Annealing Between Distributions by Averaging Moments

Roger Grosse RGROSSE@MIT.EDU
Comp. Sci. & AI Lab, MIT, Cambridge, MA, 02139, USA

Chris Maddison CMADDIS@CS.TORONTO.EDU
Dept. of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada

Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Depts. of Statistics and Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada

Abstract

Many powerful Monte Carlo techniques for estimating partition functions, such as annealed importance sampling (AIS), are based on sampling from a sequence of intermediate distributions which interpolate between a tractable initial distribution and an intractable target distribution. The near-universal practice is to use geometric averages of the initial and target distributions, but alternative paths can perform substantially better. We present a novel sequence of intermediate distributions for exponential families: averaging the moments of the initial and target distributions. We derive an asymptotically optimal piecewise linear schedule for the moments path and show that it performs at least as well as geometric averages with a linear schedule. Moment averaging performs well empirically at estimating partition functions of restricted Boltzmann machines (RBMs), which form the building blocks of many deep learning models.

1. Introduction

Many generative models are defined in terms of an unnormalized probability distribution, and computing the probability of a data point requires computing the (usually intractable) partition function. This is problematic for model selection, since one often wishes to compute the probability assigned to held-out test data. While partition function estimation is intractable in general, there has been extensive research on variational (Yedidia et al., 2005; Wainwright et al., 2005; Globerson & Jaakkola, 2007) and sampling-based (Neal, 2001; Skilling, 2006; Moral et al., 2006) approximations. In the context of model comparison, annealed importance sampling (AIS) (Neal, 2001) is especially widely used because given enough computational resources, it can provide high-accuracy estimates. AIS has enabled precise quantitative comparisons of powerful generative models in image statistics (Sohl-Dickstein & Culpepper, 2012; Theis et al., 2011) and deep learning, including restricted Boltzmann Machines and Deep Belief Networks (Salakhutdinov & Murray, 2008; Desjardins et al., 2011; Taylor & Hinton, 2009). Unfortunately, applying AIS in practice can be computationally-intensive and require laborious hand-tuning of annealing schedules. Because of this, many generative models still have not been quantitatively compared in terms of held-out likelihood (LeRoux et al., 2011).

AIS requires defining a path of intermediate distributions which interpolate between a tractable initial distribution and the intractable target distribution. Typically, one uses geometric averages of the initial and target distributions. Tantalizingly, Gelman & Meng (1998) derived the optimal paths for some toy models in the context of path sampling, and showed that they vastly outperformed geometric averages. However, as choosing an optimal path is generally intractable, geometric averages still predominate.

In this paper, we present a theoretical framework for evaluating alternative paths. We propose a novel sequence of intermediate distributions defined by averaging moments of the initial and target distributions. We show that the two sequences optimize different variational objectives, derive an asymptotically optimal piecewise linear schedule, and give strong theoretical guarantees of the performance under perfect mixing. Our proposed path often outperforms geometric averages at estimating partition functions of restricted Boltzmann machines (RBMs).

Algorithm 1 Annealed Importance Sampling

```
for $i = 1$ to $M$ do
    $\mathbf{x}_0 \leftarrow$ sample from $p_0(\mathbf{x})$
    $w^{(i)} \leftarrow \mathcal{Z}_a$
    for $k = 1$ to $K$ do
        $w^{(i)} \leftarrow w^{(i)} \frac{f_k(\mathbf{x}_{k-1})}{f_{k-1}(\mathbf{x}_{k-1})}$
        $\mathbf{x}_k \leftarrow$ sample from $T_k(\mathbf{x} \,|\, \mathbf{x}_{k-1})$
    end for
end for
return $\hat{\mathcal{Z}}_b = \sum_{i=1}^{M} w^{(i)}/M$
```

2. Estimating Partition Functions

Suppose we have a probability distribution $p_b(\mathbf{x}) = f_b(\mathbf{x})/\mathcal{Z}_b$ defined on some space $\mathcal{X}$, where $f_b(\mathbf{x})$ can be computed efficiently for a given $\mathbf{x} \in \mathcal{X}$, and we are interested in estimating the partition function $\mathcal{Z}_b$. Annealed importance sampling (AIS) is an algorithm which estimates $\mathcal{Z}_b$ by gradually changing, or "annealing," a distribution. In particular, suppose we have a sequence of $K+1$ intermediate distributions $p_k(\mathbf{x}) = f_k(\mathbf{x})/\mathcal{Z}_k$ for $k = 0, \ldots K$, where $p_a(\mathbf{x}) = p_0(\mathbf{x})$ is a tractable initial distribution, and $p_b(\mathbf{x}) = p_K(\mathbf{x})$ is the intractable target distribution. For simplicity, assume all distributions are strictly positive on $\mathcal{X}$. Suppose that for each p_k we have an MCMC transition operator T_k (e.g. Gibbs sampling) which leaves p_k invariant. AIS alternates between MCMC transitions and importance sampling updates, as shown in Alg 1.

The output of AIS is an unbiased estimate $\hat{\mathcal{Z}}_b$ of $\mathcal{Z}_b$. Remarkably, this holds even in the context of *non-equilibrium* samples along the chain (Neal, 2001; Jarzynski, 1997). However, unless the intermediate distributions and transition operators are carefully chosen, $\hat{\mathcal{Z}}_b$ may have high variance and be far from $\mathcal{Z}_b$ with high probability.

The mathematical formulation of AIS leaves much flexibility for choosing intermediate distributions. However, one typically defines a path $\gamma : [0,1] \mapsto \mathcal{P}$ through some family $\mathcal{P}$ of distributions. The intermediate distributions p_k are points along this path corresponding to a *schedule* $0 = \beta_0 < \beta_1 < \ldots < \beta_K = 1$. One typically uses the geometric path γ_{GA}, defined in terms of geometric averages of p_a and p_b:

$$p_\beta(\mathbf{x}) = f_\beta(\mathbf{x})/\mathcal{Z}(\beta) = f_a(\mathbf{x})^{1-\beta} f_b(\mathbf{x})^\beta / \mathcal{Z}(\beta). \quad (1)$$

Commonly, f_a is the uniform distribution, and (1) reduces to $p_\beta(\mathbf{x}) = f_b(\mathbf{x})^\beta/\mathcal{Z}(\beta)$. This motivates the term "annealing", and β resembles an inverse temperature parameter. As in simulated annealing, the "hotter" distributions often allow faster mixing between modes which are isolated in p_b.

AIS is closely related to a broader family of techniques for posterior inference and partition function estimation, all based on the following identity from statistical physics:

$$\log \mathcal{Z}_b - \log \mathcal{Z}_a = \int_0^1 \mathbb{E}_{\mathbf{x} \sim p_\beta} \left[\frac{d}{d\beta} \log f_\beta(\mathbf{x}) \right] d\beta. \quad (2)$$

Thermodynamic integration (Frenkel & Smit, 2002) estimates (2) using numerical quadrature, and path sampling (Gelman & Meng, 1998) does so with Monte Carlo integration. The weight update in AIS can be seen as a finite difference approximation. Tempered transitions (Neal, 1996) is a Metropolis-Hastings proposal operator which heats up and cools down the distribution, and computes an acceptance ratio by approximating (2).

