All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference

Rob Brekelmans [1 *] **Vaden Masrani** [2 *] **Frank Wood** [2] **Greg Ver Steeg** [1] **Aram Galstyan** [1]

Abstract

The recently proposed Thermodynamic Variational Objective (TVO) leverages thermodynamic integration to provide a family of variational inference objectives, which both tighten and generalize the ubiquitous Evidence Lower Bound (ELBO). However, the tightness of TVO bounds was not previously known, an expensive grid search was used to choose a "schedule" of intermediate distributions, and model learning suffered with ostensibly tighter bounds. In this work, we propose an exponential family interpretation of the geometric mixture curve underlying the TVO and various path sampling methods, which allows us to characterize the gap in TVO likelihood bounds as a sum of KL divergences. We propose to choose intermediate distributions using equal spacing in the moment parameters of our exponential family, which matches grid search performance and allows the schedule to adaptively update over the course of training. Finally, we derive a doubly reparameterized gradient estimator which improves model learning and allows the TVO to benefit from more refined bounds. To further contextualize our contributions, we provide a unified framework for understanding thermodynamic integration and the TVO using Taylor series remainders.

1. Introduction

Modern variational inference (VI) techniques are able to jointly perform maximum likelihood parameter estimation and approximate posterior inference using stochastic gradient ascent (Kingma & Welling, 2013; Rezende et al., 2014). Commonly, this is done by optimizing a tractable bound to the marginal log likelihood $\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z})\, \mathrm{d}\mathbf{z}$, obtained by introducing a divergence $D[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]$ between the variational distribution $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ and true posterior $p_\theta(\mathbf{z}\,|\,\mathbf{x})$ (Blei et al., 2017; Li & Turner, 2016; Dieng et al., 2017; Domke & Sheldon, 2018; Wang et al., 2018).

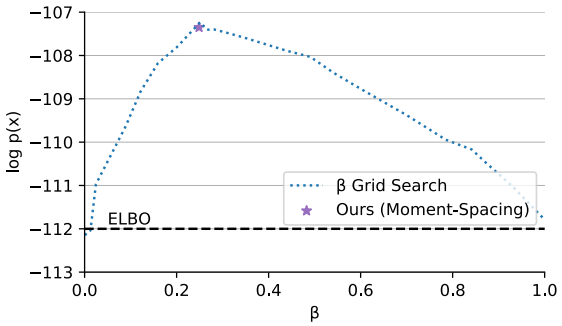


Figure 1. The original TVO paper recommended using two partition points, with a single intermediate β_1 in addition to the ELBO at $\beta_0 = 0$. We report test $\log p_\theta(\mathbf{x})$ values from training a separate VAE at each β_1, but this grid search is prohibitively expensive in practice. Our moment-spacing schedule is an adaptive method for choosing β points, which yields near-optimal performance on Omniglot and provides notable improvement over the ELBO.

The recent Thermodynamic Variational Objective (TVO) (Masrani et al., 2019) reframes likelihood estimation in terms of numerical integration along a geometric mixture path connecting $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ and $p_\theta(\mathbf{z}\,|\,\mathbf{x})$. This perspective yields a natural family of lower and upper bounds via Riemann sum approximations, with the ELBO appearing as a single-term lower bound and wake-sleep (WS) ϕ update corresponding to the simplest upper bound. The TVO generalizes these objectives by using a K-term Riemann sum to obtain tighter bounds on marginal likelihood. We refer to the discrete partition $\{\beta_k\}_{k=0}^K$ used to construct this estimator as an 'integration schedule.'

However, the gaps associated with these intermediate bounds was not previously known, an important roadblock to understanding the objective. Further, the TVO was limited by a grid search procedure for choosing the integration schedule. While TVO bounds should become tighter with more refined partitions, Masrani et al. (2019) actually observe deteriorating performance in practice with high K.

[*]Equal contribution [1]Information Sciences Institute, University of Southern California, Marina del Rey, CA [2]University of British Columbia, Vancouver, CA. Correspondence to: Rob Brekelmans <brekelma@usc.edu>, Vaden Masrani <vadmas@cs.ubc.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Our central contribution is an exponential family interpretation of the geometric mixture curve underlying the TVO and various path sampling methods (Gelman & Meng, 1998; Neal, 2001). Using the Bregman divergences associated with this family, we characterize the gaps in the TVO upper and lower bounds as the sum of KL divergences along a given path, resolving this open question about the TVO.

Further, we propose to choose intermediate distributions in the TVO based on the 'moment-averaged' path of Grosse et al. (2013), which arises naturally from the dual parameterization of our exponential family. This scheduling scheme was originally proposed in the context of annealed importance sampling (AIS), where additional sampling procedures may be required to even approximate it. We provide an efficient implementation for the TVO setting, which allows the choice of β to adapt to the shape of the integrand and degree of posterior mismatch throughout training.

In Figure 1, we observe that this flexible schedule yields near-optimal performance compared to grid search for a single intermediate distribution, so that the TVO can significantly improve upon the ELBO for minimal additional cost. However, our moments scheduler can still suffer the previously observed performance degradation as the number of intermediate distributions increases. As a final contribution, we propose a doubly reparameterized gradient estimator for the TVO, which we show can avoid this undesirable behavior and improve overall performance in continuous models.

Our exponential family analysis may be of wider interest given the prevalence of Markov Chain Monte Carlo (MCMC) techniques utilizing geometric mixture paths (Neal, 1996; 2001; Grosse et al., 2016; Syed et al., 2019; Huang et al., 2020). To this end, we also present a framework for understanding thermodynamic integration (TI) (Ogata, 1989) and the TVO using Taylor series remainders, which clarifies that the TVO is a first-order objective and provides geometric intuition for several results from Grosse et al. (2013). We hope these connections can help open new avenues for analysis at the intersection of MCMC, VI, and statistical physics.

2. Background

2.1. Thermodynamic Integration

Thermodynamic integration (TI) is a technique from statistical physics, which frames estimating ratios of partition functions as a one-dimensional integration problem. Commonly, this integral is taken over $\beta \in [0, 1]$, which parameterizes a path of geometric mixtures between a base distribution π_0, and a target distribution π_1 (Gelman & Meng, 1998)

$$\pi_\beta(\mathbf{z}) := \frac{\tilde{\pi}_\beta(\mathbf{z})}{\int \tilde{\pi}_\beta(\mathbf{z})\, d\,\mathbf{z}} = \frac{\pi_0^{1-\beta}(\mathbf{z})\pi_1^{\beta}(\mathbf{z})}{Z_\beta}. \tag{1}$$

The insight of TI is to recognize that, while the log partition function is intractable, its derivative can be written as an expectation that may be estimated using sampling or simulation techniques (Neal, 2001; Habeck, 2017)

$$\nabla_\beta \log Z_\beta = \mathbb{E}_{\pi_\beta}\left[\log \frac{\pi_1(\mathbf{z})}{\pi_0(\mathbf{z})}\right]. \tag{2}$$

In Sec. 3, we will see that this identity arises from an interpretation of the geometric mixture curve (1) as an exponential family. Applying this within the fundamental theorem of calculus,

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

$$= \int_0^1 \mathbb{E}_{\pi_\beta}\left[\log \frac{\pi_1(\mathbf{z})}{\pi_0(\mathbf{z})}\right] d\beta. \tag{4}$$

While (3) holds for any choice of path parameterized by β, we can construct efficient estimators of the integrand in (4) and estimate the partition function ratio $\log Z_1/Z_0$ using numerical integration techniques.

