A. Conjugate Duality

The Bregman divergence associated with a convex function $f : \Omega \to \mathbb{R}$ can be written as (Banerjee et al., 2005):

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

The family of Bregman divergences includes many familiar quantities, including the KL divergence corresponding to the negative entropy generator $f(p) = -\int p \log p \, d\omega$. 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 an exponential family (e.g. (13)) :

$$D_\psi[\beta_p : \beta_q] = \psi(\beta_p) - \psi(\beta_q) - \langle \beta_p - \beta_q, \nabla_\beta \psi(\beta_q) \rangle \tag{34}$$

This is a common setting in the field of information geometry (Amari, 2016), which introduces dually flat manifold structures based on the natural parameters and the mean parameters.

A.1. KL Divergence as a Bregman Divergence

For an exponential family with partition function $\psi(\beta)$ and sufficient statistics $T(\omega)$ over a random variable ω, the Bregman divergence D_ψ corresponds to a KL divergence. Recalling that $\nabla_\beta \psi(\beta) = \eta_\beta = \mathbb{E}_{\pi_\beta}[T(\omega)]$ from (16), we simplify the definition (34) to obtain

$$\begin{aligned}
D_\psi[\beta_p : \beta_q] &= \psi(\beta_p) - \psi(\beta_q) - \beta_p \cdot \eta_q + \beta_q \cdot \eta_q \\
&= \psi(\beta_p) - \psi(\beta_q) - \mathbb{E}_q[\beta_p \cdot T(\omega)] \\
&\qquad + \mathbb{E}_q[\beta_q \cdot T(\omega)] \\
&= \underbrace{\mathbb{E}_q\big[\beta_q \cdot T(\omega) - \psi(\beta_q)\big] + \mathbb{E}_q[\pi_0(\omega)]}_{\log q(\omega)} \\
&\quad - \underbrace{\mathbb{E}_q\big[\beta_p \cdot T(\omega) - \psi(\beta_p)\big] - \mathbb{E}_q[\pi_0(\omega)]}_{\log p(\omega)} \\
&= \mathbb{E}_q \log \frac{q(\omega)}{p(\omega)} \\
&= D_{KL}[q(\omega) || p(\omega)] \tag{35}
\end{aligned}$$

where we have added and subtracted terms involving the base measure $\pi_0(\omega)$, and used the definition of our exponential family from (13). The Bregman divergence D_ψ is thus equal to the KL divergence with arguments reversed.

A.2. Dual Divergence

We can leverage convex duality to derive an alternative divergence based on the conjugate function ψ^*.

$$\begin{aligned}
\psi^*(\eta) = \sup_\beta \eta \cdot \beta - \psi(\beta) \quad &\implies \quad \eta = \nabla_\beta \psi(\beta) \\
&= \eta \cdot \beta_\eta - \psi(\beta_\eta) \tag{36}
\end{aligned}$$

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 bijective mapping between η and β for minimal exponential families (Wainwright & Jordan, 2008). Thus, a distribution p may be indexed by either its natural parameters β_p or mean parameters η_p.

Noting that $(\psi^*)^* = \psi(\beta) = \sup_\eta \eta \cdot \beta - \psi^*(\eta)$ (Boyd & Vandenberghe, 2004), 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:

$$\begin{aligned}
D_{\psi^*}[\eta_p : \eta_q] &= \psi^*(\eta_p) - \psi^*(\eta_q) - \langle \eta_p - \eta_q, \nabla_\eta \psi^*(\eta_q) \rangle \\
&= \psi^*(\eta_p) - \underline{\psi^*(\eta_q)} - \eta_p \cdot \beta_q + \underline{\eta_q \cdot \beta_q} \\
&= \psi^*(\eta_p) + \underline{\psi(\beta_q)} - \eta_p \cdot \beta_q \tag{37}
\end{aligned}$$

where we have used (36) to simplify the underlined terms. Similarly,

$$\begin{aligned}
D_\psi[\beta_p : \beta_q] &= \psi(\beta_p) - \psi(\beta_q) - \langle \beta_p - \beta_q, \nabla_\beta \psi(\beta_q) \rangle \\
&= \psi(\beta_p) - \underline{\psi(\beta_q)} - \beta_p \cdot \eta_q + \underline{\beta_q \cdot \eta_q} \\
&= \psi(\beta_p) + \underline{\psi^*(\eta_q)} - \beta_p \cdot \eta_q \tag{38}
\end{aligned}$$

Comparing (37) and (38), we see that the divergences are equivalent with the arguments reversed, so that:

$$D_\psi[\beta_p : \beta_q] = D_{\psi^*}[\eta_q : \eta_p] \tag{39}$$

This indicates that the Bregman divergence D_{ψ^*} should also be a KL divergence, but with the same order of arguments. We derive this fact directly in (44), after investigating the form of the conjugate function ψ^*.

A.3. Conjugate ψ^* as Negative Entropy

We first treat the case of an exponential family with no base measure $\pi_0(\omega)$, with derivations including a base measure in App. A.4. For a distribution p in an exponential family, indexed by β_p or η_p, we can write $\log p(\omega) = \beta_p \cdot T(\omega) - \psi(\beta)$. Then, (36) becomes:

$$\begin{aligned}
\psi^*(\eta_p) &= \beta_p \cdot \eta_p - \psi(\beta_p) \tag{40} \\
&= \beta_p \cdot \mathbb{E}_p[T(\omega)] - \psi(\beta_p) \tag{41} \\
&= \mathbb{E}_p \log p(\omega) \tag{42} \\
&= -H_p(\omega) \tag{43}
\end{aligned}$$

since β_p and $\psi(\beta_p)$ are constant with respect to ω. Utilizing $\psi^*(\eta_p) = \mathbb{E}_p \log p(\omega)$ from above, the dual divergence with q becomes:

$$
\begin{aligned}
D_{\psi^*}[\eta_p : \eta_q] &= \psi^*(\eta_p) - \psi^*(\eta_q) - \langle \eta_p - \eta_q, \nabla_\eta \psi^*(\eta_q) \rangle \\
&= \mathbb{E}_p \log p(\omega) - \underline{\psi^*(\eta_q)} - \eta_p \cdot \beta_q + \underline{\eta_q \cdot \beta_q} \\
&= \mathbb{E}_p \log p(\omega) - \eta_p \cdot \beta_q + \underline{\psi(\beta_q)} \\
&= \mathbb{E}_p \log p(\omega) - \mathbb{E}_p[T(\omega) \cdot \beta_q] + \psi(\beta_q) \\
&= \mathbb{E}_p \log p(\omega) - \mathbb{E}_p \log q(\omega) \\
&= D_{KL}[p(\omega)||q(\omega)] \qquad\qquad\qquad\qquad (44)
\end{aligned}
$$

Thus, the conjugate function is the negative entropy and induces the KL divergence as its Bregman divergence (Wainwright & Jordan, 2008).

Note that, by ignoring the base distribution over ω, we have instead assumed that $\pi_0(\omega) := u(\omega)$ is uniform over the domain. In the next section, we illustrate that the effect of adding a base distribution is to turn the conjugate function into a KL divergence, with the base $\pi_0(\omega)$ in the second argument. This is consistent with our derivation of negative entropy, since $D_{KL}[p_\beta(\omega)||u(\omega)] = -H_{p_\beta}(\Omega) + const.$

A.4. Conjugate ψ^* as a KL Divergence

As noted above, the derivation of the conjugate $\psi^*(\eta)$ in (40)-(43) ignored the possibilty of a base distribution in our exponential family. We see that $\psi^*(\eta)$ takes the form of a KL divergence when considering a base measure $\pi_0(\omega)$.

$$
\begin{aligned}
\psi^*(\eta) &= \sup_\beta \beta \cdot \eta - \psi(\beta) \qquad\qquad\qquad (45) \\
&= \beta_\eta \cdot \eta - \psi(\beta_\eta) \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\beta_\eta \cdot T(\omega)] - \psi(\beta_\eta) \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\beta_\eta \cdot T(\omega)] - \psi(\beta_\eta) \pm \mathbb{E}_{\pi_{\beta_\eta}}[\log \pi_0(\omega)] \\
&= \mathbb{E}_{\pi_{\beta_\eta}}[\log \pi_{\beta_\eta(\omega)} - \log \pi_0(\omega)] \\
&= D_{KL}[\pi_{\beta_\eta}(\omega)|| \pi_0(\omega)] \qquad\qquad\qquad (46)
\end{aligned}
$$

Note that we have added and subtracted a factor of $\mathbb{E}_{\pi_{\beta_\eta}} \log \pi_0(\omega)$ in the fourth line, where our base measure $\pi_0(\omega) = q(\mathbf{z}\,|\,\mathbf{x})$ in the case of the TVO. Comparing with the derivations in (41)-(42), we need to include a term of $\mathbb{E}_p \pi_0(\omega)$ in moving to an expected log-probability $\mathbb{E}_p \log p(\omega)$, with the extra, subtracted base measure term transforming the negative entropy into a KL divergence.