The choices of a path and a schedule are central to all of these methods. Most work on adapting paths has focused on tuning schedules along a geometric path (Neal, 1996; Behrens et al., 2012; Calderhead & Girolami, 2009). Neal (1996) showed that the geometric *schedule* was optimal for annealing the scale parameter of a Gaussian, and Behrens et al. (2012) extended this result more broadly. The aim of this paper is to propose, analyze, and evaluate a novel alternative to γ_{GA} based on averaging moments of the initial and target distributions.

3. Analyzing AIS Paths

When analyzing AIS, it is common to assume *perfect transitions*, i.e. that each transition operator T_k returns an independent and exact sample from the distribution p_k (Neal, 2001). This is intended to model the situation wheres the number of intermediate distributions is much larger than the mixing time. As Neal (2001) points out, assuming perfect transitions, the Central Limit Theorem shows that the $w^{(i)}$ are approximately log-normally distributed. In this case, the variances $\mathrm{var}(w^{(i)})$ and $\mathrm{var}(\log w^{(i)})$ are both monotonically related to $\mathbb{E}[\log w^{(i)}]$. Therefore, our analysis focuses on $\mathbb{E}[\log w^{(i)}]$.

Assuming perfect transitions, the expected log weights are given by:

$$\begin{aligned} \mathbb{E}[\log w^{(i)}] &= \log \mathcal{Z}_a + \sum_{k=0}^{K-1} \mathbb{E}_{p_k}[\log f_{k+1}(\mathbf{x}) - \log f_k(\mathbf{x})] \\ &= \log \mathcal{Z}_b - \sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(p_k \| p_{k+1}). \end{aligned} \quad (3)$$

In other words, each $\log w^{(i)}$ can be seen as a biased estimator of $\log \mathcal{Z}_b$, where the bias δ is given by the sum of KL divergences $\sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(p_k \| p_{k+1})$.

Suppose $\mathcal{P}$ is a family of probability distributions parameterized by $\boldsymbol{\theta} \in \Theta$, and the $K+1$ distributions $p_0, \dots, p_K$ are chosen to be linearly spaced along a path $\gamma : [0,1] \mapsto \mathcal{P}$. Let $\boldsymbol{\theta}(\beta)$ represent the parameters of the distribution $\gamma(\beta)$. As the number of intermediate distributions K is increased, the bias δ decays like $1/K$, and the asymptotic behavior is determined by a functional $\mathcal{F}(\gamma)$:

Theorem 1. *Suppose $K+1$ distributions p_k are linearly spaced along a path γ. Under the assumption of perfect transitions, if $\boldsymbol{\theta}(\beta)$ and the Fisher information matrix $\mathbf{G}_{\boldsymbol{\theta}} = \mathrm{cov}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x}))$ are smooth, then as $K \to \infty$ the bias $\delta = \log \mathcal{Z}_b - \mathbb{E}[\log w^{(i)}]$ is determined by the functional:*

$$K\delta = K \sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(p_k \| p_{k+1}) \to \mathcal{F}(\gamma) \equiv \frac{1}{2}\int_0^1 \dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta) \dot{\boldsymbol{\theta}}(\beta) d\beta, \quad (4)$$

where $\dot{\boldsymbol{\theta}}(\beta)$ represents the derivative of $\boldsymbol{\theta}$ with respect to β. [See appendix for proof.]

This result reveals a relationship with path sampling, as Gelman & Meng (1998) showed that the variance of the path sampling estimator is proportional to the same functional. One useful result from their analysis is a derivation of the optimal schedule along a given path. In particular, the value of $\mathcal{F}(\gamma)$ using the optimal schedule is given by $\ell(\gamma)^2/2$, where ℓ is the Riemannian path length defined by

$$\ell(\gamma) = \int_0^1 \sqrt{\dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta) \dot{\boldsymbol{\theta}}(\beta)} d\beta. \quad (5)$$

Intuitively, the optimal schedule moves more slowly (i.e. assigns more intermediate distributions) along the parts of the path with high curvature. While Gelman & Meng (1998) derived the optimal paths and schedules for some simple examples, they observed that this is intractable in most cases and recommended using geometric paths in practice.

The above analysis assumes perfect transitions, which can be unrealistic in practice because many distributions of interest have separated modes between which mixing is difficult. As Neal (2001) observed, in such cases, AIS can be viewed as having two sources of variability: that caused by movement within a mode, and that caused by the allocation of samples to different modes. The former source of variability is well modeled by the perfect transitions analysis, and can be made small by adding more intermediate distributions. The latter, however, can persist even with large numbers of intermediate distributions. While our theoretical analysis focuses on perfect transitions, our proposed method often gave substantial improvement empirically in situations with poor mixing.

4. Moment Averaging

As discussed in Section 2, the typical choice of intermediate distributions for AIS is the geometric averages path γ_{GA} given by (1). In this section, we propose an alternative path for an exponential family model. An exponential family model is defined as

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^T \mathbf{g}(\mathbf{x})\right), \quad (6)$$

where $\boldsymbol{\eta}$ are the natural parameters and $\mathbf{g}$ are the sufficient statistics. Exponential families include a wide variety of statistical models as special cases, including Markov random fields.

In exponential families, geometric averages correspond to averaging the natural parameters:

$$\boldsymbol{\eta}(\beta) = (1-\beta)\boldsymbol{\eta}(0) + \beta\boldsymbol{\eta}(1) \quad (7)$$

Exponential families can also be parameterized in terms of their moments, or expected sufficient statistics, $\mathbf{s} = \mathbb{E}[\mathbf{g}(\mathbf{x})]$. For any exponential family, there is a one-to-one mapping between moments and natural parameters. We propose an alternative to γ_{GA} called the *moment averages* path, denoted γ_{MA}, and defined by averaging the moments of the initial and target distributions:

$$\mathbf{s}(\beta) = (1-\beta)\mathbf{s}(0) + \beta\mathbf{s}(1). \quad (8)$$

This path exists for any minimal exponential family model, since the set of realizable moments is convex (Wainwright & Jordan, 2008).

As an illustrative example, consider a multivariate Gaussian distribution parameterized by the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The expected sufficient statistics are $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $-\frac{1}{2}\mathbb{E}[\mathbf{x}\mathbf{x}^T] = -\frac{1}{2}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)$. By plugging these into (8), we find that γ_{MA} is given by:

$$\boldsymbol{\mu}(\beta) = (1-\beta)\boldsymbol{\mu}(0) + \beta\boldsymbol{\mu}(1) \quad (9)$$

$$\boldsymbol{\Sigma}(\beta) = (1-\beta)\boldsymbol{\Sigma}(0) + \beta\boldsymbol{\Sigma}(1) + \beta(1-\beta)(\boldsymbol{\mu}(1) - \boldsymbol{\mu}(0))(\boldsymbol{\mu}(1) - \boldsymbol{\mu}(0))^T. \quad (10)$$

In other words, the means are linearly interpolated, and the covariances are linearly interpolated and stretched out in the direction containing the two means. Intuitively, this stretching is a useful property, because it means that each intermediate distribution is more similar to the next. A comparison of the two paths is shown in Figure 1.

