Understanding Thermodynamic Variational Inference

Rob Brekelmans, Aram Galstyan, Greg Ver Steeg
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
brekelma@usc.edu; galstyan, gregv@isi.edu

Abstract

Intriguing recent work by Masrani et al. [15] utilizes thermodynamic integration, a physics-inspired technique resembling annealed importance sampling (AIS), within the context of variational inference. We interpret the Thermodynamic Variational Objective (TVO) using the conjugate duality of exponential families. This perspective allows us express the gap in TVO likelihood bounds as a sum of KL divergences along a geometric mixture curve, which corresponds to the asymptotic bias in AIS estimation. The Renyi variational inference objective also appears naturally within our framework. Finally, we draw on the AIS literature to explore strategies for choosing intermediate distributions in the TVO objective.

1 Introduction

Variational inference has become an instrumental tool in modern machine learning [3, 11, 17, 20], framing maximum likelihood parameter estimation in the presence of latent variables as a tractable optimization problem. Frequently, this involves subtracting a divergence measure from the marginal likelihood $p(x) = \int p(x,z)dz$ and maximizing the resulting lower bound. Previous work has considered the KL divergence [11, 17], χ^2-divergence [6], f-divergences [21, 22], and Renyi (α-) divergences [10, 14, 13, 18] in this setting. The variational objective might be chosen to obtain tighter bounds on likelihood [5, 15], more accurately estimate posterior variance [6], or control the mass-covering behavior of the approximating distribution [13, 21].

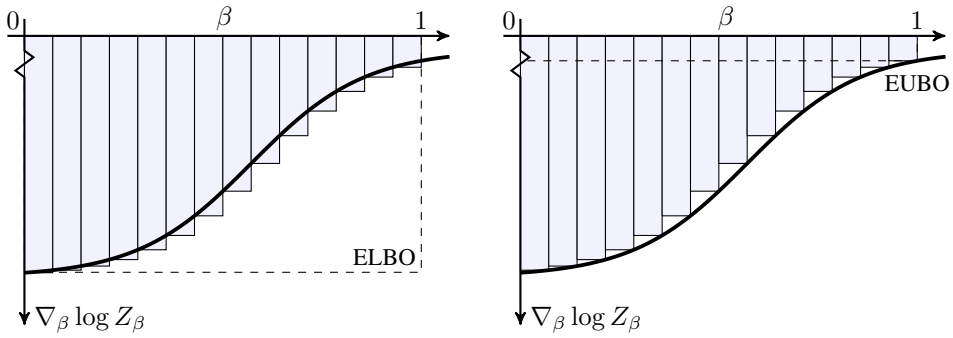
In this work, we interpret the thermodynamic variational inference framework of Masrani et al. [15] using the conjugate duality of exponential families. While Masrani et al. [15] show that their objective gives tighter lower bounds on log-likelihood than the familiar Evidence Lower Bound (ELBO), we explicitly characterize the gap in the TVO bound as a sum of KL divergences. This perspective also leads to connections with annealed importance sampling, which we use to explore 'schedules' for choosing partitions in the TVO objective.

2 Thermodynamic Variational Inference

Thermodynamic integration (TI) frames estimating the partition function or marginal likelihood $\log p(x)$ as a one-dimensional integration problem over a parameter $\beta \in [0,1]$. This parameter maps out a path of geometric mixtures between a base distribution, in our case $q(z|x)$, and an (unnormalized) target distribution $p(x,z)$ [7]. We introduce the Thermodynamic Variational Objective (TVO) directly in terms of our proposed framework, which yields several important quantities in Masrani et al. [15] from familiar properties of exponential families.

Workshop on Information Theory and Machine Learning, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Figure 1: Mean Parameters: $\nabla_\beta \log Z_\beta = \mathbb{E}_{\pi_\beta} \phi(\mathbf{x}, \mathbf{z})$



(a) Lower Bound using left-Riemann Sum. ELBO = $\log p(x) - D_{KL}[q(z|x)||p(z|x)]$ corresponds to estimator with a single $\beta = 0$.

(b) Upper Bound using right-Riemann sum. EUBO = $\log p(x) + D_{KL}[p(z|x)||q(z|x)]$ corresponds to estimator with a single $\beta = 1$.

2.1 Exponential Family Interpretation

To mirror the TVO setting, we can consider an exponential family of distributions $\pi_\beta(z|x)$ with natural parameters β, sufficient statistics $\phi(x, z) = \log p(x, z)/q(z|x)$,and a log-partition function $\psi(\beta) = \log Z_\beta(x)$ that integrates over z. We also include a base measure of $q(z|x)$. [1]

$$\pi_\beta(z|x) = q(z|x) \exp\{\beta\, \phi(x, z) - \psi(\beta)\} \qquad \text{where: } \phi(x, z) \triangleq \log \frac{p(x, z)}{q(z|x)} \tag{1}$$

$$\psi(\beta) \triangleq \log Z_\beta = \log \int \frac{p(x, z)^\beta}{q(z|x)^\beta} dq(z|x) = \log \int q(z|x)^{1-\beta} p(x, z)^\beta dz \tag{2}$$

