (see [LP17], or (2.2), below). The notation $\langle f, \mu \rangle$ used in (1.7) is

$$\langle f, \mu \rangle = \int_{\mathbb{T}^d} f \, d\mu,$$

and $\|\cdot\|_{\dot{H}^{-s}}$ denotes the norm in the homogeneous Sobolev space with index $-s$ (for instance see [AKM19], or (4.20), below).

Remark 1.2 (Effective sample size). When working with weighted points as above, it is important to ensure that the weights don't concentrate on a few particles reducing the overall efficiency. This is often measured by the *effective sample size* (see for instance Section 8.6 in [CP20]) defined by

$$(1.9) \quad \text{ESS}(w_1, \dots, w_N) \stackrel{\text{def}}{=} \frac{1}{\sum_{i=1}^N w_i^2}.$$

We will show (see Proposition 2.7 and Remark 3.8, below) that Theorem 1.1 implies that the effective sample size is bounded by

$$\mathbf{E} \text{ESS}(w_1, \dots, w_N) = \mathbf{E} \left(\frac{1}{\sum_{i=1}^N w_i^2} \right) \geq \frac{N}{8(4\tilde{C}_w + 1)}.$$

Remark 1.3 (Computational complexity). We now discuss the asymptotic behavior of the computational complexity as the temperature $\varepsilon \rightarrow 0$, and as the allowed error $\delta \rightarrow 0$. Assuming the computational cost of simulating (1.2) for time T is $O(T)$, and neglecting the discretization error, the computational cost of Algorithm 1 to achieve the error bounds (1.7) or (1.8) is $O(KTN)$. Unravelling the ε and δ dependence in (1.3)–(1.5) this implies

$$(1.10) \quad \text{complexity}(\text{Algorithm 1}) = O(KTN) \leq \frac{C_d}{\delta^2} \left(\frac{1}{\varepsilon^2} + |\ln \delta| \right),$$

for some constant $C_d = C_d(U)$. For reference we mention that the computational complexity of Langevin Monte Carlo under these conditions is $e^{O(1/\varepsilon)}$, and the complexity of rejection sampling is $O(1/\varepsilon^d)$. A more comprehensive discussion and comparison is at the end of Section 1.2, below.

Remark 1.4 (Multiple wells, and wells of unequal depth.). The assumption that U is a double well potential with wells of equal depth can be relaxed. If the wells have “nearly equal depth” so that the distribution remains truly multimodal in a temperature range, then Theorem 1.1 will still hold. The required assumptions and generalized theorem is Theorem 3.7 in Section 3, below. If there are more than two wells, then Theorem 1.1 will still hold provided we make the same non-degeneracy assumptions as in Section 10 in [HIS26].

Remark 1.5. The constants C_T and $\tilde{C}_w$ in Theorem 1.1 involve dimensional factors that arise in spectral estimates and Sobolev embedding theorems. As a result, their dimensional dependence is not explicit (see (5.12) and (5.19), below). In special cases where U has a low dimensional structure, the dimensional dependence of these constants can be controlled for an idealized model problem (similar to Proposition 3.1 in [HIS26]), but their dimensional dependence is not explicit for the full Langevin system.

Remark 1.6. In practice, one often has a target distribution of the form $\pi \propto e^{-V}$ which is hard to sample from. By the Arrhenius law, the complexity of using LMC to directly sample from $\pi \propto e^{-V}$ is $e^{O(H)}$ where H is the energy barrier. If instead if we choose $\varepsilon = 1/H$ and $U = \varepsilon V$, and use Langevin AIS to sample from $\pi = \pi_\varepsilon \propto e^{-U/\varepsilon}$, then by Remark 1.3 the time complexity is now only $O(1/\varepsilon^2) = O(H^2)$.

The proof of Theorem 1.1 consists of two steps: The first step is a general result concerning AIS. Given a family of distributions $\pi_1, \dots, \pi_{K+1}$, with corresponding Markov transition kernels $P_1, \dots, P_{K+1}$, we estimate the variance of using AIS to sample from π_{K+1} in terms of

$$(1.11) \quad C_P(T) \stackrel{\text{def}}{=} \prod_{k=1}^K \|P_k^T r_k^2\|_{L^\infty}, \quad \text{where} \quad r_k \stackrel{\text{def}}{=} \frac{\pi_{k+1}}{\pi_k}.$$

This is the content of Proposition 2.4 and Theorem 2.5, below. We will shortly examine the quantity $C_P(T)$ further and note its resemblance to the product of the χ^2 -divergences $\chi^2(\pi_{k+1}; \pi_k)$ in Section 2.2, below.

The second step in the proof of Theorem 1.1 bounds the quantity $C_P(T)$ for the specific sequence of intermediate sequence of distributions used in Theorem 1.1. We do this using certain spectral properties of the generator of (1.2). These properties are a combination of results in [Kol00, BGK05, MS14], and are collected in a convenient form in [HIS26]. The full proof is presented in Section 5, below.

1.1.2. Autonormalized Langevin AIS. We also consider an *autonormalized* version of Algorithm 1. Namely, instead of returning the unnormalized weights $\tilde{w}_K$, we run N independent realizations and normalize the weights using an empirical average. This is a commonly used idea in many *Sequential Monte Carlo* samplers (see for instance [DdFG01, Liu08, CP20]), and performs well in many situations of practical interest. For convenience, we now state the version of this algorithm precisely.

Algorithm 2 Autonormalized Langevin Annealed Importance Sampling

Tunable parameters:

- (1) Target temperature $\varepsilon > 0$.
 - (2) Annealing schedule $K \in \mathbb{N}$ and $\varepsilon_1 > \dots > \varepsilon_{K+1} = \varepsilon$
 - (3) LMC simulation time T .
 - (4) Number of particles $N \in \mathbb{N}$.
 - 1: Start with $w_0^1 = \dots = w_0^N = 1/N$, and $X_0^1, \dots, X_0^N$ arbitrary.
 - 2: **for** $k = 1, \dots, K$ **do**
 - 3: **for** $i = 1, \dots, N$ **do**
 - 4: Simulate (1.2) with temperature ε_k , starting from X_{k-1}^i , and let X_k^i be the state after time T .
 - 5: $\tilde{w}_k^i \leftarrow w_{k-1}^i \frac{\tilde{\pi}_{\varepsilon_{k+1}}(X_k^i)}{\tilde{\pi}_{\varepsilon_k}(X_k^i)}$.
 - 6: **end for**
 - 7: Let $\tilde{W}_k = \sum_{j=1}^N \tilde{w}_k^j$, and for each $i \in \{1, \dots, N\}$ set $w_k^i = \tilde{w}_k^i / \tilde{W}_k$.
 - 8: **end for**
 - 9: **return** $X_K^1, \dots, X_K^N$ and the normalized weights $w_K^1, \dots, w_K^N$.
-

Theorem 1.7. *Suppose the configuration space $\mathcal{X} = \mathbb{T}^d$, and U is a regular double-well potential with non-degenerate wells of equal depth. Given $\alpha, \nu > 0$, and $\varepsilon_1 > 0$, there exists constants C_N and $\hat{C}_T$ such that the following holds.*

For every $\delta > 0$ and $\varepsilon \in (0, \varepsilon_1)$, choose

$$(1.12) \quad K = \left\lceil \frac{1}{\varepsilon \nu} \right\rceil,$$

$$(1.13) \quad T > \hat{C}_T \left(K^{(1+\alpha)} + \frac{1}{\varepsilon} + \log\left(\frac{1}{\delta}\right) + \log N \right),$$

$$(1.14) \quad N = \left\lceil \frac{C_N}{\delta^2} \right\rceil K^2.$$

Let $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ be such that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. Run Algorithm 2 with these parameters, and let $\{X_K^i, w_K^i\}_{1 \leq i \leq N}$ be the N points and normalized weights returned. Define the empirical measure μ_N by

$$\mu_N \stackrel{\text{def}}{=} \sum_{i=1}^N w_K^i \delta_{X_K^i}.$$

Then, for every bounded measurable test function f , we have

$$\mathbf{E} \left| \langle f, \mu_N \rangle - \langle f, \pi_\varepsilon \rangle \right|^2 \leq \|f\|_{\text{osc}}^2 \delta^2.$$

Consequently, for every $s > d/2$ there exists an explicit dimensional constant $\hat{C}_s$ (independent of $\varepsilon_1, \varepsilon, \alpha, \nu, \delta$) such that

$$\mathbf{E} \left\| \mu_N - \pi_\varepsilon \right\|_{H^{-s}}^2 \leq C_s \delta^2.$$

Remark 1.8 (Effective sample size). We will also estimate the effective sample size and show that there exists a constant $C_1 = C_1(\nu, \varepsilon_1)$ such that

$$\mathbf{E} \text{ESS}(w_k^1, \dots, w_k^N) \geq \frac{N}{C_1 \bar{C}_w^2},$$

where $\bar{C}_w$ is the constant from Theorem 1.1. This is the content of Proposition 3.10, below.

Remark 1.9 (Computational complexity). We now discuss the asymptotic behavior of the computational complexity as the temperature $\varepsilon \rightarrow 0$, and as the allowed error $\delta \rightarrow 0$. As with Remark 1.3, the computational cost of Algorithm 2 is $O(KTN)$. Unravelling the ε and δ dependence in (1.12)–(1.14) this implies

$$(1.15) \quad \text{complexity}(\text{Algorithm 2}) = O(KTN) \leq \frac{C_d}{\delta^2 \varepsilon^3} \left(\frac{1}{\varepsilon^{1+\alpha}} + |\ln \delta| \right).$$

While this is much smaller than the complexity of LMC (which is $e^{O(1/\varepsilon)}$), or rejection (which is $O(1/\varepsilon^d)$), it is larger than the complexity of Algorithm 1 which is $O(1/\varepsilon^2)$. In practice, Langevin AIS and autonormalized Langevin AIS perform comparably, and the reason the bound (1.15) is worse than (1.10) may be due to suboptimality of our estimates.

The proof of Theorem 1.7 is a little more involved than Theorem 1.1 because the processes $\{X_K^i, w_K^i \mid 1 \leq i \leq N\}$ are not independent and identically distributed, but only exchangeable. As a result, we are presently unable to deduce Theorem 1.7 from a general one particle result, as we will do for Theorem 1.1. Moreover, the lack of independence introduces a few technical difficulties in the proof and this leads to bounds that are worse than those in Theorem 1.1. Explicitly, Theorem 1.7 the choice (1.12)–(1.14) requires N to grow like $1/\varepsilon^2$, where as in Theorem 1.1 one can choose N independent of ε . Moreover, the simulation time T in (1.13) now grows like $1/\varepsilon^{1+\alpha}$, where as in Theorem 1.1 it only needed to grow like $1/\varepsilon$. In spite of the difference in the provable bounds in this situation, Algorithm 2 performs well in practice and is used often [DdFG01, Liu08, CP20]. We present the proof in Section 6, below.

1.2. Literature review. Sampling from distributions is a longstanding problem that arises in many applications such as Bayesian inference, statistical Physics and machine learning. In many situations of practical interest, the state space $\mathcal{X}$ is huge (either finite, but with a computationally intractable size, or a high dimensional manifold). For any given state $x \in \mathcal{X}$, one can typically compute an energy $U(x)$ measuring how favorable the state is. Standard models (e.g. the canonical ensemble in statistical physics) dictate that the probability of finding the system in state x is proportional to $\tilde{\pi}_\varepsilon(x) = e^{-U(x)/\varepsilon}$ (as in (1.1)), where $\varepsilon > 0$ is a parameter (often the absolute temperature) controlling how fast the system transitions between states.

Practically, the normalization constant Z^ε (often called the partition function) is a high dimensional integral (or a sum over huge number of states) and is computationally intractable. Moreover, even if the normalization constant Z^ε is known, the enormous state space requires the use of algorithms that can deliver samples even though they can only inspect a miniscule fraction of all possible states. Such algorithms aren't easy to design, or rigorously analyze, and this makes such sampling problems extremely challenging.

Markov Chain Monte Carlo (MCMC) is a family of algorithms that is often used to address such problems. These work by simulating a Markov process whose stationary distribution is π_ε , and date back to the celebrated Metropolis–Hastings algorithm [MRR⁺53, Has70]. In Euclidean space the Langevin dynamics (1.2) provides a particularly convenient Markov process with stationary distribution π_ε , and this is the basis of *Langevin Monte Carlo (LMC)*, *Metropolis Adjusted Langevin Monte Carlo (MALA)* and various other sampling algorithms (see for instance [Liu08]).

One issue that requires attention when using MCMC based algorithms is the rate at which the Markov chain converges to equilibrium. If this rate is too slow, it may require simulating the Markov chain for impractical amounts of time before obtaining good samples. In some cases, one has quantitative estimates on the rate of convergence. For (1.2), it is known that if U is uniformly convex, then

$$W_2(\text{dist}(X_t^\varepsilon), \pi_\varepsilon) \leq e^{-C(U)t/\varepsilon} W_2(\text{dist}(X_0^\varepsilon), \pi_\varepsilon),$$

which means that simulating (1.2) will yield good samples of the Gibbs measure in short time, even when the temperature ε is small. Vempala and Wibisono [VW19] (see also [Che23]) showed that this is also true for the Euler–Maruyama discretization of (1.2), making LMC a practically viable sampling algorithm for log concave distributions even in high dimensions.

When U is not convex, however, the situation is different. If U is multi-modal (for instance, if π_1 is a Gaussian mixture), then the convergence rate of (1.2) is exponentially small in $1/\varepsilon$. That is, it takes time $e^{O(1/\varepsilon)}$ for the distribution of X_t^ε to become close to π_ε . This phenomenon is known as the Arrhenius law [Arr89], and occurs for the following reason. Since the drift in (1.2) pulls trajectories towards local minima of U , the noise term in (1.2) has to go against the drift for an $O(1)$ amount of time in order for trajectories to transition from the basin of attraction of one local minimum to another. This happens with exponentially small probability, leading to the Arrhenius law.

Several methods have been designed to improve the rate of convergence. Tempering methods introduced in [SW86, Nea96, MP92, Nea11] run a Markov chain on a product of the state space at various temperature levels. Other methods

include ideas based on birth-death [LLN19], optimization [PHLa20], diffusion models [CKSV25], warm starts [KLV25, LSG25]. Authors have also modified (1.2) by adding a drift [RBS15, DFY20, CFIN25], or modifying the diffusion [ERY24]. In some situations [WSH09, GLR20, Son26] polynomial convergence bounds have been rigorously proved.

The methods most closely related to this paper are known as *Sequential Monte Carlo (SMC)* algorithms. The first such instance was developed to study of the average extension of molecular chains [HM54, RR55]. SMC methods use a sequence of auxiliary distributions $\nu_1, \dots, \nu_K$ so that ν_1 is easy to sample from, ν_K is the target distribution, and then move samples between distributions using a reweighting / resampling mechanism. These methods are hugely popular, and we refer the reader to [DdFG01, Liu08, CP20, SBCCD24] for a broad overview. In the context of multi-modal distributions, rigorous convergence bounds were proved in [Sch12, PJT19, MS24, LSG24, Han25, HIS26].

The algorithms we use in this paper (Algorithms 1 and 2) are obtained using Neal’s *Annealed Importance Sampling (AIS)* algorithm [Nea01], combined with Langevin Monte Carlo. While AIS and Langevin AIS are immensely popular, there are very few rigorous convergence bounds that apply in the setting of Theorem 1.1 and 1.7. Specifically, in our setting we make no apriori assumption about the mixture decomposition, symmetry, shape of the wells, or apriori assume knowledge of how the mass distribution in wells changes as the temperature varies. We only assume non-degeneracy of critical points, and regularity of U . (The assumption that U is a double well potential with wells of equal depth can be relaxed as mentioned in Remark 1.4.) To the best of our knowledge, the only results that apply in this setting are [HIS26, Han25, Son26], and we now comment on the relationship between these results and the present paper.

In [HIS26] the authors used an SMC algorithm which resampled points at every level, instead of reweighting them as we do in Algorithms 1 and 2. The advantage of this method is that it keeps the effective sample size constant, and is extremely popular in practice [CP20]. The disadvantage is that particles are now only exchangeable, and not independent, and so theoretical bounds harder to obtain. In [HIS26] the authors show the complexity is exactly the same as that of Algorithm 2 (given by (1.15)).

