Sampling/Particle Based Inference
10708, Fall 2020
Pradeep Ravikumar

1 Inference: Sampling based Viewpoint

So far, we have considered the variational/optimization based viewpoint of inference. Typical inference tasks — computing a marginal, or the partition function — involve summing over exponential many configurations, and the variational viewpoint allows us to reduce this to an optimization problem.

An alternative approach is to reduce (approximately) computing the sums required for graphical model inference to a **sampling problem** i.e. the task of computing samples from the graphical model distribution.

This is a classical reduction, also called a Monte-Carlo approximation. Suppose we have some distribution P with density p over a random vector $X = (X_1, \ldots, X_p) \in \mathcal{X}$. Here, while we are interested in PGMs, let us consider the general distribution case. Then, suppose we are interested in marginal probabilities $P[X_S = x_s]$, for some subset $S \subseteq [p]$, and configurations $x_s \in \mathrm{Val}(X_S)$. It can be seen that:

$$\mathbb{P}[X_S = x_s] = \mathbb{E}[I[X_S = x_s]],$$

so that we are interested in the expectation of an indicator function. More generally, given any function $f : \mathcal{X} \mapsto \mathbb{R}$, suppose we are interested in the expectation:

$$\mu_f = \mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx.$$

Thus, the specific inference task — functional of the joint distribution P — we are interested in is computing some expectation of some function of the RV X.

In such a case, suppose we are given n iid samples $\{X^{(i)}\}_{i=1}^n$ drawn from P. In that case, we could compute the empirical expectation:

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}).$$

The estimate $\widehat{\mu}_f$ is often called a Monte Carlo approximation to the population expectation μ_f, and the samples $\{X^{(i)}\}_{i=1}^n$ used to approximate μ_f are often called "particles" since together they could be viewed as a discrete approximation to the joint distribution itself via the empirical distribution $P_n = \sum i = 1^n \frac{1}{n}\delta_{x^{(i)}}$. For instance, the Monte Carlo approximation of the expectation is simply the expectation of f wrt this empirical distribution.

It can be shown that:

$$\widehat{\mu}_f = \mu_f + N(0, \sigma_f^2/n) + o(1/n),$$

where

$$\sigma_f^2 = \mathrm{Var}(f(X)) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2].$$

At first blush, this seems great: $\widehat{\mu}_f$ seems only $O(\sqrt{\frac{\sigma_f}{n}})$ away from the true expectation μ_f with high probability, so that this deviation does not seem dependent on the dimension p of the RV X (note that $f(.)$ is a real-valued function). The caveat is that this variance term could in general be large for high-dimensional RVs X. To see this, note that $\mu_f = \int_{\mathcal{X}} f(x)p(x)dx$. Then, if there is a mismatch between f and p in the sense that $f(x)$ is large only where $p(x)$ is small, then it might be intuitive that to estimate μ_f well one would need a to be able to sample from the low probability regions wrt density p, for which we would need a lot of samples. Such a mis-match occurs naturally for high-dimensional X, since the corresponding configurations $\mathcal{X} \subseteq \mathbb{R}^p$ lie in a high-dimensional space.

2 Sampling: Warm Up

Typically, pseudo-random generators are used to sample from simple univariate distributions such as the uniform distribution. These apply a deterministic function $D(.)$ repeatedly to an initial seed $x^{(1)}$, so that $x^{(m+1)} = D^m(x)$ is closer to the desired say uniform distribution.

Given a sample U from Unif$[0, 1]$, one can then construct a sample from any other distribution P of some other RV $X \in \mathbb{R}$. Say the CDF of this distribution is F_X, then since $F_X(X) \sim$ Unif$[0, 1]$, we thus have that $F_X^{-1}(U) \sim P$, so that we could simply use $F_X^{-1}(U)$ as a sample from P.

This construction however is not applicable when X is a multivariate random vector in general. Note that many modern "deep" generative models for a RV X construct deterministic mappings f_θ such that $X = f_\theta(U)$ for independent Gaussian or uniform random vector U. In such a case, we can then sample from X by first sampling from the independent uniform RV U, and then simply applying the deterministic mapping f_θ. But when our generative model for X does not have such an explicit "push-forward" mapping from a uniform RV, we need more general strategies, as we discuss in the sequel.

