Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

20 Approximations

In many Bayesian inference applications, the belief $p_{\mathsf{x}|\mathsf{y}}(x|y)$ of interest is hard to evaluate or implement exactly. Given a model $p_{\mathsf{y}|\mathsf{x}}(y|x)$ and a prior $p_{\mathsf{x}}(x)$, we can construct the posterior belief

$$p_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{p_{\mathsf{y}|\mathsf{x}}(y|x)\, p_{\mathsf{x}}(x)}{p_{\mathsf{y}}(y)}, \tag{1}$$

and while the numerator terms are given (or easy to obtain), the marginal

$$p_{\mathsf{y}}(y) = \sum_{a \in \mathcal{X}} p_{\mathsf{y}|a}(y|a)\, p_{\mathsf{x}}(a) \tag{2}$$

is often difficult to compute for each $y \in \mathcal{Y}$, unless the problem of interest has special structure.

In nonBayesian modeling, computing a mixture distribution for a particular set of weights over the values of the parameter presents a similar challenge of having to sum (or integrate) over the entire domain $\mathcal{X}$.

In such situations, it is natural to seek good *approximations* for the belief $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$, or for the marginal distribution $p_{\mathsf{y}}(\cdot)$, which serves as a partition function for the belief. The most commonly used approximations can be categorized into two classes: deterministic, and stochastic. We explore both, in turn, at an introductory level.

20.1 Deterministic Approximations

In general, the goal in belief approximation is to find a "simpler" distribution $q(\cdot)$ that is sufficiently "close" to the desired belief $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

Not surprisingly, the notion of "simpler" depends somewhat on the application context. However, the key idea is that evaluating the approximation $q(\cdot)$ for different values of y should require significantly less computation than evaluating the true belief $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

The notion of "close" is easier to quantify. In particular, recall that in our initial development of inference as decision making, we expressed the target belief in the form of the optimization

$$p_{\mathsf{x}|\mathsf{y}}(\cdot|y) = \arg\min_{q} \mathbb{E}\left[C(\mathsf{x}, q)|\mathsf{y} = y\right], \tag{3}$$

where $C(\cdot,\cdot)$ is the log-loss criterion, i.e., $C(x, q) = -\log q(x)$. With this characterization, it is straightforward to incorporate complexity constraints. Specifically, let $\mathcal{Q}(y)$ denote the class of computationally feasible approximations to the belief

$p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$. Then since $D(p_{\mathsf{x}|\mathsf{y}}\|q)$ denotes the increase in expected cost (as measured by the log-loss criterion) when the belief q is used in place of the true belief $p_{\mathsf{x}|\mathsf{y}}$, the "best" belief approximation is

$$q^* = \underset{q\in\mathfrak{Q}(y)}{\arg\min}\, D(p_{\mathsf{x}|\mathsf{y}}(\cdot|y)\|q). \tag{4}$$

The information divergence therefore serves as a natural measure of closeness. But it can be difficult to work with, so working with further approximations and alternatives is common.

We have already seen one type of approximation, which is to restrict the allowable form of the prior $p_{\mathsf{x}}(x)$. For example, when we choose a conjugate prior, we know that the posterior distribution takes a particularly convenient form. This property underlies the popularity of the conjugate prior methodology. Here we explore two other deterministic approximation methods.

20.1.1 Laplace's Method

This method is particularly appropriate for approximating the partition function of distributions defined over continuous random variables. It is named after Laplace who used it first to approximate certain integrals.

Let x be a continuous scalar random variable distributed according to a probability distribution $p(\cdot)$. We know that $p(\cdot)$ satisfies $p(x) \propto p_\circ(x)$ for some computable, nonnegative function $p_\circ(\cdot)$, i.e.,

$$p(x) = \frac{p_\circ(x)}{Z_p}, \tag{5}$$

where

$$Z_p = \int p_\circ(x)\,\mathrm{d}x \tag{6}$$

is the partition function. Our goal is to approximate Z_p (and thus $p(\cdot)$). In particular, we seek to construct an approximation $\hat{p}_\circ(\cdot)$ to $p_\circ(\cdot)$, for which the partition function

$$\hat{Z}_p = \int \hat{p}_\circ(x)\,\mathrm{d}x \tag{7}$$

is easy to compute. This yields an approximation

$$\hat{p}(x) = \frac{\hat{p}_\circ(x)}{Z_{\hat{p}}}$$

to the distribution $p(\cdot)$ that is easy to work with.