2.2. The Thermodynamic Variational Objective

The TVO (Masrani et al., 2019) uses TI in the context of variational inference to provide natural upper and lower bounds on the log evidence, which can then be used as objectives for training latent variable models. In particular, the geometric mixture path interpolates between the approximate posterior $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ and the joint generative model $p_\theta(\mathbf{x}, \mathbf{z})$

$$\pi_\beta(\mathbf{z}\,|\,\mathbf{x}) = \frac{\tilde{\pi}_\beta(\mathbf{x}, \mathbf{z})}{\int \tilde{\pi}_\beta(\mathbf{x}, \mathbf{z})\, \mathrm{d}\mathbf{z}} := \frac{q_\phi(\mathbf{z}\,|\,\mathbf{x})^{1-\beta} p_\theta(\mathbf{x}, \mathbf{z})^{\beta}}{Z_\beta(\mathbf{x})}. \tag{5}$$

As distributions over $\mathbf{z}$, we can identify the endpoints as $\pi_0(\mathbf{z}\,|\,\mathbf{x}) = q_\phi(\mathbf{z}\,|\,\mathbf{x})$ and $\pi_1(\mathbf{z}\,|\,\mathbf{x}) = p_\theta(\mathbf{z}\,|\,\mathbf{x})$, with corresponding normalizing constants $Z_0 = 1$ and $Z_1 = \int p_\theta(\mathbf{x}, \mathbf{z})\, d\,\mathbf{z} = p_\theta(\mathbf{x})$.

Applying TI (3) for this set of log partition functions, Masrani et al. (2019) express the generative model likelihood using a one-dimensional integral over the unit interval

$$\log p_\theta(\mathbf{x}) = \int_0^1 \mathbb{E}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right] d\beta. \tag{6}$$

The left and right endpoints of this integrand correspond to familiar lower and upper bounds on $\log p_\theta(\mathbf{x})$. The evidence lower bound (ELBO) occurs at $\beta = 0$, while the analogous evidence upper bound (EUBO) at $\beta = 1$ uses the 'reverse' KL divergence and appears in various wake-sleep objectives (WS) (Hinton et al., 1995; Bornschein & Bengio, 2014)

$$\text{ELBO}(\theta, \phi, \mathbf{x}) = \log p_\theta(\mathbf{x}) - D_{KL}[q_\phi || p_\theta] \tag{7}$$

$$\text{EUBO}(\theta, \phi, \mathbf{x}) = \log p_\theta(\mathbf{x}) + D_{KL}[p_\theta || q_\phi]. \tag{8}$$

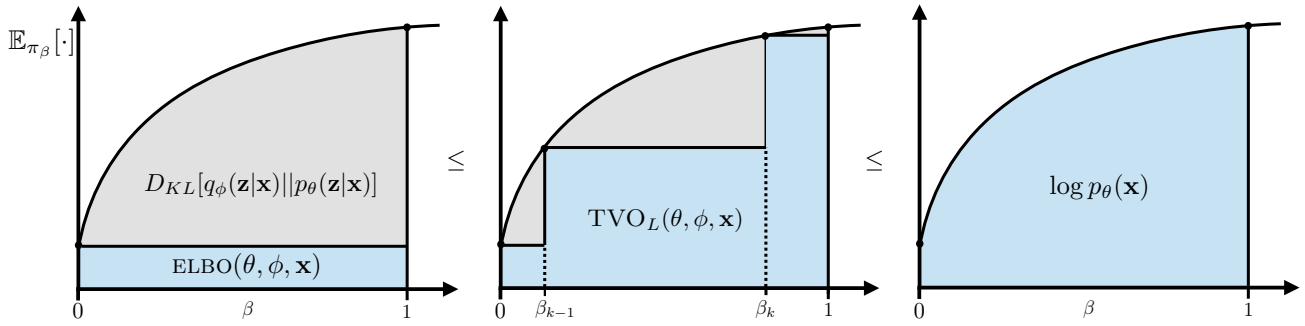


Figure 2. The TVO is a K-term Riemann sum approximation of $\log p_\theta(\mathbf{x})$, which can be expressed as a scalar integral over the unit interval in (6) and on the right. The ELBO is a single-term left Riemann approximation of the same integral using the point $\beta = 0$ with $\pi_0 = q_\phi(\mathbf{z} \mid \mathbf{x})$. Note that the integrand is negative in practice, but shown as positive for interpretability.

To arrive at the TVO, a discrete partition schedule is chosen $\mathcal{P}_\beta = \{\beta_k\}_{k=0}^K$ with $\beta_0 = 0$, $\beta_K = 1$, and $\Delta_{\beta_k} = \beta_k - \beta_{k-1}$. The integral in (6) is then approximated using a left or right Riemann sum. Since Masrani et al. (2019) show the integrand is increasing, these approximations yield valid lower and upper bounds on the marginal likelihood

$$\mathrm{TVO}_{\mathrm{L}}(\theta,\phi,\mathbf{x}) := \sum_{k=1}^{K} \Delta_{\beta_k}\, \mathbb{E}_{\pi_{\beta_{k-1}}}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\right] \quad (9)$$

$$\mathrm{TVO}_{\mathrm{U}}(\theta,\phi,\mathbf{x}) := \sum_{k=1}^{K} \Delta_{\beta_k}\, \mathbb{E}_{\pi_{\beta_k}}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\right] \quad (10)$$

with

$$\mathrm{TVO}_{\mathrm{L}}(\theta,\phi,\mathbf{x}) \leq \log p_\theta(\mathbf{x}) \leq \mathrm{TVO}_{\mathrm{U}}(\theta,\phi,\mathbf{x}). \quad (11)$$

The first term of $\mathrm{TVO}_L(\theta,\phi,\mathbf{x})$ corresponds to the ELBO, while the last term of $\mathrm{TVO}_U(\theta,\phi,\mathbf{x})$ corresponds to the EUBO. Thus, the TVO generalizes both objectives, with additional partitions leading to tighter bounds on likelihood as visualized in Fig. 2.

Although we consider thermodynamic integration over $0 \leq \beta \leq 1$ to approximate $\log p_\theta(\mathbf{x})$, note that this integral does not avoid the need for integration over $\mathbf{z}$ since each intermediate distribution must be normalized. Masrani et al. (2019) propose an efficient, self-normalized importance sampling (SNIS) scheme with proposal $q_\phi(\mathbf{z} \mid \mathbf{x})$, so that expectations at any intermediate β can be estimated by simply reweighting a single set of importance samples

$$\mathbb{E}_{\pi_\beta}[\cdot] \approx \sum_{i=1}^{S} \frac{w_i^\beta}{\sum_{i=1}^S w_s^\beta}[\cdot] \text{ where } w_i := \frac{p_\theta(\mathbf{x},\mathbf{z}_i)}{q_\phi(\mathbf{z}_i\mid\mathbf{x})}. \quad (12)$$

3. Exponential Family Interpretation

We propose a novel exponential family of distributions which, by absorbing both $p_\theta(\mathbf{x},\mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{x})$ into the sufficient statistic, corresponds to the geometric mixture path defined in (5). We provide a formal definition in Sec. 3.1, before showing in Sec. 3.2 that several key quantities in the TVO arise from familiar properties of exponential families. In Sec. 4, we leverage the Bregman divergences associated with our exponential family to naturally characterize the gap in TVO bounds as a sum of KL divergences.