In [Han25] the author shows the same SMC algorithm can be used in the non-compact setting when the state space $\mathcal{X} = \mathbb{R}^d$. The author uses a coupling argument in [MMS23] and obtains error estimates in probability. For a fixed error, the time complexity of this algorithm is $O(|\ln \varepsilon|^{10/3}/\varepsilon^7)$ as $\varepsilon \rightarrow 0$.

Finally in [Son26] revisits this sampling problem using Metropolis ball walks and parallel tempering. The author uses a soft domain decomposition [MR02, WSH09] to show that for a fixed error, the time complexity behaves like $O(1/\varepsilon^{11})$ as $\varepsilon \rightarrow 0$.

As mentioned earlier (Remark 1.3 and (1.10)), for a fixed error the time complexity of Langevin AIS (Algorithm 1) scales like $O(1/\varepsilon^2)$ as $\varepsilon \rightarrow 0$. The main new contribution of this paper of this paper over [HIS26, Han25, Son26] is twofold. First, the algorithm has smaller complexity ($O(1/\varepsilon^2)$, vs $O(1/\varepsilon^4)$ or higher). Second, the proof identifies a simple and useful quantity, $C_P(T)$, that controls the sample error of AIS (equation (1.11), see also Theorem 2.5 and Section 2.2, below) in a general setting. We presently prove Theorem 1.1 by bounding $C_P(T)$ using spectral estimates from [Kol00, BGK05, MS14, HIS26]. There may be room to bound $C_P(T)$

using different techniques, bypassing the limitation of spectral methods, but this goes beyond the scope of the present paper and is left for future study.

Plan of this paper. In Section 2 we state two results (Proposition 2.4 and Theorem 2.5) addressing convergence of AIS in a general setting, and control the sampling error in terms of $C_P(T)$. In Section 3 we study Langevin AIS and its autonormalized version, and state generalizations of Theorems 1.1 and 1.7. In Section 4 we prove Proposition 2.4 and Theorem 2.5. In Section 5 we prove the generalization of Theorem 1.1 (Theorem 3.7), and finally in Section 6 we prove the generalization of Theorem 1.7 (Theorem 3.9).

2. Convergence results for AIS

As mentioned earlier, we prove Theorem 1.1 by showing that the quantity $C_P(T)$ in (1.11) can be used to control the sampling error of AIS in a general setting. We begin by stating this precisely.

2.1. Bias and Variance estimates for AIS. Suppose $\pi = \tilde{\pi}/Z$ is a target distribution from which samples are desired. Here $\tilde{\pi}$ is an unnormalized probability distribution which is easy to compute, and $Z = \int_{\mathcal{X}} \tilde{\pi}$ is the normalization constant (which is hard to compute). The AIS algorithm [Nea01] uses an auxiliary family of distributions with densities proportional to $\tilde{\pi}_1, \dots, \tilde{\pi}_{K+1} = \tilde{\pi}$ (called a *tempering sequence*), along with reversible Markov transition kernels $P_1, \dots, P_K$. The algorithm now successively simulates a Markov chain with kernel P_k , and then reweights the points using the ration $\tilde{\pi}_{k+1}/\tilde{\pi}_k$ and is described precisely as Algorithm 3.

Algorithm 3 Annealed Importance Sampling (AIS)

Requirements:

- (1) Tempering sequence of unnormalized densities $\tilde{\pi}_1, \dots, \tilde{\pi}_K$.
 - (2) Corresponding reversible Markov transition kernels $P_1, \dots, P_K$.
 - (3) Running time T .
- 1: Start with $w_0 = 1$, X_0 arbitrary.
 - 2: **for** $k = 1, \dots, K$ **do**
 - 3: Sample X_k from $P_k^T(X_{k-1}, \cdot)$
 - 4: $\tilde{w}_k \leftarrow \tilde{w}_{k-1} \frac{\tilde{\pi}_{k+1}(X_k)}{\tilde{\pi}_k(X_k)}$.
 - 5: **end for**
 - 6: **return** X_K and the unnormalized weight $\tilde{w}_K$.
-

The goal of this section is to obtain a quantitative error bound on AIS, provided the tempering sequence satisfies the following assumptions.

Assumption 2.1. For each $k \in \{1, \dots, K+1\}$ we have a reversible Markov transition kernel P_k whose stationary distribution is $\pi_k = \tilde{\pi}_k/Z_k$, where $Z_k = \int_{\mathcal{X}} \tilde{\pi}_k$.

Assumption 2.2. There exists constants T_0, C_w such that for every $T \geq T_0$ we have

$$(2.1) \quad \prod_{k=1}^K \|P_k^T r_k^2\|_{L^\infty} \leq C_w, \quad \text{where} \quad r_k \stackrel{\text{def}}{=} \frac{\pi_{k+1}}{\pi_k}.$$

Assumption 2.3. *The uniform mixing time of P_1 is finite. Explicitly,*

$$(2.2) \quad t_{\text{mix},1}^{\infty} \stackrel{\text{def}}{=} \min \left\{ n \in \mathbb{N} \left| \sup_{x \in \mathcal{X}} \left\| \frac{P_1^n(x, \cdot)}{\pi_1(\cdot)} - 1 \right\|_{\infty} < \frac{1}{2} \right. \right\} < \infty$$

We will now show that if we use AIS with a tempering sequence satisfying Assumptions 2.1–2.3, then the bias decreases exponentially with $T/t_{\text{mix},1}^{\infty}$, and variance is bounded by $4C_w$.

Proposition 2.4 (Bias and variance bounds). *Suppose $\{P_k\}_{1 \leq k \leq K}$ are Markov transition kernels satisfying Assumptions 2.1–2.3. Let $X_K, \tilde{w}_K$ be obtained from Algorithm 3 with $K+1$ levels, running time T , and kernels $\{P_k\}$. For all bounded measurable functions f , we have the bias and variance estimates*

$$(2.3) \quad \left| \frac{\mathbf{E} \tilde{w}_K f(X_K)}{\mathbf{E} \tilde{w}_K} - \langle f, \pi_{K+1} \rangle \right| \leq 2^{2-T/t_{\text{mix},1}^{\infty}} \|f\|_{\infty}, \quad \text{provided } T \geq t_{\text{mix},1}^{\infty}$$

$$(2.4) \quad \text{Var} \left(\frac{\tilde{w}_K f(X_K)}{\mathbf{E} \tilde{w}_K} \right) \leq 4C_w \|f\|_{\infty}^2, \quad \text{provided } T \geq T_0.$$

In practice, one would take N independent samples from Algorithm 3, and estimate $\mathbf{E} \tilde{w}_K$ using the empirical mean. An immediate consequence of Proposition 2.4 is a quantitative bound on the convergence of the empirical measure.

Theorem 2.5 (Empirical measure convergence). *Suppose $\{P_k\}_{1 \leq k \leq K}$ are Markov transition kernels satisfying Assumptions 2.1–2.3. Given $\delta > 0$, choose*

$$(2.5) \quad N = \frac{64C_w}{\delta^2} \quad \text{and} \quad T = \max \{ T_0, (4 + |\log_2 \delta|) t_{\text{mix},1}^{\infty} \}.$$

Let $\{X_K^i, \tilde{w}_K^i\}_{1 \leq i \leq N}$ be N independent realizations of the points and weights returned by Algorithm 3, with $K+1$ levels, running time T , and kernels $\{P_k\}$. For every bounded measurable f , we have

$$(2.6) \quad \left\| \sum_{i=1}^N w_K^i f(X_K^i) - \langle f, \pi_{K+1} \rangle \right\|_{L^2(\mathbf{P})} \leq \|f\|_{\infty} \delta, \quad \text{where } w_K^i \stackrel{\text{def}}{=} \frac{\tilde{w}_K^i}{\sum_{j=1}^N \tilde{w}_K^j}.$$

Moreover, if $\mu_{K+1,N}$ is the empirical measure defined by

$$\mu_{K+1,N} \stackrel{\text{def}}{=} \sum_{i=1}^N w_K^i \delta_{X_K^i},$$

then for every $s > d/2$

$$(2.7) \quad \mathbf{E} \left\| \mu_{K+1,N} - \pi_{K+1} \right\|_{H^{-s}}^2 \leq C_s \delta^2.$$

Remark 2.6 (Time complexity). In order to obtain samples that satisfy (2.6), one has to simulate N realizations of T iterations of each of the chains $P_1, \dots, P_K$, and so the time complexity is

$$O(NTK) = \frac{64KC_w t_{\text{mix},1}^{\infty}}{\delta^2} (4 + |\log_2 \delta|)$$

As mentioned earlier, when working with weighted samples it is important to estimate the *effective sample size* (see (1.9)) and ensure that the weights don't concentrate on a few points. This can be done quickly from the variance bound (2.4).

Proposition 2.7 (Effective sample size). *Let $w_K^1, \dots, w_K^N$ be the normalized weights defined in (2.6). Then the sum of the squared weights satisfies*

$$(2.8) \quad \mathbf{E} \sum_{i=1}^N (w_K^i)^2 \leq \frac{8(4C_w + 1)}{N}.$$

Consequently, the expected effective sample size is bounded by

$$(2.9) \quad \mathbf{E} \text{ESS}(w_K^1, \dots, w_K^N) \geq \frac{N}{8(4C_w + 1)}.$$

Before delving into the proofs, we now briefly discuss the assumptions and implications of Proposition 2.5.

2.2. Relation to χ^2 -divergence. To understand the assumptions better, suppose momentarily we were able to choose $T = \infty$ in (2.1). In this case, the left hand side of (2.1) better can be bounded in terms of the χ^2 -divergence between each of the intermediate distributions.

Proposition 2.8. *If $T = \infty$ then the left hand side of (2.1) satisfies*

$$(2.10) \quad \prod_{k=1}^K \|P_k^\infty r_k^2\|_{L^\infty} \leq \exp \left(\sum_{k=1}^K \chi^2(\pi_{k+1}; \pi_k) \right)$$

Recall the χ^2 -divergence appearing above is a commonly used measure of the difference between two distributions. Explicitly, for two distributions with densities p, q the χ^2 -divergence is defined by

$$\chi^2(p; q) \stackrel{\text{def}}{=} \left\langle \frac{p^2}{q^2} - 1, q \right\rangle.$$

Let us now momentarily assume that π_k 's are all Gaussian, with $\pi_1 \sim \mathcal{N}(0, I_d)$ and $\pi_{K+1} \sim \mathcal{N}(0, \varepsilon I_d)$. If we choose $K = 1$ then AIS reduces to a vanilla importance sampling, and one can explicitly compute (see for instance Proposition 17.1 in [CP20]) that the right hand side of (2.10) is $O(1/\varepsilon^{d/2})$. This is comparable with a standard rejection sampling cost, and is too large to be practical.

However, if we instead choose $K = d/\varepsilon$, then the right hand side of (2.10) can be bounded *independent* of both ε and d . Thus, using AIS here with $O(d/\varepsilon)$ intermediate levels gives samples with variance $O(1)$, which is a huge improvement.

Proposition 2.9. *If the target distribution π is the Gaussian $\mathcal{N}(0, \varepsilon I_d)$, then choosing*

$$K = \left\lceil \frac{d}{\varepsilon} \right\rceil,$$

choosing the temperatures $\{\varepsilon_k\}$ so that $\varepsilon_{K+1} = \varepsilon$ and $\{1/\varepsilon_k\}_{1 \leq k \leq K}$ are linearly spaced, and choosing π_k to be the Gaussian $\mathcal{N}(0, \varepsilon_k I_d)$ will ensure

$$(2.11) \quad \prod_{k=1}^K \|P_k^\infty r_k^2\|_{L^\infty} \leq C.$$

In this case, Proposition 2.4 with $T = \infty$ will yield a variance bound that is independent of both ε and d .

The proof of Proposition 2.9 is elementary, and presented in Section 4.4, below.

3. Convergence results for Langevin AIS

3.1. Assumptions. In order to state the assumptions in Theorem 1.1 precisely, we first describe the assumptions that are required on the potential U . In short, we need the potential to be a regular, double-well function with wells of *nearly equal* depth. The criterion that the wells have nearly equal depth is required for the multimodal sampling problem to be non-degenerate in the sense that the mass in each well remains bounded away from 0. These assumptions are the same as the assumptions in [HIS26], which are explained in detail in [HIS26, Section 4.1]. We quote them here for easy reference, and refer the reader to [HIS26] for a detailed explanation and motivation.

Assumption 3.1. *The function $U \in C^{6\nu(1+\lfloor d/2 \rfloor)}(\mathbb{T}^d, \mathbb{R})$, has a nondegenerate Hessian at all critical points, and has exactly two local minima located at $x_{\min,1}$ and $x_{\min,2}$. We normalize U so that*

$$0 = U(x_{\min,1}) \leq U(x_{\min,2}).$$

Define the saddle height between $x_{\min,1}$ and $x_{\min,2}$ to be the minimum amount of energy needed to go from the global minimum $x_{\min,1}$ to $x_{\min,2}$. Explicitly, the saddle height is

$$\hat{U} = \hat{U}(x_{\min,1}, x_{\min,2}) \stackrel{\text{def}}{=} \inf_{\omega} \sup_{t \in [0,1]} U(\omega(t)).$$

Here the infimum above is taken over all continuous paths $\omega \in C([0,1]; \mathbb{T}^d)$ such that $\omega(0) = x_{\min,1}$, $\omega(1) = x_{\min,2}$.

Assumption 3.2. *The saddle height between $x_{\min,1}$ and $x_{\min,2}$ is attained at a unique critical point $s_{1,2}$ of index one. That is, the first eigenvalue of $\text{Hess } U(s_{1,2})$ is negative and the others are positive.*

The *energy barrier*, denoted by $\hat{\gamma}$, is defined to be the minimum amount of energy needed to go from the (possibly local) minimum $x_{\min,2}$ to the global minimum $x_{\min,1}$. In terms of $s_{1,2}$, the energy barrier $\hat{\gamma}$ and the saddle height are given by

$$\hat{\gamma} \stackrel{\text{def}}{=} U(s_{1,2}) - U(x_{\min,2}), \quad \text{and} \quad \hat{U} = U(s_{1,2}).$$

The ratio $\hat{\gamma}_r$ is the ratio of the saddle height $\hat{U}$ to the energy barrier $\hat{\gamma}$, given by

$$(3.1) \quad \hat{\gamma}_r \stackrel{\text{def}}{=} \frac{\hat{U}}{\hat{\gamma}}.$$

We recall the basin of attraction around $x_{\min,i}$, denoted by Ω_i , is defined by

$$\Omega_i \stackrel{\text{def}}{=} \left\{ y \in \mathbb{T}^d \mid \lim_{t \rightarrow \infty} y_t = x_{\min,i}, \text{ where } \dot{y}_t = -\nabla U(y_t) \text{ with } y_0 = y \right\}.$$

Assumption 3.3. *There exists $0 \leq \varepsilon_{\min} < \varepsilon_{\max} \leq \infty$, a constant C_m such that*

$$\inf_{\substack{\varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}] \\ 0 < \varepsilon < \infty}} \pi_{\varepsilon}(\Omega_i) \geq \frac{1}{C_m^2}.$$

Remark 3.4. For simplicity, we have assumed that U has only two wells. If U has more than two wells, the techniques we use will still apply provided we impose a non-degeneracy condition on U . The precise assumption is stated in Section 10 in [HIS26], and the required modifications to the proof are straightforward.

3.2. Convergence results for Langevin AIS. In this section we precisely state the main result of this paper (Theorem 3.7, below, which is a generalization of Theorem 1.1). We begin by stating bias and variance bounds for Algorithm 1.