With Laplace's method,[1] we make a smoothness assumption and approximate $\ln p_\circ(\cdot)$ using a Taylor series expansion:

$$\ln p_\circ(x) \cong \ln p_\circ(\hat{x}) + (x-\hat{x})\frac{\mathrm{d}}{\mathrm{d}x}\ln p_\circ(x)\Big|_{x=\hat{x}} + \frac{1}{2}(x-\hat{x})^2\frac{d^2}{\mathrm{d}x^2}\ln p_\circ(x)\Big|_{x=\hat{x}} \tag{8}$$

where $\hat{x}$ is a fixed constant. We then further choose

$$\hat{x} = \arg\max_x p(x) = \arg\max_x p_\circ(x), \tag{9}$$

which makes the linear term in (8) equal to zero. This leads to an approximation

$$\hat{p}_\circ(x) = p_\circ(\hat{x})e^{-\frac{1}{2}J(\hat{x})(x-\hat{x})^2} \tag{10}$$

and

$$Z_{\hat{p}} = p_\circ(\hat{x})\sqrt{2\pi J^{-1}(\hat{x})}, \tag{11}$$

where

$$J(\hat{x}) = -\frac{d^2}{\mathrm{d}x^2}\ln p_\circ(x)\Big|_{x=\hat{x}}. \tag{12}$$

Evidently, the resulting approximation $\hat{p}(\cdot)$ is a Gaussian distribution with mean $\hat{x}$ and variance $J^{-1}(\hat{x})$. Note the similarity of the term $J(\hat{x})$ to the Fisher Information we introduced earlier. Indeed, as we will see later in this section when we develop Laplace's approximation for the belief $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$, $J(\hat{x})$ will acquire an interpretation as a kind of Fisher Information.

Let us now consider the associated approximation loss. In particular, if μ and σ^2 denote the mean and variance of the original distribution $p(\cdot)$, then for any approximation $\hat{p}$ that is a Gaussian distribution with some mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, the loss associated with the approximation is given by

$$\begin{aligned} D(p\|\hat{p}) &= -\mathbb{E}_p\left[\ln \hat{p}(\mathsf{x})\right] - h(p) && (13)\\ &= \mathbb{E}_p\left[\frac{1}{2}\left(\ln 2\pi\hat{\sigma}^2\right) + \frac{(\mathsf{x}-\hat{\mu})^2}{2\hat{\sigma}^2}\right] - h(p) && (14)\\ &= \frac{1}{2}\ln\left(2\pi\hat{\sigma}^2\right) + \frac{(\mu-\hat{\mu})^2+\sigma^2}{2\hat{\sigma}^2} - h(p). && (15)\end{aligned}$$

Substituting $\hat{\mu} = \hat{x}$ and $\hat{\sigma}^2 = J^{-1}(\hat{x})$ above we get the divergence that corresponds to the Laplace's approximation.

If we were instead to choose $\hat{\mu}$ and $\hat{\sigma}^2$ that minimize the approximation loss (15), it is straightforward to verify that the optimizing parameters would be $\hat{\mu} = \mu$ and $\hat{\sigma} = \sigma$, from which we get the following lower bound on the Gaussian approximation loss:

$$D(p\|\hat{p}) \geq \frac{1}{2}\ln(2\pi e\sigma^2) - h(p). \tag{16}$$

[1] In statistical physics, this method is referred to as the "saddle-point approximation."

However, we emphasize that the Gaussian distribution corresponding to these optimizing parameters is not admissible. Indeed, computing μ and σ^2 requires computation of Z_p, which we assumed to be impractical. In contrast, $\hat{x}$ and $J(\hat{x})$ have the distinguishing feature that their determination does not require Z_p to be known or computed.

While the Laplace's approximation is easily computable, it is worth noting that the associated approximation loss is generally not. Both the exact expression (15) and the lower bound (16) are not easy to evaluate because they depend on the mean and the variance of the original distribution $p(\cdot)$. Thus, in practice it is difficult to know the quality of the approximation. On the other hand, in many cases, one establish the asymptotic goodness of such approximations, a point to which we will return in a later topic.