3 Sampling from PGMs

3.1 Sampling from DGMs

Conider a DGM

$$P(X) = \prod_{i=1}^{p} P(X_i \,|\, X_{\mathrm{PA}_i}),$$

over the RV $X = (X_1, \ldots, X_p)$, associated with a DAG G. Let $\sigma \in \mathcal{S}_p$ be an ordering over $[p]$ that respects the DAG, so that children come later in the ordering relative to their parents. Then, suppose we construct a sample $x = (x_1, \ldots, x_p)$ using a so-called "ancestral sampling approach" as follows:

- For $i = 1, \ldots, p$: sample $x_{\sigma(i)} \sim P(\cdot | x_{\mathrm{PA}_{\sigma(i)}})$.

Note that each step involves sampling a univariate RV, which in general is simpler, since in the conditional distribution, the values of the parents are set by the previous steps. After p steps, we generate one sample $x = (x_1, \ldots, x_p)$ from the DGM over X. It can be shown by induction that this is a sample from $P(X)$ due to the acyclicity of the DAG.

3.2 Sampling from DGMs with Evidence

However, the approach above does not work when we have evidence. Suppose we have evidence that some subset of variables E have configuration e; we are then interested in sampling from $P(X|E = e)$. Note that $P(X|E = e) = P(X, E = e)/P(E = e)$, and we have a factorized form only for the unnormalized numerator $\widetilde{P}(X) = P(X, E = e)$. If we use the ancestral sampling procedure above, then we do not have complete freedom to sample $x_i \sim P(\cdot|x_{\mathrm{PA}_i})$, for some evidence variable $X_i \in E$, since we already know its value due to the given evidence. What if we modify the ancestral sampling procedure as follows:

- Set $x_E = e$
- For $i = 1, \ldots, p$: if $X_{\sigma(i)} \notin E$, sample $x_{\sigma(i)} \sim P(\cdot|x_{\mathrm{PA}_{\sigma(i)}})$.

This is however not guaranteed to sample from $P(X|E = e)$, since we are no longer sampling from pieces of the factorized form for $P(X|E = e)$. It can be seen that we are instead sampling from the distribution $P^{\mathrm{do}(E=e)}(X)$, which as we have seen is different from $P(X|E = e)$.

3.3 Sampling from UGMs

In the case of UGMs, we again have a factorized form of the unnormalized distribution: $P(X) = \frac{1}{Z}\widetilde{P}(X)$, where $\widetilde{P}(X) = \prod_{C \in \mathcal{C}} \psi_C(X_C)$. Here there might not seem to be a procedure that has p steps, each sampling from one of p conditional distributions. There is however a procedure that is an iterative modification of the above, called Gibbs Sampling, as we will discuss in the sequel.

But if the UG is tree-structured, then we could use sum-product message passing to then derive the corresponding DAG form, and then use the ancestral sampling approach above. Suppose

$$\log P(x) \propto \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E(T)} \theta_{st}(x_s, x_t).$$

Suppose we can root this tree, at some node X_r. We can then define the set of directed edges $E_r(T)$ s.t. $(s,t) \in E_r(T)$ if $(s,t) \in E(T)$ and s is closer to the root than t. Then we know that we can reparameterize the above via the DGM:

$$\log P(x) = p(x_r) \prod_{(s,t) \in E_r(T)} p(x_s|x_t),$$

where $p(x_s|x_t) = p(x_s, x_t)/p(x_t)$, and the latter marginals can be computed via a round of sum-product message passing.

But for general UGs, there is no simple approach to reduce sampling from the UGM to sampling from univariate distributions. Let us consider the common aspect of the two unresolved cases so far: we have a tractable form for the *unnormalized* distribution that we wish to sample from. Let us consider some general approaches towards this.