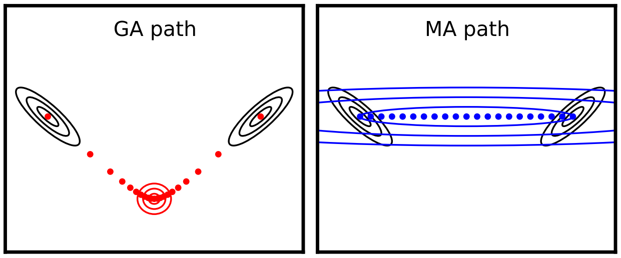

Figure 1. Comparison of γ_{GA} and γ_{MA} for multivariate Gaussians: intermediate distribution for $\beta = 0.5$, and $\boldsymbol{\mu}(\beta)$ for β evenly spaced from 0 to 1.

Next consider the example of a restricted Boltzmann machine (RBM), a widely used model in deep learning. An RBM is a Markov random field with variables $\mathbf{v}$ (the visible units) and $\mathbf{h}$ (the hidden units), and which has the distribution

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{a}^T \mathbf{v} + \mathbf{b}^T \mathbf{h} + \mathbf{v}^T \mathbf{W} \mathbf{h}\right). \tag{11}$$

The parameters of the model are the visible biases $\mathbf{a}$, the hidden biases $\mathbf{b}$, and the weights $\mathbf{W}$. Since these parameters are also the natural parameters in the exponential family representation, γ_{GA} reduces to linearly averaging the biases and the weights. The sufficient statistics of the model are the visible activations $\mathbf{v}$, the hidden activations $\mathbf{h}$, and the products $\mathbf{v}\mathbf{h}^T$. Therefore, γ_{MA} is defined by:

$$\mathbb{E}[\mathbf{v}]_\beta = (1-\beta)\mathbb{E}[\mathbf{v}]_0 + \beta\mathbb{E}[\mathbf{v}]_1 \tag{12}$$
$$\mathbb{E}[\mathbf{h}]_\beta = (1-\beta)\mathbb{E}[\mathbf{h}]_0 + \beta\mathbb{E}[\mathbf{h}]_1 \tag{13}$$
$$\mathbb{E}[\mathbf{v}\mathbf{h}^T]_\beta = (1-\beta)\mathbb{E}[\mathbf{v}\mathbf{h}^T]_0 + \beta\mathbb{E}[\mathbf{v}\mathbf{h}^T]_1 \tag{14}$$

For most models of interest, including RBMs, it is infeasible to determine γ_{MA} exactly, as it requires solving two often intractable problems: (1) estimating the moments of p_b, and (2) solving for model parameters which match the averaged moments $\mathbf{s}(\beta)$. However, much work has been devoted to practical approximations (Hinton, 2002; Tieleman, 2008), some of which we use in our experiments with intractable models. Since it would be infeasible to moment match every β_k even approximately, we introduce the moment average spline (MAS) path, denoted γ_{MAS}. We choose a set of R values $\beta_1, \ldots, \beta_R$ called *knots*, and solve for the natural parameters $\boldsymbol{\eta}(\beta_j)$ to match the moments $\mathbf{s}(\beta_j)$ for each knot. We then interpolate between the knots using geometric averages. The analysis of Section 4.2 shows that, under the assumption of perfect sampling, nothing is lost by using the piecewise geometric path γ_{MAS} in place of the exact moment averages path γ_{MA}.

4.1. Variational Interpretation

We can interpret γ_{GA} and γ_{MA} as optimizing different variational objectives, which provides additional insight into their behavior. For geometric averages, the intermediate distribution $\gamma_{GA}(\beta)$ optimizes a weighted sum of KL divergences to the initial and target distributions:

$$\arg\min_p \; (1-\beta)\mathrm{D}_{\mathrm{KL}}(p \| p_0) + \beta \mathrm{D}_{\mathrm{KL}}(p \| p_1). \tag{15}$$

On the other hand, the γ_{MA} minimizes the sum of KL divergences in the reverse direction:

$$\arg\min_p \; (1-\beta)\mathrm{D}_{\mathrm{KL}}(p_0 \| p) + \beta \mathrm{D}_{\mathrm{KL}}(p_1 \| p). \tag{16}$$

See the appendix for the derivations. The optimization problem (15) is optimized by a distribution which only puts significant mass in the "intersection" of p_0 and p_1, i.e. those regions which are likely under both the initial and target distributions. By contrast, (16) encourages the distribution to be spread out in order to capture all high probability regions of both p_0 and p_1. This interpretation helps explain why the intermediate distributions in the Gaussian example of Figure 1 take the shape that they do. In our experiments, we found that γ_{MA} often gave more accurate results than γ_{GA} because the intermediate distributions captured regions of the target distribution which were missed by γ_{GA}.

4.2. Asymptotics under Perfect Transitions

In general, we found that γ_{GA} and γ_{MA} can look very different. Intriguingly, both paths always result in the same value of the cost functional $\mathcal{F}(\gamma)$ of Theorem 1 for any exponential family model. Furthermore, nothing is lost by using the spline approximation γ_{MAS} in place of γ_{MA}:

Theorem 2. *For any exponential family model with natural parameters $\boldsymbol{\eta}$ and expected sufficient statistics $\mathbf{s}$, the functionals for the geometric and moments paths are given by:*

$$\mathcal{F}(\gamma_{GA}) = \mathcal{F}(\gamma_{MA}) = \mathcal{F}(\gamma_{MAS}) = \frac{1}{2}(\boldsymbol{\eta}(1) - \boldsymbol{\eta}(0))^T(\mathbf{s}(1) - \mathbf{s}(0)). \tag{17}$$

Proof. The two parameterizations of exponential families satisfy the relationship $\mathbf{G}_{\boldsymbol{\eta}}\dot{\boldsymbol{\eta}} = \dot{\mathbf{s}}$ (Amari & Nagaoka, 2000, sec. 3.3). Therefore, the cost functional can be rewritten as $\mathcal{F}(\gamma) = \frac{1}{2}\int_0^1 \dot{\boldsymbol{\eta}}(\beta)^T \dot{\mathbf{s}}(\beta) d\beta$. Because γ_{GA} and γ_{MA} interpolate the natural parameters and moments respectively,

$$\mathcal{F}(\gamma_{GA}) = \frac{1}{2}(\boldsymbol{\eta}(1) - \boldsymbol{\eta}(0))^T \int_0^1 \dot{\mathbf{s}}(\beta) d\beta$$
$$= \frac{1}{2}(\boldsymbol{\eta}(1) - \boldsymbol{\eta}(0))^T (\mathbf{s}(1) - \mathbf{s}(0)) \tag{18}$$
$$\mathcal{F}(\gamma_{MA}) = \frac{1}{2}(\mathbf{s}(1) - \mathbf{s}(0))^T \int_0^1 \dot{\boldsymbol{\eta}}(\beta) d\beta$$
$$= \frac{1}{2}(\mathbf{s}(1) - \mathbf{s}(0))^T (\boldsymbol{\eta}(1) - \boldsymbol{\eta}(0)). \tag{19}$$