Proposition 3.5 (Bias and variance bounds). *Let U be a double well potential that satisfies Assumptions 3.1–3.3, and let $\varepsilon_1 \in (\varepsilon_{\min}, \varepsilon_{\max}]$. Let $\nu > 0$ be a fixed constant. There exists constants $\bar{C}_w = \bar{C}_w(U, \nu)$, and $C_T = C_T(U)$ such that the following holds. For any $\varepsilon \in (\varepsilon_{\min}, \varepsilon_1]$ choose*

$$(3.2) \quad K = \left\lceil \frac{1}{\varepsilon \nu} \right\rceil \quad \text{and} \quad T > \max \left\{ t_{\text{mix}, \varepsilon_1}^\infty, C_T \left(\frac{1}{\varepsilon} + \log K \right) \right\},$$

and choose $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ so that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. Run Algorithm 1 with this choice of parameters and obtain the point X_K , and unnormalized weight $\tilde{w}_K$. Then, for all bounded measurable test functions f , we have the bias and variance estimates

$$(3.3) \quad \left| \frac{\mathbf{E} \tilde{w}_K f(X_K)}{\mathbf{E} \tilde{w}_K} - \langle f, \pi_\varepsilon \rangle \right| \leq 2^{2-T/t_{\text{mix}, 1}^\infty} \|f\|_\infty,$$

$$(3.4) \quad \text{Var} \left(\frac{\tilde{w}_K f(X_K)}{\mathbf{E} \tilde{w}_K} \right) \leq 4 \bar{C}_w \|f\|_\infty^2.$$

Remark 3.6. We will shortly see that $t_{\text{mix}, \varepsilon}^\infty < \infty$ for every $\varepsilon > 0$, but grows exponentially with $1/\varepsilon$ as $\varepsilon \rightarrow 0$. Choosing $T > t_{\text{mix}, \varepsilon}^\infty$ is of course impractical when ε is small, however, the choice of T in (3.2) only requires $T > t_{\text{mix}, \varepsilon_1}^\infty$, which is practically tractable when ε_1 is large.

Proposition 3.5 allows us to bound the error when we perform repeated independent runs of Algorithm 1. This is the main result of this paper.

Theorem 3.7. *Let U be a double well potential that satisfies Assumptions 3.1–3.3, and let $\varepsilon_1 \in (\varepsilon_{\min}, \varepsilon_{\max}]$. Given $\nu > 0$, let $\bar{C}_w, C_T$ be as in Proposition 3.5. For any $\delta > 0$ and $\varepsilon \in (\varepsilon_{\min}, \varepsilon_1]$ choose*

$$\begin{aligned} K &= \left\lceil \frac{1}{\varepsilon \nu} \right\rceil, \\ T &> \max \left\{ C_T \left(\frac{1}{\varepsilon} + \log K \right), (4 + |\log_2 \delta|) t_{\text{mix}, \varepsilon_1}^\infty \right\}, \\ N &= \left\lceil \frac{64 \bar{C}_w}{\delta^2} \right\rceil, \end{aligned}$$

and choose $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ so that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. Perform N independent runs of Algorithm 1 with these parameters and let $(X_K^1, \tilde{w}_K^1), \dots, (X_K^N, \tilde{w}_K^N)$ be the resulting points and unnormalized weights. Define the empirical measure μ_N by (1.6). Then for every bounded test function f we have (1.7). Consequently, for every $s > d/2$ there exists an explicit dimensional constant C_s (independent of $\varepsilon_1, \varepsilon, \nu, \delta$) such that (1.8) holds.

Remark 3.8 (Effective sample size). Proposition 2.7 and (3.4) immediately show that the sum of the square of the normalized weights $w_i, \dots, w_i$ (defined in (1.6)) is bounded above by

$$\mathbf{E} \left(\sum_{i=1}^N w_i^2 \right) \leq \frac{8(4 \bar{C}_w + 1)}{N}.$$

Hence the effective sample size is bounded below by

$$\mathbf{E} \text{ESS}(w_1, \dots, w_N) \geq \frac{N}{8(4\bar{C}_w + 1)}.$$

The proofs of Proposition 3.5 and Theorem 3.7 are in Section 5.

3.3. Convergence results for autonormalized Langevin AIS. We now consider Algorithm 2, which is an auto-normalized version of Algorithm 1, where we re-normalize the weights at every step. This is a generalization of Theorem 1.7 to more general potentials.

Theorem 3.9. *Let U be a double well potential that satisfies Assumptions 3.1–3.3, and let $\varepsilon_1 \in (\varepsilon_{\min}, \varepsilon_{\max}]$. Given $\alpha, \delta, \nu > 0$, there exists constants $C_N(\nu, U)$ and $\hat{C}_T(\alpha, \nu, U)$ such that the following holds. For any $\delta > 0$ and $\varepsilon \in (\varepsilon_{\min}, \varepsilon_1]$, choose*

$$(3.5) \quad K = \left\lceil \frac{1}{\varepsilon \nu} \right\rceil,$$

$$(3.6) \quad T > \hat{C}_T \left(K^{(1+\alpha)\hat{\gamma}_r} + \frac{1}{\varepsilon} + \log\left(\frac{1}{\delta}\right) + \log(N) \right),$$

$$(3.7) \quad N = \left\lceil \frac{C_N}{\delta^2} \right\rceil K^2,$$

and choose $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ so that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. (The constant $\hat{\gamma}_r$ in (3.6) is defined in (3.1).) Let $\{X_K^i, \tilde{w}_K^i\}_{1 \leq i \leq N}$ be the N points and normalized weights returned by Algorithm 2, with $K+1$ levels, running time T , and kernels $\{P_k\}$. For every bounded measurable f , we have

$$(3.8) \quad \left\| \sum_{i=1}^N w_K^i f(X_K^i) - \langle f, \pi_{K+1} \rangle \right\|_{L^2(\mathbf{P})} \leq \|f\|_{\infty} \delta, .$$

Moreover, if $\mu_{K+1,N}$ is the empirical measure defined by

$$\mu_{K+1,N} \stackrel{\text{def}}{=} \sum_{i=1}^N w_K^i \delta_{X_K^i},$$

then for every $s > d/2$,

$$(3.9) \quad \mathbf{E} \left\| \mu_{K+1,N} - \pi_{K+1} \right\|_{H^{-s}}^2 \leq C_s \delta^2.$$

As before, it is important to study the effective sample size. For AIS (Theorems 2.5 and 3.7) the bound for the effective sample size followed directly from the convergence bound (see Proposition 2.7 and Remark 3.8). For the autonormalized version, we first need to first bound the effective sample size in order to prove the convergence bound in Theorem 3.9. We state this as our next result.

Proposition 3.10 (Effective sample size). *Using the same assumptions and notation as Theorem 3.9, there exists a constant $C_1 = C_1(\nu, \varepsilon_1)$ such that for any $k \in \{1, \dots, K\}$ we have*

$$(3.10) \quad \mathbf{E} \left(\sum_{i=1}^N w_{k,i}^2 \right) \leq \frac{C_1 \bar{C}_w^2}{N},$$

where $\bar{C}_w$ is the constant from Proposition 3.5. Consequently, the effective sample size satisfies the lower bound

$$(3.11) \quad \text{E ESS}(w_{k,1}, \dots, w_{k,N}) \geq \frac{N}{C_1 \bar{C}_w^2}.$$

The proofs of Theorem 3.7 and Proposition 3.10 are presented in Section 6.

4. Proof of convergence for AIS

4.1. Bias and Variance estimates (Proposition 2.4). To prove the bias estimate (2.3) we first need an estimate on the unnormalized bias.

Lemma 4.1. *If Assumptions 2.1 and 2.3 hold, then for any bounded measurable function f , we have*

$$(4.1) \quad \left| \mathbf{E}[\tilde{w}_K f(X_K)] - \frac{Z_{K+1}}{Z_1} \langle f, \pi_{K+1} \rangle \right| \leq 2^{-T/t_{\max,1}^\infty} \frac{Z_{K+1}}{Z_1} \langle |f|, \pi_{K+1} \rangle.$$

Proof. For notational convenience, define

$$(4.2) \quad \tilde{\text{Bias}}_i(f) \stackrel{\text{def}}{=} \left| \mathbf{E}[\tilde{w}_i f(X_i)] - \frac{Z_{i+1}}{Z_1} \langle f, \pi_{i+1} \rangle \right|.$$

Notice

$$\tilde{\text{Bias}}_K(f) \leq \tilde{\text{Bias}}_K(f^+) + \tilde{\text{Bias}}_K(f^-),$$

and hence it suffices to establish (4.1) for non-negative functions f . Thus, without loss of generality, we subsequently assume that $f \geq 0$.

Observe that the mean of the estimator can be rewritten as

$$(4.3) \quad \begin{aligned} \mathbf{E}[\tilde{w}_K f(X_K)] &= \mathbf{E}[\tilde{w}_{K-1}(\tilde{r}_K f)(X_K)] \\ &= \mathbf{E}[\tilde{w}_{K-1} P_K^T(\tilde{r}_K f)(X_{K-1})]. \end{aligned}$$

On the other hand, using reversibility, the corresponding true mean satisfies

$$(4.4) \quad \frac{Z_K}{Z_1} \langle P_K^T(\tilde{r}_K f), \pi_K \rangle = \frac{Z_K}{Z_1} \langle \tilde{r}_K f, \pi_K \rangle = \frac{Z_{K+1}}{Z_1} \langle f, \pi_{K+1} \rangle.$$

Combining (4.3), (4.4), and the definition of the bias in (4.2), we obtain

$$\tilde{\text{Bias}}_K(f) = \tilde{\text{Bias}}_{K-1}(P_K^T(\tilde{r}_K f)).$$

Iterating this identity down to level 1 yields

$$(4.5) \quad \tilde{\text{Bias}}_K(f) = \tilde{\text{Bias}}_1(\tilde{f}_2),$$

where

$$\tilde{f}_i \stackrel{\text{def}}{=} S_i S_{i+1} \cdots S_K f, \quad S_k h \stackrel{\text{def}}{=} P_k^T(\tilde{r}_k h).$$

Next, observe that

$$\mathbf{E}[\tilde{w}_1 \tilde{f}_2(X_1)] = P_1^T(\tilde{r}_1 \tilde{f}_2)(X_0), \quad \langle \tilde{r}_1 \tilde{f}_2, \pi_1 \rangle = \frac{Z_2}{Z_1} \langle \tilde{f}_2, \pi_2 \rangle.$$

Since $f \geq 0$, positivity of P_k implies $\tilde{f}_2 \geq 0$. Now Assumption 2.3 implies

$$(4.6) \quad \begin{aligned} \tilde{\text{Bias}}_1(\tilde{f}_2) &= |P_1^T(\tilde{r}_1 \tilde{f}_2)(X_0) - \langle \tilde{r}_1 \tilde{f}_2, \pi_1 \rangle| \\ &\leq 2^{-T/t_{\max,1}^\infty} \langle \tilde{r}_1 \tilde{f}_2, \pi_1 \rangle = 2^{-T/t_{\max,1}^\infty} \frac{Z_2}{Z_1} \langle \tilde{f}_2, \pi_2 \rangle. \end{aligned}$$

Moreover,

$$\begin{aligned}
 \frac{Z_2}{Z_1} \langle \tilde{f}_2, \pi_2 \rangle &= \frac{Z_2}{Z_1} \langle P_2^T(\tilde{r}_2 \tilde{f}_3), \pi_2 \rangle = \frac{Z_3}{Z_1} \langle \tilde{f}_3, \pi_3 \rangle \\
 (4.7) \quad &= \dots = \frac{Z_K}{Z_1} \langle \tilde{f}_K, \pi_K \rangle = \frac{Z_{K+1}}{Z_1} \langle f, \pi_{K+1} \rangle.
 \end{aligned}$$

Combining (4.5), (4.6), and (4.7) yields (4.1), which completes the proof. $\square$

Lemma 4.1 and Assumption 2.2 quickly yield Proposition 2.4.

Proof of Proposition 2.4. Applying (4.1) with the test function $f \equiv 1$, we obtain

$$(4.8) \quad \left| \mathbf{E} \tilde{w}_K - \frac{Z_{K+1}}{Z_1} \right| \leq 2^{-T/t_{\text{mix},1}^\infty} \frac{Z_{K+1}}{Z_1}.$$

Consequently, if $T \geq t_{\text{mix},1}^\infty$, then

$$(4.9) \quad \mathbf{E} \tilde{w}_K \geq \frac{1}{2} \frac{Z_{K+1}}{Z_1}.$$

Next, we estimate

$$\begin{aligned}
 & \left| \mathbf{E} [\tilde{w}_K f(X_K)] - \mathbf{E} \tilde{w}_K \langle f, \pi_{K+1} \rangle \right| \\
 & \leq \left| \mathbf{E} [\tilde{w}_K f(X_K)] - \frac{Z_{K+1}}{Z_1} \langle f, \pi_{K+1} \rangle \right| \\
 & \quad + \left| \frac{Z_{K+1}}{Z_1} \langle f, \pi_{K+1} \rangle - \mathbf{E} \tilde{w}_K \langle f, \pi_{K+1} \rangle \right| \\
 & \stackrel{(4.1)}{\leq} 2^{-T/t_{\text{mix},1}^\infty} \frac{Z_{K+1}}{Z_1} \langle |f|, \pi_{K+1} \rangle + \left| \mathbf{E} \tilde{w}_K - \frac{Z_{K+1}}{Z_1} \right| \langle |f|, \pi_{K+1} \rangle \\
 (4.10) \quad & \stackrel{(4.8)}{\leq} 2^{1-T/t_{\text{mix},1}^\infty} \frac{Z_{K+1}}{Z_1} \langle |f|, \pi_{K+1} \rangle.
 \end{aligned}$$

Combining (4.9) and (4.10) yields the desired normalized bias estimate (2.3).

We now turn to the variance estimate (2.4). It suffices to prove the bound

$$(4.11) \quad \text{Var}(\tilde{w}_K f(X_K)) \leq \left(\frac{Z_{K+1}}{Z_1} \right)^2 C_w \|f\|_\infty^2.$$

Indeed, combining this with (4.9) immediately implies (2.4).

To prove (4.11), observe that

$$\begin{aligned}
 \text{Var}(\tilde{w}_K f(X_K)) &\leq \mathbf{E} [\tilde{w}_K^2 f(X_K)^2] \leq \mathbf{E} \tilde{w}_K^2 \|f\|_\infty^2 \\
 &= \mathbf{E} [\tilde{w}_{K-1}^2 \tilde{r}_K(X_K)^2] \|f\|_\infty^2 \\
 &= \mathbf{E} [\tilde{w}_{K-1}^2 P_K^T(\tilde{r}_K^2)(X_{K-1})] \|f\|_\infty^2 \\
 &\leq \mathbf{E} \tilde{w}_{K-1}^2 \|P_K^T \tilde{r}_K^2\|_\infty \|f\|_\infty^2.
 \end{aligned}$$

Iterating this argument over $k = 1, \dots, K$ yields

$$(4.12) \quad \text{Var}(\tilde{w}_K f(X_K)) \leq \left(\prod_{k=1}^K \|P_k^T \tilde{r}_k^2\|_\infty \right) \|f\|_\infty^2.$$

Finally, we note that

$$\prod_{k=1}^K \|P_k^T \tilde{r}_k^2\|_\infty = \left(\frac{Z_{K+1}}{Z_1}\right)^2 \prod_{k=1}^K \|P_k^T r_k^2\|_\infty \stackrel{(2.1)}{\leq} C_w \left(\frac{Z_{K+1}}{Z_1}\right)^2.$$

Combining this with (4.12) proves (4.11), and hence (2.4). This completes the proof. $\square$

4.2. Convergence of the empirical measure (Theorem 2.5). We now use Proposition 2.4 to prove Theorem 2.5.

Proof of Theorem 2.5. We first prove (2.6). Define

$$R \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^N \tilde{w}_K^i.$$

From the definition of w_K^i in (2.6), we have

$$\frac{1}{N} \tilde{w}_K^i = R w_K^i.$$

Adding and subtracting the same quantity and applying the triangle inequality yields, for any $x_i \in \mathbb{R}^d$,

$$(4.13) \quad \left| \sum_{i=1}^N w_K^i f(x_i) - \langle f, \pi_{K+1} \rangle \right| \leq I_1 + I_2,$$

where

$$\begin{aligned} I_1 &\stackrel{\text{def}}{=} \left| \sum_{i=1}^N w_K^i f(x_i) - \frac{1}{\mathbf{E} \tilde{w}_K^1} \sum_{i=1}^N R w_K^i f(x_i) \right|, \\ I_2 &\stackrel{\text{def}}{=} \left| \frac{1}{\mathbf{E} \tilde{w}_K^1} \sum_{i=1}^N \frac{1}{N} \tilde{w}_K^i f(x_i) - \langle f, \pi_{K+1} \rangle \right|. \end{aligned}$$

