Connecting the Thermodynamics Variational Objective and Annealed Importance Sampling

Thang Bui

August 2020

Abstract

This short note (i) shows learning with the thermodynamics variational objective is a form of adaptive annealed importance sampling, (ii) summarises the result of Grosse et al. (2013) on the lower-bound estimate of the log-partition function, and (iii) explores alternative interpolating paths for AIS based on α-divergences.

1 Annealed importance sampling and lower-bound on the log-partition function

Annealed importance sampling (AIS) provides an estimate of the (log)-partition function of an intractable target distribution by sampling from a distribution path that interpolates between a tractable initial distribution and the target (Neal, 2001). In detail, suppose there is a difficult density $p(z) = f(z)/\mathcal{Z}$, where $f(z)$ is the unnormalised density and $\mathcal{Z}$ is the partition function that we wish to estimate, $\mathcal{Z} = \int f(z)\mathrm{d}z$. For simplicity, we assume a tractable initial distribution, $p_0(z) = f_0(z)/\mathcal{Z}_0$. AIS first forms a sampling path of $K+1$ intermediate densities, $\{p_k(z)\}_{k=0}^K$, that slowly anneals from $p_0(z)$ to $p(z)$. The k-th intermediate density can be written as $p_k(z) = f_k(z)/\mathcal{Z}_k$, where $f_k(z)|_{k=0} = f_0(z)$ and $f_k(z)|_{k=K} = f_K(z) = f(z)$. Assuming that one can sample from an MCMC transition operator $\mathbb{T}_k(z|z_{k-1})$ that leaves p_k invariant, AIS alternates between importance sampling updates and MCMC transitions as in algorithm 1. The estimate of the partition function is the average of particle weights at the end of the sampling path.

Algorithm 1: Annealed Importance Sampling

for $m = 1$ **to** M **do**
- $z_0^{(m)} \sim p_0(z)$
- $w_0^{(m)} = \mathcal{Z}_0$
- **for** $k = 1$ **to** K **do**
  - $w_k^{(m)} = w_k^{(m)} f_k(z_{k-1}^{(m)})/f_{k-1}(z_{k-1}^{(m)})$
  - $z_k^{(m)} \leftarrow$ sample from $\mathbb{T}_k(z|z_{k-1}^{(m)})$

return $\mathcal{Z} \approx \frac{1}{M}\sum_m w_K^{(m)}$

It turns out that we can also obtain an estimate of a lower bound on the log-partition function. We next summarise the lower bounding result and proof of Grosse et al. (2013) for the log-partition function. For ease of analysis, we assume perfect transitions, i.e., $\{z_k^{(m)}\}_{m=1}^M$ are independent and exact samples from $p_k(z)$.

Consider the log of the final weight for one particle,

$$\begin{aligned}\log w_K &= \log w_{K-1} + \log f_K(z_{K-1}) - \log f_{K-1}(z_{K-1}) \\ &= \log w_{K-2} + \log f_{K-1}(z_{K-2}) - \log f_{K-2}(z_{K-2}) + \log f_K(z_{K-1}) - \log f_{K-1}(z_{K-1}) \\ &\vdots \\ &= \log w_0 + \sum_{k=1}^{K} [\log f_k(z_{k-1}) - \log f_{k-1}(z_{k-1})]\end{aligned}$$

As $w_0 = \mathcal{Z}_0$, the expected log-weight is,

$$\mathcal{F} \coloneqq \mathrm{E}[\log w_K] = \log \mathcal{Z}_0 + \sum_{k=1}^{K} \mathrm{E}_{p_{k-1}(z)} \left[\log \frac{f_k(z)}{f_{k-1}(z)} \right] \tag{1}$$