The insight of TI is that we can express $\log p(x)$ as an integral of the gradient $\nabla_\beta \log Z_\beta$ over $0 \le \beta \le 1$. This corresponds to simply evaluating $\log Z_\beta$ at the endpoints $\pi_0 = q(z|x)$, with $\log Z_0 = 0$, and $\pi_1 = p(z|x) \propto p(x, z)$, with $\log Z_1 = \log p(x)$:

$$\int_0^1 \nabla_\beta \log Z_\beta \, d\beta = \log Z_1 - \log Z_0 = \log p(x) \tag{3}$$

It is well known that the log-partition function is convex, with the first (partial) derivative $\nabla_\beta \log Z_\beta$ equal to the expectation of the sufficient statistics. In our setting, this corresponds to the expected log-importance weights under π_β, which we refer to as the mean parameter η_β [1, 20]. To obtain a numerical estimate for the integral, Masrani et al. [15] give an efficient self-normalized importance sampling scheme for evaluating expectations at any β:

$$\nabla_\beta \log Z_\beta = \mathbb{E}_{\pi_\beta} \log \frac{p(x, z)}{q(z|x)} \triangleq \eta_\beta \tag{4}$$

$$= \mathbb{E}_{q(z|x)} \frac{q(z|x)^{1-\beta} p(x, z)^\beta}{q(z|x)\, Z_\beta} \log \frac{p(x, z)}{q(z|x)}$$

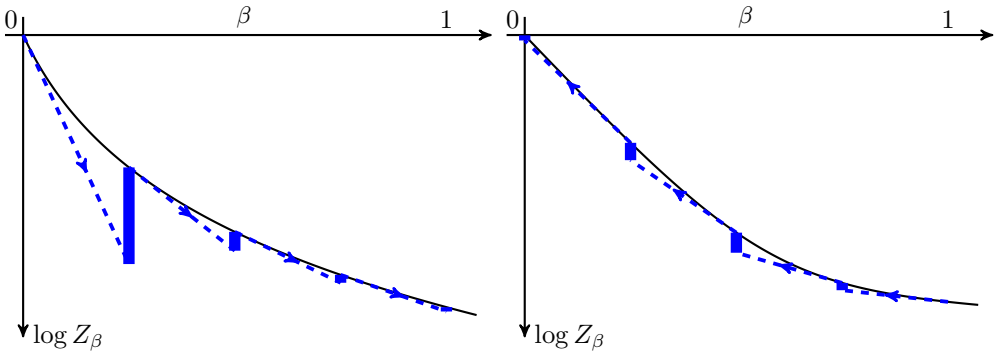
$$\approx \sum_{i=1}^{S} \frac{w_i}{\sum_i w_i} \log \frac{p(x, z_i)}{q(z_i|x)} \qquad \text{where } w_i = \frac{p(x, z_i)^\beta}{q(z_i|x)^\beta} \tag{5}$$

Notably, these weights allow the reuse of importance samples for different β by simply rescaling. Further, $\nabla_\beta \log Z_\beta$ is increasing as a function of β, since we know that $\psi(\beta)$ is convex. Thus, the left and right Riemann sums will form lower and upper bounds on the integral, respectively (see Fig. 1).

$$\mathcal{L}_{TVO_L} := \sum_{k=1}^{K} \Delta_{\beta_k} \mathbb{E}_{\pi_{\beta_{k-1}}} \log \frac{p(x, z)}{q(z|x)} \qquad \mathcal{L}_{TVO_U} := \sum_{k=1}^{K} \Delta_{\beta_k} \mathbb{E}_{\pi_{\beta_k}} \log \frac{p(x, z)}{q(z|x)} \tag{6}$$

[1] Note, the parameter β should more formally be seen as defining a mixture of the natural parameters between π_0 and π_1. We somewhat abuse notation to match the setting of [15]. See App. B for more detailed discussion.

Figure 3: Partition Function $\psi(\beta) = \log Z_\beta$



(a) Gap in $\mathcal{L}_{TVO_L}$
(Lower Bound on $\log p(x)$)
$\sum_{k=1}^{K} D_\psi[\beta_k : \beta_{k-1}]$

(b) Gap in $\mathcal{L}_{TVO_U}$
(Upper Bound on $\log p(x)$)
$\sum_{k=1}^{K} D_\psi[\beta_{k-1} : \beta_k]$

We can note that the familiar Evidence Lower Bound (ELBO) corresponds to $\nabla_\beta \psi(\beta)|_{\beta=0} = \log p(x) - D_{KL}[q(z|x)||p(z|x)]$, or the left Riemann lower bound evaluated using a single $\beta = 0$. Similarly, Masrani et al. [15] denote the right endpoint as the Evidence Upper Bound (EUBO), with $\nabla_\beta \psi(\beta)|_{\beta=1} = \log p(x) + D_{KL}[p(z|x)||q(z|x)]$.

While Masrani et al. [15] derive a reparameterization-free estimate for the gradient of each expectation with respect to the model parameters, the existence of the ELBO as a special case suggests that the reparameterization trick might also used to optimize the TVO (see App. D, [13]). Finally, we note that each intermediate partition function $\log Z_\beta$ corresponds to a scaled version of the Renyi VI objective $\mathcal{L}_\alpha$ [13], with $\log Z_\beta = \beta\, \mathcal{L}_{1-\beta}$ (see App. C).