4 Rejection Sampling

Consider the following simple approach to sampling from a DGM conditioned on evidence: $P(X|E = e)$.

- For $i = 1, \ldots, p$: if $X_{\sigma(i)} \notin E$, sample $x_{\sigma(i)} \sim P(\cdot|x_{\mathrm{PA}_{\sigma(i)}})$. If $x_{\sigma(i)} \neq e_{X_{\sigma(i)}}$, discard the entire sample, and start again from $i = 1$.

In this variant, we simply follow the standard ancestral sampling approach, but with the modification that we reject a partial sample if it conflicts with the given evidence. This is an instance of what is called **rejection sampling** since we sample from one distribution (in the case above: $P(X)$), and then reject some of those samples (more generally, with some probability), so that the non-rejected samples are drawn from the desired dstribution. In

the case above, it can be shown that the modified ancestral sampling approach samples from the conditional distribution $P(X|E=e)$. But the caveat is that we might end up rejecting a lot of the samples.

Let us consider the more general setup. Suppose we an unnormalized distribution $\widetilde{P}(X)$ that we wish to sample from. Suppose we have can sample from some simpler distribution Q, which satisfies the following condition:

$$\widetilde{P}(x) \leq kQ(x),$$

for all $x \in \mathcal{X}$, and for some $k \in \mathbb{R}_+$.

The rejection sampling approach then generates a sample from $\widetilde{P}$ by:

- sample x from Q, and accept with probability $\frac{\widetilde{P}(x)}{kQ(x)} \in [0,1]$. Keep repeating till you get an accepted sample.

It can be seen that the probability of an accepted sample is

$$\begin{aligned}\frac{Q(x)\frac{\widetilde{P}(x)}{kQ(x)}}{\int_{x\in\mathcal{X}} Q(x)\frac{\widetilde{P}(x)}{kQ(x)}} &= \frac{\widetilde{P}(x)}{\int_{x\in\mathcal{X}} \widetilde{P}(x)} \\ &= P(x),\end{aligned}$$

as required. But this requires many samples, since it is likely that many of these will be rejected. The probability that a random sample from Q will be accepted is:

$$\int_{x\in\mathcal{X}} \frac{\widetilde{P}(x)}{kQ(x)} Q(x)dx = \frac{1}{k}\int_{x\in\mathcal{X}} \widetilde{P}(x)dx.$$

Thus, a distribution Q with k as small as possible is desirable for rejection sampling.

In the case of DGMs with evidence, we have: $\widetilde{P}(x) = P(x,e) = P(x)\delta_e(E)$, and $Q(x) = P(x)$. In this case, since $\widetilde{P}(x) \leq P(x) = Q(x)$, we have that $k = 1$. The probability that a random sample will be accepted is $\widetilde{P}/P = 1/P(e)$ which could be very low for low-probability evidence, so most samples will be rejected. It thus seems very wasteful. What if we could still use all of the samples, just weight them accordingly? This leads to the procedure in the next section.

5 Importance Weighted Sampling

The high-level idea here is to approximate any integral (in our case: $\int_{x\in\mathcal{X}} f(x)p(x)dx$) via a weighted sum of function evaluations at some (random) points. One could of course simply

grid the domain $\mathcal{X}$, via a set of grid points $\{x^{(i)}\}_{i=1}^n$ and approximate the integral as:

$$\sum_{i=1}^{n} f(x^{(i)}p(x^{(i)}).$$

While there are some intelligent approaches to construct such a grid, overall, this is exponential in the dimensionality of the domain $\mathcal{X}$, and is too expensive for practical settings, and definitely for the case of PGMs, where the dimensionality of the domain i.e. the number of variables p is typically too large to allow for an exponential dependence in computational complexity.

Instead, suppose we pick these grid points randomly from some base distribution Q with density q. It can be seen that:

$$\begin{aligned}\int_{x\in\mathcal{X}} f(x)p(x)dx &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\ &\approx \frac{1}{n}\sum_{i=1}^{n} p(x^{(i)})/q(x^{(i)})f(x^{(i)}).\end{aligned}$$

The quantities $r^m = p(x^m)/q(x^m)$ are known as importance weights.

