Estimating the Partition Function by Discriminance Sampling

Qiang Liu*
CSAIL, MIT & Computer Science, Dartmouth College
qliu@cs.dartmouth.edu

Jian Peng*
Computer Science, UIUC
jianpeng@illinois.edu

Alexander Ihler
Information and Computer Science, UC Irvine
ihler@ics.uci.edu

John Fisher III
CSAIL, MIT
fisher@csail.mit.edu

Abstract

Importance sampling (IS) and its variant, annealed IS (AIS) have been widely used for estimating the partition function in graphical models, such as Markov random fields and deep generative models. However, IS tends to underestimate the partition function and is subject to high variance when the proposal distribution is more peaked than the target distribution. On the other hand, "reverse" versions of IS and AIS tend to overestimate the partition function, and degenerate when the target distribution is more peaked than the proposal distribution. In this work, we present a simple, general method that gives much more reliable and robust estimates than either IS (AIS) or reverse IS (AIS). Our method works by converting the estimation problem into a simple classification problem that discriminates between the samples drawn from the target and the proposal. We give extensive theoretical and empirical justification; in particular, we show that an annealed version of our method significantly outperforms both AIS and reverse AIS as proposed by Burda et al. (2015), which has been the state-of-the-art for likelihood evaluation in deep generative models.

1 INTRODUCTION

Probabilistic graphical models, such as Markov random fields, Bayesian networks, and deep generative models provide a powerful set of tools for machine learning (e.g., Lauritzen, 1996, Salakhutdinov and Hinton, 2009). Bayesian analysis utilizing graphical models often involves calculating the partition function, *i.e.* the normalizing constant of the distribution. Unfortunately, such computations are prohibitive (often intractable) for general loopy graphical models. As such, efficient approximations via variational inference and Monte Carlo methods are of great interest.

Importance sampling (IS) and its variants, such as annealed importance sampling (AIS) (Neal, 2001), are probably the most widely used Monte Carlo methods for estimating the partition function. IS works by drawing samples from a tractable proposal (or reference) distribution $p_0(x)$, and estimates the target partition function $Z = \int f(x)$ by averaging the importance weights $f(x)/p_0(x)$ across the samples. Unfortunately, the IS estimate often has very high variance if the choice of proposal distribution is very different from the target, especially when the proposal is more peaked or has thinner tails than the target. In addition, in practice IS often underestimates the partition function due to the heavy-tailed nature of the importance weights, leading to overly optimistic likelihood estimates when used for model evaluation (e.g., Burda et al., 2015).

On the other hand, the weighted harmonic mean method (Gelfand and Dey, 1994), which we refer to as a *reverse importance sampling* (RIS), works in an opposite way from IS. It draws samples from the *target distribution* $p(x) = f(x)/Z$, and estimates Z by taking the harmonic mean of the importance weights $f(x)/p_0(x)$ across these samples. In contrast to IS, reverse IS tends to overestimate the partition function, and gives a high variance when the target distribution is more peaked than the proposal. A reverse version of annealed importance sampling was recently proposed by Burda et al. (2015) to give conservative estimates of test likelihood, in contrast to standard AIS that (like IS) tends to overestimate the likelihood.

Given the opposing properties of IS and RIS, a natural way to improve both is to take the average, or weighted average, of their estimates, in the hope of canceling their individual biases. Unfortunately, the magnitude of bias in IS and RIS can be extremely imbalanced, and it is difficult to decide how much weight each should be given. What is worse, the average would inherent the largest variance between IS and RIS, making even more stringent requirements on the proposal, which should then be neither more peaked nor more flat than the target.

*Both authors contributed equally.

In this work, we study a more efficient and straightforward method for estimating the partition function based on samples from *both* the target distribution and the proposal distribution. The idea is to re-frame estimation of Z as a simple classification problem that discriminates between the two samples from the target $p(x)$ and proposal $p_0(x)$, respectively. Our method does not have the inherent biases observed in IS and RIS, and is much more robust in that it always has finite variance whenever the proposal and target distributions overlap, a far more mild condition that is easy to satisfy in practice. We provide extensive theoretical and empirical justification for our method. In addition, we show that an annealed importance sampling (AIS) counterpart of our method significantly outperforms AIS, reverse AIS, and their average, which are currently state-of-the-art for model evaluation in deep generative models (Salakhutdinov and Murray, 2008, Burda et al., 2015).

Outline The remainder of the paper is organized as follows. We discuss related work in Section 2 and introduce background on IS and RIS in Section 3. Section 4 discusses the proposed method followed by an annealed extension in Section 5. We give experiments in Section 6. The conclusion is provided in Section 7.

2 RELATED WORK

The same idea of estimating normalization constants by discriminating between different samples was first proposed independently by Geyer (1991, 1994), although it seems not to be well known in the machine learning community;[1] our work appears to be the first to apply the idea in graphical models, and importantly, propose the annealed version that we show is effective on challenging deep generative models. Another interesting connection can be drawn with a recent noise-matching algorithm (Gutmann and Hyvärinen, 2010) for learning graphical models with intractable partition functions, which is based on a similar idea of discriminating between the observed data (from a *unknown* target distribution) and some artificially generated noise (from the proposal distribution). In fact, our algorithm can be treated as a special noise matching algorithm on a graphical model with only a single unknown parameter Z. This connection is surprising in part because a naïve likelihood-based or Bayesian inference procedure treating Z as an unknown parameter fails to work, as discussed in Wasserman (Example 11.10, page 188, 2011) and a thread of related internet discussion (e.g., Wasserman, 2012, and links therein). An intriguing open question is to understand if there exists a principled procedure that turns any *partition function free* learning algorithm, such as Hinton (2002), Lyu (2011), Asuncion et al. (2010), Sohl-Dickstein et al. (2011), into a corresponding *partition function inference* method.