2.2 TVO Using Bregman Divergences

While the TVO is presented as a numerical integration in mean parameter space (Fig. 1), we now give an alternative interpretation in terms of Bregman divergences derived from the log partition function $\psi(\beta)$. We write the Bregman divergence D_ψ between distributions indexed by natural parameters β and β' as:

$$D_\psi[\beta : \beta'] = \psi(\beta) - \psi(\beta') - \langle \beta - \beta', \nabla_\beta \psi(\beta') \rangle$$

Geometrically, the divergence can be viewed as the difference between $\psi(\beta)$ and its linear approximation around β' in Fig. 3. While higher order Taylor series might be considered (see App. B.2), the convexity of the partition function ensures that the first order approximation will underestimate $\psi(\beta)$, yielding a nonnegative divergence.

As shown in App. A, we can leverage convex duality to obtain an alternative divergence in terms of the conjugate function $\psi^*(\eta)$ and the mean parameters η. These divergences are equivalent with the order of arguments reversed. Further, for exponential family models, ψ and ψ^* both induce the KL divergence as their Bregman divergence:

$$D_\psi[\beta : \beta'] = D_{KL}[\pi_{\beta'} || \pi_\beta] \qquad D_{\psi^*}[\eta : \eta'] = D_\psi[\beta' : \beta] = D_{KL}[\pi_\beta || \pi_{\beta'}] \tag{7}$$

These divergences give a complementary perspective on the TVO objective using the graph of the log partition function $\psi(\beta) = \log Z_\beta$ in Fig. 3. Considering a discrete partition $\gamma(\beta) = \{\beta_0, ...\beta_K\}$ with $\beta_0 = 0$, $\beta_K = 1$, we can write the gap in the left-Riemann TVO lower bound using the Bregman divergence $D_\psi[\beta_k : \beta_{k-1}]$.

$$\begin{aligned} \sum_{k=1}^{K} D_\psi[\beta_k : \beta_{k-1}] &= \log Z_1 - \log Z_0 - \sum_{k=1}^{K} (\beta_k - \beta_{k-1}) \nabla_\beta \psi(\beta_{k-1}) \\ &= \log p(x) - \sum_{k=1}^{K} \Delta_{\beta_k} \mathbb{E}_{\pi_{\beta_{k-1}}} \log \frac{p(x,z)}{q(z|x)} \end{aligned} \tag{8}$$

where intermediate log partition terms cancel due to the telescoping sum, and the second term corresponds with $\mathcal{L}_{TVO_L}$ in (6). While Masrani et al. [15] show only that $\mathcal{L}_{TVO_L}$ minimizes a quantity that is non-negative and vanishes at $q(z|x) = p(z|x)$, we can explicitly characterize the gap in the lower bound as a sum of Bregman divergences using (8), or KL divergences via (7).

$$\log p(x) - \mathcal{L}_{TVO_L} = \sum_{k=1}^{K} D_\psi[\beta_k : \beta_{k-1}] = \sum_{k=1}^{K} D_{KL}[\pi_{\beta_{k-1}} || \pi_{\beta_k}] \tag{9}$$

The dual divergence, using the KL divergence in the reverse direction, similarly characterizes the gap in the right-Riemann upper bound (see App. A.1):

$$\mathcal{L}_{TVO_U} - \log p(x) = \sum_{k=1}^{K} D_\psi[\beta_{k-1} : \beta_k] = \sum_{k=1}^{K} D_{KL}[\pi_{\beta_k} || \pi_{\beta_{k-1}}] \tag{10}$$

3 Annealing Paths

We can immediately recognize the sum of KL divergences for the lower bound gap (9) as the bias in an annealed importance sampling estimator of $\log p(x)$ under perfect transitions [9, 16]. This connection with the gap in the TVO lower bound is novel and allows application of several results from Grosse et al. [9].

In the limit of infinitesimal transitions along a geometric path with linear spacing, the scaled bias $K \sum D_{KL}[\pi_{\beta_{k-1}} || \pi_{\beta_k}]$ can be shown to approach the symmetric KL divergence between the endpoints (Thm 1 [9], with detailed derivations in App. B).

$$K \sum_{k=1}^{K} D_{KL}[\pi_{\beta_{k-1}} || \pi_{\beta_k}] \to \mathcal{F}(\gamma) = \frac{1}{2}\big(D_{KL}[\pi_0 || \pi_1] + D_{KL}[\pi_1 || \pi_0]\big) \tag{11}$$

$$= \frac{1}{2}(\beta_1 - \beta_0)(\eta_1 - \eta_0) \tag{12}$$

As in Theorem 3 of Grosse et al. [9], we can use this insight to derive an optimal linear binning schedule that minimizes the bias $\mathcal{F}$. We also consider the moment-averaging path [9] for determining the values of β at which to optimize the TVO objective.