In the TVO setting, this corresponds to

$$
\psi^*(\eta) = D_{KL}[\pi_{\beta_\eta}(\mathbf{z}\,|\,\mathbf{x})|| \, q(\mathbf{z}\,|\,\mathbf{x})] \,. \qquad (47)
$$

When including a base distribution, the induced Bregman divergence is still the KL divergence since, as in the derivation of (35), both $\mathbb{E}_p \log p(\omega)$ and $\mathbb{E}_p \log q(\omega)$ will contain terms involving the base distribution $\mathbb{E}_p \log \pi_0(\omega)$.

B. Renyi Divergence Variational Inference

In this section, we show that each intermediate partition function $\log Z_\beta$ corresponds to a scaled version of the Rényi VI objective $\mathcal{L}_\alpha$ (Li & Turner, 2016).

To begin, we recall the definition of Renyi's α divergence.

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

Note that this involves geometric mixtures similar to (14). Pulling out the factor of $\log p(\mathbf{x})$ to consider normalized distributions over $\mathbf{z}\,|\,\mathbf{x}$, we obtain the objective of Li & Turner (2016). This is similar to the ELBO, but instead subtracts a Renyi divergence of order α.

$$
\begin{aligned}
\psi(\beta) &= \log \int q(\mathbf{z}\,|\,\mathbf{x})^{1-\beta} p(\mathbf{x}, \mathbf{z})^\beta d\mathbf{z} \\
&= \beta \log p(\mathbf{x}) - (1 - \beta) D_\beta[p_\theta(\mathbf{z}\,|\,\mathbf{x})||q_\phi(\mathbf{z}\,|\,\mathbf{x})] \\
&= \beta \log p(\mathbf{x}) - \beta \, D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})] \\
&:= \beta \, \mathcal{L}_{1-\beta}
\end{aligned}
$$

where we have used the skew symmetry property $D_\alpha[p||q] = \frac{\alpha}{1-\alpha} D_{1-\alpha}[q||p]$ for $0 < \alpha < 1$ (Van Erven & Harremos, 2014). Note that $\mathcal{L}_0 = 0$ and $\mathcal{L}_1 = \log p_\theta(\mathbf{x})$ as in Li & Turner (2016) and Sec. 3.

C. TVO using Taylor Series Remainders

Recall that in Sec. 4, we have viewed the KL divergence $D_\psi[\beta : \beta']$ as the remainder in a first order Taylor approximation of $\psi(\beta)$ around β'. The TVO objectives correspond to the linear term in this approximation, with the gap in $\text{TVO}_L(\theta, \phi, \mathbf{x})$ and $\text{TVO}_U(\theta, \phi, \mathbf{x})$ bounds amounting to a sum of KL divergences or Taylor remainders. Thus, the TVO may be viewed as a first order method.

Yet we may also ask, what happens when considering other approximation orders? We proceed to show that thermodynamic integration arises from a zero-order approximation, while the symmetrized KL divergence corresponds to a similar application of the fundamental theorem of calculus in the mean parameter space $\eta_\beta = \nabla_\beta \psi(\beta)$. In App. E, we briefly describe how 'higher-order' TVO objectives might be constructed, although these will no longer be guaranteed to provide upper or lower bounds on likelihood.

We will repeatedly utilize the integral form of the Taylor remainder theorem, which characterizes the error in a k-th order approximation of $\psi(x)$ around a, with $\beta \in [a, x]^2$. This identity can be derived using the fundamental theorem of calculus and repeated integration by parts (see, e.g.

---

[2]We use generic variable x, not to be confused with data $\mathbf{x}$, for notational simplicity.

(Kountourogiannis & Loya, 2003) and references therein):

$$R_k(x) = \int_a^x \frac{\nabla_\beta^{(k+1)}\psi(\beta)}{k!}(x-\beta)^k d\beta \qquad (48)$$

C.1. Thermodynamic Integration as 0th Order Remainder

Consider a zero-order Taylor approximation of $\psi(1)$ around $a = 0$, which simply uses $\psi(0)$ as an estimator. Applying the remainder theorem, we obtain the identity (6) underlying thermodynamic integration in the TVO:

$$\psi(1) = \psi(0) + R_0(1) \qquad (49)$$

$$\psi(1) - \psi(0) = \int_0^1 \nabla_\beta \psi(\beta) d\beta \qquad (50)$$

$$\log p_\theta(\mathbf{x}) = \int_0^1 \mathbb{E}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] d\beta \qquad (51)$$

where the last line follows as the definition of $\eta = \nabla_\beta \psi(\beta) = \nabla_\beta \log Z_\beta$ in (16).

Note that this integration is symmetric, in that approximating $\psi(0)$ using $\psi(1)$ leads to an equivalent expression after reversing the order of integration.

C.2. KL Divergence as 1st Order Remainder

We can apply a similar approach to the first order Taylor approximations to reinterpret the TVO bound gaps in (9) and (10), although our remainder expressions will no longer be symmetric. We will thus distinguish between estimating $\psi(x)$ around $a < x$ and $a > x$ using $R_1^{\rightarrow}(x)$ and $R_1^{\leftarrow}(x)$, respectively, with the arrow indicating the direction of integration.

Estimating $\psi(\beta_k)$ using a first order approximation around $a = \beta_{k-1}$ as in the TVO lower bound, the remainder exactly matches the definition of the Bregman divergence in (19):

$$R_1^{\rightarrow}(\beta_k) = \psi(\beta_k) - \big(\underbrace{\psi(\beta_{k-1}) + (\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_{k-1})}_{\text{First-Order Taylor Approx}}\big)$$

$$= \int_{\beta_{k-1}}^{\beta_k} \frac{\nabla_\beta^2 \psi(\beta)}{1!}(\beta_k - \beta)^1 d\beta \qquad (52)$$

where (52) corresponds to the Taylor remainder from (48). Recall that this Bregman divergence $D_\psi[\beta_k : \beta_{k-1}]$ corresponds to a KL divergence $D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]$ and contributes to the gap in $\text{TVO}_L(\theta, \phi, \mathbf{x})$.

Simplifying the Taylor remainder expression, with $\nabla_\beta^2 \psi(\beta) = \text{Var}_{\pi_\beta} \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}$, we obtain an integral repre-

sentation of the KL divergence:

$$D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] = \int_{\beta_{k-1}}^{\beta_k} (\beta_k - \beta)\, \text{Var}_{\pi_\beta} \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\beta \qquad (53)$$

Following similar arguments in the reverse direction, we can obtain an integral form for the TVO upper bound gap $R_1^{\leftarrow}(\beta_{k-1}) = D_{KL}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$ via the first-order approximation of $\psi(\beta_{k-1})$ around $a = \beta_k$.

$$R_1^{\leftarrow}(\beta_{k-1}) = \psi(\beta_{k-1}) - \big(\psi(\beta_k) + (\beta_{k-1} - \beta_k)\nabla_\beta\psi(\beta_k)\big)$$

$$= (\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_k) - (\psi(\beta_k) - \psi(\beta_{k-1}))$$

$$= \int_{\beta_k}^{\beta_{k-1}} \frac{\nabla_\beta^2 \psi(\beta)}{1!}(\beta_{k-1} - \beta)^1 d\beta \qquad (54)$$

Note that the TVO upper bound (10) arises from the second line, with $R_1^{\leftarrow}(\beta_{k-1}) \geq 0$ and $(\beta_k - \beta_{k-1})\nabla_\beta\psi(\beta_k)$ corresponding to a right-Riemann approximation.