[1] This connection was found by the authors after acceptance.

Related to importance sampling, its use in graphical models almost always require certain variance reduction techniques. Variants of annealed importance sampling based approaches (e.g., Salakhutdinov and Murray, 2008, Theis et al., 2011, Burda et al., 2015, Ma et al., 2013) have been proposed, and are widely used for likelihood evaluation of deep generative models. Other examples of variance reduction methods include adaptive improvement of the proposal (e.g., Cheng and Druzdzel, 2000), and combining with search based methods (e.g., Gogate, 2009, Gogate and Dechter, 2012, 2011) or variational methods (e.g., Wexler and Geiger, 2007). Note that our approach is orthogonal to these developments, and can be combined with them to achieve even better results. In fact, many of our experiments are set up to demonstrate the advantages of combining our method with variational and annealing techniques.

There are also other algorithms that leverage samples from the target distribution. For example, Chib (1995), Chib and Jeliazkov (2001) calculate the marginal likelihood from the output of Gibbs sampling or Metropolis-Hastings. Also related are other generalizations of importance sampling, such as bridge sampling and path sampling (Meng and Wong, 1996, Gelman and Meng, 1998).

3 BACKGROUND

Assume we have a distribution $p(x) = f(x)/Z$, where $Z = \int f(x)d\mu(x)$ is the partition function that we are interested in calculating; here the base measure $\mu(x)$ can be the counting measure for discrete variables, or Lebesgue for continuous variables. We consider Monte Carlo methods for estimating Z. Two basic methods are the following:

Importance Sampling (IS) Assume we have a tractable distribution $p_0(x)$ which has been properly normalized, that is, $\int p_0(x)d\mu(x) = 1$. We draw samples $\{x_0^1, \ldots, x_0^n\}$ from $p_0(x)$, and estimate Z by

$$\hat{Z}_{\text{is}} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_0^i)}{p_0(x_0^i)}.$$

This is an unbiased estimator of Z, that is, $\mathbb{E}(\hat{Z}_{\text{is}}) = Z$, and its mean squared error is known to be

$$n\mathbb{E}\left[\frac{(\hat{Z}_{\text{is}} - Z)^2}{Z^2}\right] = \chi^2(p||p_0) = \int \frac{p^2}{p_0}d\mu(x) - 1, \quad (1)$$

where $\chi^2(\cdot||\cdot)$ represents the chi-square divergence.

Unfortunately, $\chi^2(p||p_0)$ is often impractically large, or even infinite, especially when $p_0(x)$ is more peaked than $p(x)$. Additionally, despite the theoretical unbiasedness of $\hat{Z}$, it often underestimates Z. This is due to the distribution of the weights $f(x)/p_0(x)$ being heavy-tailed with a resulting propensity for outliers. Consequently, the results

using IS may be more properly viewed as a probabilistic lower bound rather than an unbiased estimate (e.g., Burda et al., 2015).

Reverse Importance Sampling (RIS) The *weighted harmonic mean method* (Gelfand and Dey, 1994), which we refer to as a reverse importance sampling method, works in an opposite way to importance sampling. It draws samples $\{x_1^1, \ldots, x_1^n\}$ from the *target distribution* $p(x)$ (e.g., via MCMC when exact sampling is difficult for $p(x)$), and estimates Z by

$$\hat{Z}_{\text{ris}} = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{p_0(x_1^i)}{f(x_1^i)}\right]^{-1}.$$

Note that $1/\hat{Z}_{\text{ris}}$ can be viewed as a regular importance sampling estimate for $1/Z$, justifying $\hat{Z}_{\text{ris}}$ as a reasonable estimate of Z. Under regularity conditions (Gelfand and Dey, 1994), the asymptotic MSE of $\hat{Z}_{\text{ris}}$ (assuming it exists) is

$$n\mathbb{E}\left[\frac{(\hat{Z}_{\text{ris}} - Z)^2}{Z^2}\right] = \chi^2(p_0||p) = \int \frac{p_0^2}{p} d\mu(x) - 1.$$

The χ^2-divergence here has the opposite order as that in (1) for IS, and tends to be large or infinite when $p(x)$ is more peaked than $p_0(x)$. In addition, $\hat{Z}_{\text{ris}}$ often gives upper bounds on Z (in contrast to lower bounds by IS), which can be easily seen by viewing $1/\hat{Z}$ is a regular IS estimate for $1/Z$. The special case when $p_0(x)$ is a uniform distribution is called the *harmonic mean method* (Newton and Raftery, 1994), and sometimes the *Ogata-Tanemura method* (Ogata and Tanemura, 1985).

The fact that IS and RIS give under- and over-estimates respectively suggests the use of their average $\log \hat{Z}_{\text{avg}} = (\log \hat{Z}_{\text{is}} + \log \hat{Z}_{\text{ris}})/2$ with asymptotic MSE of

$$n\mathbb{E}\left[\frac{(\hat{Z}_{\text{avg}} - Z)^2}{Z^2}\right] = \frac{1}{4}(\chi^2(p||p_0) + \chi^2(p_0||p)). \quad (2)$$

Unfortunately, this is large whenever one of $\chi^2(p||p_0)$ or $\chi^2(p_0||p)$ is large and thus imposes more stringent constraints on p_o such that (2) is finite, *i.e.* it can neither be too peaked nor too flat compared to the target distribution. This can be significant even for simple distributions as in the following example.

