Chapter 1

Sampling and Counting

1.1 Introduction

The classical *Monte Carlo method* is an approach to estimating quantities that are hard to compute exactly. The quantity z of interest is expressed as the expectation $z = \mathbf{E}(Z)$ of a random variable (r.v.) Z over a probability space Ω, μ. It is assumed that some efficient procedure for sampling from Ω, μ is available. By taking the mean of some sufficiently large set of independent samples of Z, one may obtain an approximation to z. For example, suppose

$$S = \left\{(x, y) \in [0,1]^2 : p_i(x, y) \leq 0, \text{ for all } i\right\}$$

is some region of the unit square defined by a system of polynomial inequalities $p_i(x, y) \leq 0$. Let Z be the r.v. defined by the following experiment or trial: choose a point (x, y) uniformly at random (u.a.r.) from $[0, 1]^2$; let $Z = 1$ if $p_i(x, y) \leq 0$ for all i, and $Z = 0$ otherwise. Then the area a of S is equal to $\mathbf{E}(Z)$, and an estimate of a may be obtained from the sample mean of a sufficiently long sequence of trials. In this example, the use of the Monte Carlo method is perhaps avoidable, at the expense of a more complex algorithm; for more essential uses, see, for example, Knuth's proposal for estimating the size of a tree by taking a random path from the root to a leaf, or Rasmussen's for estimating the permanent of a 0,1-matrix.

The main focus of this book is the *Markov chain* Monte Carlo (MCMC) method which is a development of the foregoing approach, which is sometimes applicable when Z cannot be sampled "directly". Z will often be the cardinality of some combinatorially defined set S. We design a Markov Chain $\mathfrak{M}$ with state space Ω (often S itself) whose steady state distribution is μ. Efficient sampling now rests on the rapid convergence of the chain to its steady state. These ideas will be made more explicit in Chapter 2 but for the moment we focus on the relationship between near uniform generation and approximate counting.

As a first example of the approach, we consider the problem of estimating the number of independent sets of a graph G with small maximum degree Δ. In Section 1.2 we show how sampling independent sets of G, generated independently and almost uniformly, can be used to obtain an estimate for their *number*. This step of the MCMC programme—how samples are used—is often (though not always) rather routine.

We then consider the reverse process i.e. we show how good estimates of the number of independent sets can be used to generate a near uniform sample. This illustrates a sort of equivalence between the problems of generation and counting. Section 1.4 discusses a formal framework within which this can be made precise.

1.2 Approximate counting, uniform sampling and their relationship

1.2.1 An example – Independent Sets

What do we mean precisely by (efficient) approximate counting and uniform sampling?

let $N = N(G)$ denote the number of independent sets of G. A *randomised approximation scheme* for N is a randomised algorithm that takes as input a graph G and an error bound $\varepsilon > 0$, and produces as output a number Y (a random variable) such that

$$\mathbf{Pr}\big((1-\varepsilon)N \le Y \le (1+\varepsilon)N\big) \ge \frac{3}{4}. \tag{1.1}$$

A randomised approximation scheme is said to be *fully polynomial* if it runs in time polynomial in n (the input length) and ε^{-1}. We shall abbreviate the rather unwieldy phrase "fully polynomial randomised approximation scheme" to FPRAS.

There is no significance in the constant $\frac{3}{4}$ appearing in the definition, beyond its lying strictly between $\frac{1}{2}$ and 1. Any success probability greater than $\frac{1}{2}$ may be boosted to $1-\delta$ for any desired $\delta > 0$ by performing a small number of trials and taking the median of the results; the number of trials required is $O(\ln \delta^{-1})$. Indeed let $Y_1, Y_2, \ldots, Y_m$ be independent samples satisfying (1.1). Suppose that $\tilde{Y}$ is the median of $Y_1, Y_2, \ldots, Y_m$. Then

$$\mathbf{Pr}(\tilde{Y} \ge (1+\varepsilon)N) \le \mathbf{Pr}(|\{i : \ Y_i \ge (1+\varepsilon)N\}| \ge m/2) \le e^{-m/12}$$

using the Chernoff bounds. Similarly

$$\mathbf{Pr}(\tilde{Y} \ge (1-\varepsilon)N) \le e^{-m/8}.$$

Putting $m = \lceil 12 \ln(2/\delta) \rceil$ we get

$$\mathbf{Pr}\big((1-\varepsilon)N \le \tilde{Y} \le (1+\varepsilon)N\big) \ge 1-\delta. \tag{1.2}$$

For any two probability distributions π and π' on a countable set Ω, define the *total variation distance* between π and π' to be

$$D_{\mathrm{tv}}(\pi,\pi') := \max_{A\subseteq\Omega}|\pi(A)-\pi'(A)| = \frac{1}{2}\sum_{x\in\Omega}|\pi(x)-\pi'(x)|. \tag{1.3}$$

In our example $\Omega=\Omega(G)$ will be the set of independent sets of graph G and $\pi(I)=\frac{1}{|\Omega|}$ for each $I\in\Omega$ i.e. π is the uniform distribution over Ω. We will let π' be the distribution of the output of some randomised algorithm that generates a random independent subset of G.

A *good sampler* for Ω is a randomised algorithm that takes as input a graph G and a tolerance $\delta>0$, and produces an independent set I (a random variable) such that the probability distribution of I is within variation distance δ of the uniform distribution on Ω. An almost uniform sampler is said to be *fully polynomial* if it runs in time polynomial in n (the input length) and $\log\delta^{-1}$.