Note that, $\log f_k(z) = \log p_k(z) + \log \mathcal{Z}_k$ and, similarly, $\log f_{k-1}(z) = \log p_{k-1}(z) + \log \mathcal{Z}_{k-1}$, and $\mathcal{Z}_K = \mathcal{Z}$, the expected log-weight becomes,

$$\begin{aligned}\mathcal{F} &= \log \mathcal{Z}_0 + \sum_{k=1}^{K} \mathrm{E}_{p_{k-1}(z)} \left[\log \frac{p_k(z)}{p_{k-1}(z)} + \log \mathcal{Z}_k - \log \mathcal{Z}_{k-1} \right] \\ &= \log \mathcal{Z} - \sum_{k=1}^{K} \mathrm{KL}[p_{k-1}(z) || p_k(z)]. \end{aligned} \tag{2}$$

Therefore, the expected log-weight is an under-estimate of the log-partition. The gap is the sum of KL divergences between intermediate distributions. This key result of Grosse et al. (2013) does not make any assumption about the sampling path and thus holds for any paths.

2 A special case: the Thermodynamics Variational Objective

Consider a geometric averaging path, $f_k(z) = f_0(z)^{1-\beta_k} f(z)^{\beta_k}$, where $\beta_k \in [0, 1]$, $\beta_0 = 0$, and $\beta_K = 1$. This leads to,

$$\log f_k(z) - \log f_{k-1}(z) = (\beta_k - \beta_{k-1})(\log f(z) - \log f_0(z)). \tag{3}$$

If we choose the initial distribution such that $p_0(z) = f_0(z)$ and $\mathcal{Z}_0 = 1$, the expected log-weight in eq. (1) becomes,

$$\mathcal{F} = \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \mathrm{E}_{p_{k-1}(z)} \left[\log \frac{f(z)}{p_0(z)} \right]. \tag{4}$$

The expected log-weight, or lower bound to the log-partion function above in eq. (4) is exactly the Thermodynamics Variational Objective (Masrani et al., 2019). The gap between the TVO and the log-partition function was recently shown to be the sum of KL divergences between adjacent distributions in the path (Brekelmans et al., 2020). As shown earlier, we note that this property is more general and is not specific to the geometric averaging path. Learning a variational distribution with the TVO objective can thus be viewed as a form of annealed importance sampling in which (i) we adapt the initial distribution to optimise the lower bound, and (ii) self-normalised importance sampling, instead of combining MCMC and importance sampling as in algorithm 1, is used to evaluate the expectations.

3 Exploring alternative sample paths

Choosing an appropriate sequence of interpolating distributions is key to the performance of AIS. Gelman and Meng (1998) show that in a simple case of annealing between two univariate Gaussian densities, the

popular geometric averaging path is sub-optimal. However, finding an optimal path is difficult for a general target distribution. Grosse et al. (2013) propose using moment averaging path as an alternative to moment averaging, which results in higher effective sample size and potentially tighter bound estimates for Restricted Boltzmann Machines. Grosse et al. (2013) also provide variational interpretations for both geometric averaging and moment averaging paths. We build upon these to motivate alternative paths based on α-divergences.

Grosse et al. (2013) show that choosing an intermediate distribution in the geometric averaging path is equivalent to minimising a weighted sum of KL divergences to the initial and target distributions,

$$q_\beta^{\mathrm{GA}}(z) = \arg\min_{q(z)} (1-\beta)\mathrm{KL}[q(z)||p_0(z)] + \beta\mathrm{KL}[q(z)||p(z)].$$

Proof. (Grosse et al., 2013) Consider the Lagrangian of the objective function above:

$$\mathcal{L}_{\mathrm{GA}} = \lambda\left(\int q(z)\mathrm{d}z - 1\right) + (1-\beta)\mathrm{KL}[q(z)||p_0(z)] + \beta\mathrm{KL}[q(z)||p(z)].$$

Setting its the functional derivative wrt $q(z)$ to zero gives $\log q(z) = C + (1-\beta)\log q_0(z) + \beta \log p(z)$, or $q(z) \propto q_0^{1-\beta}(z)p^\beta(z)$. For exponential family densities, this result translates to $\eta_q = (1-\beta)\eta_{q_0} + \beta\eta_p$, where η denotes the natural parameters. □

