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

Rob Brekelmans

Vaden Masrani Frank Wood Greg Ver Steeg Aram Galstyan

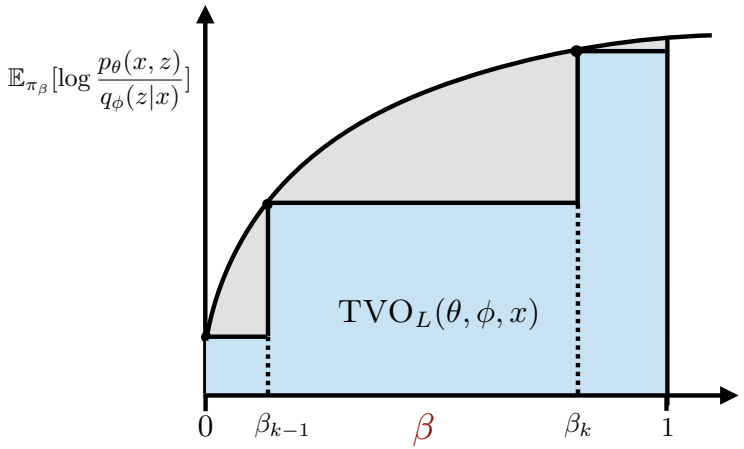
Thermodynamic Variational Objective

- TVO is a recent objective for training deep generative models
  - Generalizes and tightens the ELBO

  - $\pi_\beta(z|x) \propto q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta$

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

$$\geq \sum_{k=0}^{K-1} \mathbb{E}_{\pi_{\beta_k}}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \cdot \Delta_{\beta_k}$$



1) Masrani et. al. *"The Thermodynamic Variational Objective"*. NeurIPS 2019

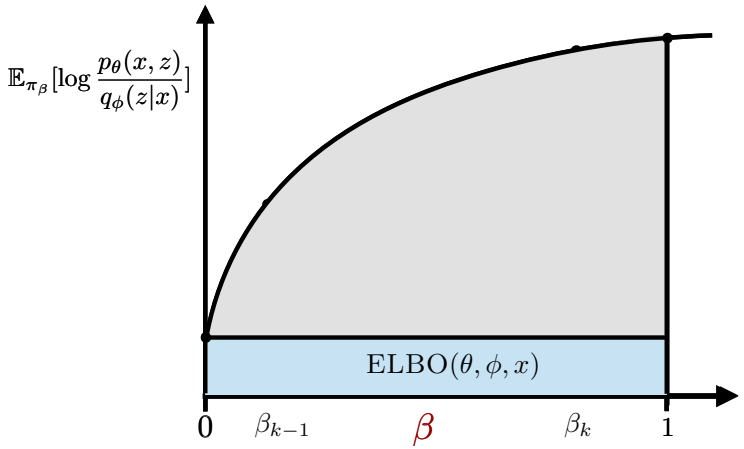
Thermodynamic Variational Objective

- TVO is a recent objective for training deep generative models
  - Generalizes and tightens the ELBO

  - $\pi_\beta(z|x) \propto q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta$

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

$$\geq \sum_{k=0}^{K-1} \mathbb{E}_{\pi_{\beta_k}}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \cdot \Delta_{\beta_k}$$