From good sampling to approximate counting

Theorem 1.2.1 *Suppose we have a good sampler for the independent sets of a graph, which works for graphs G with maximum degree bounded by Δ and suppose that the sampler has time complexity $T(n,\delta)$, where n is the number of vertices in G, and δ the allowed deviation from uniformity in the sampling distribution. Then we may construct an FPRAS for the number of independent sets of a graph, which works for graphs G with maximum degree bounded by Δ, and which has time complexity*

$$O\left(\frac{m^2}{\varepsilon^2}T\Big(n,\frac{\varepsilon}{6m}\Big)\right),$$

where m is the number of edges in G, and ε the specified error bound.

Proof Let $G=G_m>G_{m-1}>\cdots>G_1>G_0=(V,\emptyset)$ be any sequence of graphs in which each graph G_{i-1} is obtained from the previous graph G_i by removing a single edge. We may express the quantity we wish to estimate as a product of ratios:

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|}\times\frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|}\times\cdots\times\frac{|\Omega(G_1)|}{|\Omega(G_0)|}\times|\Omega(G_0)|, \tag{1.4}$$

where, it will be observed, $|\Omega(G_0)|=2^n$. Our strategy is to estimate the ratio

$$\varrho_i=\frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$$

for each i in the range $1\le i\le m$, and by substituting these quantities into identity (1.4), obtain an estimate for the number of independent sets of G:

$$|\Omega(G)| = 2^n\varrho_1\dots\varrho_m. \tag{1.5}$$

To estimate the ratio ϱ_i we use the almost uniform sampler to obtain a sufficiently large sample of independent sets from $\Omega(G_{i-1})$ and compute the proportion of samples that lie in $\Omega(G_i)$.

The following lemma gives the basic probabilistic inequality we need.

Lemma 1.2.1 *For $i = 1, 2, \ldots, m$ let $0 \le Z_i \le 1$ be independent random variables on the probability space (Ω_i, π_i) where $\mathbf{E}(Z_i) = \mu_i$ and $\mu_{\min} = \min_i \mu_i > 0$.*

For $i = 1, 2, \ldots, m$ let $\hat{Z}_i$ denote the same random variable on the probability space $(\Omega_i, \hat{\pi}_i)$ where

$$d_{TV}(\pi_i, \hat{\pi}_i) \le \delta = \frac{\varepsilon}{3m}\mu_{\min}.$$

For $i = 1, 2, \ldots, m$ let $\hat{\mu}_i = \mathbf{E}(\hat{Z}_i)$ and let $\hat{Z}_i^{(1)}, \ldots, \hat{Z}_i^{(s)}$ be a sequence of

$$s = \lceil 17m\mu_{\min}^{-2}\varepsilon^{-2}\rceil$$

independent copies of the random variable $\hat{Z}_i$ and let $\bar{Z}_i = s^{-1}\sum_{j=1}^{s}\hat{Z}_i^{(j)}$ be their mean. Let

$$W = \frac{\bar{Z}_1\bar{Z}_2\cdots\bar{Z}_m}{\mu_1\mu_2\cdots\mu_m}.$$

Then, for ϵ sufficiently small,

$$\mathbf{Pr}(|W - 1| \ge \varepsilon) \le \frac{1}{4}.$$

Proof Note first that for $i = 1, 2, \ldots, m$,

$$|\hat{\mu}_i - \mu_i| \le \delta \text{ and } \mathbf{Var}(\hat{Z}_i) \le 1. \tag{1.6}$$

Let

$$\bar{W} = \frac{\bar{Z}_1\bar{Z}_2\cdots\bar{Z}_m}{\hat{\mu}_1\hat{\mu}_2\cdots\hat{\mu}_m}.$$

Now $\mathbf{E}(\bar{W}) = 1$ and (1.6) implies

$$\left(1 - \frac{\delta}{\mu_{\min}}\right)^m \le \frac{W}{\bar{W}} \le \left(1 + \frac{\delta}{\mu_{\min}}\right)^m.$$

So,

$$\left|\frac{W}{\bar{W}} - 1\right| \le \frac{2\epsilon}{5}. \tag{1.7}$$

Furthermore

$$\begin{aligned}
\mathbf{Var}(\bar{W}) &= \mathbf{E}\left(\prod_{i=1}^{m}\frac{\bar{Z}_i^2}{\hat{\mu}_i^2}\right) - 1 \qquad (1.8)\\
&= \prod_{i=1}^{m}\left(1+\frac{\mathbf{Var}(\bar{Z}_i)}{\hat{\mu}_i^2}\right) - 1\\
&= \prod_{i=1}^{m}\left(1+\frac{\mathbf{Var}(\hat{Z}_i)}{s\hat{\mu}_i^2}\right) - 1\\
&\le \prod_{i=1}^{m}\left(1+\frac{1}{s\hat{\mu}_i^2}\right) - 1\\
&\le \left(1+\frac{\epsilon^2}{17m}\right)^m - 1\\
&\le \frac{\epsilon^2}{16}. \qquad (1.9)
\end{aligned}$$

Thus by (1.7) and (1.9),

$$\mathbf{Pr}(|W-1|\ge\epsilon) \le \mathbf{Pr}(|\bar{W}-1|\ge\tfrac{\epsilon}{2}) \le \frac{4}{\epsilon^2}\mathbf{Var}(\bar{W}) \le \frac{1}{4}.$$