This procedure is also known as *unnormalized* importance sampling. Denoting the estimator above as $\widehat{\mu}_f$, it can be seen that its mean is given as:

$$\begin{aligned}\mathbb{E}[\widehat{\mu_f}] &= \mathbb{E}_{X\sim Q}[p(X)/q(X)f(X)] \\ &= \mathbb{E}_{X\sim P}[f(X)] = \mu_f,\end{aligned}$$

so that it is unbiased. While its variance is given as:

$$\begin{aligned}\mathrm{Var}[\widehat{\mu_f}] &= \mathrm{Var}_{X\sim Q}[p(X)/q(X)f(X)] \\ &= \mathrm{Var}_{X\sim Q}[r(X)f(X)],\end{aligned}$$

where $r(X) = p(X)/q(X)$ are the importance weights. It can be shown that the variance is minimized as a function of Q when

$$q(x) \propto |f(x)|p(x).$$

There is however a caveat with the procedure above: it assumes we have access to the normalized distribution P. What if we only have access to the unnormalized distribution $\widetilde{P}$?

5.1 Normalized Importance Sampling

Suppose we only know P upto its unnormalized form $\widetilde{P}$, and Q upto unnormalized form $\widetilde{Q}$ (but can still sample from Q). Then we can approximate μ_f as:

$$\begin{aligned}\int_{x\in\mathcal{X}} f(x)p(x)dx &= \frac{Z_q}{Z_p}\int f(x)\frac{\widetilde{p}(x)}{\widetilde{q}(x)}q(x)dx \\ &\approx \frac{Z_q}{Z_p}\frac{1}{n}\sum_{i=1}^{n} r(x^{(i)})f(x^{(i)}),\end{aligned}$$

where $r(x^{(i)}) = p(x^{(i)})/q(x^{(i)})$. What about the ratio of normalization constants?

$$\begin{aligned}\frac{Z_q}{Z_p} &= \frac{\int \widetilde{p}(x)dx}{Z_q} \\ &= \int \widetilde{p}(x)\frac{q(x)}{\widetilde{q}(x)} \\ &\approx \frac{1}{n}\sum_{i=1}^{n} r(x^{(i)}),\end{aligned}$$

so that we can approximate the entire expectation μ_f as:

$$\mu_f \approx \frac{\sum_{i=1}^{n} r(x^{(i)})f(x^{(i)})}{\sum_{i=1}^{n} r(x^{(i)})}.$$

Unlike the unnormalized case, this can be seen to be biased. This can be easily seen when $n = 1$. Then the estimate is simply $f(x^{(i)})$, with expectation $\mathbb{E}_{X\sim Q}[f(X)]$ which is in general different from $\mathbb{E}_{X\sim P}[f(X)]$. But this bias converges to zero as $n \to \infty$, at the rate of $1/n$, so is not only asymptotically un-biased, but also with a very small non-asymptotic bias. Its variance can be shown to be:

$$\mathrm{Var}(\widehat{\mu_f}) = \frac{1}{n}\mathrm{Var}_{X\sim P}[f(X)](1 + \mathrm{Var}_{X\sim Q}(r(X))).$$

Recall that if we could sample from P, and compute the Monte Carlo approximation, then the variance of that estimate was $\frac{1}{n}\mathrm{Var}_{X\sim P}[f(X)]$. Thus, normalized importance sampling approach incurs an additional factor of $(1 + Var_{X\sim Q}(r(X)))$ in its variance. One could also cast this as saying that its effective number of samples is the slightly reduced number:

$$n_{\text{eff}} = n/(1 + \mathrm{Var}_{X\sim Q}(r(X)))).$$

UGMs. For UGMs, with access to the unnormalized factorized distribution $\widetilde{P}$, we could thus use normalized importance sampling with respect to some simple e.g. tree-structured approximation to P, that is easy to sample from. As an example, suppose $\log P(x) \propto \sum_{s\in V}\theta_s(x_s)+\sum_{(s,t)\in E(G)}\theta_{st}(x_s,x_t)$, then we could use $\log Q(x) \propto \sum_{s\in V}\theta_s(x_s)+\sum_{(s,t)\in E(T)}\theta_{st}(x_s,x_t)$, for some spanning tree T in the graph G.

