

18.7 Estimating the Partition Function

While much of this chapter is dedicated to describing methods that avoid needing to compute the intractable partition function $Z(\boldsymbol{\theta})$ associated with an undirected graphical model, in this section we discuss several methods for directly estimating the partition function.

Estimating the partition function can be important because we require it if we wish to compute the normalized likelihood of data. This is often important in *evaluating* the model, monitoring training performance, and comparing models to each other.

For example, imagine we have two models: model \mathcal{M}_A defining a probability distribution $p_A(\mathbf{x}; \boldsymbol{\theta}_A) = \frac{1}{Z_A} \tilde{p}_A(\mathbf{x}; \boldsymbol{\theta}_A)$ and model \mathcal{M}_B defining a probability distribution $p_B(\mathbf{x}; \boldsymbol{\theta}_B) = \frac{1}{Z_B} \tilde{p}_B(\mathbf{x}; \boldsymbol{\theta}_B)$. A common way to compare the models is to evaluate and compare the likelihood that both models assign to an i.i.d. test dataset. Suppose the test set consists of m examples $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$. If $\prod_i p_A(\mathbf{x}^{(i)}; \boldsymbol{\theta}_A) > \prod_i p_B(\mathbf{x}^{(i)}; \boldsymbol{\theta}_B)$, or equivalently if

$$\sum_i \log p_A(\mathbf{x}^{(i)}; \boldsymbol{\theta}_A) - \sum_i \log p_B(\mathbf{x}^{(i)}; \boldsymbol{\theta}_B) > 0, \quad (18.38)$$

then we say that \mathcal{M}_A is a better model than \mathcal{M}_B (or, at least, it is a better model of the test set), in the sense that it has a better test log-likelihood. Unfortunately, testing whether this condition holds requires knowledge of the partition function. Indeed, equation 18.38 seems to require evaluating the log-probability that the model assigns to each point, which in turn requires evaluating the partition function. We can simplify the situation slightly by rearranging equation 18.38 into a form in

which we need to know only the **ratio** of the two model's partition functions:

$$\sum_i \log p_A(\mathbf{x}^{(i)}; \boldsymbol{\theta}_A) - \sum_i \log p_B(\mathbf{x}^{(i)}; \boldsymbol{\theta}_B) = \sum_i \left(\log \frac{\tilde{p}_A(\mathbf{x}^{(i)}; \boldsymbol{\theta}_A)}{\tilde{p}_B(\mathbf{x}^{(i)}; \boldsymbol{\theta}_B)} \right) - m \log \frac{Z(\boldsymbol{\theta}_A)}{Z(\boldsymbol{\theta}_B)}. \quad (18.39)$$

We can thus determine whether \mathcal{M}_A is a better model than \mathcal{M}_B without knowing the partition function of either model but only their ratio. As we will see shortly, we can estimate this ratio using importance sampling, provided that the two models are similar.

If, however, we wanted to compute the actual probability of the test data under either \mathcal{M}_A or \mathcal{M}_B , we would need to compute the actual value of the partition functions. That said, if we knew the ratio of two partition functions, $r = \frac{Z(\boldsymbol{\theta}_B)}{Z(\boldsymbol{\theta}_A)}$, and we knew the actual value of just one of the two, say $Z(\boldsymbol{\theta}_A)$, we could compute the value of the other:

$$Z(\boldsymbol{\theta}_B) = rZ(\boldsymbol{\theta}_A) = \frac{Z(\boldsymbol{\theta}_B)}{Z(\boldsymbol{\theta}_A)} Z(\boldsymbol{\theta}_A). \quad (18.40)$$

A simple way to estimate the partition function is to use a Monte Carlo method such as simple importance sampling. We present the approach in terms of continuous variables using integrals, but it can be readily applied to discrete variables by replacing the integrals with summation. We use a proposal distribution $p_0(\mathbf{x}) = \frac{1}{Z_0} \tilde{p}_0(\mathbf{x})$, which supports tractable sampling and tractable evaluation of both the partition function Z_0 and the unnormalized distribution $\tilde{p}_0(\mathbf{x})$.

$$Z_1 = \int \tilde{p}_1(\mathbf{x}) d\mathbf{x} \quad (18.41)$$

$$= \int \frac{p_0(\mathbf{x})}{p_0(\mathbf{x})} \tilde{p}_1(\mathbf{x}) d\mathbf{x} \quad (18.42)$$

$$= Z_0 \int p_0(\mathbf{x}) \frac{\tilde{p}_1(\mathbf{x})}{\tilde{p}_0(\mathbf{x})} d\mathbf{x}, \quad (18.43)$$

$$\hat{Z}_1 = \frac{Z_0}{K} \sum_{k=1}^K \frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})} \quad \text{s.t. : } \mathbf{x}^{(k)} \sim p_0. \quad (18.44)$$

In the last line, we make a Monte Carlo estimator, \hat{Z}_1 , of the integral using samples drawn from $p_0(\mathbf{x})$, and then weight each sample with the ratio of the unnormalized \tilde{p}_1 and the proposal p_0 .

This approach also allows us to estimate the ratio between the partition

functions as

$$\frac{1}{K} \sum_{k=1}^K \frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})} \quad \text{s.t. : } \mathbf{x}^{(k)} \sim p_0. \quad (18.45)$$

This value can then be used directly to compare two models as described in equation 18.39.

If the distribution p_0 is close to p_1 , equation 18.44 can be an effective way of estimating the partition function (Minka, 2005). Unfortunately, most of the time p_1 is both complicated (usually multimodal) and defined over a high-dimensional space. It is difficult to find a tractable p_0 that is simple enough to evaluate while still being close enough to p_1 to result in a high-quality approximation. If p_0 and p_1 are not close, most samples from p_0 will have low probability under p_1 and therefore make (relatively) negligible contributions to the sum in equation 18.44.

Having few samples with significant weights in this sum will result in an estimator that is of poor quality because of high variance. This can be understood quantitatively through an estimate of the variance of our estimate \hat{Z}_1 :

$$\hat{\text{Var}}(\hat{Z}_1) = \frac{Z_0}{K^2} \sum_{k=1}^K \left(\frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})} - \hat{Z}_1 \right)^2. \quad (18.46)$$

This quantity is largest when there is significant deviation in the values of the importance weights $\frac{\tilde{p}_1(\mathbf{x}^{(k)})}{\tilde{p}_0(\mathbf{x}^{(k)})}$.

We now turn to two related strategies developed to cope with the challenging task of estimating partition functions for complex distributions over high-dimensional spaces: annealed importance sampling and bridge sampling. Both start with the simple importance sampling strategy introduced above, and both attempt to overcome the problem of the proposal p_0 being too far from p_1 by introducing intermediate distributions that attempt to *bridge the gap* between p_0 and p_1 .