Finally, to show that $\mathcal{F}(\gamma_{MAS}) = \mathcal{F}(\gamma_{MA})$, observe that γ_{MAS} uses the geometric path between each pair of knots $\gamma(\beta_j)$ and $\gamma(\beta_{j+1})$, while γ_{MA} uses the moments path. The above analysis shows the functionals must be equal for each segment, and therefore equal for the entire path. □

This analysis shows that all three paths result in the same expected log weights asymptotically, assuming perfect transitions. There are several caveats, however. First, we have noticed experimentally that γ_{MA} often yields substantially more accurate estimates of $\mathcal{Z}$ than γ_{GA} even when the average log weights are comparable. Second, the two paths can have very different mixing properties, which can strongly affect the results. Third, Theorem 2 assumes *linear* schedules, and there can be substantial room for improvement if one is allowed to tune the schedule.

For instance, consider moving between two Gaussians $p_a = \mathcal{N}(\mu_a, \sigma)$ and $p_b = \mathcal{N}(\mu_b, \sigma)$. The optimal schedule for the geometric path is a linear schedule with cost $\mathcal{F}(\gamma_{GA}) = O(d^2)$, where $d = |\mu_b - \mu_a|/\sigma$. Using a linear schedule, the moment path also has cost $O(d^2)$, consistent with Theorem 2. However, most of the cost of the path results from instability near the endpoints, where the variance changes suddenly. Using an optimal schedule, which places more distributions near the endpoints, the cost functional falls to $O((\log d)^2)$, which is within a constant factor of the optimal path derived by Gelman & Meng (1998). (See the appendix for the derivations.) In other words, while $\mathcal{F}(\gamma_{GA}) = \mathcal{F}(\gamma_{MA})$, they achieve this value for different reasons: γ_{GA} follows an optimal schedule along a bad path, while γ_{MA} follows a bad schedule along a near-optimal path. We speculate that, combined with the procedure of Section 4.3 for choosing a schedule, moment averages may result in large reductions in the cost functional for some models.

4.3. Optimal Binned Schedules

In general, it is hard to choose a good schedule for a given path. However, consider the set of *binned schedules*, where the path is divided into segments, some number K_j of intermediate distributions are allocated to each segment, and the distributions are spaced linearly within each segment. Under the assumption of perfect transitions, there is a simple formula for an asymptotically optimal binned schedule which requires only the parameters obtained through moment averaging:

Theorem 3. *Let γ be any path for an exponential family model defined by a set of knots β_j, each with natural parameters $\boldsymbol{\eta}_j$ and moments $\mathbf{s}_j$, connected by segments of either γ_{GA} or γ_{MA} paths. Then, under the assumption of perfect transitions, an asymptotically optimal allocation of intermediate distributions to each segment is given by:*

$$K_j \propto \sqrt{(\boldsymbol{\eta}_{j+1} - \boldsymbol{\eta}_j)^T (\mathbf{s}_{j+1} - \mathbf{s}_j)}. \tag{20}$$

Proof. By Theorem 2, the cost functional for segment j is $F_j = \frac{1}{2}(\boldsymbol{\eta}_{j+1} - \boldsymbol{\eta}_j)^T (\mathbf{s}_{j+1} - \mathbf{s}_j)$. Hence, with K_j distributions allocated to it, it contributes F_j/K_j to the total cost. The values of K_j which minimize $\sum_j F_j/K_j$ subject to $\sum_j K_j = K$ and $K_j > 0$ are given by $K_j \propto \sqrt{F_j}$. □

5. Experimental Results

In order to compare our proposed path with geometric averages, we ran AIS to estimate partition functions of several probability distributions using each path. For all of our experiments, we report two sets of results. First, we show the estimates of $\log \mathcal{Z}$ as a function of the number of intermediate distributions in order to visualize the amount of computation necessary to obtain reasonable accuracy. Second, as recommended by Neal (2001), we report the effective sample size (ESS) of the weights after a large number of intermediate distributions. This statistic roughly measures how many independent samples one obtains using AIS.[1] All results are based on 5,000 independent AIS runs, so the maximum possible ESS is 5,000.

[1] ESS is defined as ESS $= M/(1 + s^2(w_*^{(i)}))$ where $s^2(w_*^{(i)})$ is the sample variance of the normalized weights (Neal, 2001). In general, one should regard ESS estimates cautiously, as they can give misleading results in cases where an algorithm completely misses an important mode of the distribution. However, as we report the ESS in cases where the estimated partition functions are close to the true value (when known) or agree closely with each other, we believe them to be more accurate in our comparisons.

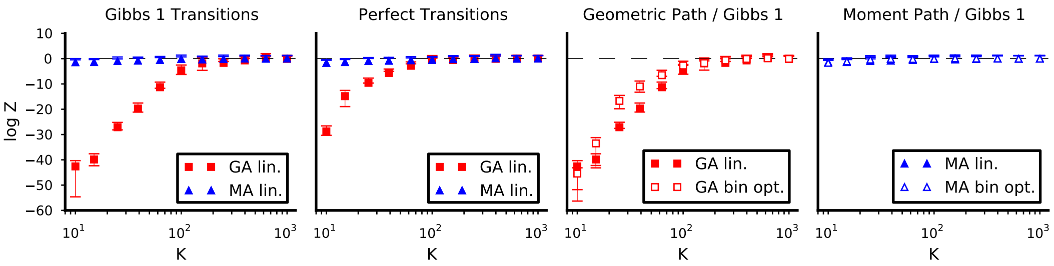

Figure 2. Estimates of the log partition function of a normalized Gaussian as the number of intermediate distributions increases. Error bars show bootstrap 95% confidence intervals. (Best viewed in color.)

5.1. Annealing Between Two Distant Gaussians

In our first experiment, the initial and target distributions were the two Gaussians shown in Fig. 1, whose parameters are given by $\mathcal{N}\left(\begin{pmatrix}-10\\0\end{pmatrix}, \begin{pmatrix}1 & -0.85\\-0.85 & 1\end{pmatrix}\right)$ and $\mathcal{N}\left(\begin{pmatrix}10\\0\end{pmatrix}, \begin{pmatrix}1 & 0.85\\0.85 & 1\end{pmatrix}\right)$. As both distributions are normalized, $\mathcal{Z}_a = \mathcal{Z}_b = 1$. We compared γ_{GA} and γ_{MA} both under perfect transitions, and using the Gibbs transition operator. We also compared linear schedules with the optimal binned schedules of Section 4.3, using 10 segments evenly spaced from 0 to 1.