□

Suppose that the graphs G_i and G_{i-1} differ in the edge $\{u,v\}$, which is present in G_i but absent from G_{i-1}. Clearly, $\Omega(G_i)\subseteq\Omega(G_{i-1})$. Any independent set in $\Omega(G_{i-1})\setminus\Omega(G_i)$ contains u and v, and may be perturbed to an independent set in G_i by deleting vertex u. (To resolve ambiguity, let u be the smaller of the two vertices.) On the other hand, each independent set in G_i can be obtained in at most one way as the result of such a perturbation; hence $|\Omega(G_{i-1})\setminus\Omega(G_i)|\le|\Omega(G_i)|$ and

$$\frac{1}{2}\le\varrho_i\le 1. \qquad (1.10)$$

To avoid trivialities, assume $0<\varepsilon\le 1$ and $m\ge 1$. Let $Z_i\in\{0,1\}$ denote the random variable which results from choosing a random independent set from G_{i-1} and returning one if the resulting independent set is also independent in G_i and zero otherwise. Note that $\mu_i=\mathbf{E}(Z_i)=\varrho_i$ for $i=1,2,\dots,m$. Let $\hat{Z}_i$ denote the random variable which results from running the postulated almost uniform sampler on the graph G_{i-1} and returning one if the resulting independent set is also independent in G_i and zero otherwise. We take $\delta=\frac{\epsilon}{6m}$ (in the sampler) and $s=\lceil 68m\epsilon^{-2}\rceil$. Let $Z_i^{(1)},\dots,Z_i^{(s)}$ be a sequence of s independent copies of the random variable $\hat{Z}_i$. As our estimator for $|\Omega(G)|$, we use the random variable $Y=2^n\,\overline{Z}_1\overline{Z}_2\dots\overline{Z}_m$. Applying Lemma 1.2.1 we see immediately that

$$\mathbf{Pr}\left(\left|\frac{Y}{|\Omega(G)|}-1\right|\ge\epsilon\right)\le\frac{1}{4}.$$

We use $s = O(m\epsilon^{-2})$ samples to estimate each ϱ_i and the time bound claimed in the theorem follows. □

From approximate counting to good sampling

Theorem 1.2.2 *Suppose that we have an FPRAS* APPROXCOUNT(G, ϵ, δ) *for the number of independent sets of a graph $G = (V, E)$ with maximum degree Δ and suppose that* APPROXCOUNT(G, ϵ, δ) *has time complexity $T(n, \epsilon, \delta)$ where $n = |V|$, ϵ is the required maximum relative error and δ is the allowed probability of failure. Then we can construct a good sampler* UGEN(G, ϵ) *for the independent sets of G with maximum degree Δ which has expected time complexity*

$$O\left(T\left(n, O\left(\frac{1}{n}\right), O\left(\frac{\delta}{n}\right)\right)\right). \tag{1.11}$$

Proof We will call our sampling procedure UGEN(G, ϵ): let

$$\delta_1 = \frac{\delta}{2n+1} \text{ and } \epsilon_1 = \frac{\log 2}{3n}.$$

UGEN(G, ϵ)
begin
  $N =$ APPROXCOUNT$(G, \epsilon_1, \delta_1)$
  Repeat until $I =$ UGENX$(G, \epsilon_1, \frac{1}{4N}) \neq \perp$
  Output I.
end

The precedure UGENX has an extra parameter ϕ which is needed to control the rate of some rejection sampling. We define UGENX recursively.

UGENX(G, ϵ_1, ϕ)
begin
  If $\phi > 1$ **then** output $I = \perp$ – failure.
  If $V = \emptyset$ **then** $I = \begin{cases} \emptyset & \text{probability } \phi \\ \perp & \text{probability } 1 - \phi \end{cases}$
  else begin
    $v = \max V$ and X is the set of neighbours of v in G.
    $G_1 = G - v - X$ and $G_2 = G - v$
    $N_1 =$ APPROXCOUNT$(G_1, \epsilon_1, \delta_1)$ and $N_2 =$ APPROXCOUNT$(G_2, \epsilon_1, \delta_1)$

    Output $I = \begin{cases} v + \text{UGENX}\left(G_1, \epsilon_1, \phi\frac{N_1+N_2}{N_1}\right) & \text{probability } \frac{N_1}{N_1+N_2} \\ \text{UGENX}\left(G_2, \epsilon_1, \phi\frac{N_1+N_2}{N_2}\right) & \text{probability } \frac{N_2}{N_1+N_2} \end{cases}$

end
end

For $I \in \Omega$ let p_I denote the probability that $Ugenx(G, \epsilon_1, \phi)$ generates I, conditional on all calls to APPROXCOUNT being successful. Then we will see that $\phi \leq p_I$ and at the bottom of the recursion, ϕ will have become ϕ/p_i and so I will be output with (conditional) probability $p_I \times \phi/p_I = \phi$ i.e. the conditional output is uniform.

Lemma 1.2.2 **(a)** *The probability that* APPROXCOUNT *gives a bad estimate during the execution of* UGEN *is at most* $(2n+1)\delta_1$.

(b *If* APPROXCOUNT *gives no bad estimates then* $\phi \leq 1$ *throughout the execution of* UGEN.

(c) *If* APPROXCOUNT *gives no bad estimates then the probability* UGEN *outputs* $\perp$ *is at most 2/3.*

(d) *If* APPROXCOUNT *gives no bad estimates then the output* I *is such that for any independent set* I_0 *of* G *we have* $\mathbf{Pr}(I = I_0) = \phi$.

(e) *Let* $\hat{\pi}$ *be the distribution of the output* I *of* UGEN *and let* π *denote the uniform distribution on* Ω. *Then* $D_{\rm tv}(\pi, \hat{p}) \leq \delta$.

Proof (a) This is clear from the fact that we call APPROXCOUNT at most $2n+1$ times during the execution of UGEN.

(b) If there is no bad estimate from APPROXCOUNT then we claim by induction on the depth of recursion d that whenever we invoke UGENX on a graph H, say, then we find the current value of ϕ, $\phi_d \leq \frac{(1+\epsilon_1)^d}{4(1-\epsilon_1)^{d+1}|\Omega(H)|}$. This is trivially true for $d = 0$ and assuming say that we recurse on H_1 we have in this call

$$\phi_{d+1} \leq \frac{(1+\epsilon_1)^d}{4(1-\epsilon_1)^{d+1}|\Omega(H)|}\frac{N_1+N_2}{N_1} \leq \frac{(1+\epsilon_1)^d}{4(1-\epsilon_1)^{d+1}|\Omega(H)|} \times \frac{(1+\epsilon_1)|\Omega(H)|}{(1-\epsilon_1)|\Omega(H_1)|}$$
$$= \frac{(1+\epsilon_1)^{d+1}}{4(1-\epsilon_1)^{d+2}|\Omega(H_1)|}$$

as required.