We observe that

$$(4.14) \quad \begin{aligned} I_1 &= \left| 1 - \frac{R}{\mathbf{E} \tilde{w}_K^1} \right| \left| \sum_{i=1}^N w_K^i f(x_i) \right| \\ &\leq \left| 1 - \frac{R}{\mathbf{E} \tilde{w}_K^1} \right| \|f\|_\infty = \left| \frac{1}{N} \sum_{i=1}^N \left(\frac{\tilde{w}_K^i}{\mathbf{E} \tilde{w}_K^1} - 1 \right) \right| \|f\|_\infty, \end{aligned}$$

$$(4.15) \quad I_2 = \left| \frac{1}{N} \sum_{i=1}^N \left(\frac{\tilde{w}_K^i f(x_i)}{\mathbf{E} \tilde{w}_K^1} - \langle f, \pi_{K+1} \rangle \right) \right|.$$

We recall that for any i.i.d. random variables $(Y_i)_{i=1}^N$ in $L^2(\mathbf{P})$,

$$(4.16) \quad \mathbf{E} \left[\left(\frac{1}{N} \sum_{i=1}^N Y_i \right)^2 \right] = \frac{1}{N} \text{Var}(Y_1) + (\mathbf{E} Y_1)^2.$$

Applying (4.13) with $x_i = X_K^i$ and combining (4.14), (4.15), and (4.16), we obtain

$$(4.17) \quad \left\| \sum_{i=1}^N w_K^i f(X_K^i) - \langle f, \pi_{K+1} \rangle \right\|_{L^2(\mathbf{P})} \leq I_3 + I_4,$$

where

$$(4.18) \quad I_3 = \frac{\|f\|_{L^\infty}}{\sqrt{N}} \operatorname{Var}\left(\frac{\tilde{w}_K^i}{\mathbf{E}\tilde{w}_K^i}\right)^{1/2} \stackrel{(2.4)}{\leq} \|f\|_\infty \sqrt{\frac{4C_w}{N}},$$

$$(4.19) \quad I_4 = \frac{1}{\sqrt{N}} \operatorname{Var}\left(\frac{\tilde{w}_K^i f(X_K^i)}{\mathbf{E}\tilde{w}_K^i}\right)^{1/2} + \left| \mathbf{E} \frac{\tilde{w}_K^i f(X_K^i)}{\mathbf{E}\tilde{w}_K^i} - \langle f, \pi_{K+1} \rangle \right|$$

$$\stackrel{(2.4), (2.3)}{\leq} \sqrt{\frac{4C_w}{N}} \|f\|_\infty + 2^{2-T/t_{\min,1}^\infty} \|f\|_\infty.$$

Finally, choosing N and T as in (2.5) and combining (4.17), (4.18), and (4.19) proves (2.6), as desired.

We now prove (2.7). The bound (2.7) follows quickly from (2.6), and the Fourier representation of the $\dot{H}^{-s}$ norm. Recall, if φ is a distribution in the homogeneous Sobolev space $H^{-s} = H^{-s}(\mathbb{T}^d)$, then the norm is equivalently defined by

$$(4.20) \quad \|\varphi\|_{\dot{H}^{-s}}^2 \stackrel{\text{def}}{=} \sum_{n \in \mathbb{Z}^d - \{0\}} \frac{|\langle \varphi, e_n \rangle|^2}{|n|^{2s}} \quad \text{where} \quad e_n(x) \stackrel{\text{def}}{=} e^{2\pi i n \cdot x}.$$

Now, we note that for any $n \in \mathbb{Z}^d - \{0\}$, the bound (2.6) implies

$$\mathbf{E} \langle \mu_{K+1,N} - \pi_{K+1}, e_n \rangle^2 \leq \delta^2.$$

Thus

$$\begin{aligned} \mathbf{E} \|\mu_{K+1,N} - \pi_{K+1}\|^2 &= \sum_{n \in \mathbb{Z}^d - \{0\}} \mathbf{E} \frac{\langle \mu_{K+1,N} - \pi_{K+1}, e_n \rangle^2}{|n|^{2s}} \leq \sum_{n \in \mathbb{Z}^d - \{0\}} \frac{\delta^2}{|n|^{2s}} \\ &\leq C_s \delta^2, \end{aligned}$$

where

$$C_s \stackrel{\text{def}}{=} \sum_{n \in \mathbb{Z}^d - \{0\}} \frac{1}{|n|^{2s}}.$$

This concludes the proof. $\square$

4.3. The effective sample size (Proposition 2.7). We now prove Proposition 2.7 bounding the effective sample size. The bound follows from a more general fact about normalized sums of nonnegative i.i.d. random variables.

Proof of Proposition 2.7. Let $\{\tilde{\zeta}_i\}_{1 \leq i \leq N}$ be non-negative i.i.d. L^2 random variables with $\tilde{\mu} = \mathbf{E}[\tilde{\zeta}_1]$ and $\tilde{\sigma}^2 = \operatorname{Var}(\tilde{\zeta}_1)$ and define

$$\zeta_i = \frac{\tilde{\zeta}_i}{S_N} \quad \text{where} \quad S_N = \sum_{i=1}^N \tilde{\zeta}_i.$$

We claim that we must have

$$(4.21) \quad \mathbf{E} \left[\sum_{i=1}^N \zeta_i^2 \right] \leq \frac{8(\tilde{\sigma}^2 + \tilde{\mu}^2)}{N\tilde{\mu}^2}.$$

Momentarily postponing the proof of (4.21), we now prove (2.8) and (2.9). Setting $\tilde{\zeta}_i = \tilde{w}_K^i$ and taking $f \equiv 1$ in (2.4) gives

$$\tilde{\sigma}^2 \leq 4C_w \tilde{\mu}^2.$$

Combining this with (4.21) yields (2.8) as desired. Using this and Jensen's inequality, implies the lower bound on the effective sample size in (2.9).

It remains to prove the claim (4.21). Let $\bar{S}_N = S_N/N$. Then

$$\begin{aligned} \mathbf{E} \left[\sum_{i=1}^N \zeta_i^2 \right] &= N \mathbf{E} \left[\frac{\tilde{\zeta}_1^2}{S_N^2} \mathbf{1}_{\{\bar{S}_N \geq \frac{\mu}{2}\}} \right] + \mathbf{E} \left[\frac{1}{S_N^2} \sum_{i=1}^N \tilde{\zeta}_i^2 \mathbf{1}_{\{\bar{S}_N < \frac{\mu}{2}\}} \right] \\ &\leq \frac{4(\tilde{\sigma}^2 + \tilde{\mu}^2)}{N\tilde{\mu}^2} + \mathbf{P} \left[|\bar{S}_N - \tilde{\mu}| > \frac{\tilde{\mu}}{2} \right] \\ &\leq \frac{8(\tilde{\sigma}^2 + \tilde{\mu}^2)}{N\tilde{\mu}^2}. \end{aligned}$$

Here we used $\sum_{i=1}^N \zeta_i^2 \leq 1$ to obtain the first inequality, and then Chebyshev's inequality to obtain the second inequality. This concludes the proof. $\square$

4.4. The χ^2 divergence (Propositions 2.8 and 2.9). The proof of Proposition 2.9 is short and direct.

Proof of Proposition 2.8. Notice $P_k^\infty r_k^2 = \langle r_k^2, \pi_k \rangle$. Thus

$$\begin{aligned} \prod_{k=1}^K \|P_k^\infty r_k^2\|_{L^\infty} &= \prod_{k=1}^K \langle r_k^2, \pi_k \rangle = \exp \left(\sum_{k=1}^K \ln \langle r_k^2, \pi_k \rangle \right) \\ (4.22) \quad &\leq \exp \left(\sum_{k=1}^K (\langle r_k^2, \pi_k \rangle - 1) \right) = \exp \left(\sum_{k=1}^K \chi^2(\pi_{k+1}; \pi_k) \right), \end{aligned}$$

which proves (2.10). $\square$

The proof of Proposition 2.9 is also a direct calculation.

Proof of Proposition 2.9. For the second assertion, the choice of temperatures described in the statement reduces to choosing

$$(4.23) \quad \varepsilon_k = \frac{\varepsilon K}{(k-1)(1-\varepsilon) + \varepsilon K}.$$

Since π_k is the Gaussian $\mathcal{N}(0, \varepsilon_k I_d)$, the χ^2 divergence can be computed explicitly. Using, for instance Proposition 17.1 in [CP20], we see

$$(4.24) \quad \chi^2(\pi_{k+1}, \pi_k) = \left(\frac{\varepsilon_k^2}{2\varepsilon_k \varepsilon_{k+1} - \varepsilon_{k+1}^2} \right)^{d/2} - 1.$$

To obtain (2.11), set

$$\delta_k \stackrel{\text{def}}{=} 1 - \frac{\varepsilon_{k+1}}{\varepsilon_k} = \left(\frac{1-\varepsilon}{\varepsilon K} \right) \varepsilon_{k+1},$$

and use (4.24) to obtain

$$(4.25) \quad \chi^2(\pi_{k+1}; \pi_k) = \left(1 + \frac{\delta_k^2}{1 - \delta_k^2} \right)^{d/2} - 1 \leq Cd \delta_{k+1}^2 \leq \frac{Cd \varepsilon_{k+1}^2}{\varepsilon^2 K^2}.$$

Hence using (4.22), (4.23) and (4.25) we obtain

$$\prod_{k=1}^K \|P_k^\infty r_k^2\| \leq \exp \left(\sum_{k=1}^K \frac{Cd}{((k-1)(1-\varepsilon) + \varepsilon K)^2} \right) \leq \exp \left(\frac{Cd}{\varepsilon K - 1} \right) \leq C,$$

proving (2.11) as desired. $\square$

5. Proof of convergence for Langevin AIS

We now prove Theorem 3.7. We assume, without loss of generality, that the first initial temperature $\varepsilon_1 = 1$.

5.1. Proofs of Proposition 3.5 and Theorem 3.7. We prove Proposition 3.5 and Theorem 3.7 by verifying that the transition kernels for (1.2) satisfy Assumptions 2.1–2.3, and then use Proposition 2.4 and Theorem 2.5. The fact that Assumption 2.1 holds is well known. Checking Assumptions 2.2 and 2.3 require a little work and we state them as the following lemmas.

Lemma 5.1. *Let $P_{\varepsilon,t}$ be the transition kernel of (1.2) at time t . If U satisfies Assumptions 3.1–3.2, then the uniform mixing time of P_1 is finite (i.e. for $\varepsilon_1 = 1$, P_{ε_1} satisfies Assumption 2.3).*

Lemma 5.2. *Let U be a double well potential that satisfies Assumption 3.1–3.3 with $\varepsilon_{\min} < 1 \leq \varepsilon_{\max}$. There exists constants $\bar{C}_w = \bar{C}_w(U, \nu)$ and $C_T = C_T(U)$ such that the following holds. For any $\varepsilon \in (\varepsilon_{\min}, 1]$, choose $K = \lceil 1/(\varepsilon\nu) \rceil$ and choose $\varepsilon_2, \dots, \varepsilon_{K+1} = \varepsilon$ so that $\{1/\varepsilon_k\}_{1 \leq k \leq K+1}$ are linearly spaced. Then, Assumption 2.2 holds with*

$$(5.1) \quad C_w = \bar{C}_w \quad \text{and} \quad T_0 = C_T \left(\frac{1}{\varepsilon} + \log K \right).$$

Given Lemmas 5.1 and 5.2, the proofs of Proposition 3.5 and Theorem 3.7 are immediate.

Proof of Proposition 3.5. It is well known (see for instance Chapter 8 in [Kol00], or Chapter 4.6 in [Pav14]) that $P_{\varepsilon,\cdot}$, the transition kernel of (1.2), is reversible, and π_ε is the unique stationary distribution. The choice of K in (3.2), and by Lemmas 5.1–5.2, will now guarantee that Assumptions 2.1–2.3 are satisfied, with the constants C_w and T_0 given in (5.1). Therefore, Proposition 2.4 applies and yields the bias and variance estimates (3.3)–(3.4). $\square$

Proof of Theorem 3.7. As with the proof of Proposition 3.5 presented above, Assumptions 2.1–2.3 hold, with C_w and T_0 as specified in (5.1). Consequently, Theorem 2.5 applies and yields (1.7)–(1.8). $\square$

It remains to prove Lemmas 5.1 and 5.2. Their proofs require certain spectral estimates, which are described in Section 5.2. Following this, we prove Lemmas 5.1, 5.2 in Section 5.3.

5.2. Spectral decomposition. Let $\mathcal{L}_\varepsilon$, defined by

$$\mathcal{L}_\varepsilon f = -\nabla U \cdot \nabla f + \varepsilon \Delta f,$$

be the generator of (1.2). We know that for any test function f , the action of $P_{\varepsilon,t}$ on f satisfies the *Kolmogorov backward equation* (see for instance [Pav14, Chapter 2]). That is, if

$$u_t(x) = P_{\varepsilon,t} f(x) \stackrel{\text{def}}{=} \int_{\mathbb{T}^d} P_{\varepsilon,t}(x, dy) f(y),$$

then

$$(5.2) \quad \partial_t u = \mathcal{L}_\varepsilon u \quad \text{and} \quad u_0 = f.$$

We now state certain spectral properties of (1.2) which will be used in the proofs of Lemmas 5.1 and 5.2. Let $L^2(\pi_\varepsilon)$ denote the weighted L^2 space with inner-product and norm defined by

$$\langle f, g \rangle_{L^2(\pi_\varepsilon)} \stackrel{\text{def}}{=} \int_{\mathbb{T}^d} f g \pi_\varepsilon dx \quad \text{and} \quad \|f\|_{L^2(\pi_\varepsilon)}^2 \stackrel{\text{def}}{=} \int_{\mathbb{T}^d} |f|^2 \pi_\varepsilon dx,$$

respectively. It is well known (see for instance [Kol00, Chapter 8]) that the operator $-\mathcal{L}_\varepsilon$ is a self-adjoint operator the weighted space $L^2(\pi_\varepsilon)$, and has a discrete spectrum with positive eigenvalues. We denote the eigenvalues by

$$(5.3) \quad 0 = \lambda_{1,\varepsilon} < \lambda_{2,\varepsilon} \leq \lambda_{3,\varepsilon} \cdots,$$

and the corresponding $L^2(\pi_\varepsilon)$ normalized eigenfunctions by $\psi_{1,\varepsilon}$, $\psi_{2,\varepsilon}$, etc. We note that the smallest eigenvalue $\lambda_{1,\varepsilon}$ is 0, and the second smallest eigenvalue $\lambda_{2,\varepsilon}$ is strictly positive (i.e. P_ε has a spectral gap). In fact, for the proof of Lemmas 5.1 and 5.2 the properties we need were collected in a convenient form in [HIS26] as Properties 4.6–4.8. We quote the portions we need here for easy reference.

Property 5.3 (Eigenvalue bounds). *For every $H > \hat{U}$, there exists a constant C_H such that for every $\varepsilon \in (0, 1]$ we have*

$$(5.4) \quad \lambda_{2,\varepsilon} \geq C_H \exp\left(-\frac{H}{\varepsilon}\right).$$

Also, there exists Λ such that for all $\varepsilon \in (0, 1]$ such that

$$(5.5) \quad \lambda_{i,\varepsilon} \geq \Lambda, \quad \text{for all } i \geq 3.$$

Property 5.4 (Eigenfunction variation). *The function $\varepsilon \mapsto \pi_\varepsilon(\Omega_1)$ is of bounded variation on the interval $(0, 1]$. Moreover, for every $\gamma < \hat{\gamma}$, there exists a constant C_γ such that for every $0 < \varepsilon' < \varepsilon \leq 1$ we have*

$$(5.6) \quad |\langle \psi_{2,\varepsilon}, \pi_{\varepsilon'} \rangle| \leq C_\gamma \left(\exp\left(-\frac{\gamma}{\varepsilon}\right) + |\pi_{\varepsilon'}(\Omega_1) - \pi_\varepsilon(\Omega_1)| \right).$$

Property 5.5 (Eigenfunction bounds). *There exists constant C_{ψ_2} , independent of ε such that*

$$(5.7) \quad \sup_{0 < \varepsilon \leq 1} \|\psi_{2,\varepsilon}\|_{L^\infty(\mathbb{T}^d)} \leq C_{\psi_2}.$$

Under Assumptions 3.1–3.1, Section 9 in [HIS26] shows that Properties 5.3–5.5 hold. To briefly discuss the significance of Properties 5.3–5.5, we note that convergence of (reversible) Markov processes can effectively be studied using the spectral decomposition (see for instance Chapter 12 in [LP17]). In particular, for reversible Markov processes the rate of convergence is controlled both above and below by the *spectral gap*, which in our case is simply $\lambda_{2,\varepsilon}$.