DGMs with evidence. Consider the task of sampling from a DGM given some evidence $P(X|E=e)$, and let us revisit the algorithm we were considering earlier:

- Set $x_E = e$
- For $i = 1, \ldots, p$: if $X_{\sigma(i)} \notin E$, sample $x_{\sigma(i)} \sim P(\cdot|x_{\mathrm{PA}_{\sigma(i)}})$.

It can be seen that this sampling distribution is simply $Q(x) = P^{\mathrm{do}(E=e)}(x)$. While this is not the same as the required conditional distribution $P(X|E=e)$, we could use normalized importance sampling given the unnormalized factorized form $\widetilde{P}(x) = P(x,e) = P(x)\delta_E(e)$. In this case, the importance weights, for any x s.t. $x_E = e$ are

$$\begin{aligned} r(x) &= \frac{\widetilde{P}(x)}{Q(x)} = \frac{P(x,e)}{P^{\mathrm{do}(E=e)}(x)} \\ &= \prod_{s \in E} P(X_s = e_{X_s} | x_{\mathrm{PA}_s}) \end{aligned}$$

This approach is called normalized likelihood weighting to compute the conditional $P(Y|E=e)$. We could also use the unnormalized importance sampling approach (called ratio likelihood weighting, for reasons which will be obvious below), by writing:

$$P(Y=y|E=e) = P(Y=y, E=e)/P(E=e),$$

and approximating the numerator and denominator separately via unnormalized importance sampling.

Which of these approaches to use? The main caveat with the ratio likelihood weighting approach is that to compute the conditional for a different value of Y i.e. $P[Y=y'|E=e]$ requires a completely different run for approximating the numerator. The advantage however is that it is easier to analyze its performance, since the numerator and denominator are both easy to analyze. Moreover, because it fixed the values of the variables in Y, it has less variance compared to the normalized likelihood weighting approach that also samples from variables in Y.

Overall, the importance weighting approaches are easy to use, and are broadly applicable, but do not leverage the structure of PGMs. Accordingly, it is more common to use a more advanced variant of the Monte Carlo techniques covered so far to sample from PGMs. We will be covering these so-called Markov Chain Monte Carlo (MCMC) techniques next.

6 Gibbs Sampling

The simplest of these is the Gibbs sampling algorithm.

- Initialize $x^{(0)}$
- For $t = 1, \ldots, T$:
  - Set $x^{(t)} = x^{(t-1)}$.
  - For $i = 1, \ldots, p$:
    - Sample $x_i^{(t)} \sim P(X_i \,|\, X_{-i})$
- Return $\{x^{(t)}\}_{t=1}^T$.

We iteratively sample from the conditional distribution of a variable conditioned on other variables. For factorized distributions such as PGMs, with $P(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(X_C)$, this has a simple closed form:

$$\begin{aligned} P(x_i|x_{-i}) &= \frac{P(x)}{\sum_{x_i'} P(x')} \\ &= \frac{\prod_{C' \text{ s.t. } X_i \notin C'} \psi_{C'}(x_{C'}) \prod_{C \text{ s.t. } X_i \in C} \psi_C(x_C)}{\prod_{C' \text{ s.t. } X_i \notin C'} \psi_{C'}(x'_{C'}) \sum_{x_i'} \prod_{C \text{ s.t. } X_i \in C} \psi_C(x'_C)} \\ &= \frac{\prod_{C \text{ s.t. } X_i \in C} \psi_C(x_C)}{\sum_{x_i'} \prod_{C \text{ s.t. } X_i \in C} \psi_C(x'_C)}. \end{aligned}$$

Note that DGMs can also be written as a product of local factors, so that the above form could be derived for DGMs as well. We thus have a simple closed form expression for the node conditional distributions $P(X_i \,|\, X_{-i})$, and Gibbs sampling just requires us to iteratively sample from these conditional distributions. But unlike ancestral sampling with no evidence DGMs, here we cycle through all the variables for many iterations. It can be seen that the resulting samples $\{x^{(t)}\}_{t=1}^T$ are now no longer independent. They do however satisfy the Markov condition that $x^{(t)} \perp\!\!\!\perp \{x^{(s)}\}_{s=1}^{t-2} \,|\, x^{(t-1)}$. They thus form a Markov chain, and are an instance of the general class of MCMC techniques that we will study in the sequel.

