Algorithms for Estimating the Partition Function of Restricted Boltzmann Machines (Extended Abstract)[*]

Oswin Krause[1†] , **Asja Fischer**[2] and **Christian Igel**[1]
[1]Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark
[2]Faculty of Mathematics, Ruhr University Bochum, 44780 Bochum, Germany
oswin.krause@di.ku.dk, asja.fischer@rub.de, igel@di.ku.dk

Abstract

Estimating the normalization constants (partition functions) of energy-based probabilistic models (Markov random fields) with a high accuracy is required for measuring performance, monitoring the training progress of adaptive models, and conducting likelihood ratio tests. We devised a unifying theoretical framework for algorithms for estimating the partition function, including Annealed Importance Sampling (AIS) and Bennett's Acceptance Ratio method (BAR). The unification reveals conceptual similarities of and differences between different approaches and suggests new algorithms. The framework is based on a generalized form of Crooks' equality, which links the expectation over a distribution of samples generated by a transition operator to the expectation over the distribution induced by the reversed operator. Different ways of sampling, such as parallel tempering and path sampling, are covered by the framework. We performed experiments in which we estimated the partition function of restricted Boltzmann machines (RBMs) and Ising models. We found that BAR using parallel tempering worked well with a small number of bridging distributions, while path sampling based AIS performed best with many bridging distributions. The normalization constant is measured w.r.t. a reference distribution, and the choice of this distribution turned out to be very important in our experiments. Overall, BAR gave the best empirical results, outperforming AIS.

1 Introduction

Markov random fields (MRFs, [Kindermann and Snell, 1980]) are undirected probabilistic graphical models that find many applications in AI, for example in computer vision [Blake *et al.*, 2011] and neural computation [Smolensky, 1986]. The distribution modelled by an MRF can be written as

$$p(\boldsymbol{x}) = \frac{1}{Z} e^{-\mathcal{E}(\boldsymbol{x})} \ ,$$

where $\mathcal{E}(\boldsymbol{x})$ is called the energy function and the normalization constant Z is referred to as the partition function. Computing and even estimating Z of an energy-based probabilistic model is typically challenging, because analytical integration is not possible and numerical integration unfeasible. But for many tasks, we would like to know Z with a high accuracy. We may need Z to assess the performance of models, to monitor maximum likelihood learning when adapting model parameters, and to perform likelihood ratio tests. In many cases, the estimation of the normalization constant is complicated by the inability to even generate samples from the distributions in question. Therefore, we would like to be able to obtain a good estimate without the requirements of exact sampling of the distribution.

This explains the popularity of algorithms such as Annealed Importance Sampling (AIS, [Neal, 2001]) which compute the ratio of normalization constants between a reference distribution p_{ref} and a target distribution p_{target}. If p_{ref} is chosen such that its normalization constant is known, the AIS-estimate of the ratio of the normalization constants gives an estimate of the partition function of p_{target}. AIS introduces so called bridging distributions, which interpolate between p_{ref} and p_{target} as well as a sampling scheme that allows to estimate the normalization constant only requiring exact samples from p_{ref}. The performance and limitations of AIS are well known in the AI community [Salakhutdinov and Murray, 2008; Schulz *et al.*, 2010]. It is very general, still, the weak assumptions come at the expense of a possibly large variance if the number of bridging distributions is too small.

There are alternative algorithms such as variants of Bennett's Acceptance Ratio (BAR, [Bennett, 1976]), which was rediscovered several times as Bridge Sampling [Meng and Wong, 1996], Reverse Logistic Regression [Geyer, 1994], and Discriminance Sampling [Liu *et al.*, 2015], which also compute estimates of the ratio of normalization constants. However, they require independent samples from all bridging distributions and consequently there are only few studies applying BAR in the context of energy-based models in the AI community [Desjardins *et al.*, 2011; Liu *et al.*, 2015]. A possible source of samples with small bias from p_{ref}, p_{target}, and all bridging distributions is Parallel Tempering (PT, [Des-

[*]This paper is an extended abstract of an article in the Artificial Intelligence Journal (AIJ) [Krause *et al.*, 2020].