1) Masrani et. al. *"The Thermodynamic Variational Objective"*. NeurIPS 2019

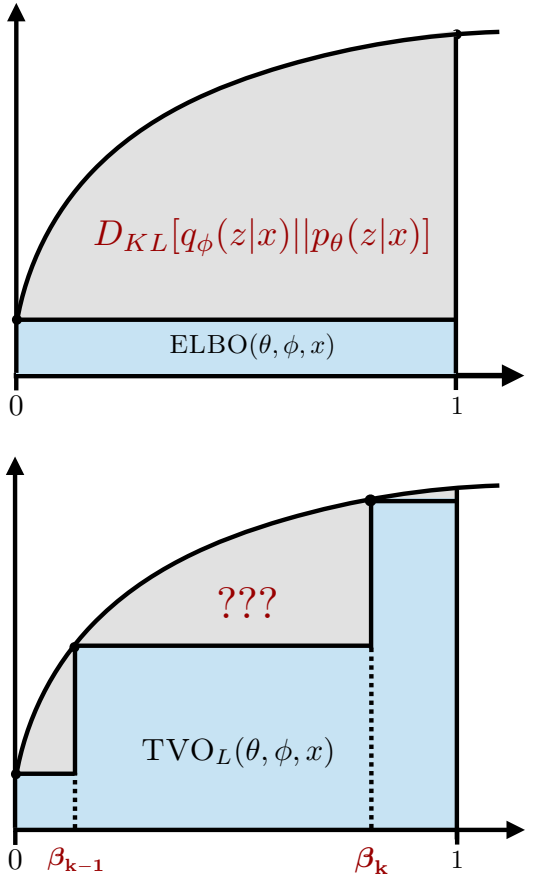
Problems with TVO

- Gap in TVO bounds previously unknown

$$\mathcal{L}_{\text{ELBO}} = \log p_\theta(x) - D_{KL}[q_\phi(z|x)||p_\theta(z|x)]$$

$$\mathcal{L}_{\text{TVO}} = \log p_\theta(x) - ???$$

- Choosing intermediate $\{\beta_k\}_{k=1}^K$
  - Log-uniform spacing, static across epochs
  - Required grid search over β_1



Exponential Family Interpretation

$$\pi_\beta(z|x) \;=\; q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta \;/\; Z_\beta(x)$$

$$\pi_\beta(z|x) = \underbrace{q_\phi(z|x)}_{\text{Base}} \exp\{\, \beta \cdot \underbrace{\log \frac{p_\theta(x,z)}{q_\phi(z|x)}}_{\text{Sufficient Statistics}} - \log Z_\beta(x)\}$$

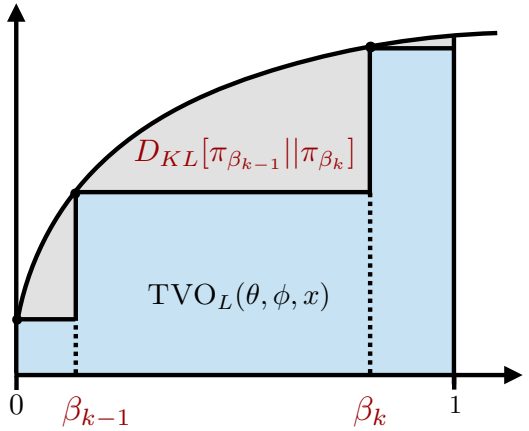
Base
(β=0)

Sufficient
Statistics

Our Contributions

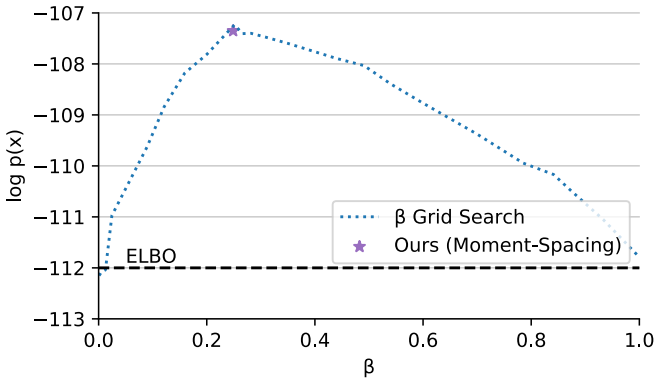
- TVO Bound Gaps using Bregman Divergences

$$\mathcal{L}_{\text{TVO}} = \log p_\theta(x) - \sum_{k=1}^{K} D_{KL}[\pi_{\beta_{k-1}}(z|x)||\pi_{\beta_k}(z|x)]$$



- Adaptively select $\{\beta_k\}_{k=1}^K$ using dual parameters of exponential family

- Single β can notably improve upon ELBO



Rest of this Talk

- Path Exponential Family
- Bregman Divergence intuition
- Moments Scheduling
- Results
- Future Directions

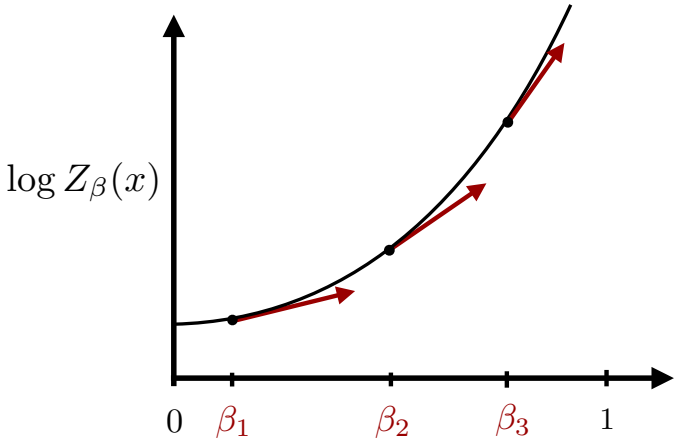
Path Exponential Family

Path Exponential Family

$$\pi_\beta(z|x) = q_\phi(z|x) \exp\{\, \beta \cdot \log \frac{p_\theta(x,z)}{q_\phi(z|x)} - \log Z_\beta(x)\}$$