For the Langevin system with a multimodal potential, the lower bound (5.4) is sharp (see [Kol00, Chapter 8, Proposition 2.2], or [MS14]), and so spectral gap is exponentially small. Thus sampling by simulating (1.2) directly, will necessarily cost $O(e^{\hat{U}/\varepsilon})$, which is not desirable.

However, in our situation the third eigenvalue is large (i.e. bounded independent of ε), as asserted by (5.5) in Property 5.3. This will give fast convergence *provided* we control the projection onto the second eigenspace. Variants of this idea have been used by several authors in many contexts to accelerate convergence [CKRZ08, KLL⁺13, FI19]. A warm start to Langevin dynamics will also

control the projection on the second eigenspace, and this was recently used by Koehler, Lee, and Vuong [KLTV25] in a related multimodal sampling problem.

In [HIS26], the authors related the projection onto the second eigenspace as a *mass imbalance*, and controlled it by resampling points with weights proportional to the ratio of the densities. This is a standard technique used in sequential Monte Carlo algorithms (see for instance [DdFG01, Liu08, CP20]), and is applicable in many situations of practical interest. The disadvantage of this approach, however, is that one loses independence of realizations, and as a result the estimates are harder to obtain and weaker than in the i.i.d. case.

In our situation, we reweight points instead of resampling, and the variance is controlled by the product (2.1) appearing in the constant C_w . That is, the reweighting controls the error in terms of how fast the kernel P_{ε_k} mixes $r_{\varepsilon_k}^2$. This in turn is controlled by the projection onto the second eigenspace, which we will bound using Properties 5.4 and 5.5 (see Lemma 5.7, below for the precise estimates).

We now use Properties 5.3–5.5 to prove Lemmas 5.1 and 5.2. We first control the high order terms in the spectral decomposition, and then control the terms in the product (2.1). These steps are stated as the next two lemmas.

Lemma 5.6. *Let $\varepsilon > 0$. Suppose there exist constants $\Lambda_1 > 0$ and $m \in \mathbb{N}$ such that $\lambda_{m,\varepsilon} \geq \Lambda_1$. Then there exists a constant $\tilde{C}_\psi = \tilde{C}_\psi(\|U\|_{C^{\lceil d/2 \rceil \vee 2}}, d, \Lambda_1)$, independent of ε , such that for $T_0(\Lambda_1) \stackrel{\text{def}}{=} (d+2)/\Lambda_1$, and any $T \geq T_0$,*

$$(5.8) \quad \sum_{i=m}^{\infty} e^{-\lambda_{i,\varepsilon} T} \|\psi_{i,\varepsilon}\|_{\infty}^2 \leq \tilde{C}_\psi Z_\varepsilon \min\{\varepsilon, 1\}^{-(d+1)} \exp\left(\frac{\|U\|_{\infty}}{\varepsilon} - \Lambda_1 T\right).$$

Lemma 5.7. *Let $\tilde{C}_\psi$ be the constant from Lemma 5.6 with $\Lambda_1 = \Lambda$ from Property 5.3. Let C_γ, C_{ψ_2} be the constants defined as in Property 5.4 and 5.5, respectively. Then, there exists a constant $C_Z(d, U)$ such that if*

$$(5.9) \quad T \geq \frac{1}{\Lambda} \max\left\{\frac{\|U\|_{\infty}}{\varepsilon} + \log K + \left(\frac{d}{2} + 1\right) |\ln \varepsilon| + \log(\tilde{C}_\psi C_Z), d + 2\right\},$$

then for each level $k \in \{1, \dots, K\}$,

$$(5.10) \quad \begin{aligned} \|P_{\varepsilon_k, T} \tilde{r}_k^2\|_{L^\infty} &\leq \frac{Z_{\varepsilon_k+2}}{Z_{\varepsilon_k}} \left(1 + \frac{1}{K}\right. \\ &\quad \left.+ C_{\psi_2} C_\gamma \left(\exp\left(-\frac{\hat{\gamma}}{2\varepsilon_k}\right) + |\pi_{\varepsilon_{k+1}}(\Omega_1) - \pi_{\varepsilon_k}(\Omega_1)|\right)\right). \end{aligned}$$

Lemmas 5.6 and 5.7 are proved in Sections 5.5 and 5.4 respectively.

5.3. Proofs of Lemmas 5.1 and 5.2. We now use Lemmas 5.6 and 5.7 to prove Lemmas 5.1 and 5.2. For the remainder of this section, we slightly abuse notation and write $Z_k, \pi_k, P_{k,t}, \psi_{i,k}, \lambda_{i,k}$ in place of $Z_{\varepsilon_k}, \pi_{\varepsilon_k}, P_{\varepsilon_k,t}, \psi_{i,\varepsilon_k}, \lambda_{i,\varepsilon_k}$.

Proof of Lemma 5.1. It suffices to prove that there exists $T_0 > 0$ such that for any $f \in L^1(\pi_1)$,

$$(5.11) \quad \|P_{1,T_0} f - \langle f, \pi_1 \rangle\|_{L^\infty} \leq \frac{1}{2} \|f\|_{L^1(\pi_1)}.$$

Indeed, set $\varepsilon = \varepsilon_1 = 1$, $\Lambda_1 = \lambda_{2,1}$, and $m = 2$ in Lemma 5.6. It follows that there exists a constant $C_{\varepsilon_1} > 0$ such that for any $T \geq (d+2)/\lambda_{2,1}$,

$$\sum_{i=2}^{\infty} e^{-\lambda_{i,1}T} \|\psi_{i,1}\|_{\infty}^2 \leq C_{\varepsilon_1} \exp(-\lambda_{2,1}T).$$

Then, using Hölder's inequality yields

$$\begin{aligned} \|P_{1,T}f - \langle f, \pi_1 \rangle\|_{L^\infty} &= \left\| \sum_{i=2}^{\infty} \exp(-\lambda_{i,1}T) \langle f, \psi_{i,1} \rangle_{L^2(\pi_1)} \psi_{i,1} \right\|_{L^\infty} \\ &\leq \sum_{i=2}^{\infty} \exp(-\lambda_{i,1}T) \|\psi_{i,1}\|_{L^\infty}^2 \|f\|_{L^1(\pi_1)} \\ &\leq C_{\varepsilon_1} \exp(-\lambda_{2,1}T) \|f\|_{L^1(\pi_1)}. \end{aligned}$$

Hence, choosing $T_0 = \max\{d+2, \log(2C_{\varepsilon_1})\}/\lambda_{2,1}$ implies (5.11) and concludes the proof. $\square$

Proof of Lemma 5.2. Choosing

$$(5.12) \quad C_T = \frac{1}{\Lambda} \max \left\{ \|U\|_{\infty} + \left(\frac{d}{2} + 1 \right) + \log(\tilde{C}_\psi C_Z), d+2 \right\},$$

and T_0 as in (5.1) implies that if $T \geq T_0$ then (5.9) holds. Thus we may apply Lemma 5.7 to obtain

$$(5.13) \quad \prod_{k=1}^K \|P_{k,T} \tilde{r}_k^2\|_{L^\infty} \leq \frac{Z_{K+1} Z_{K+2}}{Z_1 Z_2} \prod_{k=1}^K \Theta(k, k+1),$$

where

$$(5.14) \quad \Theta(k, k+1) \stackrel{\text{def}}{=} \left(1 + C_{\psi_2} C_\gamma \left(\exp\left(-\frac{\hat{\gamma}}{2\varepsilon_k}\right) + |\pi_{k+1}(\Omega_1) - \pi_k(\Omega_1)| \right) + \frac{1}{K} \right).$$

By the AM-GM inequality this implies

$$\prod_{k=1}^K \|P_{k,T} \tilde{r}_k^2\|_{L^\infty} \leq \frac{Z_{K+1} Z_{K+2}}{Z_1 Z_2} \left(\frac{1}{K} \sum_{k=1}^K \Theta(k, k+1) \right)^K.$$

Since $\{1/\varepsilon_k\}$ are linearly spaced between 1 and $1/\varepsilon$, they are given by (4.23), and hence

$$(5.15) \quad \sum_{k=1}^K \exp\left(-\frac{\hat{\gamma}}{2\varepsilon_k}\right) \leq \frac{e^{-\hat{\gamma}/2}}{1 - \exp\left(\frac{-\hat{\gamma}}{2}(\nu-1)\right)} \stackrel{\text{def}}{=} C_{\text{geom}} < \infty,$$

where we note that C_{geom} is an explicit constant that is independent of ε .

Moreover,

$$(5.16) \quad \sum_{k=1}^K |\pi_{k+1}(\Omega_1) - \pi_k(\Omega_1)| \leq \int_{\varepsilon}^1 |\partial_{\varepsilon'} \pi_{\varepsilon'}(\Omega_1)| d\varepsilon'.$$

Assumptions 3.1–3.3 and Lemma 8.2 in [HIS26] imply the existence of a constant $C_{\text{BV}} > 0$ such that

$$(5.17) \quad \int_{\varepsilon}^1 |\partial_{\varepsilon'} \pi_{\varepsilon'}(\Omega_1)| d\varepsilon' \leq C_{\text{BV}}.$$

Combining (5.15), (5.16) and (5.17) we obtain

$$(5.18) \quad \left(\frac{1}{K} \sum_{k=1}^K \Theta(k, k+1) \right)^K \leq \left(1 + \frac{C_{\psi_2} C_\gamma (C_{\text{geom}} + C_{\text{BV}}) + 1}{K} \right)^K \leq C_\Theta,$$

where

$$C_\Theta \stackrel{\text{def}}{=} \exp(C_{\psi_2} C_\gamma (C_{\text{geom}} + C_{\text{BV}}) + 1).$$

Combining (5.13) and (5.18), together with the fact that

$$\prod_{k=1}^K \|P_{k,T} \tilde{r}_k^2\|_{L^\infty} = \frac{Z_1^2}{Z_{K+1}^2} \prod_{k=1}^K \|P_{k,T} \tilde{r}_k^2\|_{L^\infty},$$

and $Z_{K+2} \leq Z_{K+1}$, proves (2.1) with

$$(5.19) \quad C_w = 2C_\Theta. \quad \square$$

5.4. Proof of Lemma 5.7. To complete the proofs of Proposition 3.5 and Theorem 3.7, it only remains to prove Lemmas 5.6 and 5.7. Since the proof of Lemma 5.7 is shorter, we present it first.

Proof of Lemma 5.7. Using (5.2) and the spectral decomposition of $\mathcal{L}_\varepsilon$, we have

$$P_{k,T} \tilde{r}_k^2 = \langle \tilde{r}_k^2, \pi_k \rangle + e^{-\lambda_{2,k} T} \langle \tilde{r}_k^2, \psi_{2,k} \rangle_{L^2(\pi_k)} \psi_{2,k} + \sum_{i=3}^{\infty} e^{-\lambda_{i,k} T} \langle \tilde{r}_k^2, \psi_{i,k} \rangle_{L^2(\pi_k)} \psi_{i,k}.$$

Notice

$$(5.20) \quad \langle \tilde{r}_k^2, \tilde{\pi}_k \rangle = \int_{\mathbb{T}^d} \exp\left(-U\left(\frac{2}{\varepsilon_{k+1}} - \frac{1}{\varepsilon_k}\right)\right) dx \stackrel{(4.23)}{=} \int_{\mathbb{T}^d} \exp\left(\frac{-U}{\varepsilon_{k+2}}\right) dx = Z_{k+2}.$$

Similarly,

$$(5.21) \quad \langle \tilde{r}_k^2, \psi_{2,k} \rangle_{L^2(\pi_k)} = \frac{Z_{k+2}}{Z_k} \langle \psi_{2,k}, \pi_{k+2} \rangle,$$

$$(5.22) \quad |\langle \tilde{r}_k^2, \psi_{i,k} \rangle_{L^2(\pi_k)}| \leq \|\tilde{r}_k^2\|_{L^1(\pi_k)} \|\psi_{i,k}\|_{L^\infty},$$

and hence

$$(5.23) \quad \|P_{k,T} \tilde{r}_k^2\|_{L^\infty} \leq \frac{Z_{k+2}}{Z_k} \left(1 + |\langle \psi_{2,k}, \pi_{k+2} \rangle| \|\psi_{2,k}\|_{L^\infty} + \sum_{i=3}^{\infty} e^{-\lambda_{i,k} T} \|\psi_{i,k}\|_{L^\infty}^2 \right).$$

We now estimate the second and third terms of the right hand side separately. For the second term, using (5.7) and (5.6) for $\gamma = \hat{\gamma}/2$ yields

$$(5.24) \quad |\langle \psi_{2,k}, \pi_{k+2} \rangle| \|\psi_{2,k}\|_{L^\infty} \leq C_{\psi_2} C_\gamma \left(\exp\left(-\frac{\hat{\gamma}}{2\varepsilon_k}\right) + |\pi_{k+2}(\Omega_1) - \pi_k(\Omega_1)| \right).$$

To estimate the third term, we first apply Lemma 5.6. Set $\varepsilon = \varepsilon_k$, $\Lambda_1 = \Lambda$, and $m = 3$ in Lemma 5.6. It follows that there exists a constant $\tilde{C}_\psi > 0$, independent of ε , such that for any

$$(5.25) \quad T \geq \frac{d+2}{\Lambda},$$

we have

$$(5.26) \quad \sum_{i=3}^{\infty} e^{-\lambda_{i,k} T} \|\psi_{i,k}\|_{L^\infty}^2 \leq \tilde{C}_\psi Z_k \varepsilon_k^{-(d+1)} \exp\left(\frac{\|U\|_\infty}{\varepsilon_k} - \Lambda T\right).$$

In particular, if T further satisfies

$$(5.27) \quad T \geq \frac{1}{\Lambda} \left(\frac{\|U\|_\infty}{\varepsilon_k} + \log \left(\frac{\tilde{C}_\psi K Z_k}{\varepsilon_k^{d+1}} \right) \right),$$

then combining this with (5.26) yields

$$(5.28) \quad \sum_{i=3}^{\infty} e^{-\lambda_{i,k} T} \|\psi_{i,k}\|_{L^\infty}^2 \leq \frac{1}{K}.$$

Finally, we estimate the normalizing constant Z_k . By the Laplace method (see for instance [Kol00, Proposition B2]),

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{(2\pi\varepsilon)^{d/2} \sqrt{|\det \nabla^2 U(x_{\min,1})|}} \int_{\Omega_1} e^{-U/\varepsilon} dx = 1.$$

This implies that there exists a constant $C_Z(d, U)$ such that

$$Z_\varepsilon \leq C_Z(d, U) \varepsilon^{\frac{d}{2}}, \quad \text{for all } \varepsilon \in (0, 1].$$

It follows that the choice of T in (5.9) ensures that T is sufficiently large to satisfy (5.25) and (5.27), so that (5.28) holds. Combining (5.23), (5.24), and (5.28) then yields (5.10). $\square$

5.5. Proof of Lemma 5.6. We now address Lemma 5.6. We will bound the left hand side of (5.30) by obtaining a uniform in i bound on $\|\psi_{i,\varepsilon}\|_\infty$, and then bound the sum using the asymptotic growth of the eigenvalues. We state this as our next two lemmas.