[†]Contact Author

jardins *et al.*, 2010]), which is often used for sampling from graphical models during training. Parallel Tempering introduces replica Markov chains to foster faster mixing, which leads to increasing sample quality in all chains. If the distributions of the replica chains are chosen to be the same as the bridging distributions of the BAR-estimator, it is possible to re-use the samples acquired during training for estimation the normalization constant at no additional cost (see, e.g., [Desjardins *et al.*, 2011]).

Still, the theoretical and practical properties of BAR compared to AIS have been little studied in the AI community. Our work contributes to closing this gap using a theoretical framework that can be used to derive AIS as well as different variants of BAR and other algorithms, allowing a concise proof of the fact that BAR is a maximum-likelihood estimator of the normalization constant [Shirts *et al.*, 2003]. In this extended abstract summarizing the work by [Krause *et al.*, 2020], we will put an emphasis on experimental results, showcasing the advantages of the different algorithms. We focus on Restricted Boltzmann Machines (RBMs, [Smolensky, 1986; Hinton, 2002; Fischer and Igel, 2014]) as a particular class of Markov random fields, but our considerations are not limited to these stochastic neural networks. [Krause *et al.*, 2020] also show results for the 2D-Ising model with external magnetic fields and we would like to point out that the results are also relevant for other types of generative models (e.g., see [Wu *et al.*, 2017] for an application scenario in the context of Generative Adversarial Networks).

2 Main Result

Let us briefly sketch our main theoretical result, a generalization of Crooks' equation [Crooks, 2000] to arbitrary sampling distributions. We refer to [Krause *et al.*, 2020] for details.

Let $p_{\text{ref}} = p_0, p_1, \ldots, p_N = p_{\text{target}}$ be a set of Gibbs distributions over some state space Ω with $p_i : \Omega \to \mathbb{R}$ and

$$p_i(x) = \frac{1}{Z_i} e^{-\mathcal{E}_i(x)} = \frac{1}{Z_i} p_i^*(x) \ ,$$

where $p_i^*(x)$ denotes the unnormalized probability distribution. Our goal is to estimate Z_N/Z_0.

We now consider a random variable $\boldsymbol{X} = (X_0, X_N, Y)$ taking values $\boldsymbol{x} = (x_0, x_N, y)$ in an extended state space $\Omega^\dagger = \Omega^2 \times \Theta$. Here y, taking values in the state space Θ, is a placeholder for any set of additional variables an actual estimation method may require. Assume that we can use the set of Gibbs distributions $p_0, p_1, \ldots, p_N$ to construct a pair of distributions p_F and p_R on $\Omega^\dagger$ with

$$\begin{aligned} p_F(\boldsymbol{x}) &= p_F(y, x_N|x_0) p_0(x_0) \quad \text{and} \\ p_R(\boldsymbol{x}) &= p_R(y, x_0|x_N) p_N(x_N) \ . \end{aligned}$$

We call p_F the *forward* distribution, because it creates samples from p_N given a sample x_0 from p_0, and we refer to p_R as the *reverse* distribution.

Consider now any function $\mathcal{F}$ on the extended state space $\Omega^\dagger$. We are interested in relating expectations of $\mathcal{F}$ under p_F to expectations of $\mathcal{F}$ under p_R. [Krause *et al.*, 2020] prove that

$$\langle \mathcal{F}(\boldsymbol{x}) \rangle_{p_R(\boldsymbol{x})} = \frac{Z_0}{Z_N} \left\langle \mathcal{F}(\boldsymbol{x}) e^{-\mathcal{W}(\boldsymbol{x})} \right\rangle_{p_F(\boldsymbol{x})} \ , \qquad (1)$$

where $\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} = \int p(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ and

$$\mathcal{W}(\boldsymbol{x}) = -\ln \frac{p_R(y, x_0|x_N) p_N^*(x_N)}{p_F(y, x_N|x_0) p_0^*(x_0)} \ .$$