Thus throughout the execution of UGEN we have $\phi \leq \frac{(1+\epsilon_1)^n}{4(1-\epsilon_1)^{n+1}} < e^{n\epsilon_1}/2 < 1$.

(c) We prove by induction on $|V|$ that $\mathbf{Pr}(I = \perp) \leq 1 - \phi|\Omega(G)|$. This is clearly true if $V = \emptyset$. Otherwise

$$\mathbf{Pr}(I = \perp) \leq$$
$$\frac{N_1}{N_1+N_2}(1 - \phi\frac{N_1+N_2}{N_1}|\Omega(G_1)|) + \frac{N_2}{N_1+N_2}(1 - \phi\frac{N_1+N_2}{N_2}|\Omega(G_2)|)$$
$$= 1 - \phi|\Omega(G)|.$$

Thus $\mathbf{Pr}(I = \perp) \leq 2/3$ as required.

(d) This is clearly true if $V = \emptyset$. If $V \neq \emptyset$ and $v = \max V \in I_0$ then, by induction

$$\mathbf{Pr}(I = I_0) = \frac{N_1}{N_1 + N_2}\phi\frac{N_1 + N_2}{N_1} = \phi$$

and similarly $\mathbf{Pr}(I = I_0) = \phi$ if $v \notin I_0$.

(e) Let $\mathcal{E}$ denote the event that some output of APPROXCOUNT is bad in the iteration that produces output. Then for $A \subseteq \Omega$,

$$\begin{aligned}\hat{\pi}(A) - \pi(A) &\leq \mathbf{Pr}(I \in A \mid \bar{\mathcal{E}}) + \mathbf{Pr}(\mathcal{E}) - \pi(A) \\ &\leq \frac{|A|}{|\Omega|} + \delta - \frac{|A|}{|\Omega|} \\ &\leq \delta.\end{aligned}$$

□

We have therefore shown that by running UGENX for *constant* expected number of times, we will with probability at least $1 - \delta$ output a randomly chosen independent set. The expected running time of UGEN is clearly as given in (1.11) which is small enough to make it a good sampler.

Having dealt with a specific example we see how to put the above ideas into a formal framework. Before doing this we enumerate some basic facts about Markov Chains.

1.3 Markov Chains

Throughout $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$, $\mathbb{Q}_+ = \{q \in \mathbb{Q} : q > 0\}$, and $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}_+$.

A Markov chain $\mathcal{M}$ on the finite state space Ω, with transition matrix P is a sequence of random variables X_t, $t = 0, 1, 2, \ldots$, which satisfy

$$\mathbf{Pr}(X_t = \sigma \mid X_{t-1} = \omega, X_{t-2}, \ldots, X_0) = P(\omega, \sigma) \qquad (t = 1, 2, \ldots),$$

We sometimes write P^{ω}_{σ}. The value of X_t is referred to as the *state* of $\mathcal{M}$ at *time* t.

Consider the digraph $D_{\mathcal{M}} = (\Omega, A)$ where $A = \{(\sigma, \omega) \in \Omega \times \Omega : P(\sigma, \omega) > 0\}$. We will by and large be concerned with chains that satisfy the following assumptions:

M1 The digraph $D_{\mathcal{M}}$ is strongly connected.

M2 $\gcd\{|C| : C \text{ is a directed cycle of } D_{\mathcal{M}}\} = 1$

Under these assumptions, $\mathcal{M}$ is *ergodic* and therefore has a unique stationary distribution π i.e.

$$\lim_{t \to \infty} \mathbf{Pr}(X_t = \omega \mid X_0 = \sigma) = \pi(\omega) \tag{1.12}$$

i.e. the limit does not depend on the starting state X_0. Furthermore, π is the unique left eigen-vector of P with eigenvalue 1 i.e. satisfying

$$P^T \pi = \pi. \tag{1.13}$$

Another useful fact is that if τ_σ denotes the expected number of steps between successive visits to state σ then

$$\tau_\sigma = \frac{1}{\pi(\sigma)}. \tag{1.14}$$

In most cases of interest, $\mathcal{M}$ is *reversible*, i.e.

$$Q(\omega, \sigma) = \pi(\omega)P(\omega, \sigma) = \pi(\sigma)P(\sigma, \omega) \qquad (\forall \omega, \sigma \in \Omega). \tag{1.15}$$

The central role of reversible chains in applications rests on the fact that π can be deduced from (1.15). If $\mu : \Omega \longrightarrow \mathbb{R}$ satisfies (1.15), then it determines π up to normalization. Indeed, if (1.15) holds and $\sum_{\omega \in \Omega} \pi(\omega) = 1$ then

$$\sum_{\omega \in \Omega} \pi(\omega)P(\omega, \sigma) = \sum_{\omega \in \Omega} \pi(\sigma)P(\sigma, \omega) = \pi(s)$$

which proves that π is a left eigenvector with eigenvalue 1.

In fact, we often *design* the chain to satisfy (1.15). Without reversibility, there is no apparent method of determining π, other than to explicitly construct the transition matrix, an exponential time (and space) computation in our setting.