Example 1. *Consider normal distributions* $p(x) = \mathcal{N}(x; 0, \sigma^2)$ *and* $p_0(x) = \mathcal{N}(x; 0, \sigma_0^2)$. *One can show that* $\text{var}(\hat{Z}_{\text{is}}) = +\infty$ *if* $\sigma_0 \le \sigma/\sqrt{2}$ *(p_0 is much more peaked than p). Conversely,* $\text{var}(\hat{Z}_{\text{ris}}) = +\infty$ *if* $\sigma \le \sigma_0/\sqrt{2}$ *(p is much more peaked than p_0). Therefore, their average Z_{avg} has finite variance only when* $\sigma/\sqrt{2} \le \sigma_0 \le \sqrt{2}\sigma$.

More advanced combinations of IS and RIS can be obtained by estimating their variances, and taking weighted averages, or selecting the better one according to their variance. Unfortunately, the variance estimates themselves are unreliable, often over- or under-estimated, making these methods ineffective. We explore and compare several of these options in our experiments; see Section 6 for details.

4 DISCRIMINANCE SAMPLING

Here we propose a new estimator of Z, termed *discriminance sampling* (evoking *importance* and *disciriminative*), based on both $\{x_1^i\} \sim p(x)$ and $\{x_0^i\} \sim p_0(x)$ jointly. The idea is to reframe estimation of Z as a classification problem between $\{x_1^i\}$ and $\{x_0^i\}$. To start, we assign a binary label $y_1^i = 1$ for each $x_1^i \sim p(x)$, and correspondingly $y_0^i = 0$ for each $x_0^i \sim p_0(x)$. Putting these samples together we get $\{x^i\} = \{x_1^i\} \cup \{x_0^i\}$ and $\{y^i\} = \{y_1^i\} \cup \{y_0^i\}$. In this way, the conditional distribution of y^i given x^i is

$$p(y^i = 0 \mid x^i) = \frac{p_0(x^i)}{f(x^i)/Z + p_0(x^i)},$$

where Z can be treated as an unknown parameter. This motivates a parameter estimation procedure where we consider a family of conditional probabilities $p(y = 1|x; c) = \frac{p_0(x)}{f(x)/c + p_0(x)}$, indexed by a parameter c, and estimate c by maximizing the conditional likelihood:

$$\hat{Z}_{\text{dis}} = \arg\max_{c:\, c \ge 0} \sum_{i=1}^{2n} \log \frac{y^i f(x^i)/c + (1 - y^i)p_0(x^i)}{f(x^i)/c + p_0(x^i)}.$$

Calculating the zero-gradient equation of the objective function, we see that the optimal c should satisfy the following ratio matching condition:

$$\frac{1}{2n}\sum_{i=1}^{2n}\frac{p_0(x^i)}{f(x^i)/c + p_0(x^i)} = \frac{1}{2}, \quad (3)$$

that is, the proportion of $y = 0$ (and $y = 1$) predicted by the model should equal $1/2$, matching the label proportions in the data. Because the LHS of (3) is an increasing function on c, Eq. (3) yields a unique solution unless $p(x^i)p_0(x^i) = 0$ for all i, that is, when $p(x)$ and $p_0(x)$ do not overlap. In practice, we can solve (3) efficiently using root finding algorithms such as the `fzero` function in MATLAB.

Proposition 1. *Assume* $\{x_1^i\}_{i=1}^n$ *and* $\{x_0^i\}_{i=1}^n$ *are i.i.d. samples from* $p(x)$ *and* $p_0(x)$, *respectively. Let* $e_1 = \sqrt{2n}(\hat{Z}_{\text{dis}}/Z - 1)$ *and* $e_2 = \sqrt{2n}(\log \hat{Z}_{\text{dis}} - \log Z)$. *Define*

$$\gamma = \mathbb{E}[\text{var}(y|x)] = \frac{1}{2}\int \frac{p(x)p_0(x)}{p(x) + p_0(x)} d\mu(x), \quad (4)$$

then if $\gamma \ne 0$, *we have* $\hat{Z}_{\text{dis}} \xrightarrow{a.s.} Z$ *as* $n \to \infty$, *and* e_1 *and* e_2 *have a normal distribution* $\mathcal{N}(0,\ (\frac{1}{4} - \gamma)/\gamma^2)$.

Proof. Apply the standard asymptotic result in DasGupta (Theorem 17.2, Page 264, 2008); the condition $\gamma \neq 0$ guarantees (3) has an unique solution as $n \to \infty$. Note that e_1 and e_2 are asymptotically equivalent because $\log \epsilon \approx \epsilon - 1$ for $\epsilon \approx 1$. □

Remarks. (i) Eq (4) above relates the accuracy of $\hat{Z}_{\text{dis}}$ with the variance of the label y given x, which is a measure of distinguishability between the two samples $\{x_1^i\}$ and $\{x_0^i\}$. Ideally, we want to choose p_0 so that it is hard to distinguish between $\{x_1^i\}$ and $\{x_0^i\}$; when $p_0 = p$, $\hat{Z}_{\text{dis}}$ equals Z exactly.

(ii) The variance of $\hat{Z}_{\text{dis}}$ is infinite only if $\gamma = 0$, that is, $\int \frac{pp_0}{p+p_0} = 0$; this is possible only if $p(x)$ and $p_0(x)$ do not overlap, that is, $p(x)p_0(x) = 0$ almost everywhere. Note that this is a much milder condition compared to that for IS, RIS and their average, since in practice it is usually easy to choose a $p_0(x)$ that shares some support with $p(x)$.