Lemma 5.8. *Define the Weyl function by*

$$N_{\mathcal{L}_\varepsilon}(\lambda) = |\{i \in \mathbb{N} \mid \lambda_{i,\varepsilon} \leq \lambda\}| = \sum_{\{i: \lambda_{i,\varepsilon} \leq \lambda\}} 1.$$

where $\lambda_{i,\varepsilon}$ are the eigenvalues of $-\mathcal{L}_\varepsilon$, ordered as in (5.3). There exists a constant $C_W(\|U\|_{C^2}, d)$ such that for any $\varepsilon > 0$ and $\lambda > 0$,

$$(5.29) \quad N_{\mathcal{L}_\varepsilon}(\lambda) \leq C_W \varepsilon^{-\frac{d}{2}} (1 + \lambda)^{\frac{d}{2}}.$$

Lemma 5.9. *For any $\Lambda_1 > 0$, there exists a constant $C_{\Lambda_1}(\|U\|_{C^{1+\lfloor d/2 \rfloor}}, d, \Lambda_1) > 0$ such that for all $\varepsilon > 0$ and any eigenvalue $\lambda_{i,\varepsilon} \geq \Lambda_1$, we have*

$$(5.30) \quad \|\psi_{i,\varepsilon}\|_{L^\infty(\mathbb{T}^d)}^2 \leq C_{\Lambda_1} Z_\varepsilon \exp\left(\frac{\|U\|_\infty}{\varepsilon}\right) \left(\frac{\lambda_{i,\varepsilon}}{\varepsilon}\right)^{1+\lfloor \frac{d}{2} \rfloor}.$$

Momentarily postponing the proofs of Lemmas 5.8 and 5.9, we now prove Lemma 5.6.

Proof of Lemma 5.6. By Lemma 5.9, for any $T \geq T_0$, we obtain

$$(5.31) \quad \begin{aligned} \sum_{i=m}^{\infty} e^{-\lambda_{i,\varepsilon} T} \|\psi_{i,\varepsilon}\|_\infty^2 &\stackrel{(5.30)}{\leq} C_{\Lambda_1} Z_\varepsilon \exp\left(\frac{\|U\|_\infty}{\varepsilon}\right) \sum_{i=m}^{\infty} \left(\frac{\lambda_{i,\varepsilon}}{\varepsilon}\right)^{1+\lfloor \frac{d}{2} \rfloor} e^{-\lambda_{i,\varepsilon} T} \\ &\leq C_{\Lambda_1} Z_\varepsilon \exp\left(\frac{\|U\|_\infty}{\varepsilon} - \Lambda_1(T - T_0)\right) \sum_{i=m}^{\infty} \left(\frac{\lambda_{i,\varepsilon}}{\varepsilon}\right)^{1+\lfloor \frac{d}{2} \rfloor} e^{-\lambda_{i,\varepsilon} T_0}. \end{aligned}$$

Next, we estimate the remaining term $\sum_{i=m}^{\infty} \lambda_{i,\varepsilon}^{1+\lfloor \frac{d}{2} \rfloor} e^{-\lambda_{i,\varepsilon} T_0}$. Define

$$f_{T_0}(\lambda) \stackrel{\text{def}}{=} \lambda^{1+\lfloor \frac{d}{2} \rfloor} e^{-\lambda T_0}, \quad g_{T_0}(\lambda) \stackrel{\text{def}}{=} \begin{cases} f_{T_0}(\frac{\Lambda_1}{2}), & \lambda \in [0, \frac{\Lambda_1}{2}), \\ f_{T_0}(\lambda), & \lambda \in [\frac{\Lambda_1}{2}, \infty). \end{cases}$$

Since $\lambda_{m,\varepsilon} \geq \Lambda_1$ by assumption, we have

$$(5.32) \quad \begin{aligned} \sum_{i=m}^{\infty} \lambda_{i,\varepsilon}^{1+\lfloor \frac{d}{2} \rfloor} e^{-\lambda_{i,\varepsilon} T_0} &= \sum_{i=m}^{\infty} g_{T_0}(\lambda_{i,\varepsilon}) \leq \sum_{\{i: \lambda_{i,\varepsilon} \geq \Lambda_1\}} g_{T_0}(\lambda_{i,\varepsilon}) \\ &\leq \int_{\frac{\Lambda_1}{2}}^{\infty} g_{T_0}(\lambda) dN_{\varepsilon}(\lambda). \end{aligned}$$

Observe that f_{T_0} is decreasing on $[\Lambda_1/2, \infty)$ (since $T_0 \geq \frac{2}{\Lambda_1}(1 + \lfloor \frac{d}{2} \rfloor)$), and therefore g_{T_0} is decreasing on $[0, \infty)$. Consequently, for each $t \in (0, g_{T_0}(0))$, there exists $\lambda_t > 0$ such that $\{g_{T_0}(\lambda) > t\} = [0, \lambda_t)$. Combining this with Lemma 5.8 gives

$$(5.33) \quad \begin{aligned} \int_0^{\infty} g_{T_0}(\lambda) dN_{\varepsilon}(\lambda) &= \int_0^{g_{T_0}(0)} N_{\varepsilon}(\{g_{T_0}(\lambda) > t\}) dt = \int_0^{g_{T_0}(0)} N_{\varepsilon}([0, \lambda_t)) dt \\ &\stackrel{(5.29)}{\leq} C_W \varepsilon^{-\frac{d}{2}} \int_0^{g_{T_0}(0)} \left(\frac{d}{2} \int_0^{\infty} \mathbf{1}_{\{\lambda < \lambda_t\}} (1 + \lambda)^{\frac{d}{2}-1} d\lambda + 1 \right) dt \\ &\leq C_W \varepsilon^{-\frac{d}{2}} \left(\frac{d}{2} \int_0^{\infty} g_{T_0}(\lambda) (1 + \lambda)^{\frac{d}{2}-1} d\lambda + g_{T_0}(0) \right) \\ &\leq C'_W \varepsilon^{-\frac{d}{2}}, \end{aligned}$$

where all constants depending on $T_0(\Lambda_1, d)$, Λ_1 , and d have been absorbed into C'_W .

Combining (5.31), (5.32), and (5.33), defining $\tilde{C}_{\psi} \stackrel{\text{def}}{=} C_{\Lambda_1} C'_W$, and using the fact that $T_0 \Lambda_1 = d + 2$, we obtain (5.8). $\square$

5.6. Proof of Lemmas 5.8 and 5.9. To finish the proof of Lemma 5.6 it remains to prove Lemmas 5.8 and 5.9, which we do here. Lemma 5.8 can be obtained directly from Weyl's law (see for instance Chapter 17.5 in [Hör07], or [Ivr16, Sog17]) In our context however, the operator $\mathcal{L}_{\varepsilon}$ depends on ε , and we need asymptotic dependence of the Weyl function as $\varepsilon \rightarrow 0$. A similar result was used in [CFIN25], and we present a proof here for convenience.

Proof of Lemma 5.8. Define the operator $\mathcal{U}_{\varepsilon}: L^2(\mathbb{T}^d, \pi_{\varepsilon}) \rightarrow L^2(\mathbb{T}^d)$ by

$$\mathcal{U}_{\varepsilon} f = \frac{1}{\sqrt{Z_{\varepsilon}}} e^{-U/2\varepsilon} f.$$

Clearly

$$\langle f, g \rangle_{L^2(\pi_{\varepsilon})} = \langle \mathcal{U}_{\varepsilon} f, \mathcal{U}_{\varepsilon} g \rangle,$$

and so $\mathcal{U}$ is an isometry. Define the operator $\mathcal{H}_{\varepsilon}$ by $\mathcal{H}_{\varepsilon} \stackrel{\text{def}}{=} \mathcal{U}_{\varepsilon} \mathcal{L}_{\varepsilon} \mathcal{U}_{\varepsilon}^{-1}$. We compute

$$-\mathcal{H}_{\varepsilon} f = -\varepsilon \Delta f + \left(\frac{1}{4} |\nabla U|^2 - \frac{1}{2} \Delta U \right) f.$$

Thus $\mathcal{L}_{\varepsilon}$ is unitarily equivalent to the operator $\mathcal{H}_{\varepsilon}$, and hence the operators $\mathcal{L}_{\varepsilon}$ and $\mathcal{H}_{\varepsilon}$ have the same spectrum.

Next, for sufficiently large $\gamma = \gamma(\|U\|_{C^2}) > 0$, we see that the operator $-\mathcal{H}_{\varepsilon, \gamma} \stackrel{\text{def}}{=} -\mathcal{H}_\varepsilon + \gamma I$ is a self-adjoint operator that satisfies the coercivity bound

$$\varepsilon \langle -\Delta f, f \rangle_{L^2} \leq \langle -\mathcal{H}_{\varepsilon, \gamma} f, f \rangle_{L^2}, \quad \forall f \in L^2.$$

Thus, by the Courant–Fischer min-max principle, we have

$$\varepsilon \lambda_k(-\Delta) \leq \lambda_k(-\mathcal{H}_\varepsilon) + \gamma,$$

and this implies

$$N_{\mathcal{H}_\varepsilon}(\lambda) \leq N_\Delta\left(\frac{\lambda + \gamma}{\varepsilon}\right) \leq C(d)\left(\frac{\lambda + \gamma}{\varepsilon}\right)^{\frac{d}{2}},$$

which completes the proof. $\square$

We now turn to the proof of Lemma 5.9. This is a standard regularity estimate for $\mathcal{L}_\varepsilon$ (see for instance [Hör07]). In our situation however, the operator $\mathcal{L}_\varepsilon$, and the weight π_ε both depend on ε , and our aim is to obtain the bound (5.30) with the right ε -dependence. As a result, we present the proof here, keeping track of the ε -dependence.

Proof of Lemma 5.9. For notational simplicity, let ψ be an eigenfunction of $-\mathcal{L}_\varepsilon$ with eigenvalue $\lambda > \Lambda_1$, normalized so that $\|\psi\|_{L^2(\pi_\varepsilon)} = 1$. We claim that for every $m \in \mathbb{N}$, there exists a constant $C_m = C_m(\|U\|_{C^m}, d, \Lambda_1)$ such that

$$(5.34) \quad \|\psi\|_{\dot{H}^m(\pi_\varepsilon)}^2 \leq C_m \left(\frac{\lambda}{\varepsilon}\right)^m.$$

Lemma 5.9 follows immediately from (5.34). Indeed, since $\langle \psi, 1 \rangle_{L^2(\pi_\varepsilon)} = 0$, the Sobolev embedding theorem and Poincaré inequality imply

$$(5.35) \quad \|\psi\|_{L^\infty(\mathbb{T}^d)}^2 \leq C(d) \|\psi\|_{\dot{H}^m}^2, \quad \text{for } m = 1 + \left\lfloor \frac{d}{2} \right\rfloor.$$

Also, since

$$(5.36) \quad \begin{aligned} \|\psi\|_{\dot{H}^m}^2 &= \sum_{|\alpha|=m} \int_{\mathbb{T}^d} |D^\alpha \psi|^2 dx \\ &= \sum_{|\alpha|=m} \int_{\mathbb{T}^d} |D^\alpha \psi|^2 \pi_\varepsilon \pi_\varepsilon^{-1} dx \leq \|\psi\|_{\dot{H}^m(\pi_\varepsilon)}^2 \|\pi_\varepsilon^{-1}\|_\infty. \end{aligned}$$

Combining (5.34), (5.35), and (5.36) implies (5.30) as desired.

It remains to prove (5.34). We do this by induction on m . The base case $m = 0$ is trivial since $\|\psi\|_{L^2(\pi_\varepsilon)} = 1$ by assumption. Suppose now (5.34) holds for some $k = m$.

Define $l_k = \{\alpha \in \mathbb{N}^d - 0 \mid |\alpha|_{\ell^1} = k\}$. Then, we see that

$$(5.37) \quad \|\psi\|_{\dot{H}^{k+1}(\pi_\varepsilon)}^2 = \sum_{\alpha \in l_k} \sum_{i=1}^d \|D^{\alpha+e_i} f\|_{L^2(\pi_\varepsilon)}^2.$$

We now fix $\alpha \in l_k$ and compute $\sum_{i=1}^d \|D^{\alpha+e_i} f\|_{L^2(\pi_\varepsilon)}^2$. Integrating by parts gives

$$(5.38) \quad \|D^{\alpha+e_i} \psi\|_{L^2(\pi_\varepsilon)}^2 = - \int_{\mathbb{T}^d} D^\alpha \psi \partial_i^2 D^\alpha \psi \pi_\varepsilon dx - I_1,$$

where

$$I_1 = \int_{\mathbb{T}^d} D^\alpha \psi \partial_i D^\alpha \psi \partial_i \pi_\varepsilon dx = \frac{1}{2} \int_{\mathbb{T}^d} \partial_i (D^\alpha \psi)^2 \partial_i \pi_\varepsilon dx = -\frac{1}{2} \int_{\mathbb{T}^d} (D^\alpha \psi)^2 \partial_i^2 \pi_\varepsilon dx.$$

Combining this with (5.38) and summing over i in (5.37) produces

$$(5.39) \quad \sum_{i=1}^d \|D^{\alpha+e_i}\psi\|_{L^2(\pi_\varepsilon)}^2 = I_2 + I_3,$$

where

$$(5.40) \quad I_2 = - \int_{\mathbb{T}^d} D^\alpha \psi D^\alpha (\Delta \psi) \pi_\varepsilon dx \quad \text{and} \quad I_3 = \frac{1}{2} \int_{\mathbb{T}^d} (D^\alpha \psi)^2 \Delta \pi_\varepsilon dx.$$

We first compute I_2 . Since ψ is an eigenfunction of $\mathcal{L}_\varepsilon$ we note

$$(5.41) \quad I_2 = \frac{\lambda}{\varepsilon} \int_{\mathbb{T}^d} (D^\alpha \psi)^2 \pi_\varepsilon dx - \frac{1}{\varepsilon} I_4,$$

where

$$I_4 = \int_{\mathbb{T}^d} D^\alpha (\nabla U \cdot \nabla \psi) D^\alpha \psi \pi_\varepsilon dx.$$

By considering the case when the differential operator D^α is completely exhausted by hitting ∇f or not, we can split the integral into two parts. Namely,

$$(5.42) \quad I_4 = I_5 + I_6$$

where

$$I_5 = \int_{\mathbb{T}^d} \nabla U \cdot \nabla (D^\alpha \psi) D^\alpha \psi \pi_\varepsilon \quad \text{and} \quad I_6 = I_4 - I_5.$$

Integrating I_5 by parts gives

$$I_5 = - \int_{\mathbb{T}^d} \Delta U (D^\alpha \psi)^2 \pi_\varepsilon dx - I_5 - \int_{\mathbb{T}^d} \nabla U \cdot \nabla \pi_\varepsilon (D^\alpha \psi)^2 dx,$$

and consequently,

$$(5.43) \quad I_5 = -\frac{1}{2} \int_{\mathbb{T}^d} (\Delta U \pi_\varepsilon + \nabla U \cdot \nabla \pi_\varepsilon) (D^\alpha \psi)^2 dx.$$

Combining (5.39), (5.40), (5.41), (5.42), and (5.43), we get

$$(5.44) \quad \sum_{i=1}^d \|D^{\alpha+e_i}\psi\|_{L^2(\pi_\varepsilon)}^2 = \frac{\lambda}{\varepsilon} \|D^\alpha \psi\|_{L^2(\pi_\varepsilon)}^2 - \frac{1}{\varepsilon} I_6 + \frac{1}{2\varepsilon} I_7,$$

where

$$I_7 = \int_{\mathbb{T}^d} (\Delta U \pi_\varepsilon + \nabla U \cdot \nabla \pi_\varepsilon + \varepsilon \Delta \pi_\varepsilon) (D^\alpha \psi)^2 dx.$$

We know π_ε is a stationary solution to the Kolmogorov forward equation

$$\mathcal{L}_\varepsilon^* \pi_\varepsilon = \Delta U \pi_\varepsilon + \nabla U \cdot \nabla \pi_\varepsilon + \varepsilon \Delta \pi_\varepsilon = 0.$$

Hence $I_7 = 0$.