Recursive Schedule: Imagine a coarse partition of the interval $[0, 1]$ using knots t_j, $0 \le j \le J$. We then seek to allocate a budget of K intermediate distributions across the segments with K_j uniformly-spaced points within the subinterval $T_j = [t_j, t_{j+1}]$. We can then use (12) to allocate K_j according to the contribution to the bias within each segment, $F_j = \frac{1}{2}(\beta_{t_{j+1}} - \beta_{t_j})(\eta_{t_{j+1}} - \eta_{t_j})$. Minimizing $\sum_j F_j / K_j$ subject to $\sum_j K_j = 1$, we obtain:

$$K_j \propto \sqrt{(\beta_{t_{j+1}} - \beta_{t_j})(\eta_{t_{j+1}} - \eta_{t_j})} \tag{13}$$

where β_{t_j} and η_{t_j} indicate the natural parameters and (estimated) moment parameters at a point t_j along the path. With given knot points, Equation (13) can thus be used to adaptively choose intermediate distributions across training.

To avoid specifying the knot points, we can also recursively allocate intermediate distributions based on evaluations of (13). After each epoch, we allocate K total points across $T_0 = (0, 0.5], T_1 = (0.5, 1.0]$, with the resulting 'left' budget K_0 split among $(0, 0.25]$, $(0.25, 0.5]$, and so on.

Moments Schedule: Since intermediate distributions can be described by either β_k or η_k, we may also use the mean parameters to construct a partition of the interval $0 \le \beta \le 1$. For a budget of K intermediate distributions, we find values of β_k which give equal spacing in the mean parameter space, so that $\eta_{\beta_k} = k/K \cdot \eta_0 + (1 - k/K) \cdot \eta_1 = k/K \cdot \text{ELBO} + (1 - k/K) \cdot \text{EUBO}$. The Legendre transform, which finds β_k corresponding to a given η_{β_k} is difficult in general, but is easily approximated here using a binary-search procedure and cheap evaluations of η_β according to (5).

Note that this schedule does not imply linear spacing in natural parameter space, so that mixing weights $(\beta_k - \beta_{k-1})$ will be non-uniform for the TVO objectives in (6).

Figure 5: Likelihood Estimates

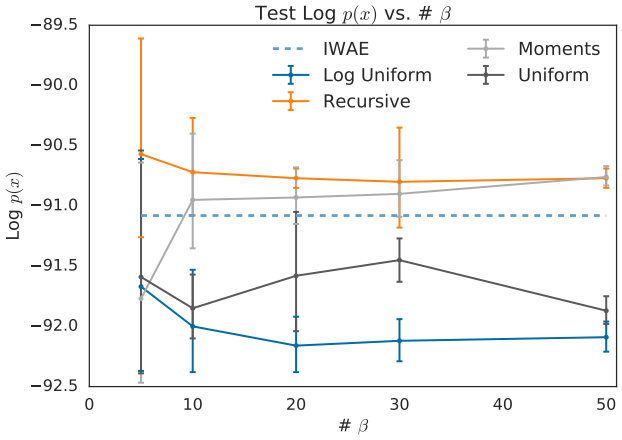
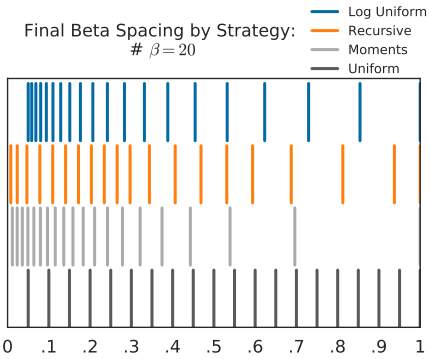


Figure 6: Intermediate Distributions:



4 Results

We compare our recursive allocation scheme and moment-averaged path from Sec. 3 against uniform linear spacing and log-uniform spacing, with the first β fixed to 0.05, in Figure 5. Using the same architecture as in [15], we train on Binary MNIST for 1000 epochs using Adam with learning rate 0.001, and estimate $\log p(x)$ on the test set using the IWAE bound [5] with 5000 samples.

We find our adaptive schemes lead to slight improvements over uniform and log-uniform methods, with the recursive linear schedule appearing to perform best. Error bars show the TVO upper and lower bounds as estimated on the test set, and we confirm that the bounds tighten as the number of partitions grows and the difference between the left- and right-Riemann evaluations decreases.

In Figure 6, we show the β values chosen by our adaptive methods at the end of training. We observed the intermediate distribution to be fairly stable after the initial training epochs. Note that equal spacing in the moment-averaged path leads to a high concentration at low β. This reflects the observation in Masrani et al. [15] that the expected sufficient statistics change quickly at low β before flattening out, requiring careful attention in approximating the integral in this region.

Finally, as discussed in Masrani et al. [15], we find that increasing the number of intermediate β need not improve performance. Although more partitions should lead to a tighter likelihood bound, we expect the bias in our self-normalized importance sampling scheme will increase as we move away from $\beta = 0$. Importance sampling schemes which reduce this bias [8] might eventually allow the TVO objective to enjoy the full benefit of finer partitions and principled schedules.

5 Conclusions

We have presented an interpretation of the Thermodynamic Variational Objective [15] in terms of a one-dimensional exponential family, allowing us to characterize the gap in the TVO bounds and make explicit connections with annealed sampling methods [7, 16, 9] and the information geometry of exponential families [1, 2]. This perspective should open new avenues for analysis of thermodynamic variational inference.