Switching the order of integration in (54), we can write the KL divergence as

$$D_{KL}^{\leftarrow}[\pi_{\beta_k}||\pi_{\beta_{k-1}}] = \int_{\beta_{k-1}}^{\beta_k} (\beta - \beta_{k-1})\, \text{Var}_{\pi_\beta} \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\beta \qquad (55)$$

While these integral expressions for the KL divergence may not be immediately intuitive, our use of the Taylor remainder theorem unifies their derivation with that of thermodynamic integration. Alternative derivations may also be found in Dabak & Johnson (2002).

C.3. Symmetrized KL Divergence

Combining the expressions for the KL divergence in Eq. (53) and (55) immediately leads to a known result relating the symmetrized KL divergence to the integral of the Fisher information along the geometric path (Amari, 2016; Dabak & Johnson, 2002).

$$D_{KL}^{\leftrightarrow} = (\beta_k - \beta_{k-1}) \int_{\beta_{k-1}}^{\beta_k} \text{Var}_{\pi_\beta}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\beta\right] \qquad (56)$$

where we have defined the symmetrized KL divergence as:

$$D_{KL}^{\leftrightarrow}[\beta_{k-1}; \beta_k] = D_{KL}^{\rightarrow}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] + D_{KL}^{\leftarrow}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$$

Our goal in this section will be to show that (56) arises from similar 'thermodynamic integration' on the graph of

the mean parameters η_β. Recall that we previously applied the fundamental theorem of calculus to $\psi(\beta) = \log Z_\beta$ to obtain the difference in log-partition functions

$$\psi(\beta_k) - \psi(\beta_{k-1}) = \int_{\beta_{k-1}}^{\beta_k} \nabla_\beta \psi(\beta) d\beta$$

We can obtain a similar expression for the mean parameters $\eta_\beta = \nabla_\beta \psi(\beta)$ by integrating over the second derivative.

$$\eta_k - \eta_{k-1} = \int_{\beta_{k-1}}^{\beta_k} \nabla_\beta^2 \psi(\beta) d\beta \tag{57}$$

Recalling that $\nabla_\beta^2 \psi(\beta) = \mathrm{Var}_{\pi_\beta} \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}$, we see that the integrands in (56) and (57) are identical. Integrating with respect to β, we obtain the 'area of a rectangle' identity for the symmetrized KL divergence (as in (30)):

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

$$= (\beta_k - \beta_{k-1}) \int_{\beta_{k-1}}^{\beta_k} \nabla_\beta^2 \psi(\beta) d\beta$$

$$= (\beta_k - \beta_{k-1}) \left(\nabla_\beta \psi(\beta) \big|_{\beta_{k-1}}^{\beta_k} \right)$$

$$= (\beta_k - \beta_{k-1})(\eta_k - \eta_{k-1}) \tag{58}$$

This identity is best understood via Fig. 5 in Sec. 4.4.

To summarize, we have given several equivalent ways of understanding the symmetrized KL divergence. The 'forward' and 'reverse' KL divergences arise as gaps in the TVO left- and right-Riemann approximations (Figure 5), or first order Taylor remainders as in (53) and (55). Summing these quantities corresponds to the area of a rectangle (58) on the graph of the TVO integrand η_β, or to the integral of a variance term via the Taylor remainder theorem (56) or fundamental theorem of calculus (57).

Note that the TVO integrand $\eta_\beta = \nabla_\beta \psi(\beta) = \mathbb{E}_{\pi_\beta}[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}]$ will be linear when its derivative, the variance of the log importance weights, is constant within $\beta \in [\beta_{k-1}, \beta_k]$. The KL divergence is actually symmetric in this case, which we treat in more detail in the next section (App. D). More generally, the curvature of the integrand indicates which direction of the KL divergence has larger magnitude, and Figure 5 reflects our empirical observations that $D_{KL}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] > D_{KL}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$.

D. Asymptotic Linear Scheduling Analysis

Grosse et al. (2013) treat a quantity identical to $\mathrm{TVO}_L(\theta, \phi, \mathbf{x})$ in the context of analysing the variance of

AIS estimators. Using the Central Limit Theorem, Neal (2001) show that the variance of an AIS estimator is monotonically related to $\mathrm{TVO}_L(\theta, \phi, \mathbf{x})$ under perfect transitions, or independent, exact samples from each intermediate β (see Grosse et al. (2013) Eq. 3). However, note that AIS estimates expectations over *chains* of MCMC samples rather than the simple reweighting used in the TVO.

In this section, we provide additional perspective on the analysis of Grosse et al. (2013), which considers the asymptotic behavior of the scaled gap in $\mathrm{TVO}_L(\theta, \phi, \mathbf{x})$, $K \cdot D_{\overset{\rightarrow}{KL}}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]$, as $K \to \infty$.

We begin by restating Theorem 1 of Grosse et al. (2013) for the case of the full TVO objective. We describe the resulting 'coarse-grained' linear binning schedule for choosing $\{\beta_k\}$ in D.1 and provide further analysis in D.2.

Theorem 1 (Grosse et al. (2013)). *Suppose $K + 1$ distributions $\{\pi_{\beta_k}\}_{k=0}^K$ are linearly spaced along a path $\mathcal{P}$. Under the assumption of perfect transitions, if the Fisher information matrix $G(\beta)$ is smooth, then as $K \to \infty$:*

$$K \sum_{k=1}^K D_{\overset{\rightarrow}{KL}}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] \to \frac{1}{2} \int_0^1 \dot{\beta}(t) \cdot G(\beta(t)) \cdot \dot{\beta}(t) dt \tag{59}$$

$$= \frac{1}{2}\left(D_{\overset{\rightarrow}{KL}}[\pi_{\beta_0}||\pi_{\beta_K}] + D_{\overset{\leftarrow}{KL}}[\pi_{\beta_K}||\pi_{\beta_0}] \right)$$

Here, we let $t \in [0, 1]$ parameterize the path $\beta(t) = (1 - t) \cdot \beta_0 + t \cdot \beta_K$, and let $\dot{\beta}(t)$ denote the derivative of the parameter β with respect to t. For linear mixing of the natural parameters as above, this is a constant: $\dot{\beta}(t) = \beta_K - \beta_0$. In the case of the full TVO integrand, $\dot{\beta}(t) = 1$.

Proof. See (Grosse et al., 2013) for a detailed proof, which proceeds by taking the Taylor expansion of $D_{KL}[\beta_k||\beta_k + \Delta_\beta]$ around each β_k for small Δ_β. In particular, $\Delta_\beta = \frac{1}{K}(\beta_K - \beta_0)$ for linearly spaced $\beta_k = (1 - \frac{k}{K}) \cdot \beta_0 + \frac{k}{K} \cdot \beta_K$. We assume w.l.o.g. $\beta_K - \beta_0 = 1$ and $\Delta_\beta = \frac{1}{K}$ as in Grosse et al. (2013) or TVO.

The zero- and first-order terms vanish, and the second-order term, with $\Delta_\beta^2 = \frac{1}{K^2}$, can be written as (see e.g. Kullback (1997) p. 26):

$$K \sum_{k=1}^K D_{KL}[\beta_k||\beta_k + \Delta_\beta] = K \cdot \frac{1}{2K^2} \sum_{k=1}^K \dot{\beta}_k \cdot G(\beta_k) \cdot \dot{\beta}_k$$
$$+ K \cdot \mathcal{O}(K^{-3}) \tag{60}$$

$$\to \frac{1}{2} \int_0^1 \dot{\beta}(t) G(\beta(t)) \dot{\beta}(t) dt \tag{61}$$

where we have absorbed $\Delta_\beta = \frac{1}{K}$ into a continuous measure dt as $K \to \infty$.