Moreover,

$$I_6 = \sum_{\beta < \alpha} \binom{\alpha}{\beta} \int_{\mathbb{T}^d} (D^{\alpha-\beta} \nabla U) \cdot (D^\beta \nabla \psi) D^\alpha \psi \pi_\varepsilon dx,$$

and by using Cauchy–Schwartz, induction hypothesis, and the fact that $|\beta| \leq k-1$,

$$|I_6| \leq \sum_{\beta < \alpha} \binom{\alpha}{\beta} \|U\|_{C^{1+|\alpha-\beta|}} \|\psi\|_{\dot{H}^{|\beta|+1}(\pi_\varepsilon)} \|\psi\|_{\dot{H}^k(\pi_\varepsilon)}$$

$$(5.45) \quad \leq C(k, \|U\|_{C^{k+1}}, d, \Lambda_1) \|\psi\|_{\dot{H}^k(\pi_\varepsilon)}^2.$$

Combining (5.37), (5.44), and (5.45) yields

$$\begin{aligned} \|\psi\|_{\dot{H}_{k+1}^2(\pi_\varepsilon)}^2 &\leq C(k, \|U\|_{C^{k+1}}, d, \Lambda_1) \left(\left(\frac{\lambda}{\varepsilon} \right)^{k+1} + \frac{1}{\varepsilon} \left(\frac{\lambda}{\varepsilon} \right)^k \right) \\ &\leq C(k, \|U\|_{C^{k+1}}, d, \Lambda_1) \left(\frac{\lambda}{\varepsilon} \right)^{k+1}, \end{aligned}$$

where we used the lower bound $\lambda \geq \Lambda_1$ and increased $C(k, \|U\|_{C^{k+1}}, d, \Lambda_1)$ appropriately to obtain the last inequality. This concludes the proof. $\square$

6. Proof of convergence for Autonormalized AIS

We will now prove Theorem 3.9. As before we assume without loss of generality that $\varepsilon_1 = 1$. Let U be a double well potential that satisfies Assumptions 3.1–3.3 for some $\varepsilon_{\min} < 1 \leq \varepsilon_{\max}$. Let $\nu > 0$ be a fixed constant, K , T_0 , and $\bar{C}_w$ be as Lemma 5.2. To prove Theorem 3.9, we will need to bound the product of both $\|P_k^T r_k^2\|_\infty$ and $\|P_k^T r_k^{-2}\|_\infty$. This is a little stronger than Assumption 2.2 which is all that was needed for the proof of Theorem 3.7. This is our first lemma.

Lemma 6.1. *There exists a constant $C_1(\nu, \varepsilon_1)$ such that for every $T \geq T_0$ we have*

$$(6.1) \quad \prod_{k=1}^K (1 \vee \|P_k^T r_k^2\|_{L^\infty} \|P_k^T r_k^{-2}\|_{L^\infty}) \leq C_1 \bar{C}_w^2.$$

Proof. Similar to (5.20), (5.21), and (5.22) we compute

$$\begin{aligned} \|\tilde{r}_k^{-2}\|_{L^1(\pi_k)} &= \langle \tilde{r}_k^{-2}, \pi_k \rangle = \frac{Z_{k-2}}{Z_k}, \\ \langle \tilde{r}_k^{-2}, \psi_{2,k} \rangle_{L^2(\pi_k)} &= \frac{Z_{k-2}}{Z_k} \langle \psi_{2,k}, \pi_{k-2} \rangle, \\ |\langle \tilde{r}_k^{-2}, \psi_{i,k} \rangle_{L^2(\pi_k)}| &\leq \|\tilde{r}_k^{-2}\|_{L^1(\pi_k)} \|\psi_{i,k}\|_{L^\infty}. \end{aligned}$$

Next we claim that the similar to (5.6), we also have a bound on $\int_{\mathbb{T}^d} \psi_{2,\varepsilon'} \pi_\varepsilon dx$. Explicitly, we claim

$$(6.2) \quad |\langle \psi_{2,\varepsilon'}, \pi_\varepsilon \rangle| \leq C_\gamma \left(\exp\left(-\frac{\gamma}{\varepsilon}\right) + |\pi_{\varepsilon'}(\Omega_1) - \pi_\varepsilon(\Omega_1)| \right)$$

for every $0 < \varepsilon' < \varepsilon \leq 1$, and some (possibly larger) constant C_γ that is independent of $\varepsilon, \varepsilon'$. The proof of this is almost identical to the proof of (5.6) presented in Section 9 of [HIS26] and we do not reproduce it here.

Following the proof of Lemma 5.7, and using (6.2) in place of (5.6) for the second estimate in that proof, we obtain a bound for $\tilde{r}_k^{-2}$ similar to (5.10). Explicitly, for the constant C_{ψ_2} in Property 5.5, and for any T satisfying (5.9), we have

$$\|P_{k,T} \tilde{r}_k^{-2}\|_{L^\infty} \leq \frac{Z_{k-2}}{Z_k} \left(1 + C_\gamma C_{\psi_2} \left(\exp\left(-\frac{\hat{\gamma}}{2\varepsilon_k}\right) + |\pi_{k+1}(\Omega_1) - \pi_k(\Omega_1)| \right) + \frac{1}{K} \right).$$

Combining this estimate with (5.10), we obtain

$$\|P_{k,T} r_k^2\|_{L^\infty} \|P_{k,T} r_k^{-2}\|_{L^\infty} \leq \frac{Z_{k-2} Z_{k+2}}{Z_k^2} \Theta(k, k+1)^2,$$

where Θ is defined in (5.14).

Since $\{1/\varepsilon_k\}$ are linearly spaced,

$$\int_{\mathbb{T}^d} e^{-U/\varepsilon_k} dx = \int_{\mathbb{T}^d} e^{-U/(2\varepsilon_{k-2})} e^{-U/(2\varepsilon_{k+2})} dx$$

and so the Cauchy-Schwartz inequality implies

$$Z_k^2 \leq Z_{k-2} Z_{k+2}.$$

Hence

$$\min \left\{ \frac{Z_{k-2} Z_{k+2}}{Z_k^2}, \Theta(k, k+1) \right\} \geq 1,$$

and so

$$\begin{aligned} \prod_{k=1}^K (1 \vee \|P_k^T r_k^2\|_{L^\infty} \|P_k^T r_k^{-2}\|_{L^\infty}) &\leq \frac{Z_{-1} Z_0}{Z_1 Z_2} \frac{Z_{K+1} Z_{K+2}}{Z_{K-1} Z_K} \left(\prod_{k=1}^K \Theta(k, k+1) \right)^2 \\ &\stackrel{(5.18)}{\leq} C_1(\nu) C_\Theta^2. \end{aligned}$$

Finally, choosing T_0 as in (5.1) (with the same choice of C_T in (5.12)), choosing $\bar{C}_w$ as in (5.19), and using the previous estimate show that (6.1) holds for any $T \geq T_0$. This concludes the proof. $\square$

Next we prove Proposition 3.10, showing that the effective sample size is at least $N/(C_1 \bar{C}_w^2)$.

Proof of Proposition 3.10. For notational convenience, in this proof we use $w_{k,i}$ to denote the weights w_k^i returned by Algorithm 2. We first note that by the definition of $w_{k,i}$ and $\tilde{W}_k$ in Algorithm 2, we can use the ratio of the *normalized* densities r_k , instead of that of the *unnormalized* densities $\tilde{r}_k$, because the same normalizing constant Z_{k+1}/Z_k appears in both the numerator and the denominator. Explicitly, we note

$$w_{k,i} = \frac{w_{k-1,i} \tilde{r}_k(X_k^i)}{\tilde{W}_k} = \frac{w_{k-1,i} r_k(X_k^i)}{W_k}, \quad \text{where} \quad W_k = \sum_{i=1}^N w_{k-1,i} r_k(X_k^i).$$

By Jensen's inequality applied to the convex function $x \mapsto 1/x^2$, we obtain

$$\frac{1}{W_k^2} = \frac{1}{\left(\sum_{i=1}^N w_{k-1,i} r_k(X_k^i) \right)^2} \leq \sum_{i=1}^N \frac{w_{k-1,i}}{r_k^2(X_k^i)}.$$

Hence,

$$\begin{aligned} \sum_{i=1}^N w_{k,i}^2 &= \frac{1}{W_k^2} \sum_{i=1}^N w_{k-1,i}^2 r_k^2(X_k^i) \\ &\leq \left(\sum_{j=1}^N \frac{w_{k-1,j}}{r_k^2(X_k^j)} \right) \left(\sum_{i=1}^N w_{k-1,i}^2 r_k^2(X_k^i) \right) \\ (6.3) \quad &= \sum_{i=1}^N w_{k-1,i}^3 + \sum_{i=1}^N \sum_{j \neq i} w_{k-1,i} w_{k-1,j} r_k^2(X_k^i) r_k^{-2}(X_k^j). \end{aligned}$$

Let $\mathbf{E}_{k-1}$ denote the conditional expectation with respect to the σ -algebra generated by $\{w_{k-1,i}, X_{k-1}\}_{1 \leq i \leq N}$. Applying $\mathbf{E}_{k-1}$ to both sides of (6.3), and using the conditional independence of X_k^i and X_k^j for $j \neq i$, we obtain

$$\begin{aligned}
\mathbf{E}_{k-1} \left[\sum_{i=1}^N w_{k,i}^2 \right] &\leq \sum_{i=1}^N w_{k-1,i}^3 + \sum_{i=1}^N \sum_{j \neq i}^N w_{k-1,i}^2 w_{k-1,j} \mathbf{E}_{k-1} [r_k^2(X_k^i)] \mathbf{E}_{k-1} [r_k^{-2}(X_k^j)] \\
&= \sum_{i=1}^N w_{k-1,i}^3 + \sum_{i=1}^N \sum_{j \neq i}^N w_{k-1,i}^2 w_{k-1,j} (P_k^T r_k^2)(X_{k-1}^i) (P_k^T r_k^{-2})(X_{k-1}^j) \\
&\leq (1 \vee \|P_k^T r_k^2\|_{L^\infty} \|P_k^T r_k^{-2}\|_{L^\infty}) \sum_{i=1}^N w_{k-1,i}^2.
\end{aligned}$$

Taking expectations on both sides and iterating gives

$$\begin{aligned}
\mathbf{E} \left[\sum_{i=1}^N w_{K,i}^2 \right] &\leq \left(\prod_{k=1}^K 1 \vee \|P_k^T r_k^2\|_{L^\infty} \|P_k^T r_k^{-2}\|_{L^\infty} \right) \sum_{i=1}^N w_{0,i}^2. \\
(6.4) \quad &= \frac{1}{N} \left(\prod_{k=1}^K 1 \vee \|P_k^T r_k^2\|_{L^\infty} \|P_k^T r_k^{-2}\|_{L^\infty} \right),
\end{aligned}$$

where the last equality holds because $w_{0,i} = 1/N$.

By increasing $\hat{C}_T$, if necessary, the choice of T in (3.6) ensures that $T \geq T_0$. Therefore, Lemma 6.1 applies, and (6.1) holds, and (6.4) implies (3.10) as desired. The bound (3.11) then follows from (3.10) and Jensen's inequality. $\square$

It remains to prove Theorem 3.9. To this end, we introduce the following notation. For every bounded test function h , and every $1 \leq k \leq K$, define

$$\begin{aligned}
\text{Err}_{k,T}(h) &= \|\langle h, \mu_{k,T} - \pi_k \rangle\|_{L^2(\mathbf{P})}, \\
\text{Err}_{k+1,0}(h) &= \|\langle h, \mu_{k+1,0} - \pi_{k+1} \rangle\|_{L^2(\mathbf{P})}
\end{aligned}$$

where

$$\mu_{k,T} \stackrel{\text{def}}{=} \sum_{i=1}^N w_{k-1}^i \delta_{X_k^i} \quad \text{and} \quad \mu_{k+1,0} \stackrel{\text{def}}{=} \sum_{i=1}^N w_k^i \delta_{X_k^i}.$$

We first state and prove the following lemma that connects the error before and after the reweighting.

Lemma 6.2. *For any $1 \leq k \leq K$,*

$$(6.5) \quad \text{Err}_{k+1,0}(h) \leq \|h\|_{L^\infty} \text{Err}_{k,T}(r_k) + \text{Err}_{k,T}(r_k h)$$

Proof. Fix $x_1, x_2, \dots, x_N \in \mathbb{T}^d$ and define the normalization constant R_k and the updated weight w_k^i by

$$R_k \stackrel{\text{def}}{=} \sum_{i=1}^N w_{k-1}^i r_k(x_i), \quad \text{and} \quad w_k^i = \frac{w_{k-1}^i r_k(x_i)}{R_k}.$$

Adding and subtracting the same term and using the triangle inequality, we obtain

$$(6.6) \quad \left| \sum_{i=1}^N w_k^i h(x_i) - \langle h, \pi_{k+1} \rangle \right| \leq \left| \sum_{i=1}^N w_k^i h(x_i) - R_k \sum_{i=1}^N w_k^i h(x_i) \right| + \left| \sum_{i=1}^N w_{k-1}^i r_k(x_i) h(x_i) - \langle h, \pi_{k+1} \rangle \right|.$$

The first term can be estimated by

$$(6.7) \quad \left| \sum_{i=1}^N w_k^i h(x_i) - R_k \sum_{i=1}^N w_k^i h(x_i) \right| \leq |1 - R_k| \|h\|_\infty,$$

and hence combining (6.6), (6.7), the fact that $\langle r_k, \pi_k \rangle = 1$ and $\langle r_k h, \pi_k \rangle = \langle h, \pi_{k+1} \rangle$, and substituting X_k^i for x_i yield (6.5). $\square$

Next, we state four lemmas required for the proof of Theorem 3.9. Their proofs are identical or closely follow those in [HIS26], and are therefore omitted. For convenience, we indicate in each statement the corresponding result in [HIS26].

Lemma 6.3 (Lemma 4.10, [HIS26]). *Assume that for each $k \in \{2, \dots, K\}$, the law of X_k^1 has density q_k . Then for any bounded test function h and $2 \leq k \leq K$,*

$$\begin{aligned} \text{Err}_{k,T}(h) &\leq e^{-\lambda_{2,k}T} |\langle h \psi_{2,k}, \pi_k \rangle| \text{Err}_{k,0}(\psi_{2,k}) \\ &\quad + E \left[\sum_{i=1}^N w_{k-1,i}^2 \right]^{\frac{1}{2}} \|h\|_\infty + \sqrt{N} \left\| \frac{q_{k-1}}{\pi_k} \right\|_\infty^{\frac{1}{2}} e^{-\Lambda T} \|h\|_\infty. \end{aligned}$$

Lemma 6.4 (Lemma 4.12, [HIS26]). *For any $\alpha > 0$, there exist constants $C_\alpha = C_\alpha(\alpha, U) > 0$ (depending on α) and $\hat{C}_N = \hat{C}_N(U, \nu) > 1$ (independent of α) such that for any $\delta > 0$, if*

$$(6.8) \quad N \geq \hat{C}_N \frac{K^2}{\delta^2}, \quad T \geq C_\alpha \left(K^{(1+\alpha)\hat{\gamma}_r} + \frac{1}{\varepsilon} + \log\left(\frac{1}{\delta}\right) + \log(N) \right),$$

then for each $2 \leq k \leq K-1$, we have

$$(6.9) \quad \text{Err}_{k+1,0}(\psi_{2,k+1}) \leq \beta_k \text{Err}_{k,0}(\psi_{2,k}) + c_k.$$

Here, the constants β_k, c_k are such that for every $k \in \{2, \dots, K-1\}$, we have

$$\begin{aligned} \prod_{j=k}^{K-1} \beta_j &\leq C_\beta, \\ c_k &\leq \frac{\delta}{K}, \end{aligned}$$

for some dimensional constant $C_\beta > 1$ (independent of α, δ).

Lemma 6.5 (Lemma 4.11, [HIS26]). *For every $2 \leq k \leq K$, let q_k be the probability density function of X_k^1 . For any $\hat{T}_0 > 0$, there exists a constant $C_q = C_q(U, \hat{T}_0)$ such that if $T \geq \hat{T}_0$, then*

$$\left\| \frac{q_{k-1}}{\pi_k} \right\|_\infty \leq C_q \exp\left(\left(\frac{1}{\varepsilon} - 1\right) \|U\|_\infty\right).$$

Lemma 6.6 (Lemma 4.13, [HIS26]). *There exists a constant $\hat{C}_1 = \hat{C}_1(U)$ such that for any $\delta > 0$ and*

$$(6.10) \quad N \geq \frac{\hat{C}_N}{\delta^2}, \quad T \geq \hat{C}_1 \left(\log \left(\frac{1}{\delta} \right) + 1 + \log(N) \right),$$

we have

$$(6.11) \quad \text{Err}_{2,0}(\psi_{2,2}) \leq \delta.$$

Here, $\hat{C}_N$ is the same constant defined in Lemma 6.4.