By choosing p_F, p_R and $\mathcal{F}$, we can now derive several estimation algorithms. By setting $\mathcal{F}(\boldsymbol{x}) = 1$, we obtain

$$\frac{Z_N}{Z_0} = \left\langle e^{-\mathcal{W}(\boldsymbol{x})} \right\rangle_{p_F(\boldsymbol{x})} \ ,$$

which for particular choices of p_F becomes equivalent to AIS or Linked Importance Sampling [Neal, 2005]. Instead of choosing an arbitrary $\mathcal{F}$, BAR uses the variant which minimizes the variance

$$\mathcal{F}(\boldsymbol{x}) = \frac{1}{1 + \frac{Z_0}{Z_N} e^{-\mathcal{W}(\boldsymbol{x})}} \ .$$

Unfortunately, this expression involves the unknown normalization constant itself. However, inserting into (1) and simplifying leads to a fixed-point problem which can be solved for $\frac{Z_0}{Z_N}$. [Krause *et al.*, 2020] show that this is indeed the maximum-likelihood estimator. A practical variant can be obtained by setting $p_F(\boldsymbol{x}) = \prod_{i=0}^N p_i(x_i)$ and $p_R(\boldsymbol{x}) = p_N(x_N) \prod_{i=1}^N p_i(x_{i-1})$. In this case, we can compute the estimator using independent samples from the N bridging distributions as produced by parallel tempering.

3 Experiments

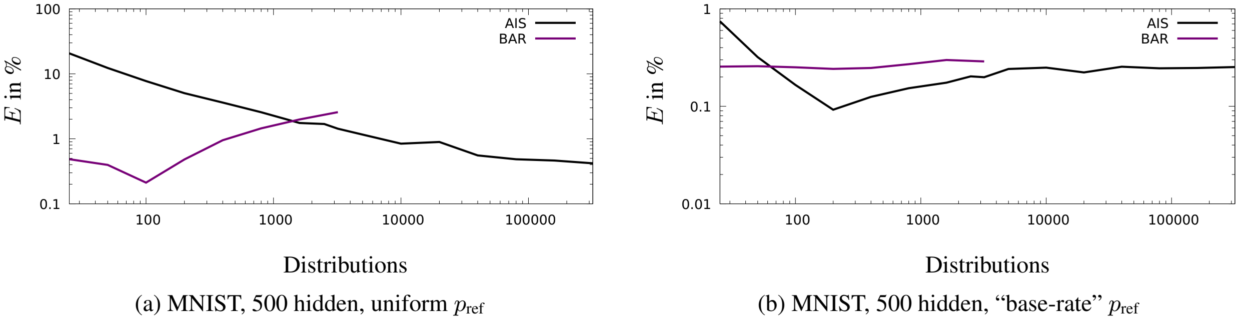
We empirically compared different algorithms for partition function estimation. Here, we will only show results for two of these algorithms, vanilla AIS and BAR (which [Krause *et al.*, 2020] refer to as BARPT-ind). We asked:

1. How accurate are the methods depending on the number of bridging distributions and the choice of the reference distribution?
2. Do the methods tend to over- or underestimate the normalization constant?
3. How do the methods perform in an online setting, where PT samples are used both for training as well as for partition function estimation?