We now show that this expression corresponds to the symmetrized KL divergence, as in (Amari, 2016; Dabak & Johnson, 2002). While this was not stated in the theorem of Grosse et al. (2013), it has also been shown by e.g. Huszar (2017). Observe that $G(\beta) = \nabla_\beta \psi(\beta) = \mathrm{Var}_{\pi_\beta}[T(\mathbf{x}, \mathbf{z})]$ as in (56) and (58). Noting that the chain rule implies $\frac{d}{dt} G(\beta(t)) = \frac{d}{d\beta} G(\beta(t)) \frac{d\beta}{dt}$, we can pull one term of $\frac{d\beta}{dt} = \dot{\beta}(t) = (\beta_K - \beta_0)$ outside the integral and perform integration by substitution. Ignoring the $1/2$ factor,

$$\frac{(\beta_K - \beta_0)}{2} \int_0^1 G(\beta(t)) \frac{d\beta}{dt} dt = \frac{(\beta_K - \beta_0)}{2} \int_{\beta_0}^{\beta_K} \nabla_\beta^2 \psi(\beta) d\beta$$

$$= \frac{1}{2}(\beta_K - \beta_0)(\eta_K - \eta_0) \qquad (62)$$

$$= \frac{1}{2}\left(D_{\overrightarrow{KL}}[\pi_{\beta_0}||\pi_{\beta_K}] + D_{\overleftarrow{KL}}[\pi_{\beta_K}||\pi_{\beta_0}]\right)$$

$$\qquad\qquad\qquad\qquad\qquad\qquad \square$$

D.1. 'Coarse-Grained' Linear Schedule

Grosse et al. (2013) then use this asymptotic condition (62) as $K \to \infty$ to inform the choice of a *discrete* partition $\mathcal{P} = \{\beta_k\}_{k=0}^K$.

More concretely, consider dividing the interval $[0, 1]$ into J equally-spaced knot points $\{\beta_j\}_{j=0}^J$. We then allocate a total budget of $K = \sum_{j=1}^J K_j$ intermediate distributions across sub-intervals $[\beta_{j-1}, \beta_j]$, with uniform linear spacing of the K_j partitions within each sub-interval.

Using (62), Grosse et al. (2013) assign a cost $F_j = (\beta_j - \beta_{j-1})(\eta_j - \eta_{j-1})$ to each 'coarse-grained' interval $[\beta_{j-1}, \beta_j]$. Minimizing $\sum_j F_j$ subject to $\sum_j K_j = K$, the allocation rule becomes:

$$K_j \propto \sqrt{(\beta_{j+1} - \beta_j)(\eta_{j+1} - \eta_j)} \qquad (63)$$

We observe that performance when using this method can be sensitive to the number of knot points used, and we found $J = 20$ to perform best in our experiments.

D.2. Additional Perspectives on Grosse et al. (2013)

Geometric Intuition for Theorem 1: To further understand Theorem 1 of Grosse et al. (2013), observe that the TVO integrand will appear linear within any interval $[\beta_{k-1}, \beta_k]$ as $K \to \infty$. For general endpoints β_0 and β_K, we let $\Delta_\beta = \beta_k - \beta_{k-1} = \frac{\beta_K - \beta_0}{K}$.

Having already visualized the symmetrized KL divergence as the area of a rectangle in Figure 5, we can see that each directed KL divergence, $D_{\overrightarrow{KL}}[\pi_{\beta_{k-1}}||\pi_{\beta_k}]$ and $D_{\overleftarrow{KL}}[\pi_{\beta_k}||\pi_{\beta_{k-1}}]$, will approach the area of *triangle* as the integrand becomes linear or $K \to \infty$, with area equal to

$1/2 \cdot \Delta_\beta \cdot \Delta_\eta$. Then, the D_{KL} scaled by K becomes

$$K \sum_{k=1}^K D_{\overrightarrow{KL}}[\pi_{\beta_{k-1}}||\pi_{\beta_k}] \to K \sum_{k=1}^K \frac{1}{2} \cdot \Delta_\beta \cdot \Delta_\eta \qquad (64)$$

$$= K \sum_{k=1}^K \frac{1}{2}(\frac{\beta_K - \beta_0}{K}) \cdot (\eta_k - \eta_{k-1})$$

$$= \frac{1}{2}(\beta_K - \beta_0) \cdot (\eta_K - \eta_0).$$

where, in the last line, we cancel factors of K and note the cancellation of intermediate η_k in the telescoping sum.

Thermodynamic Interpretation: This limiting behavior is also discussed in thermodynamics, where the LHS of (64) and (59) corresponds to the rate of entropy production in transitioning a system from π_{β_0} to π_{β_1} along a path defined by $\{\beta_k\}$. The condition that $K \to \infty$ refers to the linear response regime, with (59) related to the thermodynamic divergence (Crooks, 2007).

Exponential and Mixture Geodesics: As in the statement of Theorem 1, we can more generally consider connecting two distributions, indexed by natural parameters β_0 and β_1, using a parameter $t \in [0, 1]$. The curve $\beta_t = (1 - t) \cdot \beta_0 + t \cdot \beta_1$ then corresponds to our path exponential family (13), and is also referred to as the e-geodesic in information geometry Amari (2016).

Similarly, the moment-averaged path of Grosse et al. (2013), which also underlies our scheduling strategy in Sec. 5, can be viewed as a linear mixture in the mean parameter space. The m-geodesic then refers to the curve $\eta_t = (1 - t) \cdot \eta_0 + t \cdot \eta_1$ (Amari, 2016). Note that these mixtures reference different distribution for the same parameter t, so that $\eta_{\beta_t} \neq \eta_t$.

Grosse et al. (2013) proceed to show that the expression for the symmetric KL divergence (59) corresponds to the integral of the Fisher information along *either* the geometric or mixture paths (Theorem 2 of Grosse et al. (2013), Theorem 3.2 of Amari (2016)). The union of the intermediate distributions integrated by these two paths coincide in our one-dimensional exponential family, although this intuition does not appear to translate to higher dimensions.

E. Higher Order TVO

While the convexity of the log-partition function yields the family of Bregman divergences from the remainder in the first order Taylor approximation, we might also consider higher order terms to obtain tighter bounds on likelihood or analyse properties of the TVO integrand $\nabla_\beta \psi(\beta)$. We give an example derivation for a second-order TVO objectives, although these are no longer guaranteed to be upper or lower bounds on likelihood.

Left-to-Right Expansion We first consider expanding the approximations in the TVO left-Riemann sum to second order. We denote the resulting objective $\mathcal{L}_{\rightarrow}^{(2)}$, since we move 'left-to-right' in estimating $\psi(\beta_k)$ around β_{k-1}. We begin by writing the second-order Taylor approximation:

$$\psi(\beta_k) \approx \psi(\beta_{k-1}) + (\beta_k - \beta_{k-1}) \, \nabla_\beta \psi(\beta_{k-1})$$
$$+ \frac{1}{2}(\beta_k - \beta_{k-1})^2 \, \nabla_\beta^2 \psi(\beta_{k-1}) \qquad (65)$$

While $\text{TVO}_L(\theta, \phi, \mathbf{x})$ consists of the first-order term alone, we can also consider adding the non-negative, second-order term to form the objective $\mathcal{L}_{\rightarrow}^{(2)}$. Using successive Taylor approximations of $\psi(\beta_k)$, we obtain similar telescoping cancellations to obtain

$$\log p(\mathbf{x}) - \mathcal{L}_{\rightarrow}^{(2)} = \log p(\mathbf{x}) - \sum_{k=1}^{K}(\beta_k - \beta_{k-1}) \cdot \eta_{\beta_{k-1}}$$
$$- \sum_{k=1}^{K}\frac{1}{2}(\beta_k - \beta_{k-1})^2 \, \text{Var}_{\pi_{\beta_{k-1}}} \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z} \,|\, \mathbf{x})}$$
$$(66)$$

where $\eta_{\beta_{k-1}} = \mathbb{E}_{\pi_{\beta_{k-1}}} \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z} \,|\, \mathbf{x})}$.