Another important observation about the Laplace's approximation is that depends on the parameterization. Specifically, if we transform the random variable x according to $\mathsf{u} = g(\mathsf{x})$ where $g(\cdot)$ is some invertible mapping, then the Laplace's approximation for the distribution of x might be of different quality than the Laplace approximation for the distribution of u. In principle, one can exploit this behavior by choosing the mapping $g(\cdot)$ to yield the best possible approximation. However, since it is sufficiently difficult to even evaluate the approximation loss, this is not commonly attempted.

Our results can be extended to when there are observations y. In this case, we can exploit the Taylor series expansion

$$\ln p_{\mathsf{x}|\mathsf{y}}(x|y) \cong \ln p_{\mathsf{x}|\mathsf{y}}(\hat{x}|y) + \frac{1}{2}(x-\hat{x})^2 \frac{\partial^2}{\partial x^2} \ln p_{\mathsf{x}|\mathsf{y}}(x|y)\bigg|_{x=\hat{x}}, \tag{17}$$

where this time we have chosen

$$\hat{x} = \hat{x}_{\mathrm{MAP}}(y) = \arg\max_{x} p_{\mathsf{x}|\mathsf{y}}(x|y), \tag{18}$$

i.e., $\hat{x} = \hat{x}_{\mathrm{MAP}}(y)$ is the MAP estimate of x based on y.

In this case, we are approximating the belief $p_{\mathsf{x}|\mathsf{y}}(x|y)$ by a Gaussian with the mean given by the MAP estimate $\hat{x}_{\mathrm{MAP}}(y)$, and the inverse of the variance given by

$$\begin{aligned}
\frac{1}{\hat{\sigma}^2} &= -\frac{\partial^2}{\partial x^2} \ln p_{\mathsf{x}|\mathsf{y}}(x|y)\bigg|_{x=\hat{x}_{\mathrm{MAP}}(y)} && (19)\\
&= -\frac{\partial^2}{\partial x^2} \ln p_{\mathsf{x}}(x)\bigg|_{x=\hat{x}_{\mathrm{MAP}}(y)} - \frac{\partial^2}{\partial x^2} \ln p_{\mathsf{y}|\mathsf{x}}(y|x)\bigg|_{x=\hat{x}_{\mathrm{MAP}}(y)} && (20)\\
&= J(\hat{x}_{\mathrm{MAP}}(y)) + \tilde{J}_{\mathsf{y}=y}(\hat{x}_{\mathrm{MAP}}(y)), && (21)
\end{aligned}$$

where

$$\tilde{J}_{\mathsf{y}=y}(x) = -\frac{\partial^2}{\partial x^2} \ln p_{\mathsf{y}|\mathsf{x}}(y|x) \tag{22}$$

is referred to as the "observed" Fisher information.

We emphasize that the obvious alternative, approximating the posterior by a Gaussian distribution with mean $\mu_{\mathsf{x}|\mathsf{y}}(y)$ and variance $\sigma^2_{\mathsf{x}|\mathsf{y}}(y)$ given by those of the true posterior is not feasible, as these moments are not easily computable.

There is a variety of other related useful Gaussian approximations. For example, one can similarly approximate the posterior using a Gaussian of mean

$$\hat{x}_{\mathrm{ML}}(y) = \arg\max_{x} p_{\mathsf{y}|\mathsf{x}}(y|x), \tag{23}$$

corresponding to the ML estimate of x based on y. In this case the corresponding variance in the approximation is $\tilde{J}^{-1}_{\mathsf{y}=y}(\hat{x}_{\mathrm{ML}}(y))$, the inverse of the observed Fisher information evaluated at the ML estimate.

Finally, a further approximation is obtained by replacing the observed Fisher Information with its expected value

$$J_{\mathsf{y}}(x) = \int \tilde{J}_{\mathsf{y}=y}(x)\, p_{\mathsf{y}}(y)\, \mathrm{d}y. \tag{24}$$

For observations $\mathbf{y} = [\mathsf{y}_1, \dots, \mathsf{y}_N]$ that are i.i.d. random variables each generated according to the model $p_{\mathsf{y}|\mathsf{x}}(\cdot|x)$, this approximation becomes particularly accurate in the limit of large N.

As a final remark, extending Laplace's method to vector-valued parameters $\mathbf{x}$ is straightforward, using the corresponding multidimensional Taylor series.