7 Markov Chains

We will first begin with a brief review of Markov chains.

A sequence of random variables $\{x_t\}_{t \in \mathbb{N}}$ is said to be a Markov chain if

$$x_{t+1} \perp\!\!\!\perp x_1, \ldots, x_{t-1} \,|\, x_t.$$

It can thus be completely described by the initial state probability:

$$x_1 \sim p_0(\cdot),$$

and the *transition kernels*

$$T_t(x_t, x_{t+1}) = P(x_{t+1} \mid x_t).$$

The MC is said to be **homogeneous** when the transition kernels are the same for all $t \in \mathbb{N}$.

The marginal probability of the $(t+1)$-th variable, x_{t+1}, in a homogeneous MC can be written as:

$$p(x_{t+1}) = \sum_{x_t} T(x_t, x_{t+1}) p(x_t).$$

If these marginal distributions converge as $t \to \infty$, they would then satisfy the fixed point condition:

$$\pi(x) = \sum_{x'} T(x', x) \pi(x').$$

Such a distribution π as the stationary (since it's a fixed point of the recurrence above) or invariant distribution of the MC. If you view the transition kernel as a matrix, then the stationary distribution is simply its principal eigenvector corresponding to the eigenvalue of one.

A sufficient, but not necessary, condition for a given distribution π to be invariant with respect to the MC transition kernel is for it to satisfy the detailed balance condition:

$$\pi(x) T(x, x') = \pi(x') T(x', x).$$

It can be seen that this entails invariance of π since:

$$\sum_{x'} T(x', x) \pi(x') = \sum_{x'} T(x, x') \pi(x) = \pi(x).$$

An MC where the stationary distribution satisfies this condition is said to be reversible. In general, there need not be a unique invariant/stationary distribution. This can again be easily seen by analogy to eigenvectors of a matrix: when the eigenvalue of one has a multiplicity greater than one. For instance, if the transition kernel is simply the identity transformation, then any distribution will be invariant wrt that MC.

Suppose we wish to sample from a given distribution with density $p(\cdot)$. If we can construct an MC so that $p(\cdot)$ is a stationary distribution of the MC, then we can then simply wait for the marginal distributions to (nearly) converge to the stationary distribution, and then use the samples from the MC as samples from $p(\cdot)$.

But for this to be workable, it would be ideal if the MC marginal distributions converge to the stationary distribution no matter what initial state distribution p_0 we draw the initial state $x^{(0)}$ from. Such an MC is said to be ergodic. It is clear that if an MC is ergodic, then it has a unique stationary distribution. A sufficient condition for ergodicity is for the MC to be **regular**: for each state $x \in \mathcal{X}$, with finite domain $\mathcal{X}$ (e.g. a discrete PGM) to be reachable from any other state $x' \in \mathcal{X}$ in a bounded number of MC steps.

8 Markov Chain Monte Carlo (MCMC)

Recall that the methods we have seen so far, first sample from a surrogate distribution q, and then either weight, or reject these samples so as to shape them towards samples from p. In MCMC, we instead use MCs to sample from a given distribution with density p. We keep track of our current sample x_t, and use a MC transition kernel $q(x_{t+1}|x_t)$ to then draw the next sample x_{t+1}. This transition kernel is sometimes also called the proposal distribution, since it proposes the next sample x_{t+1}. This transition kernel can then be picked so that the stationary distribution of the MC is the desired distribution P.