We previously obtained a lower bound on log-likelihood via this construction, with $\log p(\mathbf{x}) - \text{TVO}_L(\theta, \phi, \mathbf{x}) \geq 0$. However, $\mathcal{L}_{\rightarrow}^{(2)}$ will only provide a lower bound if $\nabla_\beta \psi(\beta)$ is concave, i.e. $\nabla_\beta^3 \psi(\beta) \leq 0$. To see this, we write the Taylor remainder (48) as

$$R_2^{\rightarrow}(\beta_k) = \int_{\beta_{k-1}}^{\beta_k} \frac{1}{6}(\beta_k - \beta_t)^3 \, \nabla_\beta^3 \psi(\beta_t) d\beta_t \qquad (67)$$

with the third derivative equal to

$$\nabla_\beta^3 \psi(\beta) = \mathbb{E}_{\pi_\beta}[T(\mathbf{x}, \mathbf{z})^3] - 3 \left[\mathbb{E}_{\pi_\beta} T(\mathbf{x}, \mathbf{z}) \right] \cdot \mathbb{E}_{\pi_\beta}[T(\mathbf{x}, \mathbf{z})^2]$$
$$+ 2[\mathbb{E}_{\pi_\beta} T(\mathbf{x}, \mathbf{z})]^3$$

$$= \mathbb{E}_{\pi_\beta}[T(\mathbf{x}, \mathbf{z})^3] - [\mathbb{E}_{\pi_\beta} T(\mathbf{x}, \mathbf{z})]^3$$
$$- 3[\mathbb{E}_{\pi_\beta} T(\mathbf{x}, \mathbf{z})] \cdot [\text{Var}_{\pi_\beta} T(\mathbf{x}, \mathbf{z})]$$

In addition to indicating that $\mathcal{L}_{\rightarrow}^{(2)}$ is a lower bound on $\log p_\theta(\mathbf{x})$, testing the concavity of $\nabla_\beta \psi(\beta)$ using $\nabla_\beta^3 \psi(\beta) \leq 0$ can also indicate whether a trapezoid approximation to the TVO integral provides a valid lower bound.

We can give an identical construction for the reverse direction $\mathcal{L}_{\leftarrow}^{(2)}$ or higher order approximations. We leave a full exploration of these objectives for future work.

F. Experimental Setup

Code for all experiments can be found at https://github.com/vmasrani/tvo_all_in.

Model Following (Burda et al., 2015), we use a variational autoencoder (Kingma & Welling, 2013) with a 50-dimensional stochastic layer, $\mathbf{z} \in \mathcal{R}^{50}$

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z})$$
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \,|\, 0, \mathbf{I})$$
$$p_\theta(\mathbf{x} \,|\, \mathbf{z}) = \text{Bern}(\mathbf{x} \,|\, \text{decoder}_\theta(\mathbf{z}))$$
$$q_\phi(\mathbf{z} \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}_\phi(\mathbf{x}))$$

where the encoder and decoder are each two-layer MLPs with tanh activations and 200 hidden dimensions. The output of the encoder is duplicated and passed through an additional linear layer to parameterize the mean and log-standard deviation of a conditionally independent Normal distribution. The output of the decoder is a sigmoid which parameterizes the probabilities of the independent Bernoulli distribution. θ and ϕ refer to the weights of the decoder and encoder, respectively.

Dataset We use Omniglot (Lake et al., 2013), a dataset of 1623 handwritten characters across 50 alphabets. Each datapoint is binarized 28×28 image, i.e $\mathbf{x} \in \{0, 1\}^{784}$, where we follow the common procedure in the literature of sampling each binary-valued observation with expectation equal to the real pixel value (Salakhutdinov & Murray, 2008; Burda et al., 2015). We split the dataset into 24,345 training and 8,070 test examples.

Training Procedure All models are written in PyTorch and trained on GPUs. For each scheduler, we train for 5000 epochs using the Adam optimizer (Kingma & Ba, 2017) with a learning rate of 10^{-3}, and minibatch size of 1000. All weights are initialized with PyTorch's default initializer.

G. Implementation Details

While the Legendre transform, mapping between a target value of expected sufficient statistics $\eta = \mathbb{E}_{\pi_\beta}[T(\omega)]$ and the appropriate natural parameters β, can be a difficult problem in general, we describe how to efficiently implement our 'moments-spacing' schedule in the context of TVO.

Recall from Sec. 5 that we are interested in finding a discrete partition $\mathcal{P}_\beta = \{\beta_k\}_{k=0}^{K}$ such that:

$$\beta_k = \eta_\beta^{-1} \left((1 - \frac{k}{K}) \cdot \text{ELBO} + \frac{k}{K} \cdot \text{EUBO} \right) \qquad (68)$$

In other words, we seek to find the β_k such that $\mathbb{E}_{\pi_{\beta_k}}[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \,|\, \mathbf{x})}] \approx \eta_k$, where η_k are equally spaced between the ELBO and EUBO (see Figure 6).

More concretely, we provide pseudo-code implementing our moments spacing schedule below. Given a set of S log-importance weights per sample, and a number of intermediate distributions K:

Figure 10. Pseudo-code Implementation of Moments Scheduling for TVO

```
1  # Calculate expected sufficient statistics       1
      eta at a given beta (Eq. 12)                   2  def moments_spacing_schedule(log_iw, K,
2  def calc_eta(log_iw, beta):                              search='binary'):
3      # 1) Exponentiate/normalize over             3      # 1) Calculate target values for uniform
       importance sample dimension                          moments spacing
4      snis = torch.exp(log_iw*beta -               4      elbo = calc_eta(log_iw, 0)
5          torch.logsumexp(log_iw*beta,             5      eubo = calc_eta(log_iw, 1)
6              dim = 1, keepdim=True))              6      targets = [(1-t)*elbo+t*eubo
7      # 2) Take mean over data examples            7        for t in np.linspace(0,1,K+1)]
8      return torch.mean(snis*log_iw, dim =0)       8
9                                                    9      # 2) Find beta corresponding to each
10 def binary_search(target, log_iw, start=0,          target (including beta=0,1)
      stop=1, threshold = 0.1):                    10      beta_schedule = [0]
11                                                   11
12     beta_guess = .5*(stop-start)                 12      for _k in range(1, K):
13     eta_guess = calc_eta(log_iw,beta_guess)      13          target_eta = targets[_k]
14     if eta_guess > target + threshold:           14
15         return binary_search(                    15          beta_k = binary_search(
16             target,                              16                  target_eta,
17             log_iw,                              17                  log_iw,
18             start=beta_guess,                    18                  start = 0,
19             stop=stop)                           19                  stop = 1)
20     elif eta_guess < target - threshold:         20
21         return binary_search(                    21          beta_schedule.append(beta_k)
22             target,                              22
23             log_iw,                              23      beta_schedule.append(1)
24             start=start,                         24      # 3) Return beta_schedule: used for
25             stop=beta_guess)                         Riemann approximation points in TVO
26     else:                                            objective
27         return beta_guess                        25      return beta_schedule
```

H. Additional Results

In this section, we report wall-clock runtimes and run similar experiments as in Sec. 8 to evaluate our moments spacing schedule and reparameterized gradients on the binarized MNIST dataset (Salakhutdinov & Murray, 2008).

Wall-Clock Times We report wall clock runtimes for various scheduling methods with $S = 50$ and $K = 5$ in Fig. 11. While TVO methods require slight overhead compared with IWAE, our adaptive moments scheduler does not require significantly more computation than the log-uniform baseline.

Grid Search Comparison We evaluate our moments schedule with $K = 2$ against grid search over the choice of a single intermediate β_1 in Fig. 12. The setup is similar to that of Fig. 1 on Omniglot (see Sec. 8), but here we use reparameterization gradients instead of the original TVO. Here, we train for 1000 epochs using an Adam optimizer with learning rate 10^{-3} and batch size 100.