$$\propto q_\phi(z|x)^{1-\beta}\, p_\theta(x,z)^\beta$$

- $\log Z_\beta(x) = \log \int q_\phi(z|x)^{1-\beta}\, p_\theta(x,z)^\beta \, dz$



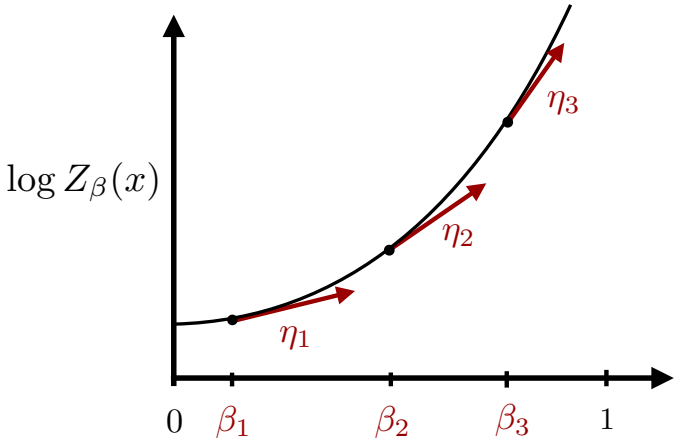
- Strictly convex in β $\implies$ $\nabla_\beta \log Z_\beta(x)$ strictly increasing

Mean Parameters

$$\pi_\beta(z|x) = q_\phi(z|x) \exp\{\, \beta \cdot \underbrace{\log \frac{p_\theta(x,z)}{q_\phi(z|x)}}_{\text{Sufficient Statistics}} - \log Z_\beta \}$$

$$\Longrightarrow \quad \eta_k := \nabla_\beta \log Z_{\beta_k}(x)$$

as a dual parameterization



$$\eta_k := \nabla_\beta \log Z_{\beta_k}(x) = \mathbb{E}_{\pi_{\beta_k}}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

Thermodynamic Integration

- Consider endpoints

$$\pi_\beta(z|x) = \frac{q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta}{Z_\beta(x)}$$

- $\pi_0(z|x) = q_\phi(z|x)$
- $\log Z_0 = 0$
- $\pi_1(z|x) \propto p_\theta(x,z)$
- $\log Z_1 = \log p_\theta(x)$

- Thermodynamic Integration / Fundamental Theorem of Calculus

$$\log Z_1 - \log Z_0 = \int_0^1 \nabla_\beta \log Z_\beta(x)\, d\beta$$

Thermodynamic Integration

- Consider endpoints

$$\pi_\beta(z|x) = \frac{q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta}{Z_\beta(x)}$$

- $\pi_0(z|x) = q_\phi(z|x)$
- $\log Z_0 = 0$
- $\pi_1(z|x) \propto p_\theta(x,z)$
- $\log Z_1 = \log p_\theta(x)$

- Thermodynamic Integration / Fundamental Theorem of Calculus

$$\log p_\theta(x) - 0 = \int_0^1 \nabla_\beta \log Z_\beta(x)\, d\beta$$

Thermodynamic Integration

- Consider endpoints

$$\pi_\beta(z|x) = \frac{q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta}{Z_\beta(x)}$$

- $\pi_0(z|x) = q_\phi(z|x)$
- $\log Z_0 = 0$
- $\pi_1(z|x) \propto p_\theta(x,z)$
- $\log Z_1 = \log p_\theta(x)$

- Thermodynamic Integration / Fundamental Theorem of Calculus

$$\log p_\theta(x) = \int_0^1 \eta_\beta \, d\beta$$

Thermodynamic Integration

- Consider endpoints

$$\pi_\beta(z|x) = \frac{q_\phi(z|x)^{1-\beta} p_\theta(x,z)^\beta}{Z_\beta(x)}$$