There is however the caveat that it requires to specify the transition kernel very carefully: there is no further degree of freedom, it either converges to the desired stationary distribution or not. It is indeed possible to construct such kernels for general distributions (e.g. Gibbs sampling), but this restricts the set of MCs we could consider.

8.1 Gibbs Sampling MC

Gibbs sampling can be seen to be an MC, with intermediate steps having transition kernels:

$$T_i(x, x') = P(x'_i|x_{-i}).$$

It might seem that this is not a homogeneous MC, since the transition kernel depends on which variable is being updated; but by viewing the aggregation of the p steps, first applying T_1, then T_2, and so on till T_p, it could be viewed as a homogenous MC i.e. with the transition kernel:

$$T(x, x') = \sum_{x_1,\ldots,x_{p-1}} T_1(x, x^1)T_2(x^1, x^2)\ldots T_p(x^{p-1}, x').$$

If each of the constituent transition kernels satisfy the detailed balance condition, it can be seen that T would satisfy the detailed balance condition as well. Each individual transition kernel need not be ergodic (since it only changes a single variable), but together it can be shown to be ergodic.

Proposition 1 *If the (potentially unnormalized) factorized distribution $\widetilde{P}$ has strictly positive factors. Then the Gibbs MC is regular, and hence ergodic.*

As we will see, the Gibbs MC is an instance of so-called Metropolis-Hastings MC, from which it will follow that this has $P(x)$ as its stationary distribution.

8.2 Metropolis MC

What if we could instead shape the transition kernel by first consider a reasonable proposal distribution $q(x'|x)$, but then also have an acceptance probability $A(x')$, so that the overall MC transition kernel is

$$T(x, x') = (q(x'|x)A(x'))I[x \neq x'] + (1 - \sum_{x' \neq x} q(x'|x)A(x'))I[x = x'].$$

How should we choose the acceptance probability? A natural choice is to prefer transitioning to a configuration x' that has higher probability than the current configuration x:

$$A(x', x) = \min\left(1, \frac{p(x')}{p(x)}\right).$$

Suppose the proposal kernel $q(x'|x)$ is symmetric so that:

$$q(x'|x) = q(x|x').$$

It can then be seen that detailed balance condition holds for such an MC with respect to the distribution $p(\cdot)$:

$$\begin{aligned} p(x)[q(x'|x)A(x', x)] &= \min(q(x'|x)p(x), q(x'|x)p(x')) \\ &= \min(q(x'|x)p(x'), q(x'|x)p(x)) \\ &= p(x')[q(x|x')A(x, x')]. \end{aligned}$$

A key facet of the above acceptance probability computation is that we could perform the computation even if we only know the target distribution upto a normalization constant, since

$$\frac{p(x')}{p(x)} = \frac{\widetilde{p}(x')}{\widetilde{p}(x)}.$$

8.3 Metropolis Hastings

A caveat with the Metropolis MC is that the proposal distribution is required to be symmetric in order for the target distribution $p(\cdot)$ to satisfy the detailed balance condition wrt the MC. But in many contexts, it might be necessary to have non-symmetric proposal distributions. This generalization is called *Metropolis-Hastings*, and uses the following natural modification of the acceptance probability:

$$A(x', x) = \min\left(1, \frac{p(x')q(x'|x)}{p(x)q(x|x')}\right).$$

It can be seen that $p(\cdot)$ satisfies the detailed balance condition wrt this Metropolis-Hastings MC even if q is not symmetric:

$$\begin{aligned} p(x)[q(x'|x)A(x',x)] &= \min(q(x'|x)p(x), q(x'|x)p(x')) \\ &= \min(q(x'|x)p(x'), q(x'|x)p(x)) \\ &= p(x')[q(x|x')A(x,x')]. \end{aligned}$$

As with the Metropolis MC, the Metropolis-Hastings MC also allows for computing the acceptance probability even if we only know the target distribution upto a normalization constant.

8.4 Using MCMC Samples

Suppose we construct an MC such that the desired distribution $P(\cdot)$ is the stationary distribution of the MC. How do we use this to sample from $P(\cdot)$?