Example 2. *To continue with Example 1, the MSE of $\hat{Z}_{\text{dis}}$ is finite for any $\sigma_0 > 0$ and $\sigma > 0$, making it far more robust than IS or reverse IS, which have finite MSE only when $\sigma_0 > \sigma/\sqrt{2}$ and $\sigma_0 < \sqrt{2}\sigma$, respectively.*

5 ANNEALED DISCRIMINANCE SAMPLING

One advantage of our method is that it can be naturally extended to cases when we have more than two distributions, in which case it is straightforward to frame a corresponding multinomial classification problem. In this section, we consider an improvement to our method by introducing a set of auxiliary distributions that serve as intermediate points between the target and reference distributions. This extension is analogous to annealed importance sampling (AIS) (Neal, 2001, Salakhutdinov and Murray, 2008) and reverse AIS (Burda et al., 2015), but with significantly better performance.

Both AIS and reverse AIS are based on a set of distributions $\{p_k = f_k(x)/Z_k \colon k = 0, \dots, m\}$ where p_0 is the normalized reference distribution ($Z_0 = 1$) and $p_m(x) = f(x)/Z$ is the target distribution; the other distributions serves as "intermediate points" between p_0 and p. A typical choice of the distributions is $f_k(x) = f(x)^{k/m} f_0(x)^{1-k/m}$, where k can be interpreted as a temperature parameter that anneals between p and p_0.

Now assume we draw m sets of samples $\{x_k^i\}_{i=1}^n \sim p_k(x)$, $k = 0, \dots, m$. Similar to Section 4, we assign each x_k^i with a label $y_k^i = k$, resulting a conditional likelihood of

$$p(y^i = k | x^i) = \frac{f_k(x^i)/Z_k}{\sum_{k=0}^m f_k(x^i)/Z_k}.$$

We then treat $\{Z_k \colon k = 1, \dots m\}$ as a set of unknown parameters, and estimate them by performing maximum conditional likelihood:

$$\{\hat{Z}_k\}_{k=0}^m = \underset{c>0 \colon c_0 = Z_0}{\arg\max} \sum_{k=0}^m \sum_{i=1}^n \log \frac{f_k(x_k^i)/c_k}{\sum_{k=0}^m f_k(x_k^i)/c_k}, \quad (5)$$

where c_0 is fixed to its known value ($Z_0 = 1$). Similar to the binary case, it is easy to show that $\{\hat{Z}_k\}$ forms a consistent estimation of $\{Z_k\}$ (although we are only interested in $\hat{Z}_m$).

Note that (5) is a convex optimization w.r.t. $\{\log c_k\}$, and can be solved efficiently. A further simplification of (5) is to construct and combine a sequence of binary classifications between the (p_k, p_{k+1}) pairs, instead of the joint multinomial classification. To be specific, we sequentially estimate the ratio $r_{k+1} = Z_{k+1}/Z_k$ between p_k and p_{k+1} by discriminating between $\{x_k^i\}$ and $\{x_{k+1}^i\}$:

$$\hat{r}_{k+1} = \underset{\substack{c_{k+1}>0 \\ c_k = 1}}{\arg\max} \sum_{k'=k}^{k+1} \sum_{i=1}^n \log \frac{f_{k'}(x_{k'}^i)/c_{k'}}{\sum_{\ell=k}^{k+1} f_\ell(x_{k'}^i)/c_\ell} \quad (6)$$

and estimate $Z = Z_m$ by chaining the ratios together:

$$\log \hat{Z} = \sum_{k=1}^m \log \hat{r}_k \approx \sum_{k=1}^m \log r_k = \log Z. \quad (7)$$

Interestingly, we find that such sequential binary classification works as well as the joint multinomial classification in our experiments, possibly because the neighboring $\{p_k, p_{k+1}\}$ are close to each other and provide more accurate estimates of their ratios.

Practical Implementation In practice, it is expensive to sample from all $p_k(x) = f_k(x)/Z_k$ at each temperature independently, especially when the number of temperatures is large. Instead, we use AIS (Neal, 2001) to sequentially generate *importance weighted samples* for each $p_k(x)$: we start with $x_0^i \sim p_0(x)$ and set $w_0^i = 1$, and sequentially update the samples using Markov chain transitions and adjust the weights accordingly:

$$x_k^i \sim T_k(\ \cdot\ | x_{k-1}^i), \quad w_k^i = w_{k-1}^i \frac{f_k(x_{k-1}^i)}{f_{k-1}(x_{k-1}^i)}, \quad (8)$$

Algorithm 1 Annealed Discriminace Sampling (Sequential Binary Version)

Draw $x_0^i \sim p_0(x)$. Set $w_0^i = 1$ and $Z_0 = 1$.
for $k = 1$ to m **do**
  Generate weighted sample $\{x_k^i, w_k^i\}_{i=1}^n$ by the AIS update in (8).
  update $\hat{Z}_k = Z_{k-1}\hat{r}_k$, where $\hat{r}_k$ maximizes the weighted conditional likelihood (9).
end for
Return: $\hat{Z}_m$ is an estimate of the partition function Z.

Algorithm 2 Annealed Discriminace Sampling (Multinomial Version)

1. Use AIS to generate $\{x_k^i, w_k^i\}_{i=1}^n$ for $\forall\ 0 \le k \le m$.
2. Estimate $\hat{Z}_k$ by maximizing

$$\{\hat{Z}_k\}_{k=0}^m = \underset{c>0:\ c_0=Z_0}{\arg\max} \sum_{k=0}^{m}\sum_{i=1}^{n} \tilde{w}_k^i \log \frac{f_k(x_k^i)/c_k}{\sum_{k=0}^m f_k(x_k^i)/c_k},$$

where $\tilde{w}_k^i = w_k^i / \sum_i w_k^i$ are the normalized weights.