3.1. Definition

To match the TVO setting in (5), we consider an exponential family of distributions with natural parameter β, base measure $q_\phi(\mathbf{z} \mid \mathbf{x})$, and sufficient statistics equal to the log importance weights as in (9)-(10)

$$\pi_\beta(\mathbf{z}\mid\mathbf{x}) := \pi_0(\mathbf{z}\mid\mathbf{x}) \exp\{\,\beta \cdot T(\mathbf{x},\mathbf{z}) - \psi(\mathbf{x};\beta)\} \quad (13)$$

$$\text{where } T(\mathbf{x},\mathbf{z}) := \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})} \quad \pi_0(\mathbf{z}\mid\mathbf{x}) := q_\phi(\mathbf{z}\mid\mathbf{x})$$

This induces a log-partition function $\psi(\mathbf{x};\beta)$, which normalizes over $\mathbf{z}$ and corresponds to $\log Z_\beta(\mathbf{x})$ in (5)

$$\begin{aligned}\psi(\mathbf{x};\beta) &:= \log \int q_\phi(\mathbf{z}\mid\mathbf{x}) \exp\{\,\beta \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\}\,\mathrm{d}\mathbf{z}, \\ &= \log \int q_\phi(\mathbf{z}\mid\mathbf{x})^{1-\beta} p_\theta(\mathbf{x},\mathbf{z})^\beta\,\mathrm{d}\mathbf{z} \quad (14) \\ &= \log Z_\beta(\mathbf{x}). \quad (15)\end{aligned}$$

The log-partition function will play a key role in our analysis, often written as $\psi(\beta)$ to omit the dependence on $\mathbf{x}$.

We emphasize that we have made no additional assumptions on $p_\theta(\mathbf{x},\mathbf{z})$ or $q_\phi(\mathbf{z} \mid \mathbf{x})$, and do not assume they come from exponential families themselves. This 'higher-order' exponential family thus maintains full generality and may be constructed between arbitrary distributions.

3.2. TVO using Exponential Families

We now show that a number of key quantities, which were manually derived in the original TVO work, may be directly obtained from our exponential family.

TI Integrates the Mean Parameters It is well known that the log-partition function $\psi(\beta)$ is convex, with its first (partial) derivative equal to the expectation of the sufficient statistics under π_β (Wainwright & Jordan, 2008).

$$\eta_\beta := \nabla_\beta \psi(\beta) = \mathbb{E}\left[T(\mathbf{x}, \mathbf{z})\right] = \mathbb{E}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] \quad (16)$$

This quantity is known as the *mean parameter* η_β, which provides a dual coordinate system for indexing intermediate distributions (Wainwright & Jordan (2008) Sec. 3.5.2). Comparing with (2) and (6), we observe that the ability to trade derivatives of the log-partition function for expectations in TI arises from this property of exponential families.

We may then interpret the TVO as integrating over the mean parameters $\eta_\beta = \nabla_\beta \log Z_\beta$ of our path exponential family, which can be seen by rewriting (3)

$$\psi(1) - \psi(0) = \int_0^1 \eta_\beta \, d\beta = \int_0^1 \mathbb{E}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] d\beta.$$

TVO Likelihood Bounds The convexity of the log partition function arises from the fact that entries in its matrix of second partial derivatives with respect to the natural parameters correspond to the (co)variance of the sufficient statistics (Wainwright & Jordan, 2008). In our 1-d case, this corresponds to the variance of the log importance weights

$$\nabla^2_\beta \psi(\beta) = \mathrm{Var}[T(\mathbf{x}, \mathbf{z})] = \mathrm{Var}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]. \quad (17)$$

We can see that the TVO integrand $\nabla_\beta \psi(\beta)$ is increasing from non-negativity of $\nabla^2_\beta \psi(\beta) \geq 0 \; \forall \beta$, which ensures that the left and right Riemann sums will yield valid lower and upper bounds on the marginal log likelihood.

ELBO on the Graph of η_β Inspecting Fig. 2, we see that the gap in the TVO bounds corresponds to the amount by which a Riemann approximation under- or over-estimates the area under the curve (AUC). We can solidify this intution for the case of the ELBO, a single-term approximation of $\log p_\theta(\mathbf{x})$ using $\beta = 0$ for the entire interval $\beta_1 - \beta_0 = 1 - 0$

$$\begin{aligned}
\text{GAP} &= \underbrace{\left[\int_0^1 \nabla_\beta \psi(\beta) d\beta\right]}_{\text{AUC}} - \underbrace{(1-0)}_{\text{WIDTH}} \underbrace{\mathbb{E}_{\pi_0}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]}_{\text{HEIGHT}} \\
&= \log p_\theta(\mathbf{x}) - \text{ELBO}(\theta, \phi, \mathbf{x}) \\
&= D_{KL}[q_\phi(\mathbf{z} \mid \mathbf{x}) || p_\theta(\mathbf{z} \mid \mathbf{x})]. \quad (18)
\end{aligned}$$

In the next section, we generalize this reasoning to more refined partitions, showing that the gap in arbitrary TVO bounds corresponds to a sum of KL divergences between adjacent π_{β_k} along a given path $\{\beta_k\}_{k=0}^K$.

4. TVO Likelihood Bound Gaps

In previous work, it was shown only that $\text{TVO}_L(\theta, \phi, \mathbf{x})$ minimizes a quantity that is non-negative and vanishes at $q_\phi(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{z} \mid \mathbf{x})$ (Masrani et al., 2019). Using the Bregman divergences associated with our path exponential family, we can now provide a unified characterization of the gaps in TVO bounds.

4.1. Bregman Divergences

We begin with a brief review of the Bregman divergence, which can be visualized on the graph of the TVO integrand in Fig. 3 or the log partition function in Fig. 4.

A Bregman divergence D_ψ is defined with respect to a convex function ψ (Banerjee et al., 2005) which, in our case, takes distributions indexed by natural parameters β and β' as its arguments

$$D_\psi[\beta : \beta'] = \psi(\beta) - \underbrace{\left(\psi(\beta') + (\beta - \beta')\nabla_\beta \psi(\beta')\right)}_{\text{First Order Taylor Approx}}. \quad (19)$$

Geometrically, the Bregman divergence corresponds to the gap in a first-order Taylor approximation of $\psi(\beta)$ around the second argument β', as depicted in Fig. 4. Note that this difference is guaranteed to be nonnegative, since we know that the tangent will everywhere underestimate a convex function (Boyd & Vandenberghe, 2004).