As a canonical example of a reversible chain we have a *random walk* on a graph. A random walk on the undirected graph $G = (V, E)$ is a Markov chain with state space V associated with a particle that moves from vertex to vertex according to the following rule: the probability of a transition from vertex i, of degree d_i, to vertex j is $\frac{1}{d_i}$ if $\{i, j\} \in E$, and 0 otherwise. Its stationary distribution is given by

$$\pi(v) = \frac{d_v}{2|E|} \qquad\qquad v \in V. \tag{1.16}$$

To see this note that $Q(v, w) = Q(w, v)$ if v, w are not adjacent and otherwise

$$Q(v, w) = \frac{1}{2|E|} = Q(w, v),$$

verifying the *detailed balance* equations (1.15).

Note that if G is a regular graph then the steady state is uniform over V.

If G is bipartite then the walk as described is not ergodic. This is because all cycles are of even length. This is usually handled by adding d_v loops to vertex v for each vertex v. (Each loop counts as a single exit from v.) The net effect of this is to make the particle stay put with probability $\frac{1}{2}$ at each step. The steady state is unaffected. The chain is now *lazy*.

A chain is lazy if $P(\omega,\omega) \geq \frac{1}{2}$ for all $\omega \in \Omega$.

If $p_0(\omega) = \mathbf{Pr}(X_0 = \omega)$, then $p_t(\sigma) = \sum_\omega p_0(\omega)P^t(\omega,\sigma)$ is the distribution at time t. As a measure of convergence, the natural choice in this context is variation distance.

The *mixing time* of the chain is then

$$\tau(\varepsilon) = \max_{p_0} \min_t \{D_{\rm tv}(p_t,\pi) \leq \varepsilon\},$$

and it is easy to show that the maximum occurs when $X_0 = \omega_0$, with probability one, for some state ω_0. This is because $D_{\rm tv}(p_t,\pi)$ is a convex function of p_0 and so the maximum of $D_{\rm tv}(p_t,\pi)$ occurs at an extreme point of the set of probabilities p_0.

We now provide a simple lemma which indicates that variation distance $D_{\rm tv}(p_t,\pi)$ goes to zero exponentially.

We define several related quantities: $p_t^{(i)}$ denotes the t-fold distribution, conditional on $X_0 = i$.

$$d_i(t) = D_{\rm tv}(p_t^{(i)},\pi),\ d(t) = \max_i d_i(t),\ \bar{d}(t) = \max_{i,j} D_{\rm tv}(p_t^{(i)},p_t^{(j)}).$$

Lemma 1.3.1 *For all $s,t \geq 0$,*

(a) $\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t)$.

(b) $d(s+t) \leq 2d(s)d(t)$.

(c) $d(s) \leq 2\bar{d}(s)$.

(d) $d(s) \leq d(t)$ *for* $s \leq t$.

Proof We will use the characterisation of variation distance as

$$D_{\rm tv}(\mu_1,\mu_2) = \min \mathbf{Pr}(X_1 \neq X_2) \tag{1.17}$$

where the minimum is taken over pairs of random variables X_1, X_2 such that X_i has distribution $\mu_i, i = 1,2$.

Fix states i_1, i_2 and times s, t and let Y^1, Y^2 denote the chains started at i_1, i_2 respectively. By (1.17) we can construct a joint distribution for (Y_s^1, Y_s^2) such that

$$\mathbf{Pr}(Y_s^1 \neq Y_s^2) = D_{\mathrm{tv}}(p_s^{(i_1)}, p_s^{(i_2)}) \leq \bar{d}(s).$$

Now for each pair j_1, j_2 we can use (1.17) to construct a joint distribution for (Y_{s+t}^1, Y_{s+t}^2) such that

$$\mathbf{Pr}(Y_{s+t}^1 \neq Y_{s+t}^2 \mid Y_s^1 = j_1, Y_s^2 = j_2) = D_{\mathrm{tv}}(p_t^{(j_1)}, p_t^{(j_2)}).$$

The RHS is 0 if $j_1 = j_2$ and otherwise at most $\bar{d}(t)$. So, unconditionally,

$$\mathbf{Pr}(Y_{s+t}^1 \neq Y_{s+t}^2) \leq \bar{d}(s)\bar{d}(t)$$

and (1.17) establishes part (a) of the lemma.

For part (b), the same argument, with Y^2 now being the stationary chain shows

$$d(s+t) \leq d(s)\bar{d}(t) \tag{1.18}$$

and so (b) will follow from (c), which follows from the triangular inequality for variation distance. Finally note that (d) follows from (1.18). □

We will for the most part use carefully defined Markov chains as our good samplers. As an example, we now define a simple chain with state space Ω equal to the collection of independent sets of a graph G. The chain is ergodic and its steady state is uniform over Ω. So, running the chain for sufficiently long will produce a near uniformly chosen independent set, see (1.12). Unfortunately, this chain does not have a small enough mixing time for this to qualify as a good sampler, unless $\Delta(G) \leq 4$.

We define the chain as follows: suppose $X_t = I$. Then we choose a a vertex v of G uniformly at random. If $v \in I$ then we put $X_{t+1} = I \setminus \{v\}$. If $v \notin I$ and $I \cup \{v\}$ is an indepedent set then we put $X_{t+1} = I \cup \{v\}$. Otherwise we let $X_{t+1} = X_t = I$. Thus the transition matrix can be described as follows: $n = |V|$ and I, J are independent sets of G.

$$P(I, J) = \begin{cases} \frac{1}{n} & |I \Delta J| = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here $I \Delta J$ denotes the symmetric difference $(I \setminus J) \cup (J \setminus I)$.

The chain satisfies M1 and M2: In $D_{\mathcal{M}}$ every vertex can reach and is reachable from $\emptyset$, implying M1 holds. Also, $D_{\mathcal{M}}$ contains loops unless G has no edges. In both cases M2 holds trivially.

Note finally that $P(I, J) = P(J, I)$ and so (1.15) holds with $\pi(I) = \frac{1}{|\Omega|}$. Thus the chain is reversible and the steady state is uniform.