We again find that our moments spacing schedule arrives at an optimal choice of β_1, and can even outperform the best static value due to its ability to adaptively update at each epoch. It is interesting to note that the final choice of β_1, which reflects the shape of the TVO integrand, is nearly identical at $\beta_1 \approx 0.30$ across both MNIST and Omniglot.

Comparison with IWAE We compare TVO using our moments scheduling against the IWAE and IWAE DREG as in Fig. 9 of the main text. We find that our TVO reparameterized gradient estimator achieves nearly identical model learning performance as IWAE and IWAE DREG, with notably improved posterior inference for all values of S.

Evaluating Scheduling Strategies In the Fig. 14-18 below, we reproduce the setting of Fig. 8 to evaluate our scheduling strategies by K, for TVO with both REINFORCE and reparameterized gradient estimators, on Omniglot and MNIST. We also report posterior inference results as measured by test $D_{KL}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]$. In general, we find comparable performance between our moments schedule and the log-uniform baseline, although our approach performs best with $K = 2$ and does not require grid search. Further, on MNIST with batch size 1000 and low K, log-uniform, linear, and coarse-grained schedules suffer from poor performance due to instability in training, which is avoided by our moments schedule. Training can be stabilized by using smaller batch sizes as in Fig. 12.
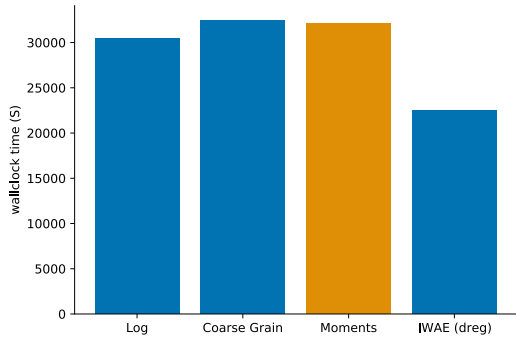
Figure 11. Omniglot Runtimes ($S = 50$, $K = 5$, 5k epochs)

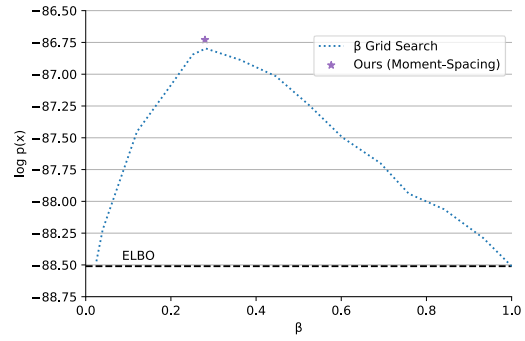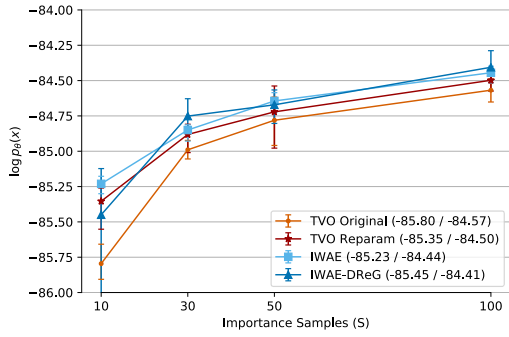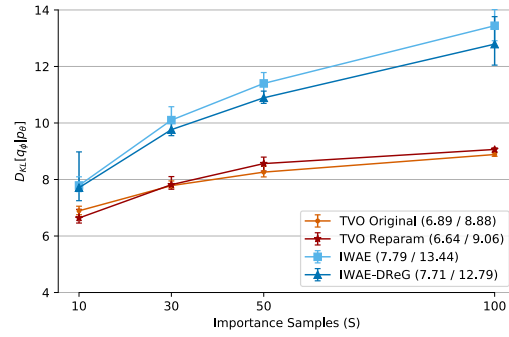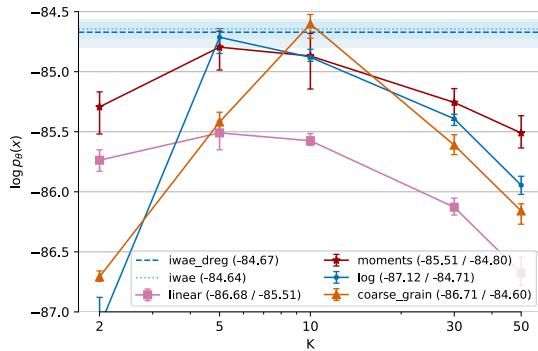Figure 12. MNIST $K = 2$, with reparameterization gradients.
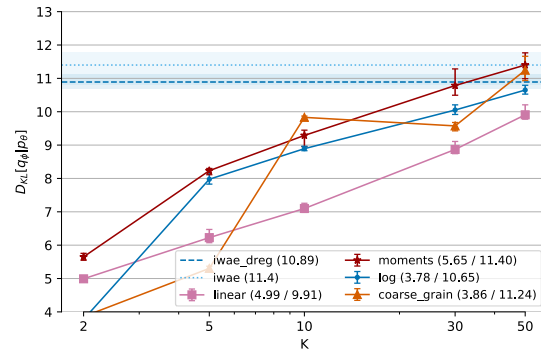

(a) MNIST Test $\log p_\theta(\mathbf{x})$

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

Figure 13. Model Learning and Inference by S (with $K = 5$)
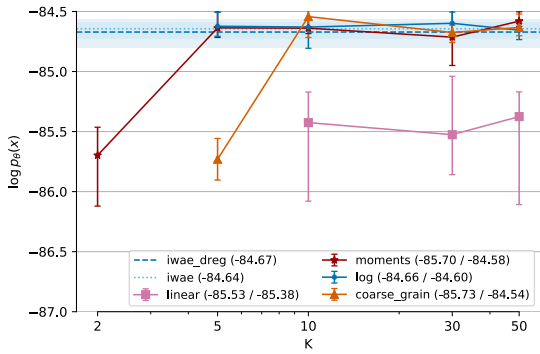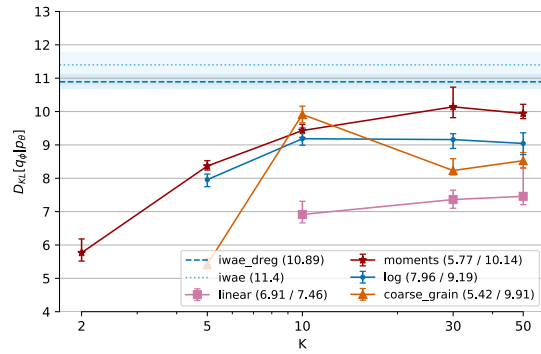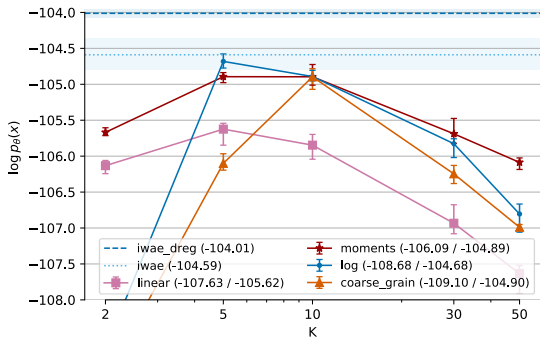

(a) MNIST Test $\log p_\theta(\mathbf{x})$

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

Figure 14. TVO with REINFORCE Gradients: Model Learning and Inference by K (with $S = 50$)

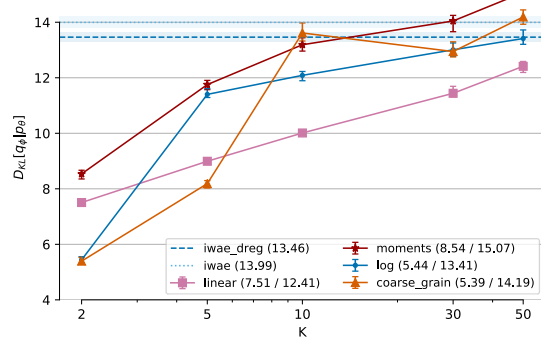(a) MNIST Test $\log p_\theta(\mathbf{x})$