The Bregman divergence D_ψ for the exponential family in (13) is also equivalent to the KL divergence, with the order of the arguments reversed (also see App. A). Applying (16) and adding and subtracting a base measure term,

$$\begin{aligned}
D_\psi[\beta : \beta'] &= \psi(\beta) - \psi(\beta') - (\beta - \beta')\nabla_\beta \psi(\beta') \quad (20) \\
&= \psi(\beta) \; - \beta \; \cdot \mathbb{E}_{\pi_{\beta'}}[T] - \mathbb{E}_{\pi_{\beta'}}[\log \pi_0] \quad (21) \\
&\quad - \psi(\beta') + \beta' \cdot \mathbb{E}_{\pi_{\beta'}}[T] + \mathbb{E}_{\pi_{\beta'}}[\log \pi_0] \\
&= \mathbb{E}_{\pi_{\beta'}}[\log \pi_{\beta'} - \log \pi_\beta], \quad (22)
\end{aligned}$$

where in the third line, we use the fact that $\mathbb{E}[\log \pi_\beta] = \mathbb{E}[\log \pi_0(\mathbf{z} \mid \mathbf{x}) + \beta \cdot T(\mathbf{x}, \mathbf{z})] - \psi(\mathbf{x}; \beta)$ from (13). We then obtain our desired result, with

$$D_\psi[\beta : \beta'] = D_{KL}[\pi_{\beta'} || \pi_\beta]. \quad (23)$$

KL Divergence on the Graph of η_β We can also visualize the Bregman divergence on the graph of the integrand $\eta_\beta = \nabla_\beta \psi(\beta)$ in Fig. 3, which leads to a natural expression for the gaps in TVO upper and lower bounds.

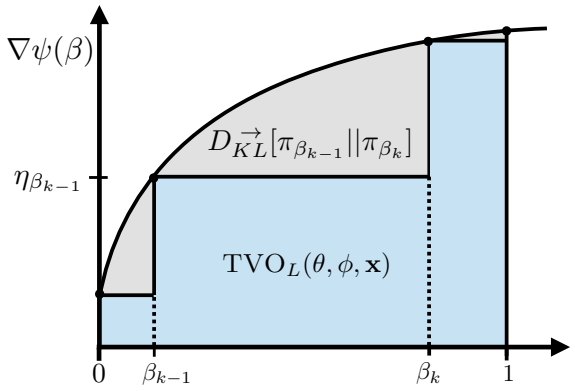


Figure 3. The Bregman divergence $D_{KL}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]$ can be visualized as the area under the curve minus the left-Riemann sum via (24). This term contributes to the gap in the likelihood bound $\mathrm{TVO_L}$. We also derive an integral form for the KL divergence in App. C.2. Note that both the integrand and $\psi(\beta)$ are negative in practice.

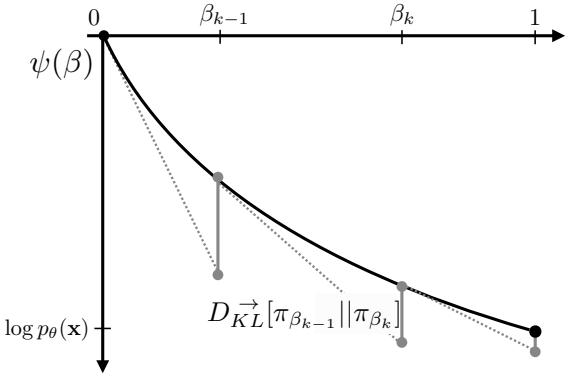


Figure 4. $\mathrm{TVO}_L(\theta,\phi,\mathbf{x})$ may be viewed as constructing successive first-order Taylor approximations to intermediate $\psi(\beta_k)$, with the accumulated error corresponding to the gap in the bound. The upper bound takes KL divergences in the reverse direction, with the first argument decreasing along the path.

To begin, we consider a single subinterval $[\beta_{k-1},\beta_k]$ and follow the same reasoning as for the ELBO in Sec. 3.2. In particular, the area under the integrand in this region is $\mathrm{AUC} = \int_{\beta_{k-1}}^{\beta_k} \nabla_\beta\psi(\beta)d\beta = \psi(\beta_k) - \psi(\beta_{k-1})$, with the left-Riemann approximation corresponding to $(\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_k)$. Taking the difference between these expressions, we obtain the definition of the Bregman divergence in (19)

$$\mathrm{GAP} = \underbrace{\psi(\beta_k) - \psi(\beta_{k-1})}_{\mathrm{AUC}} - \underbrace{(\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_{k-1})}_{\text{Term in } \mathrm{TVO}_L(\theta,\phi,\mathbf{x})}$$

$$= D_\psi[\beta_k : \beta_{k-1}] = D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]. \tag{24}$$

where arrows indicate whether the first argument of the KL divergence is increasing or decreasing along the path. For the gap in the right-Riemann upper bound, we follow similar derivations with the order of the arguments reversed in Sec. 4.3. This results in a gap of $D_{KL}^{\leftarrow}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$, with expectations $\eta_{\beta_k} = \nabla_\beta\psi(\beta_k)$ taken under π_{β_k}.

4.2. TVO Lower Bound Gap

Extending the above reasoning to the entire unit interval, we can consider any sorted partition $\mathcal{P}_\beta = \{\beta_k\}_{k=0}^K$ with $\beta_0 = 0$ and $\beta_K = 1$. Summing (24) across intervals, note that intermediate $\psi(\beta_k)$ terms cancel in telescoping fashion

$$\sum_{k=1}^K D_\psi[\beta_k : \beta_{k-1}] \tag{25}$$

$$= \psi(1) - \psi(0) - \sum_{k=1}^K (\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_{k-1})$$

where the last term matches the $\mathrm{TVO_L}$ objective in (9).

Writing D_ψ as a KL divergence as in (23) and recalling that $\psi(1) - \psi(0) = \log p_\theta(\mathbf{x})$, we obtain

$$\log p(\mathbf{x}) - \mathrm{TVO}_L(\theta,\phi,\mathbf{x}) = \sum_{k=1}^K D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]. \tag{26}$$

We therefore see that the gap in the TVO lower bound is the sum of KL divergences between adjacent π_{β_k} distributions.

Alternatively, we can view (25) as constructing successive first-order Taylor approximations to intermediate $\psi(\beta_k)$ in Fig. 4. The likelihood bound gap of $\sum_{k=1}^K D_\psi[\beta_k : \beta_{k-1}]$ measures the accumulated error along the path. While the ELBO estimates $\psi(1) = \log p_\theta(\mathbf{x})$ directly from $\beta = 0$, more refined partitions can reduce the error and improve our bounds. As $K \to \infty$, $\mathrm{TVO}_L(\theta,\phi,\mathbf{x})$ becomes tight as our π_{β_k} are infinitesimally close, and the Riemann integral estimate would become exact given exact estimates of η_{β_k}.

4.3. TVO Upper Bound Gap

To characterize the gap in the upper bound, we first leverage convex duality to obtain a Bregman divergence in terms of the conjugate function $\psi^*(\eta)$ and the mean parameters η. As shown in App. A, this divergence, D_{ψ^*}, is equivalent to D_ψ with the order of arguments reversed

$$\begin{aligned} D_{\psi^*}[\eta_k : \eta_{k-1}] &= D_\psi[\beta_{k-1} : \beta_k] \\ &= D_{KL}^{\leftarrow}[\,\pi_{\beta_k}||\,\pi_{\beta_{k-1}}]. \end{aligned} \tag{27}$$

As in (25), we expand the dual divergences along a path as

$$\sum_{k=1}^K D_\psi[\beta_{k-1} : \beta_k] \tag{28}$$

$$= \psi(0) - \psi(1) - \sum_{k=1}^K (\beta_{k-1} - \beta_k)\nabla_\beta\psi(\beta_k)$$

Since the last term corresponds to a right-Riemann sum, we can similarly characterize the gap in $\text{TVO}_U(\theta, \phi, \mathbf{x})$ using a sum of KL divergences in the reverse direction

$$\text{TVO}_U(\theta, \phi, \mathbf{x}) - \log p(\mathbf{x}) = \sum_{k=1}^{K} D_{KL}^{\leftarrow}[\pi_{\beta_k} || \pi_{\beta_{k-1}}]. \quad (29)$$

4.4. Integral Forms and Symmetrized KL

To further contextualize the developments in this section, we show in App. C that both thermodynamic integration and the TVO may be understood using the integral form of the Taylor remainder theorem. In particular, the expression (4) underlying TI corresponds to the gap in a zero-order approximation, whereas we have previously shown that the KL divergence arises from a first-order remainder.