20.1.2 Variational Methods

A different type of deterministic approximations is obtained by modifying the optimization criterion (4). In general, the objective function in (4) is difficult to work with for the same reason that $p_{\mathsf{x}|\mathsf{y}}(x|y)$ itself is difficult to work with. An alternative approach is to replace (4) with the optimization

$$\hat{p} = \arg\min_{q \in \mathcal{Q}(y)} D(q \| p_{\mathsf{x}|\mathsf{y}}(\cdot|y)), \tag{25}$$

which has the advantage that the expectation involved in the objective function is now with respect to simpler distributions q. This is termed the *variational method*, or "ensemble learning" or "variational Bayes" in some communities.

Let us consider an illustration of the application of the method in statistical physics, where it is equivalent to the method of variational free-energy minimization. In particular, consider the state distribution given by the linear exponential family

$$p(x) = \frac{1}{Z(\beta)} e^{-\beta \mathcal{E}(x)}, \quad x \in \mathcal{X}, \tag{26}$$

where β is the natural parameter and $Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta \mathcal{E}(x)}$ is the partition function. The natural statistic $\mathcal{E}(x)$ is the energy associated with state x, which is presumed to be a complicated function of the state. Typically the alphabet $\mathcal{X}$ is very large, corresponding to a high-dimensional state distribution.

In such problems, the partition function is of a particular interest (recall from our development of exponential families that it can generate all moments of $\mathcal{E}(\mathsf{x})$, for example), and is equivalently expressed as the free energy $F(\beta) \triangleq -\ln Z(\beta)$. However, the free energy is typically difficult to calculate, so instead it is approximated by the variational free energy

$$\tilde{F}(\beta; q) \triangleq \mathbb{E}_q \left[\ln \frac{q(x)}{e^{-\beta \mathcal{E}(x)}} \right] = \sum_{x \in \mathcal{X}} q(x) \ln \frac{q(x)}{e^{-\beta \mathcal{E}(x)}}, \tag{27}$$

which is parameterized by a distribution $q(\cdot)$ constrained to lie in a set of tractable distributions $\mathcal{Q}$. Note that the $\tilde{F}(\beta; q)$ is typically straightforward to evaluate from its definition (27) for any particular β of interest, since $\mathcal{E}(x)$ is assumed to be easy to evaluate.

To understand the nature of the free energy approximation obtained by the variational method, observe from (27) that the variational free energy can be equivalently expressed in the form

$$\begin{aligned}\tilde{F}(\beta; q) &= \sum_{x \in \mathcal{X}} q(x) \ln \frac{q(x)}{p(x)} - \ln Z(\beta) \\ &= D(q \| p) + F(\beta),\end{aligned} \tag{28}$$

and thus minimizing $D(q\|p)$ with respect to $q \in \mathcal{Q}$ for a given β is equivalent to minimizing an upper bound on $F(\beta)$. Consequently,

$$\tilde{Z}(\beta; q) = e^{-\tilde{F}(\beta; q)} \leq Z(\beta).$$

Returning to our inference problem, we can interpret this approach as maximizing a lower bound on the partition function Z_p of our distribution of interest. Clearly, the maximal value of the lower bound represents the best approximation of the partition function achieved over the class of distributions $\mathcal{Q}$. One appealing property of the variational method is that it is invariant to re-parameterization.

The quality of variational approximation can be good provided that the class of feasible distributions $\mathcal{Q}$ is in a sufficiently small neighborhood of the true distribution. To be more concrete, let us consider a linear exponential family that connects $p(\cdot)$ and $q(\cdot)$, so that $p(\cdot; 0) = p(\cdot)$ and $p(\cdot; \delta) = q(\cdot)$. Using the Taylor series expansion, we obtain

$$\ln p(x; \delta) = \ln p(x; 0) + \delta \frac{\partial}{\partial \theta} \ln p(x; \theta)|_{\theta=0} + \frac{1}{2} \delta^2 \frac{\partial^2}{\partial \theta^2} \ln p(x; \theta)|_{\theta=0}, \tag{29}$$

which implies

$$\begin{aligned} D(p\|q) &= D(p(\cdot;0)\|p(\cdot;\delta)) && (30)\\ &= \mathbb{E}_{p(\cdot;0)}\left[\ln p(\cdot;0) - \ln p(\cdot;\delta)\right] && (31)\\ &= -\frac{1}{2}\delta^2\ \mathbb{E}_{p(\cdot;0)}\left[\frac{\partial^2}{\partial\theta^2}\ln p(x;\theta)\bigg|_{\theta=0}\right] && (32)\\ &\cong \frac{1}{2}\delta^2 J_{\mathsf{x}}(0), && (33) \end{aligned}$$

where $J_{\mathsf{x}}(\theta)$ is the Fisher Information of the model $p(\cdot;\theta)$ evaluated at θ. Similarly, we can show that

$$D(q\|p) = D(p(\cdot;\delta)\|p(\cdot;0)) \cong \frac{1}{2}\delta^2 J_{\mathsf{x}}(\delta). \tag{34}$$