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

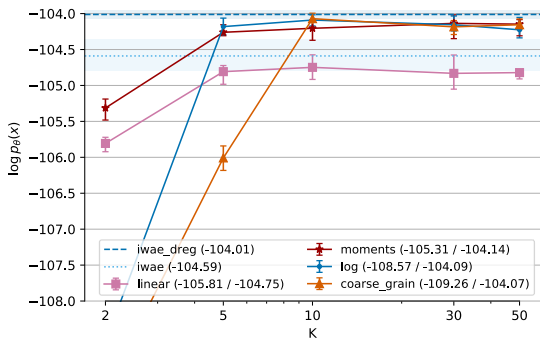*Figure 15.* TVO with Reparameterized Gradients: Model Learning and Inference by K (with $S=50$)
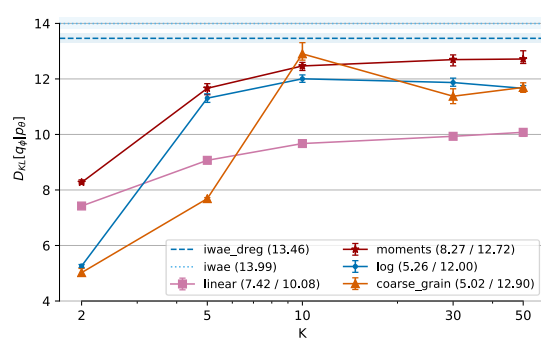

(a) Omniglot Test $\log p_\theta(\mathbf{x})$

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

Figure 17. TVO with REINFORCE Gradients: Model Learning and Inference by K (with $S=50$)

(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 18. TVO with Reparameterized Gradients: Model Learning and Inference by K (with $S=50$)

I. Reparameterization Gradients for the TVO Integrand

Recall that the TVO objective involves terms of the form

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] \quad \text{where} \quad \pi_\beta(\mathbf{z}\,|\,\mathbf{x}) = \frac{q_\phi(\mathbf{z}\,|\,\mathbf{x})^{1-\beta}p_\theta(\mathbf{x},\mathbf{z})^\beta}{Z_\beta} \quad \text{and} \quad f(\mathbf{z}) = \log\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})} \tag{69}$$

While Masrani et al. (2019) derive a REINFORCE-style gradient estimator for the TVO, we seek to apply the reparameterization trick when possible, and thus differentiate with respect to only the inference network parameters ϕ. Note that, for $\mathbf{z}_i \sim q_\phi(\mathbf{z}\,|\,\mathbf{x})$ reparameterizable with $\mathbf{z} = z(\epsilon, \phi)$ and $\epsilon_i \sim p(\epsilon)$, any expectation under π_β can be written as

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] = \frac{1}{Z_\beta}\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[w^\beta f(\mathbf{z})\right] = \frac{1}{Z_\beta}\mathbb{E}_\epsilon\left[w^\beta f(\mathbf{z})\right] \quad \text{where} \quad w = \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})} \tag{70}$$

In differentiating (70), we will frequently encounter terms of the form $\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{d}{d\phi}\log\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right]$ for generic $f(\mathbf{z})$. Noting that the total derivative contains score function partial derivatives, we apply the reparameterization trick to these terms in an approach similar to the 'doubly-reparameterized' estimator of Tucker et al. (2018). The following lemma summarizes these calculations, rewritten using expectations under π_β as in (70).

Lemma 1. *Let* $f(\mathbf{z}) : \mathbb{R}^M \mapsto \mathbb{R}$, $\pi_\beta(\mathbf{z}\,|\,\mathbf{x})$, *and* $w = \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}$ *all depend on* ϕ. *When* $\mathbf{z} \sim q_\phi(\mathbf{z}\,|\,\mathbf{x})$ *is reparameterizable via* $\mathbf{z} = z(\epsilon, \phi)$, $\epsilon \sim p(\epsilon)$, *the following identity holds for expectations under* π_β

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{d}{d\phi}\log w\right] = \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left((1-\beta)f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}} - \frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right)\right]. \tag{71}$$

Proof. See Appendix I.3. □

Corollary 1.1. *For the choice of* $f(\mathbf{z}) = 1$ *we obtain*

$$\mathbb{E}_{\pi_\beta}\left[\frac{d}{d\phi}\log w\right] = (1-\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]. \tag{72}$$

The following lemma will allow us to apply reparameterization within the normalization constant.

Lemma 2. *Let the same conditions hold as in Lemma 1, with* $Z_\beta = \int q_\phi(\mathbf{z}\,|\,\mathbf{x})^{1-\beta}p_\theta(\mathbf{x},\mathbf{z})^\beta\,d\mathbf{z}$. *Then*

$$\frac{d}{d\phi}Z_\beta = \beta(1-\beta)\mathbb{E}_\epsilon\left[w^\beta\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]. \tag{73}$$

Proof. See Appendix I.4. □

We now proceed to differentiate the TVO integrand given by (69).

I.1. Reparameterized TVO Gradient Estimator

For generic $f(\mathbf{z}) : \mathbb{R}^M \mapsto \mathbb{R}$ and reparameterizable $\mathbf{z} \sim q_\phi(\mathbf{z}\,|\,\mathbf{x})$ as above, the gradient with respect to ϕ can be written as

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] = \mathbb{E}_{\pi_\beta}\left[\left(\frac{d}{d\phi}f(\mathbf{z})\right) - \beta\left(\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right)\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[f(\mathbf{z}), \frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]. \tag{74}$$

The gradient of the TVO integrand is of particular interest. For $f(\mathbf{z}) = \log w$ with $w = \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}$, (74) simplifies to

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}\left[\log w\right] = (1-2\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[\log w, \frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]. \tag{75}$$

Proof. We track changes between lines in blue, and begin by applying the product rule.

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] = \frac{d}{d\phi}\left(Z_\beta^{-1}\mathbb{E}_\epsilon\left[w^\beta f(\mathbf{z})\right]\right) \tag{76}$$

$$= \left(\frac{d}{d\phi}Z_\beta^{-1}\right)\mathbb{E}_\epsilon\left[w^\beta f(\mathbf{z})\right] + Z_\beta^{-1}\mathbb{E}_\epsilon\left[f(\mathbf{z})\left(\frac{d}{d\phi}w^\beta\right)\right] + Z_\beta^{-1}\mathbb{E}_\epsilon\left[w^\beta\left(\frac{d}{d\phi}f(\mathbf{z})\right)\right] \tag{77}$$

$$= \left(\frac{d}{d\phi}Z_\beta\right)\left(\frac{-1}{Z_\beta^2}\right)\mathbb{E}_\epsilon\left[w^\beta f(\mathbf{z})\right] + Z_\beta^{-1}\mathbb{E}_\epsilon\left[\beta w^\beta f(\mathbf{z})\left(\frac{d}{d\phi}\log w\right)\right] + Z_\beta^{-1}\mathbb{E}_\epsilon\left[w^\beta\left(\frac{d}{d\phi}f(\mathbf{z})\right)\right] \tag{78}$$

$$= \underbrace{\left(\frac{d}{d\phi}Z_\beta\right)\left(\frac{-1}{Z_\beta}\right)\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right]}_{①} + \underbrace{\beta\,\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{d}{d\phi}\log w\right]}_{②} + \mathbb{E}_{\pi_\beta}\left[\frac{d}{d\phi}f(\mathbf{z})\right] \tag{79}$$

We proceed to simplify only the first two terms, applying Lemma 2 to ① and Lemma 1 to ②.

$$① + ② = \underbrace{\beta(1-\beta)\mathbb{E}_\epsilon\left[w^\beta\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]}_{\text{Lemma 2}}\left(\frac{-1}{Z_\beta}\right)\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] + \beta\underbrace{\left((1-\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}f(\mathbf{z})\right] - \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right]\right)}_{\text{Lemma 1}} \tag{80}$$

$$= \beta(1-\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right](-1)\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] + \beta(1-\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}f(\mathbf{z})\right] - \beta\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right] \tag{81}$$

$$= \beta(1-\beta)\left(\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}f(\mathbf{z})\right] - \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right]\right) - \beta\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right] \tag{82}$$

$$= \beta(1-\beta)\left(\text{Cov}_{\pi_\beta}\left[f(\mathbf{z}),\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]\right) - \beta\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right]. \tag{83}$$

By plugging (83) back into (79) we arrive at the reparameterized gradient for general $f(\mathbf{z})$ (74).