Return: $\hat{Z}_m$ is an estimate of the partition function Z.

for $\forall\ 1 \le k \le m$, where $T_k(\cdot|\cdot)$ is a Gibbs or Metropolis-Hastings transition kernel of p_k. By the augmented variable space argument of Neal (2001), we can show the weighted sample $(x_k^i, w_k^i)_{i=1}^n$ follows $p_k(x)$ in the sense that

$$\mathbb{E}(w_k^i h(x_k^i)) = const \cdot \mathbb{E}_{x\sim q_k}(h(x))$$

for any $0 \le k \le m$ and integrable function $h(x)$; this allows us to estimate the partition functions Z_k using weighted versions of the multinomial (5) or sequential binary (6) classifications based on the weighted samples (x_k^i, w_k^i). For example, we can estimate the ratio $r_{k+1} = Z_{k+1}/Z_k$ between p_{k+1} and p_k by maximizing a weighted version of the conditional likelihood in (6),

$$\hat{r}_{k+1} = \underset{\substack{c_{k+1}>0\\ c_k=1}}{\arg\max} \sum_{k'=k}^{k+1}\sum_{i=1}^{n} \tilde{w}_{k'}^i \log \frac{f_{k'}(x_{k'}^i)/c_{k'}}{\sum_{\ell=k}^{k+1} f_\ell(x_{k'}^i)/c_\ell}, \quad (9)$$

where $\tilde{w}_k^i = w_k^i / \sum_{i=1}^n w_k^i$ are the normalized weights under each temperature k. See Algorithm 1 for the full algorithm of the sequential binary version of our method; the corresponding multinomial version is shown in Algorithm 2. Note that both Algorithm 1 and 2 can recycle the samples and weights generated by AIS and can be implemented conveniently based on AIS. Alternatively, we can also base our estimator on other methods that draw samples jointly from different temperatures, such as simulated tempering and parallel tempering (see Liu, 2008, and references therein).

6 EXPERIMENTS

We present experimental results on a toy Gaussian example, pairwise Markov random fields (10×10 grids), and deep generative models trained on real world data. Our contributions are threefold: (1) We demonstrate the advantage of our method compared to IS, reverse IS and their combinations, and show that our method yields significantly smaller bias and variance across all our experiments. (2) We illustrate the benefits of combining deterministic variational methods and the Monte Carlo based methods discussed in this paper; we show that Monte Carlo methods can provide tighter (but "probabilistic") bounds than deterministic variational methods, and can be further improved by using variational methods to provide better reference distributions p_0. (3) We test the annealed version of our algorithm in real-world deep learning models, including restricted Boltzmann machines (RBM) and deep Boltzmann machines (DBM), and show that it significantly outperforms the state-of-the-art AIS and reverse AIS methods (Burda et al., 2015).

Setting We compare our algorithm with IS, RIS and three different methods that combine IS and RIS in hopes of offsetting their relative biases:

(1) *Naïve Averaging*:

$$\log \hat{Z}_{\text{avg}} = (\log \hat{Z}_{\text{is}} + \log \hat{Z}_{\text{ris}})/2.$$

Here the average is taken on the log domain; note that averaging in the Z domain, i.e., $(\hat{Z}_{\text{is}} + \hat{Z}_{\text{ris}})/2$ does not make sense since we almost always have $\hat{Z}_{\text{is}} \ll \hat{Z}_{\text{ris}}$, and the result will be dominated by $\hat{Z}_{\text{ris}}$.

(2) *Weighted Averaging*:

$$\log \hat{Z}_{\text{w}} = (\hat{v}_{is}^{-1} \log \hat{Z}_{\text{is}} + \hat{v}_{ris}^{-1} \log \hat{Z}_{\text{ris}})/(\hat{v}_{is}^{-1} + \hat{v}_{ris}^{-1}),$$

where $\hat{v}_{is}$, $\hat{v}_{ris}$ are empirical estimates of $\text{var}(\log \hat{Z}_{\text{is}})$ and $\text{var}(\log \hat{Z}_{\text{ris}})$,

$$\hat{v}_{is} = \widehat{\text{var}}(\{\frac{f(x_0^i)}{p_0(x_0^i)}\})/\hat{Z}_{\text{is}}^2,$$

$$\hat{v}_{ris} = \widehat{\text{var}}(\{\frac{p_0(x_1^i)}{f(x_1^i)}\})\hat{Z}_{\text{ris}}^2,$$

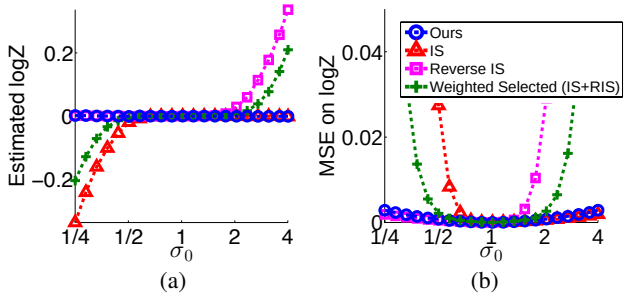
where $\widehat{\text{var}}(\cdot)$ represents the empirical variance estimate. Note that the weights defined above should minimize the variance of the combination if the variance estimates are accurate. Unfortunately, variance estimates for the weighted averaging approach are typically unreliable and have the same under- / over-estimation problem as IS and RIS, causing the weighted average to perform poorly.

(3). *Weighted Selection*:

$$\log \hat{Z}_{\text{s}} = [\hat{v}_{is} < \hat{v}_{ris}] \log \hat{Z}_{\text{is}} + [\hat{v}_{ris} < \hat{v}_{is}] \log \hat{Z}_{\text{ris}},$$

where we select the estimator with smaller estimated variance; here $[\cdot]$ is the indicator function.