While both divergences decrease as δ^2 with $\delta \to 0$, their difference decreases even faster. In particular, using the Taylor series expansion for $J_{\mathsf{x}}(\delta)$, we obtain

$$D(q\|p) - D(p\|q) \cong \frac{1}{2}\delta^3 J'_{\mathsf{x}}(0). \tag{35}$$

Therefore, for likelihood models whose Fisher Information has a bounded derivative, the approximate loss $D(q\|p)$ is close to the original approximation loss $D(p\|q)$.

We conclude the section with a very simple example.

Example 1. Let $\mathsf{x}_1, \mathsf{x}_2$ be zero-mean, jointly Gaussian random variables with some correlation between them and generally different variances σ_1^2 and σ_2^2. Now suppose we seek a variational approximation such that x_1 and x_2 are still zero-mean and jointly Gaussian, but now i.i.d. Our goal is to determine the variance σ^2 of the best approximating distribution of this form.

With a bit of straightforward calculation we obtain that the σ^2 that minimizes the appropriate information divergence (25) must satisfy

$$\frac{1}{\sigma^2} = \frac{1}{2}\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right], \tag{36}$$

which for $\sigma_1 \gg \sigma_2$ is roughly $\sigma^2 \cong 2\sigma_2^2$. By contrast, were we to minimize the true approximation loss (4), we would obtain

$$\sigma^2 = \frac{1}{2}\left[\sigma_1^2 + \sigma_2^2\right], \tag{37}$$

which for $\sigma_1 \gg \sigma_2$ is roughly $\sigma^2 \cong (1/2)\sigma_1^2$. Evidently, the variational result can be quite different.

20.2 Stochastic Approximations

Stochastic approximation methods, also called *Monte-Carlo* methods, or sampling methods, also aim to approximate a belief $p_{\mathsf{x}|\mathsf{y}}$, or its features, such as mean, variance and others. But in contrast to the deterministic approximations, these methods construct estimates from data samples.

As before, we remove explicit conditioning on y from the problem and assume that the distribution $p(\cdot)$ is proportional to an easily computable function $p_\circ(x)$,

$$p(x) = \frac{p_\circ(x)}{Z_p}, \tag{38}$$

where

$$Z_p = \int p_\circ(x)\,\mathrm{d}x \tag{39}$$

is the partition function. We seek to approximate Z_p or, more generally, the expectation $\mathbb{E}_p\left[f(\mathsf{x})\right]$ for some function $f(\cdot)$ of interest. For example, for $f(x) = x$, the expectation of interest is the mean of the distribution p, i.e., $\mathbb{E}\left[f(\mathsf{x})\right] = \mathbb{E}\left[\mathsf{x}\right]$. As another example, for[2] $f(x) = \mathbb{1}_{x<x'}$ for some x', the expectation of interest corresponds to the cumulative distribution at x', i.e., $\mathbb{E}\left[f(\mathsf{x})\right] = \mathbb{P}\left(\mathsf{x} < x'\right)$.

The basic idea underlying sampling methods is to generate n i.i.d. samples $\mathsf{x}_1, \ldots, \mathsf{x}_n$ from the distribution p, then form the approximation

$$\hat{\bar{\mathsf{f}}} \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathsf{x}_i) \tag{40}$$

for the quantity

$$\bar{f} \triangleq \mathbb{E}\left[f(\mathsf{x})\right]. \tag{41}$$

Such an approximation has the desirable properties

$$\mathbb{E}\left[\hat{\bar{\mathsf{f}}}\right] = \bar{f} \tag{42}$$

and

$$\operatorname{var}(\hat{\bar{\mathsf{f}}}) = \frac{1}{n}\operatorname{var}(f) \to 0 \quad \text{as } n \to \infty. \tag{43}$$

While the accuracy of the approximation does not depend explicitly on the alphabet size $|\mathcal{X}|$ of x, we will see later in this section that the difficulty of getting i.i.d. samples grows with the alphabet size. This reduces the convergence rate of the approximation.