Figure 2 shows the estimates of $\log \mathcal{Z}_b$ for numbers of intermediate distributions ranging from 10 to 1,000. Observe that by 1,000 intermediate distributions, all paths yield accurate estimates of $\log \mathcal{Z}$. However, γ_{MA} gave accurate estimates with smaller numbers of intermediate distributions. For example the moment path achieves an average within one nat of $\log \mathcal{Z}_b$ after 25 intermediate distributions, while the geometric path is still off by 27 nats.

When comparing γ_{GA} and γ_{MA} using perfect transitions, Theorem 2 implies that both paths have exactly the same cost functional $\mathcal{F}$, and therefore the log weights should be roughly the same.[2] In fact, the average log weights were close throughout; e.g., at 25 intermediate distributions, the average was -27.15 for γ_{MA} and -28.04 for γ_{GA}. However, the log weights for γ_{MA} had much larger variance, 1437.89, compared to 58.4 for γ_{GA}. This is because the intermediate distributions on γ_{MA} were broader, as predicted by the analysis of Section 4.1. Most of the particles died out, but enough of them landed in high probability regions to yield reasonable estimates of $\log \mathcal{Z}_b$.

5.2. Partition Function Estimation for RBMs

Our next set of experiments focused on RBMs, a building block of many deep learning models (see Section 4). We considered RBMs trained with three different algorithms: contrastive divergence (CD) with one step (CD1) (Hinton, 2002), CD with 25 steps (CD25), and persistent contrastive divergence (PCD) (Tieleman, 2008). All of the RBMs were trained on the MNIST handwritten digits dataset (LeCun et al., 1998), which has long served as a benchmark for deep learning algorithms. We experimented both on small, tractable RBMs and and full-size, intractable RBMs.

		CD1(20)		PCD(20)	
		$\log \mathcal{Z} = 279.59$		$\log \mathcal{Z} = 178.06$	
$p_a(\mathbf{v})$	path	$\log \hat{\mathcal{Z}}$	ESS	$\log \hat{\mathcal{Z}}$	ESS
uni.	GA lin.	279.60	248	177.99	204
uni.	GA bin opt.	279.51	124	177.92	142
uni.	MAS lin.	279.59	**2686**	178.09	**289**
uni.	MAS bin opt.	279.60	**2619**	178.08	**934**

Table 1. Comparing estimates of the log partition function of toy RBMs under different paths using 100,000 intermediate distributions annealing from uniform with 5,000 chains and Gibbs transitions. ESS is the effective sample size of the 5,000 chains, bolded values are ESS estimates that are *not* significantly different (bootstrap hypothesis test with 1,000 samples at $\alpha = 0.05$ significance level) under path with highest ESS.

Since it is hard to compute γ_{MA} exactly for RBMs, we used the moments spline path γ_{MAS} of Section 4 with the 9 knot locations $0.1, 0.2, \ldots, 0.9$. We considered the two initial distributions discussed by Salakhutdinov & Murray (2008): (1) the uniform distribution, equivalent to an RBM where all the weights and biases are set to 0, and (2) the *base rate RBM*, where the weights and hidden biases are set to 0, and the hidden biases are set to match the averages of the pixels over the MNIST training set.

Small, Tractable RBMs: To better understand the behavior of γ_{GA} and γ_{MAS}, we first evaluated the paths on RBMs with only 20 hidden units. Here, it is possible to compute the exact partition function and moments and generate exact samples by exhaustively summing over all 2^{20} hidden configurations. In this ex-

[2] As Theorem 1 is an asymptotic result, the average log weights need not agree closely for finite K. In this example, however, the values were close even for small K.

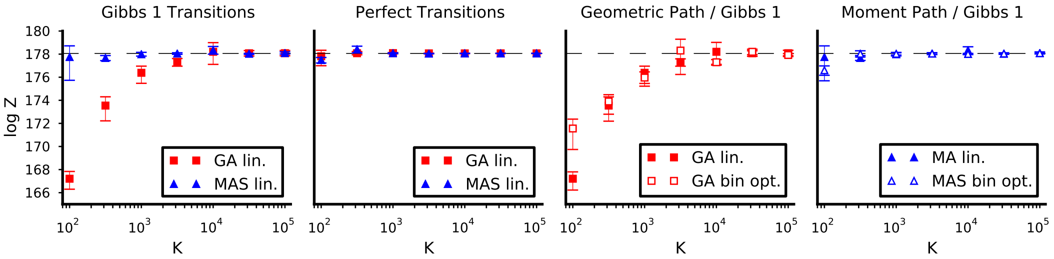

Figure 3. Estimates of the log partition function of PCD(20) toy RBM (CD1(20) analogous) as the number of intermediate distributions increases. Error bars show bootstrap 95% confidence intervals. (Best viewed in color.)

periment, the moments of the target RBMs were computed exactly, and moment matching was performed using conjugate gradient with the exact gradients computed by exhaustive enumeration.

The results are shown in Figure 3 and Table 1. As expected under Theorem 2, both γ_{GA} and γ_{MA} were able to accurately estimate the partition function using as few as 100 intermediate distributions under exact sampling. However, when limited to a single Gibbs step, γ_{MA} achieved good accuracy using much fewer intermediate distributions (Figure 3), and a higher ESS at 100,000 distributions. To check that the improved performance didn't rely on accurate moments of p_b, we repeated the experiment with highly biased moments[3], and found that the results did not change substantially.

Full-size, Intractable RBMs: For intractable RBMs, moment averaging required approximately solving two intractable problems: moment estimation for the target RBM, and moment matching. We estimated the moments from 1,000 independent Gibbs chains, using 10,000 Gibbs steps with 1,000 steps of burn-in. The moment averaged RBMs were trained using persistent contrastive divergence (PCD) (Tieleman, 2008). (We used 50,000 updates with a fixed learning rate of 0.01 and no momentum.) We also ran a cheap, inaccurate moment matching scheme (denoted MAS cheap) where visible moments were estimated from the empirical MNIST base rate and the hidden moments from the conditional distributions of the hidden units given the MNIST digits. Intermediate RBMs were fit using 1,000 PCD updates and 100 particles, for a total computational cost far smaller than that of AIS itself. Performance was comparable under this approximation, suggesting that γ_{MA} can be approximated cheaply and effectively. As with the tractable RBMs, we found that optimal binned schedules made little difference in performance, so we focus here on linear schedules.

The most serious failure was γ_{GA} for CD1(500) with uniform initialization, which under-estimated our best estimates of the log partition function (and hence over-estimated held-out likelihood) by nearly 20 nats. The geometric path from uniform to PCD(500) and the moments path from uniform to CD1(500) also resulted in underestimates, though less drastic. The rest of the paths agreed closely with each other on their partition function estimates, although some methods achieved substantially higher ESS values on some RBMs. One conclusion is that it's worth exploring multiple initializations and paths for a given RBM in order to ensure accurate results.

Figure 4 compares samples along γ_{GA} and γ_{MA} for the PCD(500) RBM, starting from the base rate RBM. For a wide range of β values, the γ_{GA} RBMs assigned most of their probability mass to an all-black image. As discussed in Section 4.1, γ_{GA} prefers configurations which are probable under both the initial and target distributions. In this case, the hidden activations were closer to uniform conditioned on the black image than on a digit, so γ_{GA} preferred the black image. By contrast, γ_{MA} yielded diverse, blurry digits which gradually coalesced into crisper ones.