- $\pi_0(z|x) = q_\phi(z|x)$
- $\log Z_0 = 0$
- $\pi_1(z|x) \propto p_\theta(x,z)$
- $\log Z_1 = \log p_\theta(x)$

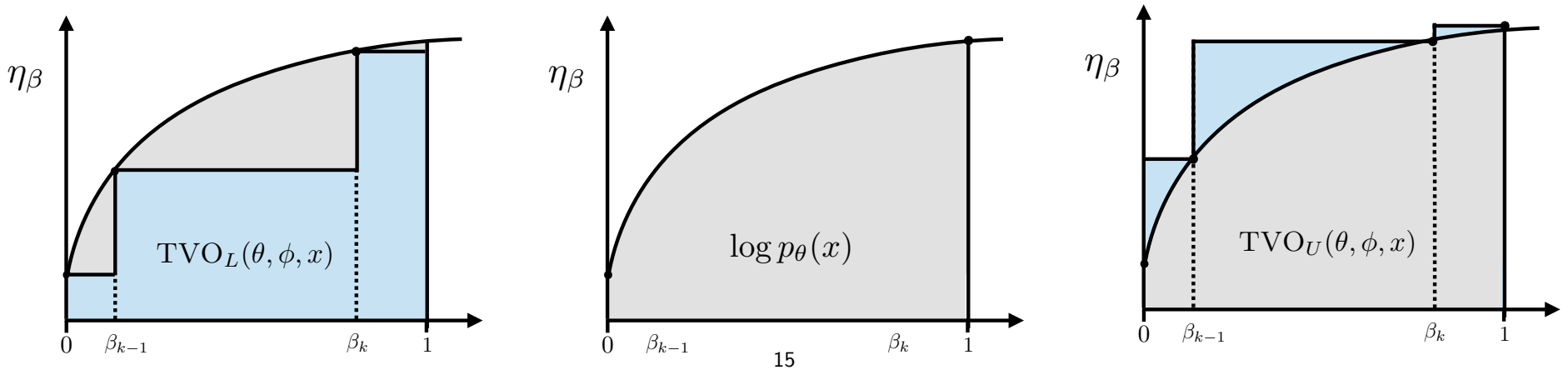
- Thermodynamic Integration / Fundamental Theorem of Calculus

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

Thermodynamic Variational Objective

- Lower or Upper bounds via Riemann Sum approximations
  - Since integrand $\eta_\beta = \nabla_\beta \log Z_\beta$ is increasing,

Left Riemann Sum $\leq$ $\log p_\theta(x)$ $\leq$ Right Riemann Sum



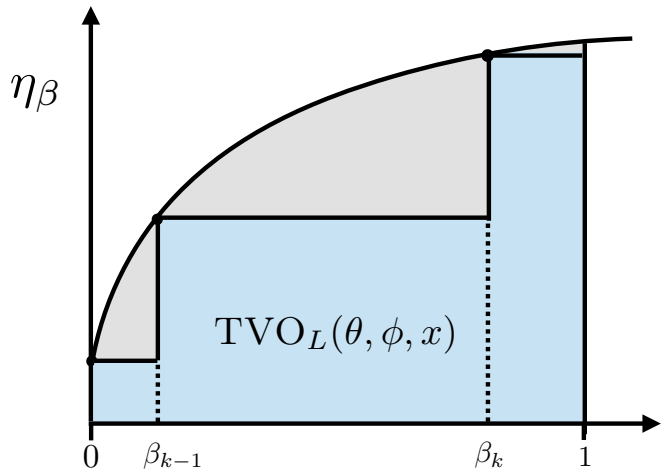
Thermodynamic Variational Objective

- Log likelihood as an integral

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

- Left Riemann Lower Bound

$$\log p_\theta(x) \geq \sum_{k=0}^{K-1} (\beta_{k+1} - \beta_k)\, \mathbb{E}_{\pi_{\beta_k}}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$



- Self-normalized importance sampling for each $\mathbb{E}_{\pi_{\beta_k}}\left[\cdot\right]$

Bregman Divergence

Bregman Divergence

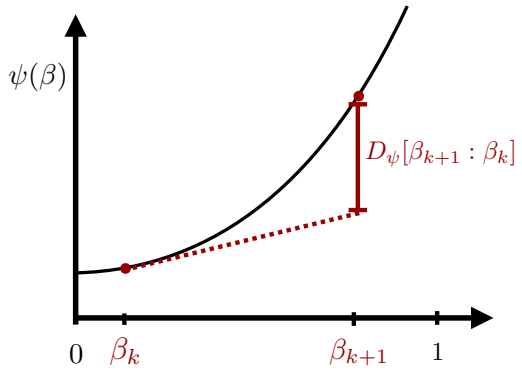
- Let $\psi(\beta) := \log Z_\beta(x)$, which is convex