To use an approximation of the form (40), we need methods for generating samples from the distribution p, which is assumed to be difficult. Thus, inherent in the problem of approximating expectations via (40) is the problem of generating samples from

[2] The function $\mathbb{1}_{\mathcal{A}}$ is one if $\mathcal{A}$ is true, and zero otherwise. Note that this additional Kronecker notation relates to our other notation via $\mathbb{1}_{x=a} = \mathbb{1}_x(a)$.

a target distribution. Most stochastic approximation methods take an approach of obtaining samples from a different, easy to sample, distribution q, and then modifying either the samples or the approximation (40) to produce the correct estimate of the expectation. The methods typically require that $q(x) > 0$ for all $x \in \mathcal{X}$ such that $p(x) > 0$, i.e., the support of the distribution q includes the support of the distribution p. Intuitively, this is a necessary requirement as it ensures that we have a chance to sample all relevant values of x. Below, we introduce three substantially different ways to construct such approximations.

20.2.1 Importance Sampling

The primary goal of importance sampling is to produce an approximation to the expectation $\mathbb{E}_p[f(\mathsf{x})]$ of interest; obtaining samples from p is not of direct interest. We assume that we have a distribution q from which we can readily generate samples; q is termed the *sampler distribution*. Examples of sampler distributions that are easy to work with include the uniform and Gaussian distributions. We need not assume that we have a normalized version of q. Rather, it suffices to have access to some nonnegative function $q_\circ$ such that

$$q(x) = \frac{q_\circ(x)}{Z_q} \tag{44}$$

with

$$Z_q = \sum_{x \in \mathcal{X}} q_\circ(x). \tag{45}$$

Given n i.i.d. samples $\mathbf{x} = [x_1, \ldots, x_n]$ from the distribution q, the importance sampling method constructs an estimate

$$\hat{\hat{f}}(\mathbf{x}) \triangleq \sum_{i=1}^{n} \frac{w(x_i)}{\sum_{i'=1}^{n} w(x_{i'})} f(x_i), \tag{46}$$

where

$$w(x) = \frac{p_\circ(x)}{q_\circ(x)} \tag{47}$$

are the *importance weights*. In other words, the approximation computes a convex combination of the observed values of the function f, with the weights proportional to the ratio of the distributions p and q at the corresponding values of x.

Let us analyze the quality of this approximation. We first re-arrange the terms in the approximation (46) to make it easier to analyze:

$$\hat{\hat{f}}(\mathbf{x}) \triangleq \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w(x_i) f(x_i)}{\dfrac{1}{n}\displaystyle\sum_{i'=1}^{n} w(x_{i'})}. \tag{48}$$

Now since

$$\mathbb{E}_q\left[w(\mathsf{x})\right] = \sum_{x\in\mathcal{X}} \frac{p_\circ(x)}{q_\circ(x)} q(x) = \frac{Z_p}{Z_q} \sum_{x\in\mathcal{X}} p(x) = \frac{Z_p}{Z_q}, \tag{49}$$

the strong law of large numbers implies that the denominator of (48) converges to Z_p/Z_q as $n \to \infty$. This result also suggests a way to estimate the partition function Z_p if Z_q is known:

$$\hat{Z}_p(\mathbf{x}) = \frac{Z_q}{n} \sum_{i=1}^{n} w(x_i). \tag{50}$$

Similarly, we have

$$\mathbb{E}_q\left[w(\mathsf{x})\, f(\mathsf{x})\right] = \sum_{x\in\mathcal{X}} \frac{p_\circ(x)}{q_\circ(x)} f(x)\, q(x) = \frac{Z_p}{Z_q} \sum_{x\in\mathcal{X}} f(x)\, p(x) = \frac{Z_p}{Z_q} \bar{f}, \tag{51}$$

and by the strong law of large numbers, the numerator of (48) converges to $\bar{f}\, Z_p/Z_q$ as $n \to \infty$. Combining these two results, we obtain that $\hat{\bar{\mathsf{f}}} \to \bar{f}$ as $n \to \infty$.