6. Conclusion

We presented a theoretical analysis of the performance of AIS paths and introduced a novel path for exponential families based on averaging moments. We gave a variational interpretation of this path and an asymptotically optimal piecewise linear schedule, and showed that it asymptotically outperforms geometric averages with a linear schedule. Moment averages performed well empirically at estimating partition functions of RBMs. Many widely used sampling algorithms are also based on paths, and our contributions are potentially relevant to any of these algorithms.

[3] In particular, we computed the biased moments from the conditional distributions of the hidden units given the MNIST training examples, where each example of digit class i was counted $i + 1$ times.

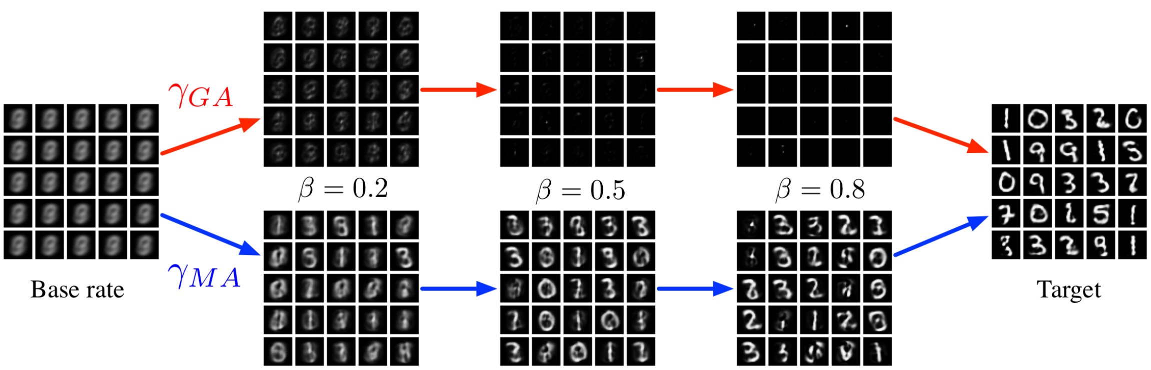

Figure 4. Visible activations for samples from PCD(500) RBM. **(left)** base rate RBM, $\beta = 0$ **(top)** geometric path **(bottom)** moments path **(right)** target RBM, $\beta = 1$.

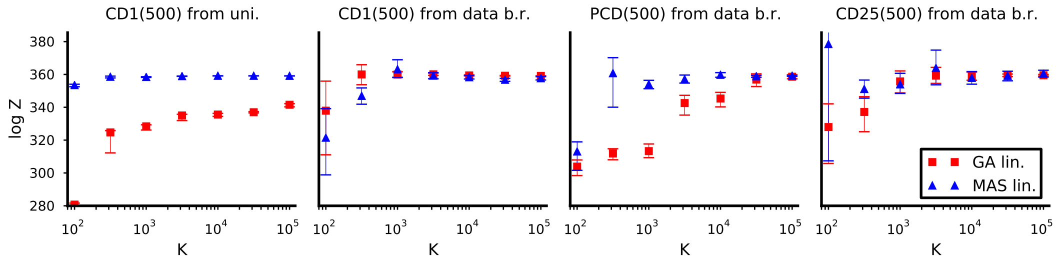

Figure 5. Estimates of the log partition function of RBMs as the number of intermediate distributions increases on some representative paths. Error bars show bootstrap 95% confidence intervals. (Best viewed in color.)

References

Amari, S. and Nagaoka, H. *Methods of Information Geometry*. Oxford University Press, 2000.

Behrens, G., Friel, N., and Hurn, M. Tuning tempered transitions. *Statistics and Computing*, 22:65–78, 2012.

Calderhead, B. and Girolami, M. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, 53(12):4028–4045, 2009.

Desjardins, G., Courville, A., and Bengio, Y. On tracking the partition function. In *NIPS 24*. MIT Press, 2011.

Frenkel, D. and Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2 edition, 2002.

Gelman, A. and Meng, X.-L. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science*, 13 (2):163–186, 1998.

Globerson, A. and Jaakkola, T. Approximate Inference Using Conditional Entropy Decompositions. In *11th International Workshop on AI and Statistics (AISTATS'2007)*, 2007.

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

Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56:5018–5035, 1997.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeRoux, N., Heess, N., Shotton, J., and Winn, J. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 2011.

Moral, P. Del, Doucet, A., and Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Methodology)*, 68(3):411–436, 2006.

Neal, R. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

		CD1(500)		PCD(500)		CD25(500)	
$p_a(\mathbf{v})$	path	$\log \hat{Z}$	ESS	$\log \hat{Z}$	ESS	$\log \hat{Z}$	ESS
uni.	GA lin.	341.53	4	417.91	169	451.34	13
uni.	MAS lin.	359.09	3076	418.27	**620**	449.22	**12**
uni.	MAS cheap lin.	359.09	**3773**	418.33	5	450.90	**30**
data b.r.	GA lin.	359.10	**4924**	418.20	159	451.27	**2888**
data b.r.	MAS lin.	359.07	2203	418.26	**1460**	451.31	304
data b.r.	MAS cheap lin.	359.09	2465	418.25	359	451.14	244

Table 2. Comparing estimates of the log partition function of RBMs under different paths using 100,000 intermediate distributions annealing from uniform with 5,000 chains and Gibbs transitions. ESS is the effective sample size of the 5,000 chains, bolded values are ESS estimates that are *not* significantly different (bootstrap hypothesis test with 1,000 samples at $\alpha = 0.05$ significance level) under path from the same initial distribution with highest ESS.

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

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *Int'l Conf. on Machine Learning*, pp. 6424–6429, 2008.

Skilling, J. Nested sampling for Bayesian computations. In *ISBA World Meeting on Bayesian Statistics*, 2006.

Sohl-Dickstein, J. and Culpepper, B. J. Hamiltonian annealed importance sampling for partition function estimation. Technical report, Redwood Center, UC Berkeley, 2012.

Taylor, G. and Hinton, G. Products of hidden markov models: It takes n¿1 to tango. In *Uncertainty in Artificial Intelligence*, 2009.

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

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Intl. Conf. on Machine Learning*, 2008.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

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

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Inf. Theory*, 51(7):2282–2312, 2005.

Appendix

Proof of Theorem 1 from Section 3

Theorem 1. *Suppose $K+1$ distributions p_k are linearly spaced along a path γ. Under the assumption of perfect transitions, if $\boldsymbol{\theta}(\beta)$ and the Fisher information matrix $\mathbf{G}_{\boldsymbol{\theta}} = \mathrm{cov}_{\mathbf{x}\sim p_{\boldsymbol{\theta}}}(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x}))$ are smooth, then as $K \to \infty$ the bias $\delta = \log \mathcal{Z}_b - \mathbb{E}[\log w^{(i)}]$ is determined by the functional:*

$$\begin{aligned} K\delta = K \sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(p_k \| p_{k+1}) &\to \mathcal{F}(\gamma) \\ &\equiv \frac{1}{2}\int_0^1 \dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta)\dot{\boldsymbol{\theta}}(\beta) d\beta, \qquad (21) \end{aligned}$$

where $\dot{\boldsymbol{\theta}}(\beta)$ represents the derivative of $\boldsymbol{\theta}$ with respect to β.