- We can then define the Bregman Divergence

$$D_\psi[\beta_{k+1} : \beta_k] = \psi(\beta_{k+1}) - \psi(\beta_k) - (\beta_{k+1} - \beta_k) \cdot \nabla\psi(\beta_k)$$

Bregman Divergence

- Let $\psi(\beta) := \log Z_\beta(x)$, which is convex
- We can then define the Bregman Divergence



$$D_\psi[\beta_{k+1} : \beta_k] = \psi(\beta_{k+1}) \underbrace{- \psi(\beta_k) - (\beta_{k+1} - \beta_k) \cdot \nabla_\beta \psi(\beta_k)}_{\text{First order Taylor Approximation}}$$

- For $\log Z_\beta(x)$, Bregman divergence = KL divergence (w/arguments reversed)

$$D_\psi[\beta_{k+1} : \beta_k] = D_{KL}[\pi_{\beta_k} || \pi_{\beta_{k+1}}]$$

Gap in TVO Lower Bound

$$\underbrace{\psi(\beta_{k+1}) - \psi(\beta_k)}_{} \quad - \quad \underbrace{(\beta_{k+1} - \beta_k) \cdot \nabla_\beta \, \psi(\beta_k)}_{} \quad = \quad \underbrace{D_\psi[\beta_{k+1} : \beta_k]}_{}$$

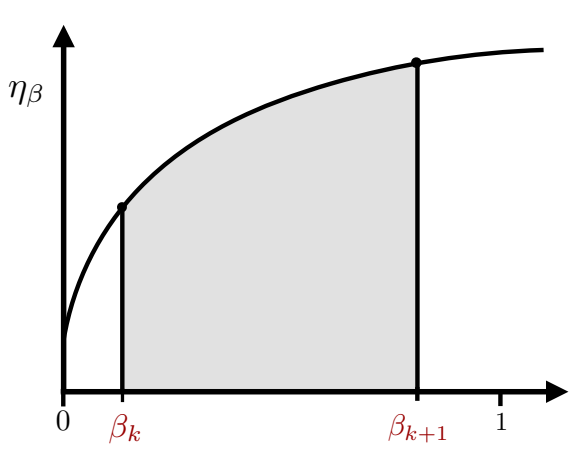
Gap in TVO Lower Bound

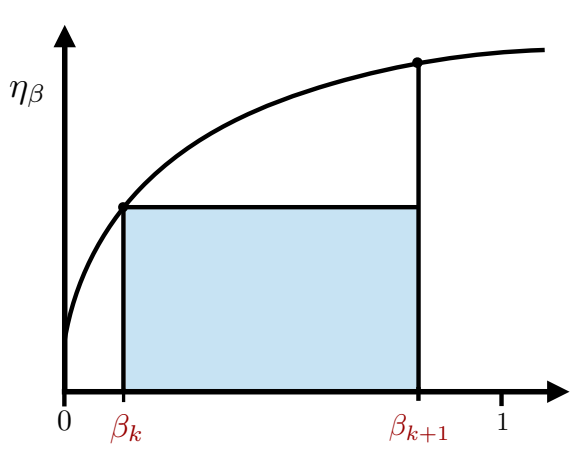
$$\underbrace{\log Z_{\beta_{k+1}} - \log Z_{\beta_k}}_{\int_{\beta_k}^{\beta_{k+1}} \nabla_\beta \log Z_\beta \, d\beta} - \underbrace{(\beta_{k+1} - \beta_k) \cdot \nabla_\beta \log Z_{\beta_k}}_{(\beta_{k+1} - \beta_k) \cdot \eta_{\beta_k}} = \underbrace{D_\psi[\beta_{k+1} : \beta_k]}_{D_{KL}[\pi_{\beta_k} || \pi_{\beta_{k+1}}]}$$