We can thus obtain integral expressions for the KL divergence, lending further intution for its interpretation as the area of a region in Fig. 3 or Fig. 5

$$D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}} || \pi_{\beta_k}] = \int_{\beta_{k-1}}^{\beta_k} (\beta_k - \beta) \, \text{Var}_{\pi_\beta} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \,|\, \mathbf{x})} \right] d\beta.$$

Combining the remainders in each direction, we recover a known identity relating the symmetrized KL divergence to the integral of the Fisher information (App. C.3 (56), Dabak & Johnson (2002)).

Similarly, we can visualize (twice) the symmetrized KL as the area of a rectangle in Fig. 5, by adding the gaps in the left- and right-Riemann approximations for a single interval

$$\begin{aligned} D_{KL}^{\leftrightarrow}[\pi_{\beta_{k-1}}; \pi_{\beta_k}] &= D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}} || \pi_{\beta_k}] + D_{KL}^{\leftarrow}[\pi_{\beta_k} || \pi_{\beta_{k-1}}] \\ &= (\beta_k - \beta_{k-1})(\eta_k - \eta_{k-1}) \,. \end{aligned} \quad (30)$$

From the Taylor remainder perspective, we note that (30) can be derived using a further application of TI, or the fundamental theorem of calculus, to the function $\nabla_\beta \psi(\beta)$, with $\eta_k - \eta_{k-1} = \int_{\beta_{k-1}}^{\beta_k} \nabla_\beta^2 \psi(\beta) \, d\beta$ (App. C.3 (58)).

For $\beta_0 = 0$ and $\beta_1 = 1$, we can confirm from (7)-(8) that $\eta_1 - \eta_0 = \text{EUBO} - \text{ELBO} = D_{KL}[q_\phi || p_\theta] + D_{KL}[p_\theta || q_\phi]$.

Before presenting our proposed approach for choosing β in the next section, we note that App. D.1 describes a 'coarse-grained' linear binning schedule from Grosse et al. (2013), which allocates intermediate distributions based on the identity (30) and is evaluated as a baseline in Sec. 8.

5. Moment-Spacing Schedule

Masrani et al. (2019) observe that TVO performance can depend heavily on the choice of partition schedule $\mathcal{P}_\beta$, and propose log-uniform spacing of $\{\beta_2, ...\beta_{K-1}\}$ with grid search over the initial β_1.

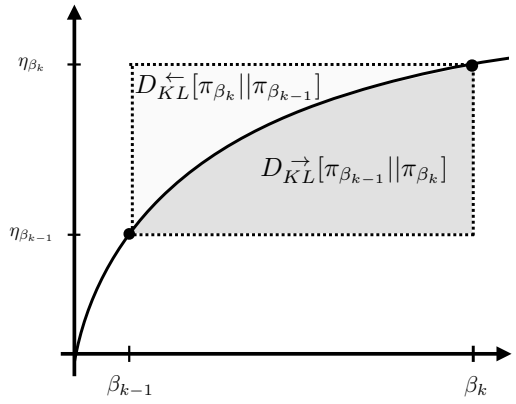


Figure 5. Adding the KL divergences in each direction, we can visualize the symmetrized KL divergence as the area of a rectangle. The curvature of the TVO integrand suggests which direction of the KL divergence is larger, with the divergence becoming symmetric when η_β is linear in β (see App. D.2).

Instead, we propose choosing β_k to yield equal spacing in the y-axis of the TVO integrand $\eta_\beta = \mathbb{E}_{\pi_\beta}[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \,|\, \mathbf{x})}]$, which corresponds to Lebesgue integration rather than Riemann integration in Fig. 6. This scheduling arises naturally from our exponential family in Sec. 3, with the mean parameters η_β corresponding to the dual parameters (Wainwright & Jordan (2008) Sec. 3.5.2). Equal spacing in the mean parameters also corresponds to the 'moment-averaged' path of Grosse et al. (2013), which was shown to yield robust estimators and natural generative samples from intermediate π_β in the context of AIS.

Given a budget of intermediate distributions $K = |\mathcal{P}_\beta|$, we seek β_k such that η_{β_k} are uniformly distributed between the endpoints $\eta_0 = \text{ELBO}$ and $\eta_1 = \text{EUBO}$ (see (7)-(8))

$$\beta_k = \eta_\beta^{-1} \left((1 - \frac{k}{K}) \cdot \text{ELBO} + \frac{k}{K} \cdot \text{EUBO} \right). \quad (31)$$

We use $\eta_\beta^{-1}(\mu)$ to indicate the value of the natural parameter β such that the expected sufficient statistics η_β match a desired target μ. This mapping between parameterizations is known as the Legendre transform and can be a difficult optimization in its own right (Wainwright & Jordan, 2008).

However, in the TVO setting, estimating moments η_β for a given β simply involves reweighting and normalizing the importance samples using SNIS in (12). Equipped with this cheap evaluation mechanism, we can apply binary search to find the β_k with a given expectation value η_{β_k}, as in (31). We update our choice of schedule at the end of each epoch, and provide further implementation details in App. G.

We visualize an example of our moments schedule in Fig. 6. Note that uniform spacing in η does not imply uniform spacing in β, since the Legendre transform is non-linear.

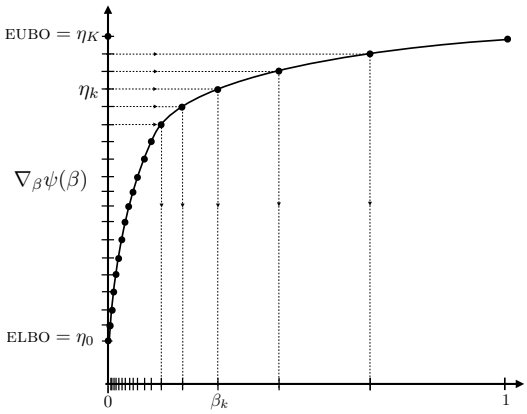


Figure 6. By enforcing equal spacing in the mean parameter space, our moments schedule naturally 'adapts' by allocating more partitions to regions where the integrand is changing quickly.

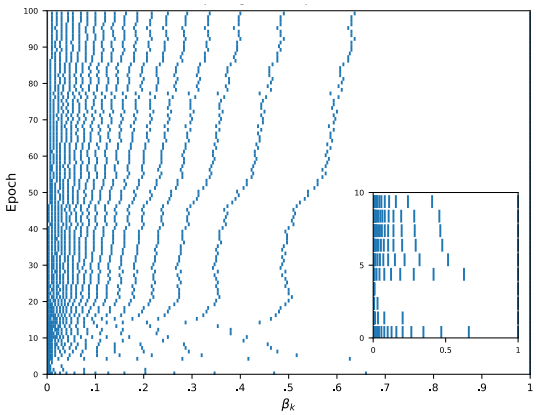


Figure 7. We visualize placement of β_k for our moments-spacing schedule across the first 100 epochs, with $K = 20$. Most β_k concentrate near 0 in early epochs, but spread out as training proceeds and the integrand becomes flatter as a function of β.