To address these questions, we added several partition function estimation methods to the open-source machine learning library Shark [Igel *et al.*, 2008]. In this extended abstract, we present only results for MNIST data set [LeCun *et al.*, 1998]. In all experiments, we considered the task of estimating $\ln \frac{Z_N}{Z_0}$ with a fixed budget of sampling steps. We measured the mean relative error of an estimate C defined as

$$E = \left\langle \left| \frac{C}{\ln \frac{Z_N}{Z_0}} - 1 \right| \right\rangle_{p(C)} \ .$$

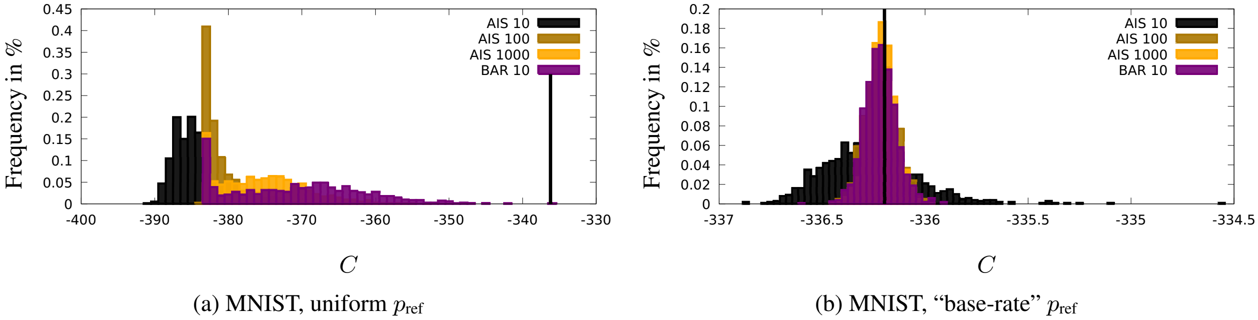
All our experiments were based on the same setup. For a given RBM we took two sets of samples, one created by PT as a way to get approximately independent samples and one using the AIS sampling-scheme. Given a reference distribution p_{ref} and the target distribution $p(x) = \frac{1}{Z_N} e^{-\mathcal{E}(x)}$ specified by



(a) MNIST, 500 hidden, uniform p_{ref}

(b) MNIST, 500 hidden, "base-rate" p_{ref}

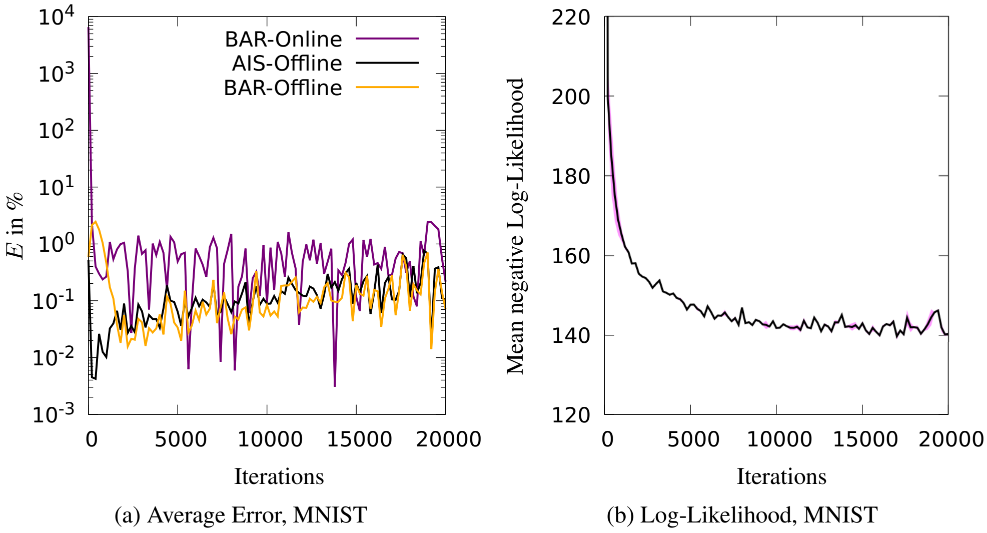
Figure 1: Mean relative error in percent for the different algorithms for different numbers of bridging distributions while keeping the total amount of samples constant, for RBMs with 500 hidden units trained on MNIST, using different reference distributions.