References

[1] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. *Iclr-2015*, pages 1–12, 2015. URL `http://arxiv.org/abs/1509.00519`.

[6] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via $\backslash\chi$ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

[7] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[8] Adam Goliński, Frank Wood, and Tom Rainforth. Amortized monte carlo integration. *International Conference on Machine Learning*, 2019.

[9] Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, pages 2769–2777, 2013.

[10] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard Turner. Black-box α-divergence minimization. 2016.

[11] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. (Ml):1–14, 2013. ISSN 1312.6114v10. URL `http://arxiv.org/abs/1312.6114`.

[12] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[13] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[14] Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. Stochastic expectation propagation. In *Advances in neural information processing systems*, pages 2323–2331, 2015.

[15] Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. *arXiv preprint arXiv:1907.00031*, 2019.

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

[17] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[18] Abhijoy Saha, Karthik Bharath, and Sebastian Kurtek. A geometric variational approach to bayesian inference. *Journal of the American Statistical Association*, pages 1–25, 2019.

[19] Tim Van Erven and Peter Harremos. Renyi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[20] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[21] Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pages 5737–5747, 2018.

[22] Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, and David Barber. Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*, 2019.

A Conjugate Duality

A Bregman divergence associated with a convex function $f : \Omega \to \mathbb{R}$ can be written as [2, 1]:

$$D_{B_f}[p : q] = f(p) + f(q) - \langle p - q, \nabla f(q) \rangle$$

The family of Bregman divergences includes many familiar quantities, including the KL divergence corresponding to the negative entropy $f(p) = -\int p \log p \, d\omega$. Geometrically, the divergence can be viewed as the difference between $f(p)$ and its linear approximation around q. Since f is convex, we know that a first order estimator will lie below the function, yielding $D_f[p : q] \geq 0$. For our purposes, we can let $f \triangleq \psi(\beta) = \log Z_\beta$ over the domain of probability distributions indexed by natural parameters of the exponential family in (1):

$$D_\psi[\beta_p : \beta_q] = \psi(\beta_p) - \psi(\beta_q) - \langle \beta_p - \beta_q, \nabla_\beta \psi(\beta_q) \rangle$$

This is a common setting in the field of information geometry [1], which introduces dually flat manifold structures based on the natural parameters and the mean parameters. In particular, we can leverage convex duality to derive an alternative divergence based on the conjugate function ψ^*:

$$\begin{aligned} \psi^*(\eta) &= \sup_\beta \, \eta \cdot \beta - \psi(\beta) \qquad \implies \qquad \eta = \nabla_\beta \, \psi(\beta) \\ &= \eta \cdot \beta_\eta - \psi(\beta_\eta) \end{aligned} \tag{14}$$

The conjugate measures the maximum distance between the line $\eta \cdot \beta$ and the function $\psi(\beta)$, which occurs at the unique point β_η where $\eta = \nabla_\beta \psi(\beta)$. This yields a one-to-one correspondence between η and β for minimal exponential families [20]. Thus, a distribution p may be indexed by either its mean parameters η_p or natural parameters β_p.

Noting that $\psi^{**} = \psi(\beta) = \sup_\eta \, \eta \cdot \beta - \psi^*(\eta)$ [4], we can use a similar argument as above to write this correspondence as $\beta = \nabla_\eta \psi^*(\eta)$. We can then write the dual divergence D_{ψ^*} as:

$$\begin{aligned} D_{\psi^*}[\eta_p : \eta_q] &= \psi^*(\eta_p) - \psi^*(\eta_q) - \langle \eta_p - \eta_q, \nabla_\eta \, \psi^*(\eta_q) \rangle \\ &= \psi^*(\eta_p) - \underline{\psi^*(\eta_q)} - \eta_p \cdot \beta_q + \underline{\eta_q \cdot \beta_q} \\ &= \psi^*(\eta_p) + \underline{\psi(\beta_q)} - \eta_p \cdot \beta_q \end{aligned} \tag{15}$$

Similarly,

$$D_\psi[\beta_p : \beta_q] = \psi(\beta_p) - \psi(\beta_q) - \langle \beta_p - \beta_q, \nabla_\beta \psi(\beta_q) \rangle \tag{16}$$

$$\begin{aligned} &= \psi(\beta_p) - \underline{\psi(\beta_q)} - \beta_p \cdot \eta_q + \underline{\beta_q \cdot \eta_q} \\ &= \psi(\beta_p) + \underline{\psi^*(\eta_q)} - \beta_p \cdot \eta_q \end{aligned} \tag{17}$$

Comparing (15) and (17), we see that the divergences are equivalent with the arguments reversed, so:

$$D_\psi[\beta_p : \beta_q] = D_{\psi^*}[\eta_q : \eta_p] \tag{18}$$

For an exponential family with partition function $\psi(\beta)$ and sufficient statistics ϕ, we can make further statements about $\psi^*(\eta)$ and the induced divergence measures.