1.4 A formal computational framework

The sample spaces we have in mind are sets of combinatorial objects. However, in order to discuss the computational complexity of generation, it is necessary to consider a sequence of instances of increasing size. We therefore work within the following formal framework. The models of computation are the Turing Machine (TM) for deterministic computations and the Probabilistic Turing Machine (PTM) for randomized computations. (A PTM is a TM with a source of uniform and independent random bits.) We must confine ourselves to some class of distributions which are "easily described", from a computational viewpoint, in large instances. We identify this below with a class of unnormalized measures which we call "weight functions".

Let Σ be a fixed alphabet of at least two symbols, and $W : \Sigma^* \times \Sigma^* \longrightarrow \mathbb{N}$ be such that, for some polynomial b, $W(\sigma,\omega) = 0$ unless $|\omega| \le b(|\sigma|)$. Moreover $W(\sigma,\omega)$ must be computable in time polynomial in $|\sigma|$ whenever $W(\sigma,\omega) > 0$. (If the TM for W may ignore part of its input, this implies that W is *always* computable in polynomial time.) Let us call W a *weight function*. Here σ may be thought of as an encoding of an instance of some combinatorial problem, and the ω of interest are encodings of the structures we wish to generate.

Let $\Omega_\sigma = \{\omega : W(\sigma,\omega) > 0\}$. Then the sequence of discrete probability spaces determined by W is $(\Omega_\sigma, \pi_\sigma)$, where π_σ is the *density*

$$\pi_\sigma(\omega) = W(\sigma,\omega)/Z(\sigma), \quad \text{with } Z(\sigma) = \sum_{\omega' \in \Omega_\sigma} W(\sigma,\omega')$$

being the corresponding *normalising function*. It is easy to see that the class of normalising functions so defined is essentially Valiant's class $\#\mathsf{P}$. The definition implies that, for some fixed $c \in \mathbb{N}$, $|\Omega_\sigma| \le Z(\sigma) \le 2^{|\sigma|^c}$. If $Z(\sigma) = 0$, then $\Omega_\sigma = \emptyset$ and π_σ is the unique (improper) measure on Ω_σ.

In our definition, two distinct weight functions may define the same sequence of spaces. Therefore let us say weight functions W_1, W_2 are *equivalent* if there exists $\kappa : \Sigma^* \longrightarrow \mathbb{Q}_+$ so that $W_2(\sigma,\omega) = \kappa(\sigma) W_1(\sigma,\omega)$ $(\forall \sigma, \omega \in \Sigma^*)$. Then there is a bijection between sequences of probability spaces $(\Omega_\sigma, \pi_\sigma)$ and equivalence classes of weight functions. Thus, if we write $\widetilde{W}$ for the equivalence class containing W, we may identify it with the sequence $(\Omega_\sigma, \pi_\sigma)$.

We insist that sample spaces are discrete, and weight functions are integer valued. Computationally, discrete spaces are essential. If we wish to work with continuous spaces, then approximations must be made to some predetermined number of bits. The same is true if we are interested in real-valued densities (as in some statistical applications). However, the effect of such approximations can be absorbed into the variation distance of the sampling procedure. The reader may still wonder why we require W to have codomain $\mathbb{N}$ rather than $\mathbb{Q}$, which would seem more natural. This is because we use unnormalised measures, and we wish to avoid the following technical difficulty. In a large sample space it is possible to specify polynomial size rationals for the unnormalised measure which result in exponential size rationals for the probabilities. An example is the set $[2^n]$, with the measure assigning probability proportional to $1/i$ to $i \in [2^n]$. In such spaces there is no possibility of *exact* sampling in sub-exponential expected time, and we must accept approximations. We prefer not to deal with these anomalous spaces, but to insist that these approximations be made explicit. Thus, in this example we could use weights $\lfloor K/i \rfloor$ for some suitably large integer K.

A *fully polynomial approximate sampler* (which we shorten to *good sampler*) for $(\Omega_\sigma, \pi_\sigma)$ is a PTM which, on inputs σ and $\varepsilon \in \mathbb{Q}_+$ $(0 < \varepsilon \leq 1)$, outputs $\omega \in \Sigma^*$, according to a measure μ_σ satisfying $D_{\mathrm{tv}}(\mu_\sigma, \pi_\sigma) \leq \varepsilon$, in time bounded by a bivariate polynomial in $|\sigma|, \log \varepsilon^{-1}$. We allow $\omega \notin \Omega_\sigma$. If $\Omega_\sigma = \emptyset$, the algorithm does not terminate within its time bound. However, this can be detected, and we may construct a polynomial time algorithm which terminates either with a random ω or a proof that Ω_σ is empty.

Our real interest here is in combinatorial Markov chains, which we define as follows. Let $M : \Sigma^* \times \Sigma^* \times \Sigma^* \longrightarrow \mathbb{N}$ and define

$$\mathcal{R}_\sigma = \{(\omega, \omega') : \ M(\sigma, \omega, \omega') > 0\}, \quad \Omega_\sigma = \{\omega : \ \exists w' \text{ with } (\omega, \omega') \in \mathcal{R}_\sigma\}.$$

Let M have the following properties.

(a) There is a polynomial b such that $|\omega|, |\omega'| \leq b(|\sigma|)$ if $M(\sigma, \omega, \omega') > 0$, and M is computable in time polynomial in $|\sigma|$ whenever $M(\sigma, \omega, \omega') > 0$.

(b) There exist constants $K(\sigma) \in \mathbb{N}_+$, of polynomial size, such that

$$\sum_{\omega' \in \Sigma^*} M(\sigma, \omega, \omega') = K(\sigma) \qquad (\forall \omega \in \Omega_\sigma).$$

(c) The transitive closure of $\mathcal{R}_\sigma$ is $\Omega_\sigma \times \Omega_\sigma$, and for some ω, $(\omega, \omega) \in \mathcal{R}_\sigma$.