Note that IS (resp. RIS) uses only $\{x_0^i\}_{i=1}^n \sim p_0(x)$ (resp. $\{x_1^i\}_{i=1}^n \sim p(x)$), while our method uses both $\{x_0^i\}$ and $\{x^i\}$. To make a conservative comparison, we use *only the first half of the samples* $\{x_1^i\}_{i=1}^{n/2}$ and $\{x_0^i\}_{i=1}^{n/2}$ in our method. However, we do not halve the samples when calculating the averages $\hat{Z}_{\text{avg}}$, $\hat{Z}_{\text{w}}$ and $\hat{Z}_{\text{s}}$, allowing them to use *exactly twice the information* as our method. Despite these unfavorable conditions, our method still consistently outperforms IS, RIS and all the averaging based methods.



(a) (b)

Figure 1: Gaussian toy example. The estimated values (a) and mean square errors (b) on $\log Z$ by different methods (the true value is $\log Z = 0$). IS performs poorly when σ_0 is small (p_0 is too peaked), while reverse IS is poor when σ_0 is large (p is too peaked). Our method performs much better and is robust for all values of σ_0. The result is averaged over 1000 random trials.

Gaussian Toy Example As in Example 1, we consider $p(x) = \mathcal{N}(x; 0, \sigma^2)$ with fixed $\sigma = 1$ and $p_0(x) = \mathcal{N}(x; 0, \sigma_0^2)$ with different values of σ_0. We are interested in calculating the normalization constant of $p(x)$, which is trivially $Z = 1$ in this case. We use $n = 1000$ samples from both the target $p(x)$ and reference $p_0(x)$.

Figure 1 reports the bias and MSE on $\log Z$ as returned by our methods, IS, reverse IS, and the weighted selection $\log \hat{Z}_{\text{s}}$ (the naïve average $\log \hat{Z}_{\text{avg}}$ and weighted average $\log \hat{Z}_{\text{w}}$ are worse than $\log \hat{Z}_{\text{s}}$ and not shown in the figure for clarity). We find that our method significantly outperforms all the other algorithms, despite using fewer samples than the averaging based methods.

The performance of IS and reverse IS in Figure 1(a) is consistent with the theoretical analysis: IS tends to give a lower bound, and degenerates quickly when σ_0 is small (p_0 is more peaked than p), while reverse IS gives an upper bound, and degenerates when σ_0 is large (p is more peaked than p_0). In this case, it is interesting to see that the performances of IS and reverse IS are extremely imbalanced and anti-correlated (whenever IS performs well, RIS performs poorly, and vice versa), which explains why weighted selection is better than naïve averaging in this case.

MRF on 10×10 Grid We consider Markov random fields (MRFs) on a 10×10 grid

$$p(x) = \frac{1}{Z} \exp\Big(\sum_{ij} \theta_{ij}(x_i, x_j) + \sum_i \theta_i(x_i)\Big),$$

where $x_i \in \{0, 1\}$. We generate each $\theta_i(k)$ randomly by $\mathcal{N}(0, \sigma_s^2)$, with fixed $\sigma_s = 0.1$ and each $\theta_{ij}(k, l)$ from $\mathcal{N}(0, \sigma_p^2)$, where σ_p characterizes the interaction strength in the MRF. We also explore different choices of reference distribution p_0 for the MRF, including

(1) *Uniform distribution* as shown in Figure 2(a).

(2) *Mean field* approximation as shown in Figure 2(b).

(3) *Mixture of trees* constructed from the reparameterization obtained from tree reweighted belief propagation (TRBP) (Wainwright et al., 2005) (Figure 2(c)). The edge appearance probabilities in TRBP are set by assigning uniform weights to a random set of spanning trees.

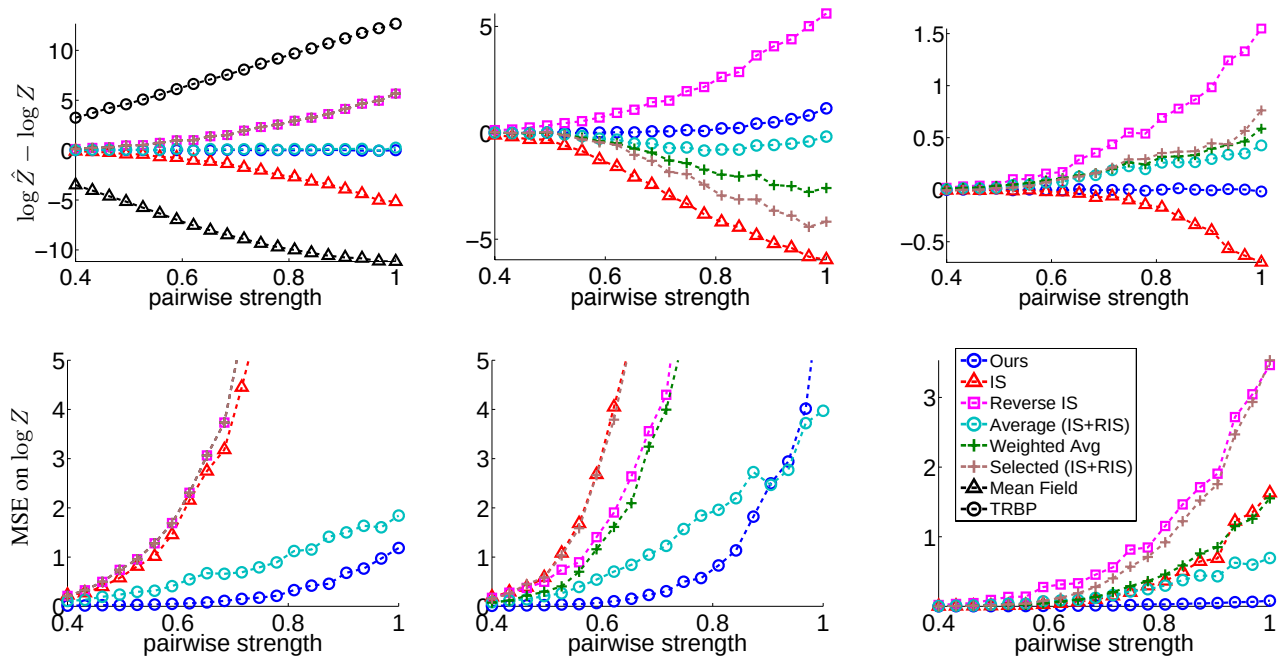
The samples from p_0 are drawn exactly, while those from p are drawn using Gibbs sampling with 500 burn-in steps. We use $n = 1000$ samples from each distribution in all cases, and average the results over 500 random trials.