In particular, note that, for a distribution p indexed by β_p and η_p, we can write $\log p(x) = \beta_p \cdot \phi(x) - \psi(\beta)$. Then, (14) becomes:

$$\begin{aligned} \psi^*(\eta_p) &= \eta_p \cdot \beta_p - \psi(\beta_p) \\ &= \mathbb{E}_p[\phi(x) \cdot \beta_p] - \psi(\beta_p) \\ &= \mathbb{E}_p \log p(x) \\ &= -H_p(X) \end{aligned}$$

since $\psi(\beta_p)$ is constant with respect to x. The dual divergence with q then becomes:

$$\begin{aligned} D_{\psi^*}[\eta_p : \eta_q] &= \psi^*(\eta_p) - \psi^*(\eta_q) - \langle \eta_p - \eta_q, \nabla_\eta \, \psi^*(\eta_q) \rangle \\ &= \mathbb{E}_p \log p(x) - \underline{\psi^*(\eta_q)} - \eta_p \cdot \beta_q + \underline{\eta_q \cdot \beta_q} \\ &= \mathbb{E}_p \log p(x) - \eta_p \cdot \beta_q + \underline{\psi(\beta_q)} \\ &= \mathbb{E}_p \log p(x) - \mathbb{E}_p[\phi(x) \cdot \beta_q] + \psi(\beta_q) \\ &= \mathbb{E}_p \log p(x) - \mathbb{E}_p \log q(x) \\ &= D_{KL}[p(x) || q(x)] \end{aligned}$$

Thus, the conjugate function is the negative entropy and induces the KL divergence as its Bregman divergence. Using (18), we see that the $\psi(\beta)$ will yield the reverse KL by a similar set of arguments. These results are well known from [1, 20], for example, but help to unify a geometric analysis of the gap in the TVO with characterizations of the bias in annealed importance sampling estimates [9] (see Sec. 2.1. More generally, this correspondence with the information geometry of exponential families provides justification for use of the KL divergence in these models, which is often assumed a priori (see Ch. 3 and 6 of [1]).

A.1 TVO Upper Bound Gap

Similarly to (8), we can use the KL divergence in the other direction $D_{KL}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$ to obtain an upper bound on $\log p(x)$:

$$\begin{aligned}\sum_{k=1}^{K} D_\psi[\beta_{k-1} : \beta_k] &= \log Z_0 - \log Z_1 - \sum_{k=1}^{K}(\beta_{k-1} - \beta_k)\nabla_\beta \psi(\beta_k) \\ &= \sum_{k=1}^{K} \Delta_{\beta_k} \mathbb{E}_{\pi_{\beta_k}} \log \frac{p(x,z)}{q(z|x)} - \log p(x) \end{aligned} \tag{19}$$

This expression could be equivalently written using the dual divergence $D_{\psi^*}[\eta_k : \eta_{k-1}]$, and corresponds to the amount by which the right-Riemann sum $\mathcal{L}_{TVO_U}$ over-estimates $\log p(x)$.

$$\mathcal{L}_{TVO_U} - \log p(x) = \sum_{k=1}^{K} D_\psi[\beta_{k-1} : \beta_k] = \sum_{k=1}^{K} D_{KL}[\pi_{\beta_k}||\pi_{\beta_{k-1}}] \tag{20}$$

B Annealing Paths and Fisher Information

B.1 Exponential and Mixture Geodesics

In Sec. 2.1, we mention our slight abuse of notation in using β to denote the natural parameters of the exponential family. More precisely, we should consider natural parameters θ, with $\theta = 0$ corresponding to $\pi_0 = q(z|x)$ and $\theta = 1$ with $\pi_1 = p(z|x)$. Then, $\beta \in [0, 1]$ can be used to define the exponential mixture path between endpoints θ_a and θ_b, so that $\theta(\beta) = (1 - \beta) \cdot \theta_a + \beta \cdot \theta_b$. This corresponds to an additive mixture of the natural parameters or a geometric mixture of the probability densities, and is known as the e-geodesic in information geometry [1]. For the TVO setting, we have endpoints $\theta_a = 0$ and $\theta_b = 1$, so that $\theta(\beta) = \beta$. We have used this directly for notational ease, but our results for $\theta(\beta) = \beta$ naturally translate to geodesic paths $\theta(\beta)$ between arbitrary endpoints θ_a, θ_b.

Grosse et al. [9] also consider AIS paths using additive mixtures of the mean parameters η, which they denote the moment-averaged path $\gamma_{MA}(\eta)$. This also known as the m-geodesic, a dual connection between base and target distribution.

B.2 Annealing Path Bias

The bias in an AIS estimation of $\log p(x)$ along a geometric mixture path $\gamma_{GA}(\beta)$ can easily be seen to be $\sum D_{KL}[\pi_{\beta_{k-1}} : \pi_{\beta_k}]$ by expanding the expected log importance weights [9, 16].