The resulting spacing in the x-axis reflects how quickly η_β is changing as a function of β, matching the intuition that we should allocate more points in regions where the integrand is changing quickly. Our moment-spacing schedule thus adapts to the shape of the TVO integrand, which can change significantly across training (Fig. 7). The integrand itself reflects the degree of posterior mismatch, since the curve will be flat when $q_\phi(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{z} \mid \mathbf{x})$, with $\eta_\beta = \log p_\theta(\mathbf{x})$ $\forall \beta$. On the other hand, an integrand rising sharply away from $\beta = 0$ indicates a poor proposal, with only several importance samples dominating the SNIS weights.

6. Doubly-Reparameterized TVO Gradient

To optimize the TVO, Masrani et al. (2019) derive a REINFORCE-style gradient estimator (see their App. F), which provides lower variance gradients and improved performance with discrete latent variables. Writing λ to denote $\{\phi, \theta\}$, with $w = p_\theta(\mathbf{x}, \mathbf{z})/q_\phi(\mathbf{z} \mid \mathbf{x})$ and $\tilde{\pi}_\beta(\mathbf{x}, \mathbf{z})$ as in (5), we obtain gradients for expectations of arbitrary $f(\mathbf{z})$ under π_β, with the TVO integrand corresponding to $f(\mathbf{z}) = \log w$,

$$\frac{d}{d\lambda}\mathbb{E}_{\pi_\beta}[f(\mathbf{z})] = \mathbb{E}_{\pi_\beta}[\frac{d}{d\lambda}f(\mathbf{z})] + \mathrm{Cov}_{\pi_\beta}\left[f(\mathbf{z}), \frac{d}{d\lambda}\log\tilde{\pi}_\beta(\mathbf{x}, \mathbf{z})\right] \quad (32)$$

However, when $\mathbf{z}_i \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ can be reparameterized via $\mathbf{z}_i = z(\epsilon_i, \phi)$, $\epsilon_i \sim p(\epsilon)$, we can improve the estimator in (32) by more directly incorporating $f(\mathbf{z})$ gradient information. To this end, we derive a doubly-reparameterized gradient estimator in App. I

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}[f(\mathbf{z})] = \mathbb{E}_{\pi_\beta}\left[\frac{d}{d\phi}f(\mathbf{z}) - \beta \cdot \frac{\partial \mathbf{z}}{\partial \phi}\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}\right] + (1-\beta)\,\mathrm{Cov}_{\pi_\beta}\left[f(\mathbf{z}),\, \beta \cdot \frac{\partial \mathbf{z}}{\partial \phi}\frac{\partial \log w}{\partial \mathbf{z}}\right]. \quad (33)$$

Doubly-reparameterized gradient estimators avoid a known signal-to-noise ratio issue for inference network gradients (Rainforth et al., 2018), using a second application of the reparameterization trick within the expanded total derivative (Tucker et al., 2018). We use a simplified form of (33) (see App. I (75)) for learning ϕ and (32) for learning θ.

Comparing the covariance terms of (32) and (33), note that $\frac{d}{d\lambda}\log\tilde{\pi}_\beta(\mathbf{x}, \mathbf{z})$ and $\beta\,\frac{\partial \mathbf{z}}{\partial \phi}\frac{\partial \log w}{\partial \mathbf{z}}$ differ by their differentation operator and a factor of $\log q_\phi$ due to reparameterization, with $\log\tilde{\pi}_\beta = \log q_\phi + \beta \log w$.

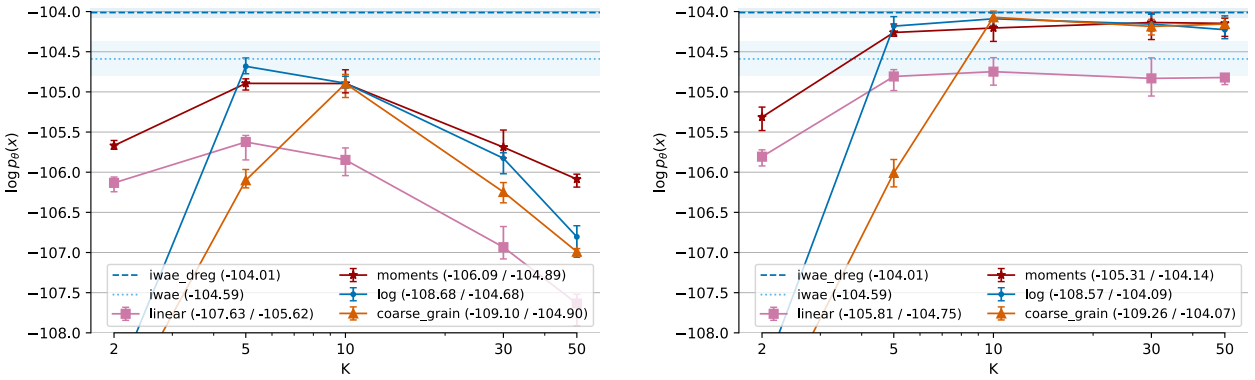
Further, the effect of the partial derivative $\frac{\partial \mathbf{z}}{\partial \phi}\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}$ in the first term of (33) linearly decreases as $\beta \to 1$ and $\pi_\beta(\mathbf{z} \mid \mathbf{x})$ has less dependence on ϕ.

Finally, we see that (33) passes two basic sanity checks, with the covariance correction term vanishing at both endpoints. At $\beta = 0$, we recover the gradient of the ELBO, $\frac{d}{d\phi}\mathbb{E}_{\pi_0}[f(\mathbf{z})] = \mathbb{E}_{z(\epsilon,\phi)}[\frac{d}{d\phi}f(\mathbf{z})]$. At $\beta = 1$, note that the $\frac{\partial \mathbf{z}}{\partial \phi}\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}$ term cancels when expanding $\frac{d}{d\phi}f(\mathbf{z})$, leaving $\frac{d}{d\phi}\mathbb{E}_{\pi_1}[f_\phi(\mathbf{z})] = \mathbb{E}_{p_\theta}[\frac{\partial}{\partial \phi}f_\phi(\mathbf{z})]$. This is to be expected for expectations under $p_\theta(\mathbf{z} \mid \mathbf{x})$, since the derivative with respect to ϕ passes inside the expectation and $\frac{\partial \mathbf{z}}{\partial \phi} = 0$.

7. Related Work

Thermodynamic integration (TI) is a strategy for estimating partition function ratios or free energy differences in simulations of physical systems (Ogata, 1989; Gelman & Meng, 1998; Frenkel & Smit, 2001), and also finds applications in model selection for phylogenetics (Lartillot & Philippe, 2006; Xie et al., 2011).

Physics applications of TI often involve sampling forward and reverse state trajectories (Frenkel & Smit, 2001; Habeck,



(a) TVO with REINFORCE Gradients

(b) TVO with Doubly-Reparameterized Gradients

Figure 8. Scheduling Performance by K on Omniglot, with $S = 50$. Legend shows (min / max) test $\log p_\theta(\mathbf{x})$ across K.

2017), as might be done using MCMC transition operators. Indeed, similar upper and lower bounds to the TVO are used to evaluate bidirectional Monte Carlo (Grosse et al., 2016). A body of recent work 'bridging the gap' between VI and MCMC (Salimans et al., 2015; Wolf et al., 2016; Hoffman, 2017; Li et al., 2017; Caterini et al., 2018; Huang et al., 2018; Ruiz & Titsias, 2019; Lawson et al., 2019) may thus provide a basis for practical improvements in thermodynamic variational inference.