When the path consists of exponential family densities, reversing the direction of the KL divergences above gives an objective function for the moment averaging path,

$$q_\beta^{\mathrm{MA}}(z) = \arg\min_{q(z)} (1-\beta)\mathrm{KL}[p_0(z)||q(z)] + \beta\mathrm{KL}[p(z)||q(z)].$$

Proof. (Grosse et al., 2013) When $q(z)$ is in the exponential family, it can be written as $q(z) = \exp(\eta^\intercal g(z))/\mathcal{Z}_\eta$, where $g(z)$ denotes the sufficient statistics. The variational objective above can be rewritten as,

$$\mathcal{L}_{\mathrm{MA}} = C + \log \mathcal{Z}_\eta - \int [(1-\beta)q_0(z) + \beta p(z)]\eta^\intercal g(z)\mathrm{d}z.$$

Note that $\frac{\mathrm{d}\log \mathcal{Z}_\eta}{\mathrm{d}\eta} = \mathrm{E}_{q(z)}[g(z)]$. Setting the gradient of the above objective and rearranging give,

$$\mathrm{E}_{q(z)}[g(z)] = (1-\beta)\mathrm{E}_{q_0(z)}[g(z)] + \beta\mathrm{E}_{p(z)}[g(z)].$$

□

We can generalise the above variational objectives by considering the α-divergence in place of KL. In detail, consider the following objective,

$$\mathcal{L}_{\mathrm{AA}} = (1-\beta)\mathrm{D}_\alpha[p_0(z)||q(z)] + \beta\mathrm{D}_\alpha[p(z)||q(z)]. \tag{5}$$

Minimising this objective wrt q(z) results in the following stationary condition:

$$\mathrm{E}_{q(z)}[g(z)] = (1-\beta)\mathrm{E}_{q_0'(z)}[g(z)] + \beta\mathrm{E}_{p'(z)}[g(z)], \tag{6}$$

$$\text{where} \quad q_0'(z) = q_0^\alpha(z)q^{1-\alpha}(z), \tag{7}$$

$$p'(z) = p^\alpha(z)q^{1-\alpha}(z). \tag{8}$$

The derivation for this result uses the proof for the moment averaging path above, and the relationship between stationary points of α- and KL divergences (Minka, 2005, Theorem 3). When $\alpha = 1$ or $\alpha \to 0$, we arrive at moment averaging and geometric averaging, respectively.

We note again that this section is merely an exercise to explore alternative paths and we hypothesize that the optimal path by no means should be a result of minimising the above objective for a particular α value. However, we suggest that for a fixed $\{\beta_k\}_{k=0}^K$ schedule, by carefully choosing α, there are potentially better paths than the moment averaging or geometric averaging path.

4 Some toy examples

We first consider various paths interpolating between $\mathcal{N}(z; -4, 1)$ and $\mathcal{N}(z; 4, 0.2)$, with the number of densities $K + 1 = 25$. The paths given by various averaging approaches are shown in fig. 1. We also show the elements of the sum in the AIS bound in eq. (1), as well as the sum which is the estimate of the log-partition function. Since both the initial and target densities are normalised, the ground-truth log-partition function is $\log(1) = 0$. We note that the averaging path corresponding to $\alpha = 0.05$ is better than the geometric averaging path by about 3 nats and than the moment averaging path by about 5 nats.