In the limit of infinitesimal transitions with linear spacing, the bias can be shown to approach the symmetric KL divergence between the endpoints $\theta_0 = 0$ and $\theta_K = 1$ ([9], Thm 1). Letting β denote the mixing path between base and target distributions, so that $\theta(\beta) = (1 - \beta) \cdot \theta_0 + t \cdot \theta_K$, and $G_\theta(\beta)$ be the Fisher information matrix evaluated at $\theta(\beta)$,

$$K \sum_{k=1}^{K} D_{KL}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] \to \mathcal{F}(\gamma) = \frac{1}{2} \int_0^1 \dot{\theta}(\beta)^T G_\theta(\beta) \dot{\theta}(\beta) d\beta \tag{21}$$

$$= \frac{1}{2}\big(D_{KL}[\pi_0||\pi_1] + D_{KL}[\pi_1||\pi_0]\big) \tag{22}$$

Here, $\dot{\theta}(\beta)$ indicates the derivative of the natural parameters w.r.t. β, which is a constant w.r.t. β: $\dot{\theta}(\beta) = \theta_K - \theta_0$ and equals 1 in the case of TVO. The proof proceeds by taking the Taylor

approximation of the KL divergence $D_{KL}[\beta_k||\beta_k + \Delta_\beta]$ around β_k for small Δ_β, where (21) corresponds the second order term [9, 12].

The expression in (21) then corresponds to the integral of the Fisher information along either the e- or m- geodesic paths (Theorem 3.2 of Amari [1]), and is equal to the symmetric KL divergence. Equivalently, Theorem 2 of Grosse et al. [9] shows that both the geometric- and moment-averaged paths yield the same asymptotic bias, matching (22):

$$\begin{aligned}
\mathcal{F}(\gamma_{GA}) = \mathcal{F}(\gamma_{MA}) &= \frac{1}{2}(\eta_1 - \eta_0)^T(\beta_1 - \beta_0) && (23)\\
&= \frac{1}{2}\big(\mathbb{E}_{\pi_{\beta_1}}[\phi_1 \cdot \beta_1] - \mathbb{E}_{\pi_{\beta_1}}[\phi_1 \cdot \beta_0] + -\mathbb{E}_{\pi_{\beta_0}}[\phi_0 \cdot \beta_1] + \mathbb{E}_{\pi_{\beta_0}}[\phi_0 \cdot \beta_0]\big) \\
&= \frac{1}{2}\big(\mathbb{E}_{\pi_{\beta_1}} \log \frac{\pi_{\beta_1}}{\pi_{\beta_0}} + \psi(1) - \psi(0) + \mathbb{E}_{\pi_{\beta_0}} \log \frac{\pi_{\beta_0}}{\pi_{\beta_1}} - \psi(1) + \psi(0)\big) && (24)\\
&= \frac{1}{2}\big(D_{KL}[\pi_0||\pi_1] + D_{KL}[\pi_1||\pi_0]\big) && (25)
\end{aligned}$$

where we have added and subtracted the partition functions in the second to last line to normalize the log probabilities.

C Connections with Renyi Divergence Variational Inference

C.1 Renyi Divergences as Partition Functions

We proceed to show that the partition function $\psi(\beta) = \log Z_\beta$ of this exponential family is related to a scaled Renyi divergence for intermediate β. Using the definition of [19], the Renyi divergence of order α is defined as:

$$D_\alpha[p||q] = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\omega$$

for distributions p and q over some measure $d\omega$. We can immediately notice the similarity of this expression with (2), implying that:

$$\log Z_\beta = (\beta - 1) D_\beta[p(x, z)||q(z|x)] \tag{26}$$

Van Erven and Harremos [19] indeed show that this is a convex function of β on $[0, \infty]$, matching what we know about the log partition function. We can obtain further interesting forms for $\log Z_\beta$ by removing $\log p(x)$ from the expression in (26):

$$\begin{aligned}
\log Z_\beta &= \log \int q(z|x)^{1-\beta} p(x, z)^\beta dz \\
&= \beta \log p(x) - (1 - \beta) D_\beta[p(z|x)||q(z|x)] && (27)\\
&= \beta \log p(x) - \beta\, D_{1-\beta}[q(z|x)||p(z|x)] && (28)
\end{aligned}$$

where in the third line we have used the skew symmetry property $D_\alpha[p||q] = \frac{\alpha}{1-\alpha} D_{1-\alpha}[q||p]$ of the Renyi divergence for $0 < \alpha < 1$ [19]. We note several special orders of the Renyi divergence can be found in [19, 13], and confirm that $\log Z_0 = 0$ and $\log Z_1 = \log p(x)$ as above, subject to the support of q being contained in that of p.

C.2 Renyi Divergence Variational Inference

We can recognize the expression in (28) as corresponding to the Renyi divergence variational inference framework of [13], which constructs a lower bound on likelihood by subtracting an α-divergence:

$$\begin{aligned}
\mathcal{L}_\alpha &= \log p(x) - D_\alpha[q(z|x)||p(z|x)] \\
&= \frac{1}{1 - \alpha} \log p(x)^{1-\alpha} + \frac{1}{\alpha - 1} \log \int q(z|x)^\alpha \frac{p(x, z)^{1-\alpha}}{p(x)^{1-\alpha}} dz \\
&= \frac{1}{1 - \alpha} \log \int q(z|x)^\alpha p(x, z)^{1-\alpha} dz && (29)
\end{aligned}$$

Using (28) and $\beta = 1 - \alpha$, we can immediately relate the partition function $\log Z_\beta$ and objective of [13].

$$\psi(\beta) = \log Z_\beta = \beta \mathcal{L}_{1-\beta} = \beta\big(\log p(x) - D_{1-\beta}[q(z|x)||p(z|x)]\big) \tag{30}$$