(d) Writing $M_\omega(\sigma, \omega') = M(\sigma, \omega, \omega')$ $(\omega \in \Sigma^*)$, it follows from (a) that M_ω is a weight function. We require that there is a good sampler for $\widetilde{M}_\omega$ $(\forall \omega)$.

We call M a *density matrix*, and associate with it a sequence of Markov chains $\mathcal{M}_\sigma = (\Omega_\sigma, P_\sigma)$, with transition matrices

$$P_\sigma(\omega_1, \omega_2) = M(\sigma, \omega_1, \omega_2)/K(\sigma) \qquad (\omega_1, \omega_2 \in \Omega_\sigma).$$

Properties (a) and (c) ensure that $\mathcal{M}_\sigma$ is finite and ergodic. Property (d) ensures that we can efficiently simulate $\mathcal{M}_\sigma$ to a close approximation for any given number of steps. Property (b) ensures that polynomial powers of the transition matrix cannot generate rationals of superpolynomial size, and hence the state probabilities at any polynomial time cannot be rationals of superpolynomial size. We include this property since we do not wish to preclude exact generation using Markov chains. In any case, this condition can always be satisfied to any desired approximation, and is usually satisfied naturally. There is little loss in restricting $K(\sigma)$ to be a power of 2. If any such $K(\sigma)$ exist, it is easy to show that there is a chain with the same stationary distribution and K a power of 2, simply by increasing the "self-loop" probability on all states. Since we are interested in the stationary distribution, we can use this slightly slower chain. Thus we may insist on K being a power of 2 where convenient.

Density matrices M_1, M_2 are *equivalent* if there exists $\kappa : \Sigma^* \longrightarrow \mathbb{Q}_+$ such that $M_2(\sigma, \omega, \omega') = \kappa(\sigma)M_1(\sigma, \omega, \omega')$ for all $\sigma, \omega, \omega' \in \Sigma^*$. We can identify the equivalence class $\widetilde{M}$ with the sequence $\mathcal{M}_\sigma$. We say that $\mathcal{M}_\sigma$ is a *rapidly mixing Markov chain* if its mixing time $\tau_\sigma(\varepsilon)$ is bounded by a polynomial in $|\sigma|, \log \varepsilon^{-1}$.

If $\mathcal{M}_\sigma$ is a Markov chain sequence, let π_σ denote the stationary distribution of $\mathcal{M}_\sigma$. Then, if W is a weight function, $\mathcal{M}_\sigma$ is a *Monte Carlo Markov chain* (MCMC) for $\widetilde{W}$ if both $\widetilde{W}$, $\mathcal{M}_\sigma$ determine the same sequence of probability spaces $(\Omega_\sigma, \pi_\sigma)$. (This slight overloading of the MCMC abbreviation should not cause confusion.) The usual way to establish this is by reversibility, i.e. if $W(\sigma,\omega)M(\sigma,\omega,\omega') = W(\sigma,\omega')M(\sigma,\omega',\omega)$ for all $\sigma \in \Sigma^*$ and $\omega, \omega' \in \Omega_\sigma$. Clearly we have a good sampler for $\widetilde{W}$ if $\mathcal{M}_\sigma$ is a rapidly mixing Markov chain.

One of the main applications of sampling is to *approximate integration*. In our setting this means estimating $Z(\sigma)$ to some specified relative error. In the important case where W is a characteristic function, we call the approximate integration problem *approximate counting*. Specifically, a *fully polynomial randomized approximation scheme* (fpras) for $Z(\sigma)$ is a PTM which on input σ, ϵ outputs $\hat{Z}$ so that

$$\mathbf{Pr}(1/(1+\epsilon) \le \hat{Z}/Z \le 1+\epsilon) \ge \tfrac{3}{4},$$

and which runs in time polynomial in $|\sigma|$ and $1/\epsilon$.

The success probability can be increased to $1-\delta$ by taking the median of $O(\log \delta)$ samples, see (1.2).

Let size : $\Sigma^* \longrightarrow \mathbb{N}$ be such that size(σ) is polynomially bounded in $|\sigma|$, and if size$(\sigma') <$ size(σ) then $|\sigma'|$ is polynomially bounded in $|\sigma|$. If size$(\sigma) = 0$, we call the problem a *base problem*. For the class of base problems, we assume the existence of a good sampler and a fpras for $Z(\sigma)$.

For all σ, let $\Xi(\sigma)$ be a polynomial time computable set such that

(a) size$(\xi) <$ size(σ) $(\forall \xi \in \Xi)$.

(b) There exist polynomial time computable constants $k_\xi(\sigma) \in \mathbb{Q}_+$ and injections $\phi_\xi(\sigma) : \Omega_\xi \longrightarrow \Omega_\sigma$ $(\forall \xi \in \Xi)$, such that

$$k_\xi W(\xi,\omega) \le W(\sigma, \phi_\xi(\omega)) \qquad (\forall \omega \in \Omega_\xi).$$

Both $\phi_\xi(\omega)$ and ${\phi_\xi}^{-1}(\omega)$ must be computable in polynomial time, given $\omega \in \Omega_\xi$ and $\omega \in \Omega_\sigma$, respectively.

(c) For some $\zeta \in \Xi$, $Z(\sigma)/\left(k_\zeta(\sigma)Z(\zeta)\right)$ is polynomially bounded in $|\sigma|$.