In general, the convergence rate deteriorates as the alphabet size $L = |\mathcal{X}|$ increases. Furthermore, the rate of convergence depends on the choice of the sampler distribution q. Unfortunately, evaluating the quality of particular sampler distribution for approximating expectations with respect to the given distribution p can itself be hard. Qualitatively, if $q_\circ(x)$ is small for values of x where $|f(x)p_\circ(x)|$ is large, the number of samples n may need be be very large before we get samples in such regions. This suggests using a heavy-tailed distribution for q.

Example 2. Suppose we use uniform distribution as the sampler distribution. In this case, $q_\circ(x) = 1$ for all $x \in \mathcal{X}$, $Z_q = |\mathcal{X}|$. The importance weights are defined as $w(x) = p_\circ(x)$ and the estimate of the partition function takes the following form:

$$\hat{Z}_p(\mathbf{x}) = \frac{|\mathcal{X}|}{n} \sum_{i=1}^{n} p_\circ(x_i). \tag{52}$$

We substitute these specific expressions into (48) to instantiate the approximation of $\bar{f}$ for this case:

$$\hat{\bar{\mathsf{f}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} w(x_i) f(x_i)}{\frac{1}{n}\sum_{i'=1}^{n} w(x_{i'})} = \frac{\frac{1}{n}\sum_{i=1}^{n} p_\circ(x_i) f(x_i)}{\frac{\hat{Z}_p(\mathbf{x})}{|\mathcal{X}|}} = \frac{|\mathcal{X}|}{n} \sum_{i=1}^{n} \frac{p_\circ(x_i)}{\hat{Z}_p} f(x_i). \tag{53}$$

Such an approximation can be problematic even when Z_p is computable exactly, thus obviating the need for $\hat{Z}_p$. To show this, we analyze the approximation for the case when Z_p is known:

$$\hat{\bar{\mathsf{f}}} = \frac{|\mathcal{X}|}{n} \sum_{i=1}^{n} p(x_i) f(x_i). \tag{54}$$

Since the samples are generated by the uniform distribution, the variance of the approximation

$$\mathrm{var}_{\mathcal{U}}\left(\hat{\bar{f}}\right) = \frac{|\mathcal{X}|^2}{n}\,\mathrm{var}_{\mathcal{U}}\left(f(\mathsf{x})\,p(\mathsf{x})\right), \tag{55}$$

can be quite large when the alphabet $\mathcal{X}$ is large. Intuitively, this happens when we have many realizations x where $p(x)$ is low and undersample the regions where $p(x)$ is high.

20.2.2 Rejection Sampling

Unlike importance sampling that modified the approximation itself, rejection sampling aims to generate samples from the original distribution p. These samples can then be used to approximate expectations of the form (40).

As before, we assume that we have an unnormalized distribution $q_\circ$ from which we can generate samples. We refer to this distribution as the *proposal distribution*. Moreover, we assume to know a constant c such that

$$cq_\circ(x) > p_\circ(x) \quad \text{for all } x. \tag{56}$$

To construct a set of i.i.d. samples from the distribution p, we proceed as follows. First, we generate an observation x from q, then we generate an observation u from a uniform distribution over the interval $[0, cq_\circ(x)]$. If $u \leq p_\circ(x)$, we keep the sample x. Otherwise, we discard it.

To convince ourselves that the resulting set of samples represents a set of i.i.d. observations of x, let us consider the area under the curve $cq_\circ(\cdot)$. With x being the horizontal coordinate in the plane, let u be the vertical coordinate. To verify that the retained samples have the desired distribution, it suffices to note that the retained pairs $(\mathsf{x}_i, \mathsf{u}_i)$ are uniformly distributed over the area below the $p_\circ(\cdot)$ curve.

The retained samples can then be used effectively in conjunction with (40). And while the variance of the approximation (40) decays inversely with the number of retained samples n, the decay with the number of *generated* samples can be much slower, depending on how "close" $cq_\circ(\cdot)$ is to $p_\circ(\cdot)$. Indeed, the rejection rate is proportional to the area $cZ_q - Z_p$ between these curves, which grows with c. It is also worth emphasizing that c grows in direct proportion to the alphabet size $|\mathcal{X}|$, so the acceptance rate in practice can be quite small in high-dimensional problems.