Several recent VI objectives also naturally appear within the TVO framework. As we show in App. B, each log-partition function $\log Z_\beta(\mathbf{x})$ (14) in our exponential family corresponds to a Renyi divergence VI objective (Li & Turner, 2016) with order $\alpha = 1 - \beta$. The CUBO objectives of Dieng et al. (2017) correspond to upper bounds on $\log p_\theta(\mathbf{x})$ and log partition functions with $\beta \in [1, 2]$. From our exponential family perspective, there is no explicit restriction that our natural parameters β remain in the unit interval, with the χ^2-divergence at $\beta = 2$ of notable interest (Cortes et al., 2010). Bamler et al. (2017; 2019) also apply a Taylor series approach to obtain tighter bounds on $\log p_\theta(\mathbf{x})$, although the expansion is with respect to the importance weights $T(\mathbf{x}, \mathbf{z}) = \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}$ rather the natural parameter β.

8. Experiments

We investigate the effect of our moment-spacing schedule and reparameterization gradients using a continuous latent variable model on the Omniglot dataset. We estimate test $\log p_\theta(\mathbf{x})$ using the IWAE bound (Burda et al., 2015) with 5k samples, and use $S = 50$ samples for training unless noted. In all plots, we report averages over five random seeds, with error bars indicating min and max values. We describe our model architecture and experiment design in App. F, [1] with runtimes and additional results on binary MNIST in App. H.

[1] https://github.com/vmasrani/tvo_all_in

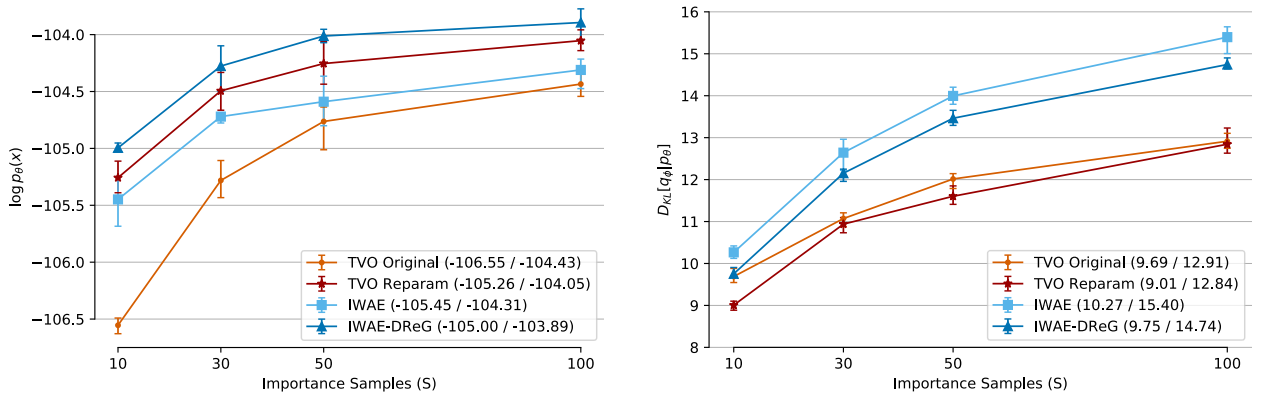
Moment Spacing Dynamics We seek understand the dynamics of our moment spacing schedule in Fig. 7, visualizing the choice of β points across training epochs with $K = 20$. Our intermediate distributions concentrate near $\beta = 0$ at the beginning of training, since $q_\phi(\mathbf{z} \mid \mathbf{x})$ and $p_\theta(\mathbf{z} \mid \mathbf{x})$ are mismatched and the TVO integrand rises sharply away from $q_\phi(\mathbf{z} \mid \mathbf{x})$. This effect is particularly dramatic within the first five epochs.

While the curve is still fairly noisy within the first twenty epochs, it begins flatten as training progresses and $q_\phi(\mathbf{z} \mid \mathbf{x})$ learns to match $p_\theta(\mathbf{z} \mid \mathbf{x})$. This is reflected in the β_k achieving a given proportion of the moments difference (EUBO-ELBO) moving to higher values. We found the moment-scheduling partitions to be relatively stable after 100 epochs.

Grid Search Comparison Next, we fix $K = 2$ with only β_1 chosen by the moment spacing schedule. We compare against grid search in Fig. 1 and Fig. 12 (App. H), and plot test $\log p_\theta(\mathbf{x})$ as a function of $\beta_1 \in [0, 1]$ across 25 static values. We report the value of β_1 for our moments schedule at the final epoch, which indicates where η_{β_1} is halfway between our estimated ELBO and EUBO.

We find that our adaptive scheduling matches the best performance from grid search, with the optimal intermediate distribution occurring at $\beta_1 \approx 0.3$ on both datasets. With a single, properly chosen intermediate distribution, we find that the TVO can achieve notable improvements over the ELBO at minimal additional cost.

Evaluating Scheduling Strategies From a numerical integration perspective, the TVO bounds should become arbitrarily tight as $K \to \infty$. However, Masrani et al. (2019) observe that additional partitions can be detrimental for learning in practice. We thus investigate the performance of our moment spacing schedule with a varying number of partitions. We plot test log likelihood at $K = \{2, 5, 10, 30, 50\}$,



(a) Omniglot Test $\log p_\theta(\mathbf{x})$

(b) Omniglot Test $D_{KL}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]$

Figure 9. Model Learning and Inference by S with $K = 5$. Legend shows (min / max) values across S.

and compare against three scheduling baselines: linear, log-uniform spacing, and the 'coarse-grained' schedule from Grosse et al. (2013) (see App. D.1). We begin the log-uniform spacing at $\beta_1 = 0.025$, a choice which results from grid search over β_1 for $K > 2$ in Masrani et al. (2019).

We observe in Fig. 8a that the moment scheduler provides the best performance at high and low K, while the log-uniform schedule can perform best for particular K. As previously observed, all scheduling mechanisms still suffer degradation in performance at large K.

Reparameterized TVO Gradients While our scheduling techniques do not address the detrimental effect of using many intermediate β, we now investigate the use of our reparameterization gradient estimator from Sec. 6. Repeating the previous experiment in Fig. 8b, we find that reparameterization helps preserve competitive performance for high K and improves overall model likelihoods. Our moments schedule is still particularly useful at low K, while the various scheduling methods converge to similar performance with many partition points. All scheduling techniques will be equivalent in the limit, as discussed in App. D.2.

Comparison with IWAE Finally, we compare TVO with moments scheduling against the importance weighted autoencoder (IWAE) (Burda et al., 2015) and doubly reparameterized IWAE DREG (Tucker et al., 2018) for model learning and posterior inference. It is interesting to note that IWAE corresponds to a direct estimate of $\psi(1)$, with the SNIS normalizer $\sum_{i=1}^{S} w_i^1$ in TVO (12) appearing inside the log.

In Fig. 9, we observe that TVO with reparameterization gradients achieves model learning performance in between that of IWAE and IWAE DREG, with lower KL divergences across all values of S. We repeat this experiment for MNIST in App. H Fig. 13, where TVO matches IWAE DREG model learning with better inference. Although we tend to obtain lower D_{KL} with lower model likelihood, we do not observe strong evidence of the signal-to-noise ratio issues of Rainforth et al. (2018) on either dataset. TVO with reparameterization thus appears to provide a favorable tradeoff between model learning and posterior inference.