If $\widetilde{W}$ satisfies these conditions, we call the problem *self-contractible*. Summing over $\omega \in \Omega_\xi$ and using the injectivity of ϕ_ξ shows that (b) implies $k_\xi Z(\xi) \leq Z(\sigma)\ \forall \xi \in \Xi$. Now, suppose we have a good sampler for $\widetilde{W}$. Then we may estimate $k_\xi Z(\xi)/Z(\sigma)$ by rejection sampling. We sample ω from $W(\sigma,\cdot)$, and accept with probability $k_\xi W(\xi,\phi^{-1}(\omega))/W(\sigma,\omega)$ if $\phi^{-1}(\omega) \neq \emptyset$. The overall acceptance probability is

$$\sum_{w\in\phi(\Omega_\xi)} \frac{k_\xi W(\xi,\phi^{-1}(\omega))}{W(\sigma,\omega)}\frac{W(\sigma,\omega)}{Z(\sigma)} = \frac{k_\xi Z(\xi)}{Z(\sigma)}.$$

Moreover, from (c) there is some $\zeta \in \Xi$ such that we can estimate this ratio to sufficient relative accuracy in polynomial time. Since $\text{size}(\zeta) < \text{size}(\sigma)$, we may repeat this process with ζ replacing σ. Then, letting $\sigma_0 = \sigma$, $\sigma_1 = \zeta, \ldots$, we may iterate until $\text{size}(\sigma_r) = 0$. Now $|\sigma_i|$ is polynomially bounded in $|\sigma|$ for all $i = 0, 1, \ldots, r$. For σ_r we can approximate $Z(\sigma_r)$ in polynomial time. Then we may multiply estimates together to approximate

$$Z(\sigma_r)\prod_{i=1}^{r}\frac{Z(\sigma_{i-1})}{k_{\sigma_i}(\sigma_{i-1})Z(\sigma_i)} = \frac{Z(\sigma)}{\prod_{i=1}^{r}k_{\sigma_i}(\sigma_{i-1})}$$

to the required relative error, and hence $Z(\sigma)$. A converse result may be obtained under rather stronger conditions. Suppose that the base problems are such that $Z(\sigma)$ may be determined *exactly* and $\Omega(\sigma)$ can be sampled *perfectly*. Suppose that (b) and (c) are strengthened to

(b)$'$ There exist polynomial time computable constants $k_\xi(\sigma) \in \mathbb{Q}_+$ and injections $\phi_\xi(\sigma) : \Omega_\xi \longrightarrow \Omega_\sigma$ $(\forall \xi \in \Xi)$, such that

$$k_\xi W(\xi,\omega) = W(\sigma,\phi_\xi(\omega)) \qquad (\forall \omega \in \Omega_\xi).$$

Both $\phi_\xi(\omega)$ $(\omega \in \Omega_\xi)$ and ${\phi_\xi}^{-1}(\omega)$ $(\omega \in \Omega_\sigma)$ must be computable in polynomial time.

(c)$'$ The sets $\phi_\xi(\Omega_\xi)$ form a partition of Ω_σ.

Let us call such a problem *self-partitionable*. Clearly (b)$'$ implies (b). Also, from (b)$'$ and (c)$'$, since

$$\begin{aligned}\sum_{\xi\in\Xi} k_\xi Z(\xi) &= \sum_{\xi\in\Xi}\sum_{\omega\in\Omega_\xi} k_\xi W(\xi,\omega) &&= \sum_{\xi\in\Xi}\sum_{\omega\in\Omega_\xi} W(\sigma,\phi_\xi(\omega))\\ &= \sum_{\omega\in\Omega_\sigma} W(\sigma,\omega) &&= Z(\sigma),\end{aligned} \tag{1.19}$$

and the polynomial size of Ξ now implies (c). We sketch the generation procedure, skipping details. Suppose we can estimate $Z(\sigma)$ by $\hat{Z}(\sigma)$ within relative error ϵ to high enough probability. We branch to $\xi \in \Xi$ with probability $k_\xi \hat{Z}(\xi)/(1+\epsilon)\hat{Z}(\sigma)$. If the total of these probabilities over $\Xi(\sigma)$ is more than 1 we "fail", i.e. we abandon this whole sampling "trial". If the total is less than 1, as we would expect, then we fail with the (small) unassigned probability. Otherwise we repeat, getting $\sigma = \sigma_0, \sigma_1, \dots, \sigma_r$ until we reach a base case, and then we generate ω' from $W(\sigma_r, .)$. Then ω is determined from $\omega_{i-1} = \phi_{\sigma_i}(\omega_i)$ $(i = 1, \dots, r)$, with $\omega_0 = \omega$, $\omega_r = \omega'$. Then under the assumption that our approximations $\hat{Z}(\sigma_i)$ are always within bounds, the probability that ω is generated is

$$\frac{k_{\sigma_1}\hat{Z}(\sigma_1)}{(1+\epsilon)\hat{Z}(\sigma_0)}\frac{k_{\sigma_2}\hat{Z}(\sigma_2)}{(1+\epsilon)\hat{Z}(\sigma_1)}\cdots\frac{k_{\sigma_r}Z(\sigma_r)}{(1+\epsilon)\hat{Z}(\sigma_{r-1})}\frac{W(\sigma_r,\omega_r)}{Z(\sigma_r)} = \frac{W(\sigma,\omega)}{(1+\epsilon)^r\hat{Z}(\sigma)},$$

after an easy induction. This is equivalent to the desired weight function. Provided that ϵ is sufficiently small, the failure probability and the variation distance can be kept small on a single trial. Then we may output an arbitrary ω if we fail after some large enough number of trials. Hence the overall variation distance is small. The running time of the algorithm will depend polyonomially on the *logarithm* of the error ϵ, since it is linked to the failure probability of the approximation algorithm. It follows that for self-partitionable problems, approximate integration and good sampling are equivalent. It is easy to see that self-reducible problems are self-partitionable, but the converse is not necessarily true. An example is the volume approximation problem.

We can show that approximate integration implies good sampling under rather weaker conditions than self-partitionability. We do not develop this here, however, since we have no example of a problem satisfying these conditions which is not self-partitionable. In any case, the usual direction in applications is to go from sampling to integration.