Example 3. As a simple continuous-alphabet example, let $p \sim \mathtt{N}(\mathbf{0}, \sigma_p^2\mathbf{I})$ and $q \sim \mathtt{N}(\mathbf{0}, \sigma_q^2\mathbf{I})$, be d-dimensional Gaussian distributions. In this case, the smallest bounding constant c ensures that the two distributions touch at $\mathbf{x} = \mathbf{0}$:

$$c = \frac{(2\pi\sigma_q^2)^{d/2}}{(2\pi\sigma_p^2)^{d/2}}. \tag{57}$$

Suppose $\sigma_q^2 = (1+\epsilon)\sigma_p^2$ for some $\epsilon > 0$, which can be arbitrarily small. Then $c = (1+\epsilon)^{d/2} \to \infty$ as $d \to \infty$ and thus the area $c - 1$ between the curves is also unbounded in d.

20.2.3 Markov Chain Monte Carlo

The Markov Chain Monte Carlo (MCMC) methods are considered the most powerful among the sampling methods. Like rejection sampling, these methods aim to generate samples from p. MCMC methods avoid the requirement that the proposal distribution q be "close" to the target distribution p, which makes them viable even in high dimensions. All MCMC methods *asymptotically* produce *correlated* samples from the target distribution p.

To implement an MCMC method, we construct a reversible, ergodic Markov chain whose steady state probabilities are the probabilities of the different symbols in $\mathcal{X}$. Hence, the chain has $|\mathcal{X}|$ states, which we denote $1, 2, \ldots, |\mathcal{X}|$. When a sample path is generated from such a chain starting from any initial state, the sequence of states $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_i$ has the property that

$$\frac{1}{n}\sum_{i=1}^{n} f(\mathsf{x}_i) \to \bar{f} \quad \text{as } n \to \infty, \tag{58}$$

though the rate, of course, depends on the statistical dependence between the samples.

Many MCMC methods have been developed over the years. We briefly describe one example, corresponding to the original Metropolis-Hastings method.

We let $q(\cdot; x_i)$ be a proposal distribution that we can evaluate and from which we can sample. We use this distribution to construct the transition distribution for the Markov chain. In our notation, x_i is the current state; $q(\cdot; x_i)$ generates a candidate next state. The simplest example of such proposal distribution is the uniform distribution over the alphabet $\mathcal{X}$. For continuous x, $q(\cdot; x_i)$ could be a Gaussian distribution with mean x_i and fixed variance σ^2. We emphasize that $q(\cdot; x_i)$ need not look similar to $p(\cdot)$ for any x_i.

Given a current state x_i, we generate a candidate next state x' from $q(\cdot; x_i)$, then compute an acceptance probability

$$\alpha(x', x_i) = \min\left\{1, \frac{p_\circ(x')\, q(x_i; x')}{p_\circ(x_i)\, q(x'; x_i)}\right\}. \tag{59}$$

With probability $\alpha(x', x_i)$, we transition to the state x', otherwise we remain in the current state. Specifically,

$$x_{k+1} = \begin{cases} x' & \text{with probability } \alpha(x', x_k) \\ x_k & \text{with probability } 1 - \alpha(x', x_k). \end{cases} \tag{60}$$

It can be shown that the transition probability distribution $p_{\mathsf{x}_{k+1}|\mathsf{x}_k}(\cdot|\cdot) \triangleq p(\cdot|\cdot)$ for this chain takes the form

$$p(x_{k+1}|x_k) = q(x_{k+1}; x_k)\, \alpha(x_{k+1}, x_k) + \mathbb{1}_{x_k}(x_{k+1}) \left[1 - \sum_a q(a; x_k)\, \alpha(a, x_k)\right]. \tag{61}$$

We can also check that $p(\cdot;\cdot)$ satisfies the local, or detailed, balance equations

$$p(m)\,p(m'|m) = p(m')\,p(m|m'), \tag{62}$$

where p is the target distribution, to verify that the Markov chain is reversible.

In turn, when combined with ergodicity, this implies that p is the steady-state distribution. Indeed, summing (62) over m' gives the balance equations

$$p(m) = p(m)\sum_{m'} p(m'|m) = \sum_{m'} p(m')\,p(m|m'). \tag{63}$$

And as a final remark, we point out that this construction works for both discrete and continuous distributions, as do the other belief approximation techniques we have discussed, though sometimes the continuous case requires additional care in the analysis.

20.3 Further Reading

MacKay provides a good introduction to the concepts we discussed here. Both Pawitan and Berger have some additional discussion of Laplace-like approximations to the posterior distributions. The tutorial paper by Jaakkola offers a more in-depth discussion of the variational methods.