Proof of Theorem 3.9. Let C_β be as in Lemma 6.4, and define

$$C_r = 1 \vee \left(\max_{1 \leq k \leq K} \|r_k\|_\infty \right).$$

Fix $\delta > 0$ and set $\tilde{\delta} \stackrel{\text{def}}{=} \frac{\delta}{8C_\beta C_r}$. Fix $\alpha > 0$ and suppose that

$$(6.12) \quad N \geq \frac{\max \left\{ \hat{C}_N, C_1 \bar{C}_w^2 C_\beta^{-2} \right\}}{\tilde{\delta}^2} K^2,$$

$$(6.13) \quad T \geq \max \left\{ C_\alpha \left(K^{(1+\alpha)\hat{\gamma}_r} + \frac{1}{\varepsilon} + \log \frac{1}{\tilde{\delta}} + \log N \right), \right. \\ \left. \hat{C}_1 \left(\log \frac{1}{\tilde{\delta}} + 1 + \log N \right), C_T \left(\frac{1}{\varepsilon} + \log K \right), \right. \\ \left. \frac{1}{2\Lambda} \frac{\|U\|_\infty}{\varepsilon} + \frac{1}{\Lambda} \log \left(\frac{C_q^{1/2}}{\tilde{\delta}} \sqrt{N} \right), 1 \right\},$$

where $\bar{C}_w, C_T$ are as in Lemma 5.2, C_1 is as in Lemma 6.1, C_α and $\hat{C}_N$ are as in Lemma 6.4, $C_q = C_q(U, 1)$ is as in Lemma 6.5, and $\hat{C}_1$ is as in Lemma 6.6. In particular, choosing $\hat{C}_T$ and C_N in (3.6) and (3.7) sufficiently large ensures that the lower bounds in (6.12) and (6.13) are satisfied. Moreover, with this choice of N and T , N is sufficiently large to satisfy (6.8) and (6.10) with $\delta = \tilde{\delta}$ and T is sufficiently large to satisfy $T \geq \max\{T_0, 1\}$, (6.8), and (6.10) with $\delta = \tilde{\delta}$. Hence, Proposition 3.10 holds, Lemmas 6.4 and 6.6 hold with $\delta = \tilde{\delta}$, and Lemma 6.5 holds with $\hat{T}_0 = 1$.

Using Lemmas 6.2, 6.3, and Proposition 3.10, we obtain that for any bounded test function h ,

$$(6.14) \quad \text{Err}_{K+1,0}(h) \leq \left(\text{Err}_{K,0}(\psi_{2,K}) + \frac{C_1^{\frac{1}{2}} \bar{C}_w}{\sqrt{N}} + \sqrt{N} \left\| \frac{q_{K-1}}{\pi_K} \right\|_\infty^{1/2} e^{-\Lambda T} \right) \|r_K\|_\infty \|h\|_\infty.$$

We estimate the terms on the right hand side separately. Applying (6.9) for every $2 \leq k \leq K-1$, together with (6.11), yields

$$(6.15) \quad \text{Err}_{K,0}(\psi_{2,K}) \leq \left(\prod_{j=2}^{K-1} \beta_j \right) \text{Err}_{2,0}(\psi_{2,2}) + \sum_{k=2}^{K-2} c_k \left(\prod_{j=k+1}^{K-1} \beta_j \right) + c_{K-1} \\ \leq C_\beta \tilde{\delta} + \sum_{k=2}^{K-2} C_\beta \frac{\tilde{\delta}}{K} + \frac{\tilde{\delta}}{K} \leq 2C_\beta \tilde{\delta} = \frac{\delta}{4C_r}.$$

Moreover,

$$(6.16) \quad \frac{C_1^{\frac{1}{2}} \bar{C}_w}{\sqrt{N}} \stackrel{(6.12)}{\leq} C_\beta \tilde{\delta} = \frac{\delta}{8C_r}.$$

Finally, using Lemma 6.5, we obtain

$$(6.17) \quad \sqrt{N} \left\| \frac{q_{K-1}}{\pi_K} \right\|_\infty^{1/2} e^{-\Lambda T} \leq \sqrt{N} C_q^{1/2} \exp\left(\frac{\|U\|_\infty}{2\varepsilon} - \Lambda T\right) \stackrel{(6.13)}{\leq} \tilde{\delta} \leq \frac{\delta}{8C_r}.$$

Combining (6.14), (6.15), (6.16), and (6.17), we obtain

$$\text{Err}_{K+1,0}(h) \leq \frac{\delta}{C_r} \|h\|_\infty \|r_K\|_\infty \leq \delta \|h\|_\infty,$$

which completes the proof of (3.8). Equation (3.9) directly follows from (3.8) as in the proof of (2.7) in Theorem 2.5. $\square$

References

- [AKM19] S. Armstrong, T. Kuusi, and J.-C. Mourrat. *Quantitative stochastic homogenization and large-scale regularity*, volume 352 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2019. doi:[10.1007/978-3-030-15545-2](https://doi.org/10.1007/978-3-030-15545-2).
- [Arr89] S. Arrhenius. On the reaction velocity of the inversion of cane sugar by acids. *Zeitschrift für physikalische Chemie*, 4:226ff, 1889. doi:[10.1016/B978-0-08-012344-8.50005-2](https://doi.org/10.1016/B978-0-08-012344-8.50005-2).
- [BGK05] A. Bovier, V. Gaynard, and M. Klein. Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7(1):69–99, 2005. doi:[10.4171/JEMS/22](https://doi.org/10.4171/JEMS/22).
- [CFIN25] A. Christie, Y. Feng, G. Iyer, and A. Novikov. Speeding up Langevin dynamics by mixing. *Multiscale Model. Simul.*, 23(4):1696–1743, 2025. doi:[10.1137/24M168115X](https://doi.org/10.1137/24M168115X).
- [Che23] S. Chewi. *Log-Concave Sampling*. 2023. URL <https://chewisinho.github.io/main.pdf>.
- [CKRZ08] P. Constantin, A. Kiselev, L. Ryzhik, and A. Zlatoš. Diffusion and mixing in fluid flow. *Ann. of Math. (2)*, 168(2):643–674, 2008. doi:[10.4007/annals.2008.168.643](https://doi.org/10.4007/annals.2008.168.643).
- [CKSV25] O. Chehab, A. Korba, A. J. Stromme, and A. Vacher. Provable convergence and limitations of geometric tempering for Langevin dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=DZcmz9wU0i>.
- [CP20] N. Chopin and O. Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer Series in Statistics. Springer, Cham, 2020. doi:[10.1007/978-3-030-47845-2](https://doi.org/10.1007/978-3-030-47845-2).
- [DdFG01] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 3–14. Springer, New York, 2001. doi:[10.1007/978-1-4757-3437-9_1](https://doi.org/10.1007/978-1-4757-3437-9_1).
- [DFY20] M. Damak, B. Franke, and N. Yaakoubi. Accelerating planar Ornstein-Uhlenbeck diffusion with suitable drift. *Discrete Contin. Dyn. Syst.*, 40:4093, 2020. doi:[10.3934/dcds.2020173](https://doi.org/10.3934/dcds.2020173).
- [ERY24] B. Engquist, K. Ren, and Y. Yang. Sampling with adaptive variance for multimodal distributions, 2024, [2411.15220](https://arxiv.org/abs/2411.15220). URL <https://arxiv.org/abs/2411.15220>.
- [FI19] Y. Feng and G. Iyer. Dissipation enhancement by mixing. *Nonlinearity*, 32(5):1810–1851, 2019. doi:[10.1088/1361-6544/ab0e56](https://doi.org/10.1088/1361-6544/ab0e56).
- [GLR20] R. Ge, H. Lee, and A. Risteski. Simulated tempering langevin Monte Carlo II: An improved proof using soft markov chain decomposition, 2020, [1812.00793](https://arxiv.org/abs/1812.00793). URL <https://arxiv.org/abs/1812.00793>.
- [Han25] R. Han. Convergence of a sequential monte carlo algorithm towards multimodal distributions on rd, 2025, [2511.22564](https://arxiv.org/abs/2511.22564). URL <https://arxiv.org/abs/2511.22564>.
- [Has70] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi:[10.1093/biomet/57.1.97](https://doi.org/10.1093/biomet/57.1.97).

- [HIS26] R. Han, G. Iyer, and D. Slepčev. Time-complexity of sampling from a multimodal distribution using sequential monte carlo, 2026, 2508.02763. URL <https://arxiv.org/abs/2508.02763>.
- [HM54] J. M. Hammersley and K. W. Morton. Poor man’s Monte Carlo. *J. Roy. Statist. Soc. Ser. B*, 16:23–38; discussion 61–75, 1954. URL [http://links.jstor.org/sici?sici=0035-9246\(1954\)16:1<23:PMMC>2.0.CO;2-0&origin=MSN](http://links.jstor.org/sici?sici=0035-9246(1954)16:1<23:PMMC>2.0.CO;2-0&origin=MSN).
- [Hör07] L. Hörmander. *The analysis of linear partial differential operators. III*. Classics in Mathematics. Springer, Berlin, 2007. Pseudo-differential operators, Reprint of the 1994 edition.
- [Ivr16] V. Ivrii. 100 years of Weyl’s law. *Bull. Math. Sci.*, 6(3):379–452, 2016. doi:[10.1007/s13373-016-0089-y](https://doi.org/10.1007/s13373-016-0089-y).
- [KLL⁺13] T. C. Kwok, L. C. Lau, Y. T. Lee, S. Oveis Gharan, and L. Trevisan. Improved cheeger’s inequality: analysis of spectral partitioning algorithms through higher order spectral gap. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC ’13*, page 11–20, New York, NY, USA, 2013. Association for Computing Machinery. doi:[10.1145/2488608.2488611](https://doi.org/10.1145/2488608.2488611).
- [KLV25] F. Koehler, H. Lee, and T.-D. Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3264–3326. PMLR, 30 Jun–04 Jul 2025. URL <https://proceedings.mlr.press/v291/koehler25a.html>.
- [Kol00] V. N. Kolokoltsov. *Semiclassical analysis for diffusions and stochastic processes*, volume 1724 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. doi:[10.1007/BFb0112488](https://doi.org/10.1007/BFb0112488).
- [Liu08] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, New York, 2008.
- [LLN19] Y. Lu, J. Lu, and J. Nolen. Accelerating Langevin sampling with birth-death, 2019, 1905.09863.
- [LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2017. doi:[10.1090/mbk/107](https://doi.org/10.1090/mbk/107). Second edition of [MR2466937], With contributions by Elizabeth L. Wilmer, With a chapter on “Coupling from the past” by James G. Propp and David B. Wilson.
- [LSG24] H. Lee and M. Santana-Gijzen. Convergence bounds for sequential Monte Carlo on multimodal distributions using soft decomposition, 2024, 2405.19553. URL <https://arxiv.org/abs/2405.19553>.
- [LSG25] H. Lee and M. Santana-Gijzen. Sampling from multimodal distributions with warm starts: Non-asymptotic bounds for the reweighted annealed leap-point sampler, 2025, 2512.17977. URL <https://arxiv.org/abs/2512.17977>.
- [MMS23] J. Marion, J. Mathews, and S. C. Schmidler. Finite-sample complexity of sequential Monte Carlo estimators. *Ann. Statist.*, 51(3):1357–1375, 2023. doi:[10.1214/23-aos2295](https://doi.org/10.1214/23-aos2295).
- [MP92] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. 19(6):451–458, 1992. doi:[10.1209/0295-5075/19/6/002](https://doi.org/10.1209/0295-5075/19/6/002).
- [MR02] N. Madras and D. Randall. Markov chain decomposition for convergence rate analysis. *Ann. Appl. Probab.*, 12(2):581–606, 2002. doi:[10.1214/aoap/1026915617](https://doi.org/10.1214/aoap/1026915617).
- [MRR⁺53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi:[10.1063/1.1699114](https://doi.org/10.1063/1.1699114).
- [MS14] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 2014. doi:[10.1214/14-AOP908](https://doi.org/10.1214/14-AOP908).
- [MS24] J. Mathews and S. C. Schmidler. Finite sample complexity of sequential Monte Carlo estimators on multimodal target distributions. *Ann. Appl. Probab.*, 34(1B):1199–1223, 2024. doi:[10.1214/23-aap1989](https://doi.org/10.1214/23-aap1989).
- [Nea96] R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6:353–366, 1996. doi:[10.1007/BF00143556](https://doi.org/10.1007/BF00143556).
- [Nea01] R. M. Neal. Annealed importance sampling. *Stat. Comput.*, 11(2):125–139, 2001. doi:[10.1023/A:1008923215028](https://doi.org/10.1023/A:1008923215028).

- [Nea11] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 113–162. CRC Press, Boca Raton, FL, 2011.
- [Pav14] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. doi:[10.1007/978-1-4939-1323-7](https://doi.org/10.1007/978-1-4939-1323-7). Diffusion processes, the Fokker-Planck and Langevin equations.
- [PHLa20] E. Pompe, C. Holmes, and K. Łatuszyński. A framework for adaptive MCMC targeting multimodal distributions. *Ann. Statist.*, 48(5):2930–2952, 2020. doi:[10.1214/19-AOS1916](https://doi.org/10.1214/19-AOS1916).
- [PJT19] D. Paulin, A. Jasra, and A. Thiery. Error bounds for sequential Monte Carlo samplers for multimodal distributions. *Bernoulli*, 25(1):310–340, 2019. doi:[10.3150/17-bej988](https://doi.org/10.3150/17-bej988).
- [RBS15] L. Rey-Bellet and K. Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081, may 2015. doi:[10.1088/0951-7715/28/7/2081](https://doi.org/10.1088/0951-7715/28/7/2081).
- [RR55] M. N. Rosenbluth and A. W. Rosenbluth. Monte Carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23(2):356–359, 02 1955, https://pubs.aip.org/aip/jcp/article-pdf/23/2/356/18806837/356_1_online.pdf. doi:[10.1063/1.1741967](https://doi.org/10.1063/1.1741967).
- [SBCCD24] S. Syed, A. Bouchard-Côté, K. Chern, and A. Doucet. Optimised annealed sequential Monte Carlo samplers, 2024, [2408.12057](https://arxiv.org/abs/2408.12057). URL <https://arxiv.org/abs/2408.12057>.
- [Sch12] N. Schweizer. Non-asymptotic error bounds for sequential MCMC methods in multimodal settings, 2012, [1205.6733](https://arxiv.org/abs/1205.6733). URL <https://arxiv.org/abs/1205.6733>.
- [Sog17] C. D. Sogge. *Fourier integrals in classical analysis*, volume 210 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, second edition, 2017. doi:[10.1017/9781316341186](https://doi.org/10.1017/9781316341186).
- [Son26] S. Son. Rapid convergence of tempering chains to multimodal gibbs measures, 2026, [2604.04823](https://arxiv.org/abs/2604.04823). URL <https://arxiv.org/abs/2604.04823>.
- [SW86] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. 57(21):2607 – 2609, 1986. doi:[10.1103/PhysRevLett.57.2607](https://doi.org/10.1103/PhysRevLett.57.2607).
- [VW19] S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. doi:[10.48550/arXiv.1903.0856](https://doi.org/10.48550/arXiv.1903.0856).
- [WSH09] D. B. Woodard, S. C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.*, 19(2):617–640, 2009. doi:[10.1214/08-AAP555](https://doi.org/10.1214/08-AAP555).

PRINCETON UNIVERSITY, PRINCETON NJ 08544.

Email address: akshat.agarwal@princeton.edu

DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PA 15213.

Email address: gautam@math.cmu.edu

UNIVERSITY OF UTAH, SALT LAKE CITY, UT 84112

Email address: u0680511@utah.edu

DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PA 15213.

Email address: seungjas@andrew.cmu.edu

BRIGHAM YOUNG UNIVERSITY, PROVO, UT 84602.

Email address: wyattwimmer@gmail.com