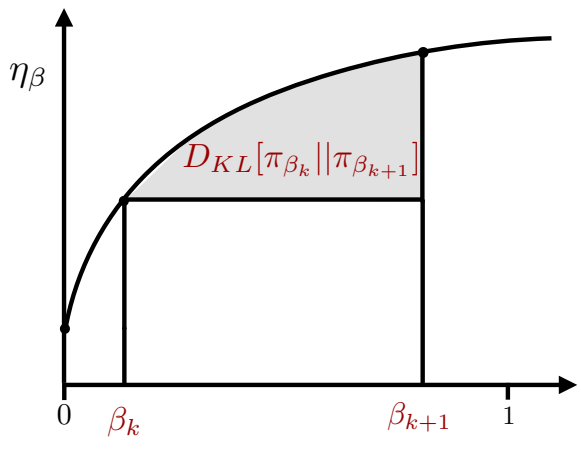
Area under curve

Width · Height

Left Riemann term







Bregman Divergences in TVO

- TVO Lower Bound (Left Riemann):

$$\mathcal{L}_{\mathrm{TVO_L}} = \log p_\theta(x) - \sum_{k=0}^{K-1} D_{KL}[\pi_{\beta_k}(z|x)||\pi_{\beta_{k+1}}(z|x)]$$

- TVO Upper Bound (Right Riemann):

$$\mathcal{L}_{\mathrm{TVO_U}} = \log p_\theta(x) + \sum_{k=0}^{K-1} D_{KL}[\pi_{\beta_{k+1}}(z|x)||\pi_{\beta_k}(z|x)]$$

- Additional analysis:
  - Symmetrized KL divergence
  - Taylor Remainder theorem
  - Renyi Divergence VI (Li and Turner 2016)

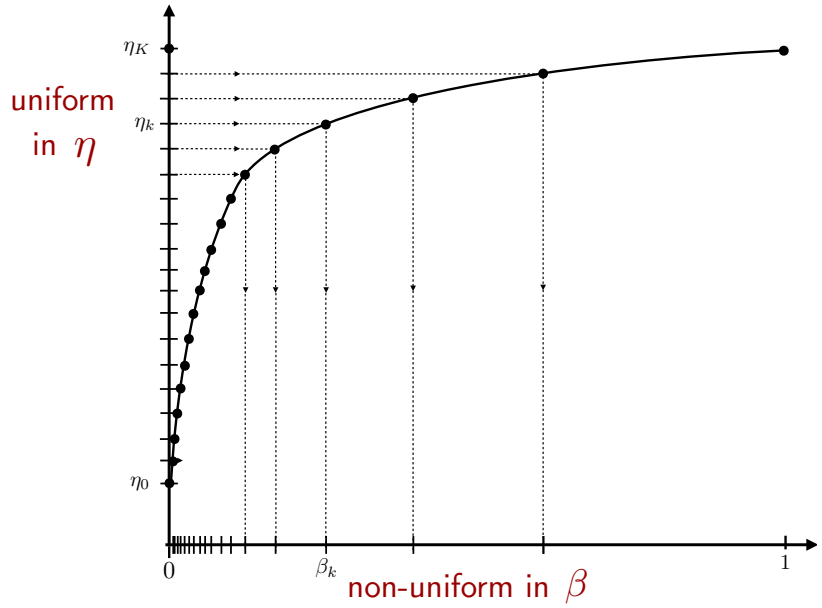
Moments Scheduling

Choosing β "Schedules"

- Budget of K intermediate points as a hyper parameter
- Goal: *Assign more intermediate* β_k *where integrand is changing quickly*
- Shape of the TVO integrand related to posterior mismatch
  - Adaptive choice of $\{\beta_k\}_{k=1}^{K}$ based on training progress

Moment Spacing Schedule

- Find β_k to yield equal spacing in the mean parameters η_k [2]
  - Legendre transform
    - Efficient with SNIS, binary search
  - Corresponds to Lebesgue integration