9. Conclusion

In this work, we interpret the geometric mixture curve found in thermodynamic integration (TI), annealed importance sampling (AIS), and the Thermodynamic Variational Objective (TVO), using the Bregman duality of exponential families. We leveraged this approach to characterize the gap in TVO lower and upper bounds as a sum of KL divergences along a given path, and presented an adaptive scheduling technique based on the mean parameterization of our exponential family. Finally, we derived a doubly-reparameterized gradient estimator for terms in the TVO integrand.

The use of self-normalized importance sampling (SNIS) to estimate expectations under π_β may still be a key limitation of the TVO (see Masrani et al. (2019)), although we relied on the efficiency of SNIS for our moment-spacing schedule. Improved MCMC estimators that can be integrated with end-to-end learning of $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ and $p_\theta(\mathbf{x}, \mathbf{z})$ remain an intriguing direction for future work. In this study, we did not observe performance gains using equal spacing in either the KL or symmetrized KL divergence, but alternative schedules might also be motivated via physical interpretations (Andresen & Gordon, 1994; Salamon et al., 2002; Sivak & Crooks, 2012). We thus hope that our work can encourage further contributions in thermodynamic variational inference (TVI), a class of methods combining insights from VI, MCMC, and statistical physics.

Acknowledgements

The authors would like to thank Tuan Anh Le for clarifying the interpretation of the symmetrized KL divergence, and an anonymous reviewer for suggesting the connection with Lebesgue integration. RB thanks Kyle Reing, Artemy Kolchinsky, and Frank Nielsen for helpful discussions. VM thanks David Dehaene for his suggestion of reparameterized TVO gradients. This paper builds closely upon a workshop paper by RB, GV, and AG.

RB and GV acknowledge support from the Defense Advanced Research Projects Agency (DARPA) under awards FA8750-17-C-0106 and W911NF-16-1-0575. VM acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under award number PGSD3-535575-2019 and the British Columbia Graduate Scholarship, award number 6768. VM/FW acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and the Intel Parallel Computing Centers program. This material is based upon work supported by the United States Air Force Research Laboratory (AFRL) under the Defense Advanced Research Projects Agency (DARPA) Data Driven Discovery Models (D3M) program (Contract No. FA8750-19-2-0222) and Learning with Less Labels (LwLL) program (Contract No.FA8750-19-C-0515). Additional support was provided by UBC's Composites Research Network (CRN), Data Science Institute (DSI) and Support for Teams to Advance Interdisciplinary Research (STAIR) Grants. This research was enabled in part by technical support and computational resources provided by WestGrid (https://www.westgrid.ca/) and Compute Canada (www.computecanada.ca).

References

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Andresen, B. and Gordon, J. M. Constant thermodynamic speed for minimizing entropy production in thermodynamic processes and simulated annealing. *Physical Review E*, 50(6):4346, 1994.

Bamler, R., Zhang, C., Opper, M., and Mandt, S. Perturbative black box variational inference. In *Advances in Neural Information Processing Systems*, pp. 5079–5088, 2017.

Bamler, R., Zhang, C., Opper, M., and Mandt, S. Tightening bounds for variational inference by revisiting perturbation theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124004, 2019.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

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

Bornschein, J. and Bengio, Y. Reweighted wake-sleep. 1, 2014. URL http://arxiv.org/abs/1406.2751.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *Iclr-2015*, pp. 1–12, 2015. URL http://arxiv.org/abs/1509.00519.

Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian variational auto-encoder. In *Advances in Neural Information Processing Systems*, pp. 8167–8177, 2018.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pp. 442–450, 2010.

Crooks, G. E. Measuring thermodynamic length. *Physical Review Letters*, 99(10):100602, 2007.

Dabak, A. G. and Johnson, D. H. Relations between kullback-leibler distance and fisher information. 2002.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2732–2741, 2017.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In *Advances in neural information processing systems*, pp. 4470–4479, 2018.

Frenkel, D. and Smit, B. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

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

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

Grosse, R. B., Ancha, S., and Roy, D. M. Measuring the reliability of mcmc inference with bidirectional monte carlo. In *Advances in Neural Information Processing Systems*, pp. 2451–2459, 2016.

Habeck, M. Model evidence from nonequilibrium simulations. In *Advances in Neural Information Processing Systems*, pp. 1753–1762, 2017.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

Hoffman, M. D. Learning deep latent gaussian models with markov chain monte carlo. In *International conference on machine learning*, pp. 1510–1519, 2017.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, pp. 9701–9711, 2018.

Huang, S., Makhzani, A., Cao, Y., and Grosse, R. Evaluating lossy compression rates of deep generative models. *International Conference on Machine Learning*, 2020.

Huszar, F. Grosse's challenge: Duality and exponential families, Nov 2017. URL https://www.inference.vc/grosses-challenge/.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. (Ml):1–14, 2013. ISSN 1312.6114v10. URL http://arxiv.org/abs/1312.6114.

Kountourogiannis, D. and Loya, P. A derivation of taylor's formula with integral remainder. *Mathematics magazine*, 76(3):217–219, 2003.

Kullback, S. *Information theory and statistics*. Courier Corporation, 1997.

Lake, B. M., Salakhutdinov, R. R., and Tenenbaum, J. One-shot learning by inverting a compositional causal process. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2526–2534. Curran Associates, Inc., 2013.

Lartillot, N. and Philippe, H. Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2): 195–207, 2006.

Lawson, J., Tucker, G., Dai, B., and Ranganath, R. Energy-inspired models: Learning with sampler-induced distributions. In *Advances in Neural Information Processing Systems*, pp. 8499–8511, 2019.

Li, Y. and Turner, R. E. Renyi divergence variational inference. In *Advances in Neural Information Processing Systems*, pp. 1073–1081, 2016.

Li, Y., Turner, R. E., and Liu, Q. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. *arXiv preprint arXiv:1907.00031*, 2019.

Neal, R. M. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

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

Ogata, Y. A monte carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pp. 4277–4285, 2018.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286, 2014.

Ruiz, F. and Titsias, M. A contrastive divergence for combining variational inference and mcmc. In *International Conference on Machine Learning*, pp. 5537–5545, 2019.

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pp. 872–879. ACM, 2008.

Salamon, P., Nulton, J., Siragusa, G., Limon, A., Bedeaux, D., and Kjelstrup, S. A simple example of control to minimize entropy production. *Journal of Non-Equilibrium Thermodynamics*, 27(1):45–55, 2002.

Salimans, T., Kingma, D. P., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. *Proceedings of the 32nd International Conference on Machine Learning*, (Mcmc):1218–1226, 2015. URL http://arxiv.org/abs/1410.6460.

Sivak, D. A. and Crooks, G. E. Thermodynamic metrics and optimal paths. *Physical review letters*, 108(19):190602, 2012.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Non-reversible parallel tempering: A scalable highly parallel mcmc scheme. *arXiv preprint arXiv:1905.02939*, 2019.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. *International Conference on Learning Representations*, 2018.

Van Erven, T. and Harremos, P. Renyi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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.

Wang, D., Liu, H., and Liu, Q. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pp. 5737–5747, 2018.

Wolf, C., Karl, M., and van der Smagt, P. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2): 150–160, 2011.