One sample per chain. One approach is to run the chain till it converges to its stationary distribution $P(\cdot)$, collect one sample; and then repeat this process, running a fresh chain everytime we need to collect an additional sample. This is obviously time-consuming.

Burn-in. More commonly, one first runs the chain till it is near convergence, the so-called *burn in* phase. Say $\{x_t\}_{t=1}^T$ are the burn-in samples. Then the next n samples $\{x_t\}_{t=1}^T$ are then used to coompute the estimate:

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_{T+i}).$$

The caveat here is that the samples $\{x_t\}_{t=1}^T$ though all have marginal distributions (close to) the target distribution $P(\cdot)$, they are dependent. Nonetheless, it can be seen that this is unbiased:

$$\mathbb{E}[\widehat{\mu}_f] = \mu_f,$$

due to the linearity of expectation.

Moreover, its variance could be written as:

$$\sigma_f^2 = \mathrm{Var}_{X\sim P}[f(X)] + 2\sum_{\ell=0}^{\infty} \mathrm{Cov}[f(X_{T+m}), f(X_{T+m+\ell})].$$

The latter term is a sum of so-called auto-covariance terms, each of which measure the covariance between the function evaluated at different lags along the MC. Note that if the

MC samples were independent, these auto-covariance terms would be equal to zero. This additional component of the variance could thus be viewed as reducing the effective sample size due to the sample dependence.

But should we then simply pick samples separated from each other in the MC, i.e. compute the average of the function over the samples $\{x_{T+1+i*d}\}_{i=0}^{d-1}$, that are separated by a lag of d even after the burn-in of T samples. However it can be shown that using all of the samples $\{x_{T+1+i}\}_{i=0}^{d-1}$ rather than just x_{T+1} always has a lower variance even if the samples are dependent. Thus, it is almost never a good idea to throw away samples. The only exception is where evaluating $f(\cdot)$ is expensive: in that case, evaluating on highly correlated samples might be a poor choice of our evaluation budget.

8.5 Mixing Time

Something that is critical to the practical use of MCMC is for the burn-in time T as discussed above to not be too large. A close and more formally specified notion is that of *mixing time*.

Definition 2 *For any MC $\mathcal{T}$, let $P_{\mathcal{T}}^{(T)}$ denote the marginal distribution of the samples after time T, and let P be its stationary distribution. Then, its ϵ-mixing time $T_{\epsilon;\mathcal{T}}$ is defined as:*

$$T_{\epsilon;\mathcal{T}} = \min\{T \mid D_{TV}(P^{(T)}, P) \leq \epsilon\}.$$

Thus, after the mixing time, the marginal distribution of the MC samples are close in TV-distance to the target/stationary distribution. There is a long literature, still unsatisfactory, and definitely so in the context of PGM target distributions, on characterizing mixing times for MCs. Loosely, the mixing times are at least exponential in key structural complexity terms, so that we cannot just finesse hard problems via MCMC.

Most of these theoretical results are based on the notion of the conductance of a chain.

Definition 3 *For any MC $\mathcal{T}$, let P be its stationary distribution. Then its conductance $\tau_{\mathcal{T}}$ is defined as:*

$$\tau_{\mathcal{T}} = \min_{\mathcal{S} \subset \mathcal{X} \mid P(\mathcal{S}) \in (0,1/2)} \frac{P(\mathcal{S} \to \mathcal{S}^c)}{P(\mathcal{S})},$$

where

$$P(\mathcal{S} \to \mathcal{S}^c) = \sum_{x \in \mathcal{S}, x' \in \mathcal{S}^c} T(x, x').$$

$P(\mathcal{S} \to \mathcal{S}^c)$ measures the likelihood (loosely, how easy is it) to move from configurations in $\mathcal{S}$ to configurations in $\mathcal{S}^c$. If the conductance is too low, this entails that there are some sets

of configurations where the MC might get almost *trapped*, and hence would require a lot of steps to get out of, thus increasing the mixing time. Thus, the mixing time can be shown to be inversely associated with the conductance.