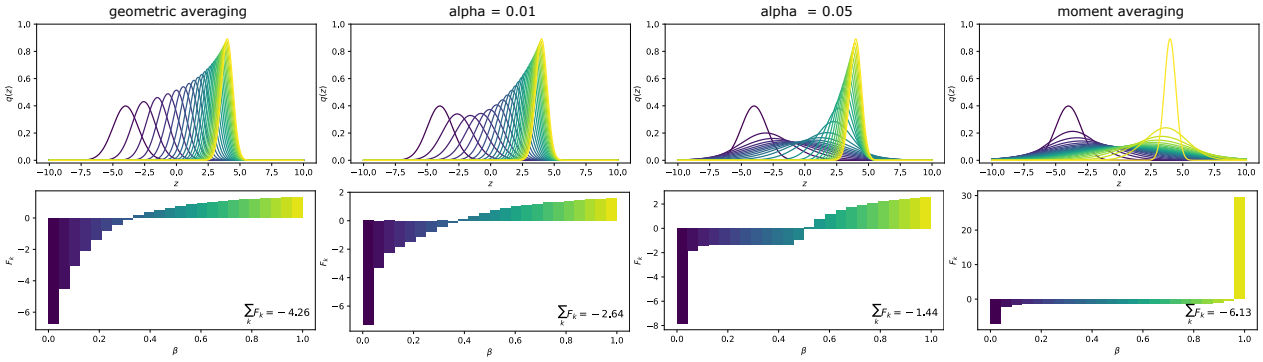

Figure 1: [Top] Intermediate densities interpolating between $\mathcal{N}(-4, 1)$ and $\mathcal{N}(4, 0.2)$ for various paths and [Bottom] corresponding elements of the sum in eq. (1).

We show the difference between the estimates of the log-partition function provided by various paths in fig. 2. We observe that for a single example, the relative ranking of different paths remains stable. However, this ranking varies across examples, for instance: $\alpha = 0.05$ seems to be best when annealing from $\mathcal{N}(-4, 1)$ to $\mathcal{N}(4, 0.2)$, but moment averaging ($\alpha = 1$) seems superior when annealing from $\mathcal{N}(-4, 0.2)$ to $\mathcal{N}(4, 1.0)$.

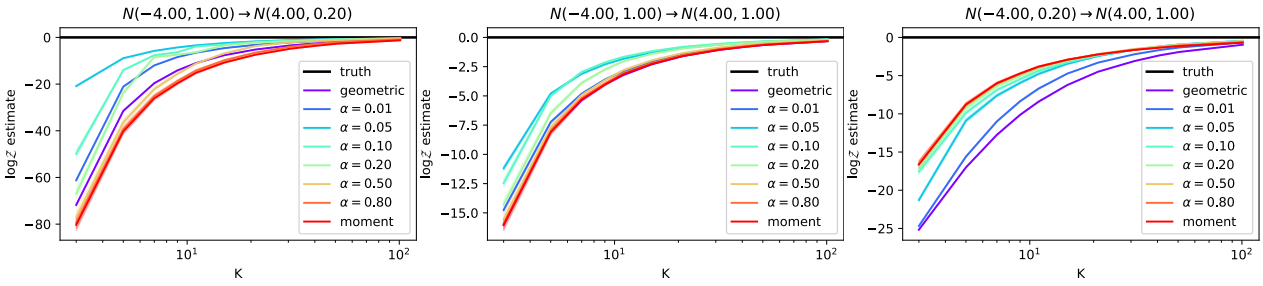

Figure 2: The lower-bounds on the log-partition function eq. (1) provided by various paths, with various numbers of regularly spaced β's.

5 Summary

We have shown learning using the Thermodynamics Variational Objective is a special case of adaptive Annealed Importance Sampling. The KL gap result recently presented by Brekelmans et al. (2020) is thus also true for more general sample paths, thanks to the results of Grosse et al. (2013). Some toy examples were provided to illustrate alternative paths based on α-divergences. Using these sample paths for Restricted Boltzmann Machines is an interesting future direction.

References

Brekelmans, R., Masrani, V., Wood, F., Steeg, G. V., and Galstyan, A. (2020). All in the exponential family: Bregman duality in thermodynamic variational inference. *arXiv preprint arXiv:2007.00642*.

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

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. (2013). Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, pages 2769–2777.

Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, pages 11525–11534.

Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research.

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