Proof. First, consider a second-order Taylor expansion of $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}(\beta)\|\boldsymbol{\theta}(\beta+h))$ around $h=0$. The constant and first order terms are zero. For the second order term,

$$\nabla^2_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}\|\boldsymbol{\theta}_0)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{G}_{\boldsymbol{\theta}},$$

so the second-order Taylor expansion is given by:

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}(\beta)\|\boldsymbol{\theta}(\beta+h)) = \frac{1}{2}h^2 \dot{\boldsymbol{\theta}}^T(\beta)\mathbf{G}_{\boldsymbol{\theta}}(\beta)\dot{\boldsymbol{\theta}}(\beta) + \epsilon,$$

where

$$|\epsilon| \le \frac{h^3}{6}\max_{\beta}\left|\frac{d^3}{dh^3}\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}(\beta)\|\boldsymbol{\theta}(\beta+h))\right|.$$

Assuming a linear schedule, the bias is given by

$$\begin{aligned}\delta &= \sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(p_k \| p_{k+1}) \\ &= \sum_{k=0}^{K-1} \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}(k/K) \parallel \boldsymbol{\theta}((k+1)/K)) \\ &= \frac{1}{2K^2} \sum_{k=0}^{K-1} \dot{\boldsymbol{\theta}}(\beta_k)^T \mathbf{G}_{\boldsymbol{\theta}(\beta_k)} \dot{\boldsymbol{\theta}}(\beta_k) + \sum_{k=0}^{K-1} \epsilon_k\end{aligned}$$

The second term decays like $1/K^2$, so it approaches zero even when scaled by K. The asymptotic bias, therefore, is determined by the first term. When scaled by K, this approaches

$$\mathcal{F}(\gamma) \equiv \frac{1}{2} \int_{\beta=0}^{1} \dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta) \dot{\boldsymbol{\theta}}(\beta) d\beta.$$

Therefore, $K\delta \to \mathcal{F}(\gamma)$. □

Derivation of variational interpretation from Section 4.1

Geometric averages

For simplicity of notation, assume the state space $\mathcal{X}$ is discrete. Consider solving for a distribution q to minimize the weighted sum of KL divergences

$$(1-\beta)\mathrm{D}_{\mathrm{KL}}(q \| p_0) + \beta \mathrm{D}_{\mathrm{KL}}(q \| p_1) \qquad (22)$$

with the constraint that $\sum_{\mathbf{x}} q(\mathbf{x}) = 1$. The Lagrangian is given by:

$$\begin{aligned}\mathcal{L}(q) &= \lambda \sum_{\mathbf{x}} q(\mathbf{x}) + (1-\beta) \sum_{\mathbf{x}} q(\mathbf{x}) \left(\log q(\mathbf{x}) - \log p_a(\mathbf{x})\right) \\ &\quad + \beta \sum_{\mathbf{x}} q(\mathbf{x}) \left(\log q(\mathbf{x}) - \log p_b(\mathbf{x})\right) \\ &= \sum_{\mathbf{x}} \lambda q(\mathbf{x}) + q(\mathbf{x}) \log q(\mathbf{x}) \\ &\quad - q(\mathbf{x}) \left[(1-\beta) \log p_a(\mathbf{x}) - \beta \log p_b(\mathbf{x})\right]\end{aligned}$$

Differentiating with respect to $q(\mathbf{x})$,

$$\frac{\partial \mathcal{L}(q)}{\partial q(\mathbf{x})} = \lambda + 1 + \log q(\mathbf{x}) - (1-\beta) \log p_a(\mathbf{x}) - \beta \log p_b(\mathbf{x}).$$

Setting this to zero gives:

$$q(\mathbf{x}) \propto p_a(\mathbf{x})^{1-\beta} p_b(\mathbf{x})^{\beta}.$$

This is the optimum over the probability simplex. If p_a and p_b belong to an exponential family $\mathcal{P}$, with natural parameters $\boldsymbol{\eta}_{p_a}$ and $\boldsymbol{\eta}_{p_b}$, the optimum is achieved within $\mathcal{P}$ using $\boldsymbol{\eta}_\beta = (1-\beta)\boldsymbol{\eta}_{p_a} + \beta \boldsymbol{\eta}_{p_b}$.

Moment averages

Suppose we wish to minimize

$$(1-\beta)\mathrm{D}_{\mathrm{KL}}(p_0 \| q) + \beta \mathrm{D}_{\mathrm{KL}}(p_1 \| q).$$

with respect to the natural parameters $\boldsymbol{\eta}$ of an exponential family distribution q. We expand the cost function to get

$$\begin{aligned}J(\boldsymbol{\eta}) &= (1-\beta) \sum_{\mathbf{x}} p_a(\mathbf{x})(\log p_a(\mathbf{x}) - \log q(\mathbf{x})) \\ &\quad + \beta \sum_{\mathbf{x}} p_b(\mathbf{x})(\log p_b(\mathbf{x}) - \log q(\mathbf{x})) \\ &= \text{const} - \sum_{\mathbf{x}} \left[(1-\beta) p_a(\mathbf{x}) + \beta p_b(\mathbf{x})\right] \log q(\mathbf{x}) \\ &= \text{const} + \log \mathcal{Z}(\boldsymbol{\eta}) \\ &\quad - \sum_{\mathbf{x}} \left[(1-\beta) p_a(\mathbf{x}) + \beta p_b(\mathbf{x})\right] \boldsymbol{\eta}^T \mathbf{g}(\mathbf{x})\end{aligned}$$

The partial derivatives are given by:

$$\begin{aligned}\frac{\partial J}{\partial \eta_i} &= \sum_{\mathbf{x}} q(\mathbf{x}) g_i(\mathbf{x}) - \sum_{\mathbf{x}} \left[(1-\beta) p_a(\mathbf{x}) + \beta p_b(\mathbf{x})\right] g_i(\mathbf{x}) \\ &= \mathbb{E}_q[g_i(\mathbf{x})] - (1-\beta)\mathbb{E}_{p_a}(g_i(\mathbf{x})) - \beta \mathbb{E}_{p_b}(g_i(\mathbf{x}))\end{aligned}$$

Setting this to zero, we see that the optimum solution is given by averaging the moments of p_a and p_b:

$$\mathbb{E}_q[g_i(\mathbf{x})] = (1-\beta)\mathbb{E}_{p_a}(g_i(\mathbf{x})) + \beta \mathbb{E}_{p_b}(g_i(\mathbf{x}))$$

Intuitively, this can be thought of as a maximum likelihood estimate of $\boldsymbol{\eta}$ for a dataset with $(1-\beta)$ fraction of the points drawn from p_a and β fraction drawn from p_b.