(a) MNIST, uniform p_{ref}

(b) MNIST, "base-rate" p_{ref}

Figure 2: Distribution of estimates of $\ln(Z_N/Z_0)$ for RBMs with 16 hidden neurons trained on MNIST. Histograms are created for 1000 estimates. Numbers after the algorithm names specify the amount of bridging distributions used. The black vertical lines indicate the ground truth. Estimates are computed using different reference distributions.



(a) Average Error, MNIST

(b) Log-Likelihood, MNIST

Figure 3: Results of the Online-Experiments on MNIST using "base-rate" p_{ref} as the reference distribution of PT. Left: Average error of the estimation of Z_N/Z_0 as in Figure 1. We compare the quality of using the same samples as the training algorithm (Online) with using separate samples (Offline) Right: True log-likelihood values during training (black) with confidence interval (purple) based on the error of BAR-Online (left).

the RBM parameters, we constructed the bridging distributions as

$$p_i(x) = \frac{1}{Z_i}(p^*_{\text{ref}}(x))^{1-\beta_i} e^{-\beta_i \mathcal{E}(x)} \ ,$$

with $\beta_i = \frac{i}{N}$, for $i = 1, 2 \dots, N - 1$. Thus, we chose a geometric mean of distributions with a uniform spacing of β-values. This is often done in practice and fits the theoretical results by [Fischer and Igel, 2015]. In our experiments, we considered two different choices of reference distribution p_{ref}, the uniform distribution, which is used often in AIS, and a distribution which uses the pixel-wise marginal probabilities as computed from the MNIST training set. The second distribution is also called “base-rate” by [Salakhutdinov and Murray, 2008].

To answer the first question raised above, we used a larger RBM with 500 hidden units trained with CD-25 by [Salakhutdinov and Murray, 2008]. We compared standard AIS with BAR using samples generated by PT. We varied the number of bridging distributions while keeping the overall budget of sampling steps fixed. For estimation, we used 1,200,000 samples in total where 50 % were used as burn-in time by PT. The results can be seen in Figure 1. In our experiments, BAR performed better with a small amount of bridging distributions, while AIS required a very large amount. A large amount of distributions can be detrimental for BAR as performing initial burn-in of PT becomes very expensive. Furthermore, all algorithms performed better when using the “base-rate” reference distribution, even outperforming our pre-computed ground truth estimate of $\ln Z_n/Z_0$ (which was computed by AIS using a uniform distribution and a very large budget).

In the next experiment, we investigated the second question. For this, we trained a smaller RBM with 16 hidden neurons where the normalization constant can be computed exactly. We took the RBM obtained at the end of the training, estimated $\ln(Z_N/Z_0)$ 1000 times for the different algorithms (i.e., BAR with 10 and AIS with 10,100, or 1000 bridging distributions), and computed histograms over these estimates. The resulting histograms are given in Figure 2. The results show that the distribution of estimates obtained by AIS when using the uniform reference distribution is skewed towards small values. All algorithms underestimated the true value, but BAR was slightly better, most likely because PT-samples had a longer time to converge to the true distribution. When using the “base-rate” reference distribution, all algorithms performed significantly better. However, AIS with 10 bridging distributions still showed a larger variance than BAR as well as a visible skew. This is in line with the general knowledge that AIS tends to underestimate the normalization constant, even though it is unbiased in expectation.

In our third experiment, we investigated the online performance. Again, we trained an RBM with 16 hidden units and computed the normalization constant exactly. To generate the samples required for training, we used PT with 50 Markov-chains and re-used the samples to compute an estimate of the normalization constant. We did not use any burn-in, which is in line with usual training procedures. We compared our results to AIS and BAR using separate samples (and using burn-in for PT). The results can be seen in Figure 3. While Online-BAR performed slightly worse than the offline-variants, it was still able to accurately track the progress of training. To visualize this, we plotted the true log-likelihood curve and added shaded regions depicting the error-ranges as obtained using the BAR-estimator.