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\right] = \beta(1-\beta)\left(\text{Cov}_{\pi_\beta}\left[f(\mathbf{z}),\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]\right) - \beta\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right] + \mathbb{E}_{\pi_\beta}\left[\frac{d}{d\phi}f(\mathbf{z})\right] \tag{84}$$

$$= \mathbb{E}_{\pi_\beta}\left[\left(\frac{d}{d\phi}f(\mathbf{z})\right) - \beta\left(\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right)\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[f(\mathbf{z}),\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right]. \tag{85}$$

Finally, to optimize the TVO integrand, we can substitute $f(\mathbf{z}) = \log w$ for various terms in (85). We then use Corollary 1.1 to apply the reparameterization trick within the total derivative in the first term.

$$\frac{d}{d\phi}\mathbb{E}_{\pi_\beta}\left[\log w\right] = \mathbb{E}_{\pi_\beta}\left[\left(\frac{d}{d\phi}\log w\right) - \beta\left(\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[\log w,\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right] \tag{86}$$

$$= \mathbb{E}_{\pi_\beta}\left[\underbrace{(1-\beta)\left(\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right)}_{\text{Corollary 1.1}} - \beta\left(\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[\log w,\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right] \tag{87}$$

$$= (1-2\beta)\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right] + \beta(1-\beta)\text{Cov}_{\pi_\beta}\left[\log w,\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\mathbf{z}}\right] \tag{88}$$

This establishes (75) and is the expression that we use to optimize the TVO with reparameterization in the main text. $\square$

I.2. REPARAM / REINFORCE Equivalence for π_β

It is well known (Tucker et al., 2018) that the reparameterization trick and REINFORCE estimator are equivalent for expectations under $q_\phi(\mathbf{z}\,|\,\mathbf{x})$, which allows us to trade high variance REINFORCE gradients for reparameterization gradients which directly consider derivatives of the function $f(\mathbf{z})$.

$$\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[f(\mathbf{z})\frac{\partial}{\partial\phi}\log q_\phi(\mathbf{z}\,|\,\mathbf{x})\right] = \mathbb{E}_\epsilon\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right]. \tag{89}$$

We use this equivalence to show a similar result for expectations under π_β, which we will then use in the proofs of Lemma 1 in I.3 and Lemma 2 in I.4.

Lemma 3. *Let the same conditions hold as in Lemma 1. Then*

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{\partial}{\partial\phi}\log q_\phi(\mathbf{z}\,|\,\mathbf{x})\right] = \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + \beta f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}}\right)\right]. \tag{90}$$

Proof.

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{\partial}{\partial\phi}\log q_\phi(\mathbf{z}\,|\,\mathbf{x})\right] = \frac{1}{Z_\beta}\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[w^\beta f(\mathbf{z})\frac{\partial}{\partial\phi}\log q_\phi(\mathbf{z}\,|\,\mathbf{x})\right] \qquad \text{Using (70)} \tag{91}$$

$$= \frac{1}{Z_\beta}\mathbb{E}_\epsilon\left[\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial(w^\beta f(\mathbf{z}))}{\partial\mathbf{z}}\right] \qquad \text{Using (89)} \tag{92}$$

$$= \frac{1}{Z_\beta}\mathbb{E}_\epsilon\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(w^\beta\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + f(\mathbf{z})\frac{\partial w^\beta}{\partial\mathbf{z}}\right)\right] \tag{93}$$

$$= \frac{1}{Z_\beta}\mathbb{E}_\epsilon\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(w^\beta\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + f(\mathbf{z})\beta w^\beta\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] \tag{94}$$

$$= \frac{1}{Z_\beta}\mathbb{E}_\epsilon\left[w^\beta\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + f(\mathbf{z})\beta\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] \tag{95}$$

$$= \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + \beta f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] \qquad \text{Using (70)} \tag{96}$$

$\square$

I.3. Proof of Lemma 1

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{d}{d\phi}\log w\right] = \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left((1-\beta)f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}} - \frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right)\right]. \tag{97}$$

Proof. Using the fact that $\partial_\phi\log w = -\partial_\phi\log q_\phi(\mathbf{z}\,|\,\mathbf{x})$,

$$\mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\frac{d}{d\phi}\log w\right] = \mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\left(\frac{\partial\log w}{\partial\phi} + \frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\phi}\right)\right] \tag{98}$$

$$= \mathbb{E}_{\pi_\beta}\left[f(\mathbf{z})\left(-\frac{\partial\log q_\phi(\mathbf{z}\,|\,\mathbf{x})}{\partial\phi} + \frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\phi}\right)\right] \tag{99}$$

$$= -\mathbb{E}_{\pi_\beta}\left[\left(f(\mathbf{z})\frac{\partial\log q_\phi(\mathbf{z}\,|\,\mathbf{x})}{\partial\phi}\right) - \left(f(\mathbf{z})\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\phi}\right)\right] \tag{100}$$

$$= -\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + \beta f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}}\right) - \left(f(\mathbf{z})\frac{\partial\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\phi}\right)\right] \qquad \text{Using Lemma 3} \tag{101}$$

$$= -\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + \beta f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}} - f(\mathbf{z})\frac{\partial\log w}{\partial\phi}\right)\right] \tag{102}$$

$$= -\mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left(\frac{\partial f(\mathbf{z})}{\partial\mathbf{z}} + (\beta-1)f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}}\right)\right] \tag{103}$$

$$= \mathbb{E}_{\pi_\beta}\left[\frac{\partial\mathbf{z}}{\partial\phi}\left((1-\beta)f(\mathbf{z})\frac{\partial\log w}{\partial\mathbf{z}} - \frac{\partial f(\mathbf{z})}{\partial\mathbf{z}}\right)\right] \qquad . \tag{104}$$

$\square$

I.4. Proof of Lemma 2

$$\frac{d}{d\phi}Z_\beta = \beta(1-\beta)\mathbb{E}_\epsilon\left[w^\beta \frac{\partial\,\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\,\mathbf{z}}\right]. \tag{105}$$

Proof. Noting that we can use reparameterization inside the integral $Z_\beta = \int q_\phi(\mathbf{z}\,|\,\mathbf{x})^{1-\beta}p_\theta(\mathbf{x},\mathbf{z})^\beta dz = \mathbb{E}_{q_\phi}[w^\beta] = \mathbb{E}_\epsilon[w^\beta]$, we obtain

$$\frac{d}{d\phi}Z_\beta = \frac{d}{d\phi}\mathbb{E}_\epsilon[w^\beta] \tag{106}$$

$$= \mathbb{E}_\epsilon\left[\beta w^\beta \frac{d}{d\phi}\log w\right] \tag{107}$$

$$= \beta Z_\beta \mathbb{E}_{\pi_\beta}\left[\frac{d}{d\phi}\log w\right] \tag{108}$$

$$= \beta(1-\beta)Z_\beta\mathbb{E}_{\pi_\beta}\left[\frac{\partial\,\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\,\mathbf{z}}\right] \quad \text{Using Corollary 1.1} \tag{109}$$

$$= \beta(1-\beta)\mathbb{E}_\epsilon\left[w^\beta \frac{\partial\,\mathbf{z}}{\partial\phi}\frac{\partial\log w}{\partial\,\mathbf{z}}\right] \tag{110}$$

$$\square$$