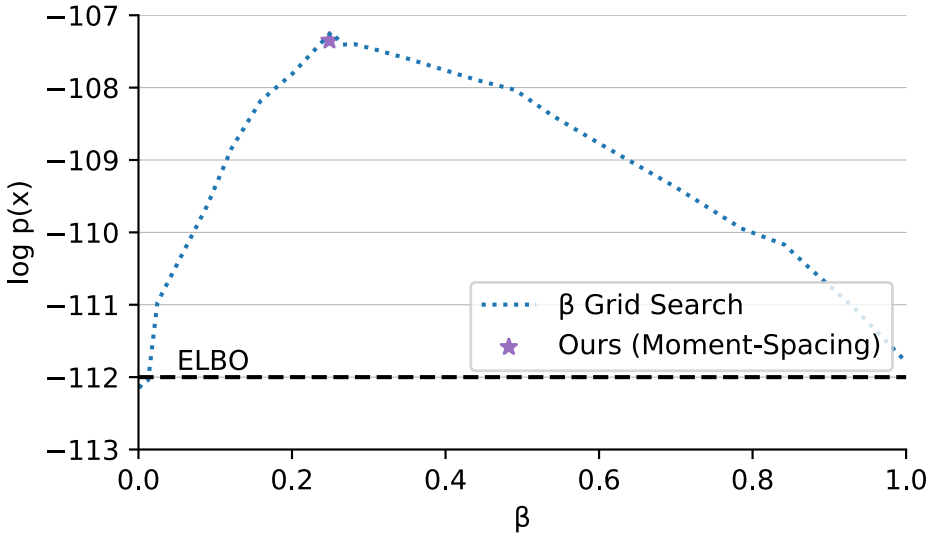


2) Grosse et. al. *"Annealing between Distributions by Averaging Moments"*. NeurIPS 2013

Results

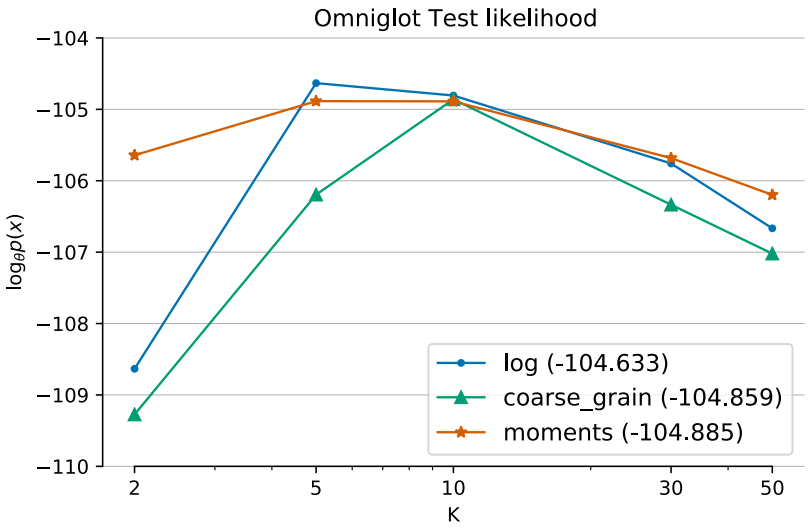
Results

- TVO with single intermediate β_1
  - (2 term Riemann approx)
  - Compare with:
    - Grid search over β_1
    - ELBO (single term $\beta_0 = 0$)



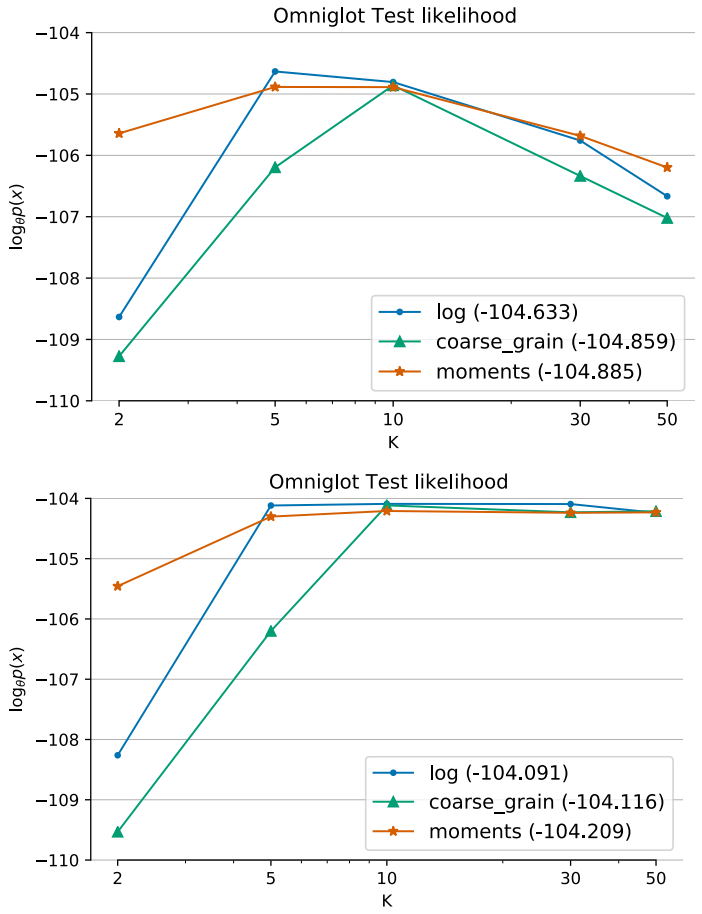
Results

- TVO with REINFORCE
  - Moments schedule
  - *Deteriorating* performance with *higher* K (i.e. # β)