From Figure 2, we find our method significantly outperforms IS and reverse IS, and all the versions of their combinations under all three choices of p_0. The deterministic bounds returned by TRBP and MF are shown in Figure 2(a), and are significantly looser than the sampling based bounds in these cases (which, however, provides probabilistic, instead of deterministic bound guarantees as the variational methods). We note that the proposal produced by MF is only as good as the uniform proposal. On the other hand, the p_0 produced by a mixture of TRBP trees gives significantly better results (note that the y-axes are not on the same scale). This result demonstrates the potential of combining variational methods and sampling methods, with carefully designed choices for p_0 and estimation methods (such as our method).

Interestingly, we find the performance of IS and reverse IS are relatively balanced in the MRF examples, making the naïve average of IS and reverse IS outperform both the weighted average and weighted selection. This is in contrast to the Gaussian toy example, where IS and reverse IS are extremely imbalanced. Unfortunately, there is no general method to tell whether IS and reverse IS will be balanced or not beforehand.

Deep Generative Models We compare our annealed discriminance sampling (ADS) with the AIS and reverse AIS estimator (RAISE) as introduced in Burda et al. (2015) for partition function estimation in deep generative models, including a restricted Boltzmann machine (RBM) and a deep Boltzmann machine (DBM). We implement AIS and RAISE following Algorithm 1 and Algorithm 3[2] in Burda et al. (2015), respectively. We then take the samples and weights generated by AIS and run our sequential binary ADS in Algorithm 1 and multinomial ADS in Algorithm 2. In principle, we can also reuse the same samples generated by AIS to construct a version of a reverse AIS estimator. Unfortunately, we find this works poorly in practice, and it seems to be important to follow Algorithm 3 in Burda et al. (2015) to generate new samples specifically for the

[2] Algorithm 3 of Burda et al. (2015) was designed for calculating the testing likelihood; we adopt it for calculating the partition function by replacing its conditional kernel $\tilde{T}_k^{(\mathbf{v}_{\text{test}})}(\cdot|h'_{k-1})$ in the forward step with the unconditional kernel $\tilde{T}_k(\cdot|x'_{k-1})$.



(a) Uniform proposal (b) Mean field proposal (c) Mixture of TRBP trees

Figure 2: MRFs on a 10×10 grid. The three columns represent the results when using different reference distributions p_0, including the uniform distribution (a), mean field approximation (b) and a mixture of trees constructed from TRBP (c). Our algorithm consistently performs best. Note that the comparison is again in favor of the averaging methods since they use twice as many samples as our method.

reverse AIS estimates. Note that implemented in this way, the average of AIS and RAISE uses twice the number of samples as our method. In addition, we emphasize that the RAISE as proposed in Algorithm 3 in Burda et al. (2015) includes both a forward and backward sampling step, requiring twice the computational cost of AIS. In contrast, our method has roughly the same time complexity as AIS, because the cost of the discriminance analysis step in our method, especially the sequential binary version, is negligible compared to the sample generation steps of AIS as used in Algorithm 1 in Burda et al. (2015).

In both our experiments for RBM and DBM and for all the annealing-based algorithms, we use $2^1 \sim 2^{10}$ linearly spaced intermediate temperatures (or distributions) and $n = 1000$ samples (corresponding to 1000 separate MCMC chains in AIS). The reference distribution p_0 is taken to be the *data base rate* (DBR) distribution as suggested by Salakhutdinov and Hinton (2009), which is constructed based the marginal statistics of the image dataset. In all cases, we repeat the estimates 10 times and report the average bias and MSE results.

To obtain the true partition function of both RBM and DBM, we calculate the average of AIS and RAISE with an extremely large number (in our case, 100,000) of temperatures, until their estimates coincide to within 0.1 nats, i.e., $|\log \hat{Z}_{\text{is}} - \log \hat{Z}_{\text{ris}}| \leq 0.1$. This gives a high confidence estimate of the true partition function, since AIS and RAISE are probabilistic lower and upper bounds, respectively.

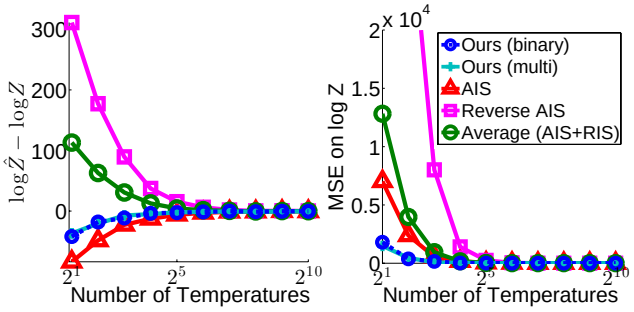


Figure 3: Estimates of log-partition function of a restricted Boltzmann machine trained on MNIST with different numbers of intermediate temperatures. While all methods converge to the same value, our method significantly outperforms other methods when the intermediate temperatures are few.