Analysis of Gaussian example in Section 4.2

Here we evaluate the cost functionals for the Gaussian example of Section 4.2 under γ_{GA} and γ_{MA} using both linear and optimal schedules. Recall that $p_a = \mathcal{N}(\mu_0, \sigma)$ and $p_b = \mathcal{N}(\mu_1, \sigma)$. The natural parameters of the Gaussian are the information form representation, with precision $\lambda = 1/\sigma^2$ and potential $h = \lambda\mu$. The sufficient statistics are the first and (rescaled) second moments given by $\mathbb{E}[x] = \mu$ and $-\frac{1}{2}\mathbb{E}[x^2] = -\frac{1}{2}s \equiv -\frac{1}{2}(\sigma^2 + \mu^2)$.

To simplify calculations, let β range from $-1/2$ to $1/2$ (rather than 0 to 1), and assume $\mu_0 = -1/2$ and $\mu_1 = 1/2$. The general case can be obtained by rescaling μ_0, μ_1, and σ.

Geometric averages

Geometric averages correspond to averaging the natural parameters:

$$\begin{aligned}\lambda(\beta) &= 1/\sigma^2\\ h(\beta) &= \beta/\sigma^2\end{aligned}$$

Solving for the moments,

$$\begin{aligned}\mu(\beta) &= \beta\\ s(\beta) &= \sigma^2 + \beta^2.\end{aligned}$$

The derivatives are given by:

$$\begin{aligned}\dot\lambda(\beta) &= 0\\ \dot h(\beta) &= 1/\sigma^2\\ \dot\mu(\beta) &= 1\\ \dot s(\beta) &= t\end{aligned}$$

Ignoring the constant, the cost functional is given by:

$$\begin{aligned}\mathcal{F}(\gamma) &= \frac{1}{2}\int_{-1/2}^{1/2} \dot h(\beta)\dot\mu(\beta) - \frac{1}{2}\dot\lambda(\beta)\dot s(\beta) d\beta\\ &= \frac{1}{2}\int_{-1/2}^{1/2}\frac{1}{\sigma^2}d\beta\\ &= \frac{1}{2\sigma^2}.\end{aligned}$$

We can also compute the cost under the optimal schedule by computing the path length (see Section 3):

$$\begin{aligned}\ell(\gamma) &= \int_{-1/2}^{1/2}\sqrt{\dot h(\beta)\dot\mu(\beta) - \frac{1}{2}\dot\lambda(\beta)\dot s(\beta)d\beta}\\ &= \int_{-1/2}^{1/2}\sqrt{1/\sigma^2}d\beta\\ &= \frac{1}{\sigma}.\end{aligned}$$

Since the functional under the optimal schedule is given by $\ell^2/2$, these two answers agree with each other, i.e. the linear schedule is optimal.

We assumed for simplicity that $\mu_0 = -1/2$ and $\mu_1 = 1/2$. In general, we can rescale σ and $\mu_1 - \mu_0$ by the same amount without changing the functional. Therefore, $\mathcal{F}(\gamma_{GA})$ is given by:

$$\frac{(\mu_1-\mu_0)^2}{2\sigma^2} \equiv \frac{d^2}{2}.$$

Moment averaging

Now let's look at moment averaging. The parameterizations are given by:

$$\begin{aligned}\mu(\beta) &= \beta\\ s(\beta) &= \sigma^2 + \frac{1}{4}\\ \lambda(\beta) &= \left(\sigma^2 + \frac{1}{4} - \beta^2\right)^{-1}\\ h(\beta) &= \left(\sigma^2 + \frac{1}{4} - \beta^2\right)^{-1}\beta\end{aligned}$$

with derivatives

$$\begin{aligned}\dot\mu(\beta) &= 1\\ \dot s(\beta) &= 0\\ \dot\lambda(\beta) &= 2\left(\sigma^2 + \frac{1}{4} - \beta^2\right)^{-2}\beta\\ \dot h(\beta) &= \lambda(\beta)\dot\mu(\beta) + \mu(\beta)\dot\lambda(\beta)\\ &= \left(\sigma^2 + \frac{1}{4} - \beta^2\right)^{-1} + 2\left(\sigma^2 + \frac{1}{4} - \beta^2\right)^{-2}\beta^2\end{aligned}$$

The cost functional is given by:

$$\begin{aligned}\mathcal{F}(\gamma_{MA}) &= \frac{1}{2}\int_{-1/2}^{1/2}\dot\mu(\beta)\dot h(\beta) - \frac{1}{2}\dot s(\beta)\dot\lambda(\beta)\\ &= \frac{1}{2}\int_{-1/2}^{1/2}\dot h(\beta)d\beta\\ &= \frac{1}{2}[h(1/2) - h(-1/2)]\\ &= \frac{1}{2\sigma^2}.\end{aligned}$$

This agrees exactly with $\mathcal{F}(\gamma_{GA})$, consistent with Theorem 2.

However, we can see by inspection that for small σ, most of the mass of this integral is concentrated near the endpoints, where the variance changes suddenly. This suggests that the optimal schedule would place more intermediate distributions near the endpoints.

We can bound the cost under the optimal schedule by bounding the path length $\ell(\gamma_{MA})$:

$$\begin{aligned}
\ell(\gamma_{MA}) &= \int_{-1/2}^{1/2} \sqrt{\dot{\mu}(\beta)\dot{h}(\beta) - \frac{1}{2}\dot{s}(\beta)\dot{\lambda}(\beta)} d\beta \\
&= \int_{-1/2}^{1/2} \sqrt{\dot{h}(\beta)} d\beta \\
&= \int_{-1/2}^{1/2} \sqrt{\lambda(\beta)\dot{\mu}(\beta) + \mu(\beta)\dot{\lambda}(\beta)} d\beta \\
&\leq \int_{-1/2}^{1/2} \sqrt{|\lambda(\beta)\dot{\mu}(\beta)|} d\beta + \int_{-1/2}^{1/2} \sqrt{|\mu(\beta)\dot{\lambda}(\beta)|} d\beta \\
&= \int_{-1/2}^{1/2} \frac{1}{\sqrt{\sigma^2 + \frac{1}{4} - \beta^2}} d\beta + \sqrt{2} \int_{-1/2}^{1/2} \frac{|\beta|}{\sigma^2 + \frac{1}{4} - \beta^2} d\beta \\
&= 2 \sin^{-1} \left(\frac{1}{\sqrt{4\sigma^2 + 1}} \right) + \sqrt{2} \log \left(1 + \frac{1}{4\sigma^2} \right) \\
&\leq \pi + \sqrt{2} \log \left(1 + \frac{1}{4\sigma^2} \right)
\end{aligned}$$

The path length has dropped from linear to logarithmic! Since $\mathcal{F}$ grows like ℓ^2, the cost drops from quadratic to log squared.

This shows that even though Theorem 2 guarantees that both γ_{GA} and γ_{MA} have the same functional under a linear schedule, one path may do substantially better than the other if one is allowed to change the schedule.