Each partition function thus corresponds to a lower bound on log-likelihood constructed using $D_{1-\beta}[q(z|x)||p(z|x)]$, scaled by β. We will proceed to reference the divergence order using β, even when referring to results in the α notation from [13], [19], etc. We can rewrite the Renyi VI objective and its Monte Carlo estimator [13] as:

$$\begin{aligned}\mathcal{L}_{1-\beta} &= \frac{1}{\beta} \log \mathbb{E}_q[(\frac{p(x,z)}{q(z|x)})^\beta] \\ &\approx \frac{1}{\beta} \log \frac{1}{S} \sum_{s=1}^{S} \frac{p(x,z_s)^\beta}{q(z_s|x)^\beta} \triangleq \hat{\mathcal{L}}_{1-\beta,S} = \frac{1}{\beta} \log \hat{Z}_\beta \end{aligned} \tag{31}$$

Note that we have $\mathcal{L}_0 = \log p(x)$, so that the importance-weighted autoencoder (IWAE) [5] is recovered for $\hat{\mathcal{L}}_{0,S}$. Further, the estimate is non-increasing in α for a fixed number of samples S, and increasing in S for fixed α [13]. For any β and finite samples, Li and Turner [13] show that, if there exists a point where $\hat{\mathcal{L}}_{1-\beta,S} = \log p(x)$, it will occur with $\beta_S \geq 1$ (i.e. $\alpha_S \leq 0$). While we have yet to fully characterize the role of intermediate $\log Z_\beta$ in the TVO setting, this could motivate considering $\beta \geq 1$.

D Gradient Estimation

Li and Turner [13] also provide a gradient estimator based on the reparameterization trick [11] that is applicable for any β, although they consider only a single order of the Renyi divergence at a time. This amounts to a self-normalized importance sampling estimate [5, 13] that uses the same weights as (5) and [15].

$$\nabla_\theta \hat{\mathcal{L}}_{1-\beta} = \frac{1}{\beta} \nabla_\theta \log \hat{Z}_\beta = \sum_{s=1}^{S} \frac{w_s}{\sum_s w_s} \nabla_\theta \log \frac{p(x,z_s)}{q(z_s|x)} \qquad \text{where } w_s = \frac{p(x,z_s)^\beta}{q(z_s|x)^\beta} \tag{32}$$

where θ is used to denote the parameters of both p and q. Thus, we can use the same self-normalized importance sampling scheme as in (5) to estimate the gradient of $\log Z_\beta$ with respect to parameters θ. This involves taking the expectation of the gradient of the importance weights, and, given the correspondence shown in (31), matches the estimator derived in [13].

In order to optimize the TVO with respect to q and p, Masrani et al. [15] derive an expression for $\nabla_\theta \nabla_\beta \log Z_\beta = \nabla_\theta \mathbb{E}_{\pi_\beta} \phi(x,z)$, where θ denotes the parameters of both p and q :

$$\nabla_\theta \nabla_\beta \log Z_\beta = \nabla_\theta \mathbb{E}_{\pi_\beta} \log \frac{p(x,z)}{q(z|x)} \tag{33}$$

$$= \mathbb{E}_{\pi_\beta} \nabla_\theta \log \frac{p(x,z)}{q(z|x)} + \int \log \frac{p(x,z)}{q(z|x)} \nabla_\theta \pi_\beta(z|x) dz \tag{34}$$

$$= \mathbb{E}_{\pi_\beta} \nabla_\theta \log \frac{p(x,z)}{q(z|x)} + \mathbb{E}_{\pi_\beta} \log \frac{p(x,z)}{q(z|x)} \nabla_\theta \log \pi_\beta(z|x) \tag{35}$$

$$= \mathbb{E}_{\pi_\beta} \nabla_\theta \log \frac{p(x,z)}{q(z|x)} + \mathrm{Cov}_{\pi_\beta}[\log \frac{p(x,z)}{q(z|x)}, \nabla_\theta\big((1-\beta)\log q(z|x) + \beta \log p(x,z)\big)] \tag{36}$$

where we have used the product rule in the second line and $\nabla_\theta \pi_\beta = \pi_\beta \nabla_\theta \log \pi_\beta$ in the third line. See App. D of [15] for more detailed derivations.

However, note that the first term may be optimized via the reparameterization trick since we will use the self-normalized importance sampling scheme above to translate samples from $q(\epsilon) \to q(z|x) \to \pi_\beta(z|x)$ using an encoding transformation $z = g(\epsilon)$:

$$\mathbb{E}_{\pi_\beta} \nabla_\theta \log \frac{p(x,z)}{q(z|x)} \approx \sum_{s=1}^{S} \frac{w_s}{\sum_s w_s} \nabla_\theta \log \frac{p(x,g(\epsilon_s))}{q(g(\epsilon_s)|x)} \qquad \text{where } w_s = \frac{p(x,g(\epsilon_s))^\beta}{q(g(\epsilon_s)|x)^\beta} \tag{37}$$

Thus, the reparameterization-free gradient estimator found in Masrani et al. [15] simply contains an additional term measuring the covariance of the sufficient statistics and gradients of the unnormalized log-mixture distribution. This expectation is again estimated using the same importance sampling scheme as above [15].