4 Conclusions

[Krause *et al.*, 2020] derived a generalized form of Crooks’ equality, which can be used to devise generalizations of known estimators for the partition functions of Markov Random Fields (MRFs), including Annealed Importance Sampling (AIS) and Bennett’s Acceptance Ratio method (BAR). Various methods for generating samples are covered such as Parallel Tempering (PT), path sampling (used by AIS), and Linked Importance Sampling [Neal, 2005].

When empirically comparing PT-based estimators with vanilla AIS for estimation the normalization constants we found different regimes: PT based estimators (as BAR shown here) worked well with a small number of bridging distributions but required many samples, while AIS required many bridging distributions but only a few samples of those. This makes BAR a particularly good candidate for monitoring training process when PT is used during training.

Another important result is that choosing a reference distribution which is close to the target distribution can have a major impact on the quality of the estimate.

Acknowledgements

This work was supported by the Innovation Fund Denmark through the *Danish Center for Big Data Analytics Driven Innovation* (DABAI).

References

[Bennett, 1976] Charles H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245 – 268, 1976.

[Blake *et al.*, 2011] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[Crooks, 2000] Gavin E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Physical Review E*, 61:2361–2366, 2000.

[Desjardins *et al.*, 2010] Guillaume Desjardins, Aaron Courville, Yoshua Bengio, Pascal Vincent, and Olivier Delalleau. Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 145–152, 2010.

[Desjardins *et al.*, 2011] Guillaume Desjardins, Aaron C. Courville, and Yoshua Bengio. On tracking the partition function. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2501–2509, 2011.

[Fischer and Igel, 2014] Asja Fischer and Christian Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47:25–39, 2014.

[Fischer and Igel, 2015] Asja Fischer and Christian Igel. A bound for the convergence rate of parallel tempering for sampling restricted Boltzmann machines. *Theoretical Computer Science*, 598:102–117, 2015.

[Geyer, 1994] Charles J. Geyer. Estimating normalizing constants and reweighting mixtures. Technical Report 568, School of Statistics, University of Minnesota, 1994.

[Hinton, 2002] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[Igel *et al.*, 2008] Christian Igel, Tobias Glasmachers, and Verena Heidrich-Meisner. Shark. *Journal of Machine Learning Research*, 9:993–996, 2008.

[Kindermann and Snell, 1980] Ross Kindermann and J. Laurie Snell. *Markov Random Fields and Their Applications*. AMS, 1980.

[Krause *et al.*, 2020] Oswin Krause, Asja Fischer, and Christian Igel. Algorithms for estimating the partition function of restricted Boltzmann machines. *Artificial Intelligence Journal*, 278, 2020.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Liu *et al.*, 2015] Qiang Liu, Jian Peng, Alexander Ihler, and John Fisher III. Estimating the partition function by discriminance sampling. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 514–522. AUAI Press, 2015.

[Meng and Wong, 1996] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical explanation. *Statistica Sinica*, 6:831–860, 1996.

[Neal, 2001] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[Neal, 2005] Radford M. Neal. Estimating ratios of normalizing constants using linked importance sampling. *ArXiv Mathematics e-prints*, 2005.

[Salakhutdinov and Murray, 2008] Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 872–879, New York, NY, USA, 2008. ACM.

[Schulz *et al.*, 2010] Hannes Schulz, Andreas C. Müller, and Sven Behnke. Investigating convergence of restricted Boltzmann machine learning. In *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[Shirts *et al.*, 2003] Michael R. Shirts, Eric Bair, Giles Hooker, and Vijay S. Pande. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Physical Review Letters*, 91:140601, 2003.

[Smolensky, 1986] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*, pages 194–281. MIT Press, 1986.

[Wu *et al.*, 2017] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations (ICLR)*, 2017.