We first consider a restricted Boltzmann machine (RBM) with 500 hidden nodes trained on MNIST using contrastive divergence with 25 steps. Figure 3 shows the results of different algorithms. When there are many intermediate

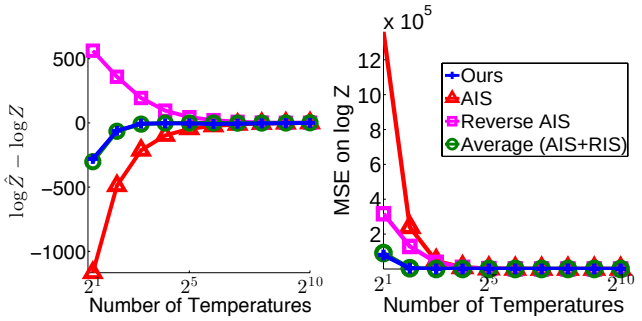


Figure 4: Estimates of log-partition function of a 784-500-1000 deep Boltzmann machine trained on MNIST with different numbers of intermediate temperatures. While all methods converge to the same value, our method outperforms both AIS and RAISE when the intermediate temperatures are few. Our multinomial ADS performs the same as our binary version, and is omitted in the figure for clarity.

temperatures, all algorithms give accurate estimates. When there are fewer intermediate temperatures, our ADS is able to compute significantly more accurate estimates than AIS and RAISE, or even their average. In addition, we find that the binary and multinomial versions of our ADS algorithm work similarly (almost identically) in all our experiments.

We then experiment on a more complex deep Boltzmann machine, trained with a 784-500-1000 structure on MNIST closely following the procedure in Salakhutdinov and Hinton (2009): we initially train the first layer RBM for 100 epochs, then the second layer with 200 epochs and then fine-tune the two layers jointly for 300 epochs. The results of the different algorithms are shown in Figure 4. Similarly to the results in the RBM experiment, ADS significantly outperforms both AIS and RAISE when the number of temperatures is small. In this case, we find that the average of AIS and RAISE is quite accurate for this DBM model, almost as good as ADS. However, our algorithm still gives significant advantages in practice: again, we use only half the number of samples that are used by the average of AIS and RAISE and have only $1/3$ of the total computational cost (because RAISE is twice as expensive as AIS or our ADS method, as discussed previously).

7 CONCLUSION

In this paper, we introduced *discriminance* sampling, a novel and efficient method for estimating log-partition functions of probabilistic distributions. Using samples drawn from both the target and proposal distributions, we formulated the estimation problem into a discriminant analysis problem that classifies samples into their corresponding distributions. Our approach does not under- / over-estimate the true values like IS and reverse IS, and places much less stringent requirements on the proposal distributions. In addition, we also extend our method to define annealed discriminance sampling (ADS) and demonstrate that ADS significantly outperform AIS, reverse AIS, and performs as well or better than their average, which are currently state-of-the-art methods for model evaluation in deep generative models.

Acknowledgement This work is supported in part by VITALITE, which receives support from Army Research Office (ARO) Multidisciplinary Research Initiative (MURI) program (Award number W911NF-11-1-0391); NSF grants IIS-1065618 and IIS-1254071; and by the United States Air Force under Contract No. FA8750-14-C-0011 under the DARPA PPAML program.

References

Asuncion, A., Liu, Q., Ihler, A., and Smyth, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *AISTATS*.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Accurate and conservative estimates of MRF log-likelihood using reverse annealing. In *AISTATS*.

Cheng, J. and Druzdzel, M. (2000). AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research*.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Geyer, C. (1991). Reweighting monte carlo mixtures. Technical Report 568, School of Statistics, University of Minnesota.

Geyer, C. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, University of Minnesota.

Gogate, V. (2009). *Sampling Algorithms for Probabilistic Graphical Models with Determinism*. PhD thesis, University of California, Irvine.

Gogate, V. and Dechter, R. (2011). SampleSearch: Importance sampling in presence of determinism. *Artificial Intelligence*, 175(2):694–729.

Gogate, V. and Dechter, R. (2012). Importance sampling-based estimation over AND/OR search spaces for graphical models. *Artificial Intelligence*, 184–185(0):38 – 77.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *ICML*.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press.

Liu, J. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Lyu, S. (2011). Unifying non-maximum likelihood learning objectives with minimum KL contraction. In *NIPS*.

Ma, J., Peng, J., Wang, S., and Xu, J. (2013). Estimating the partition function of graphical models using Langevin importance sampling. In *AISTATS*.

Meng, X. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, pages 421–433.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann machines. In *AISTATS*.

Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In *ICML*.

Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. (2011). Minimum probability flow learning. In *ICML*.

Theis, L., Gerwinn, S., Sinz, F., and Bethge, M. (2011). In all likelihood, deep belief is not enough. *The Journal of Machine Learning Research*, 12:3071–3096.

Wainwright, M., Jaakkola, T., and Willsky, A. (2005). A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335.

Wasserman, L. (2011). *All of statistics*. Springer Science & Business Media.

Wasserman, L. (2012). The normalizing constant paradox. https://normaldeviate.wordpress.com/2012/10/05/the-normalizing-constant-paradox/.

Wexler, Y. and Geiger, D. (2007). Importance sampling via variational optimization. In *UAI*.