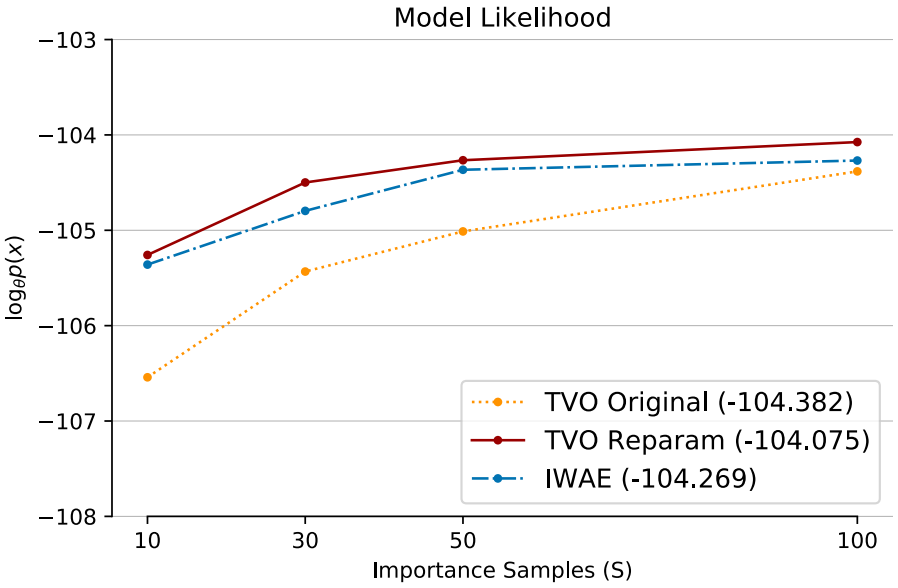
Results

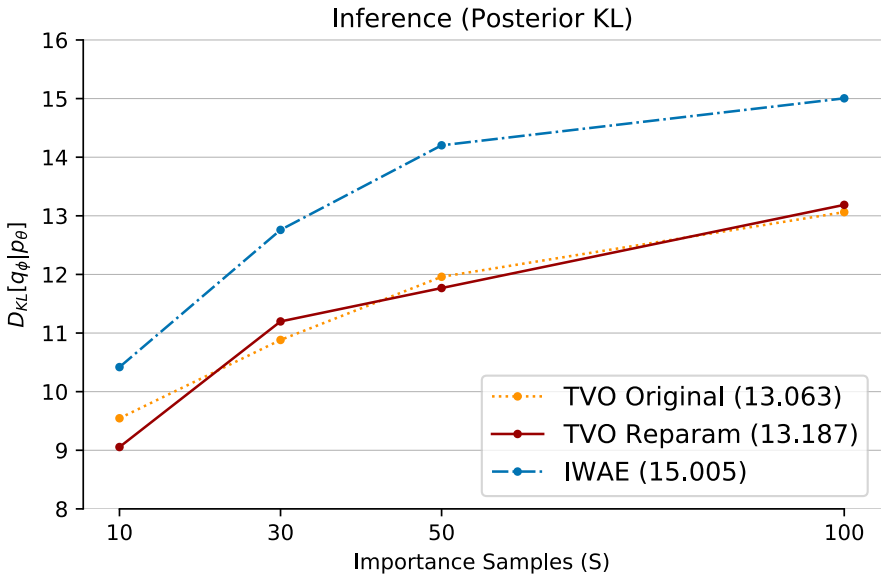
- TVO with REINFORCE
  - Moments schedule
  - *Deteriorating* performance with *higher* K (i.e. # β)
- TVO with reparameterization (Ours)
  - Schedules perform similarly



Results

- Comparing TVO with reparamerization, original TVO with IWAE
  - By number of importance samples S





Summary

- Path exponential family
    - Clarifies analysis of TVO
    - General construction with rich geometric structure
- Future Directions in "Thermodynamic Variational Inference"
    - Improve SNIS sampler with MCMC methods
    - Explore deeper connections with thermodynamics