Importance Weighting and Variational Inference

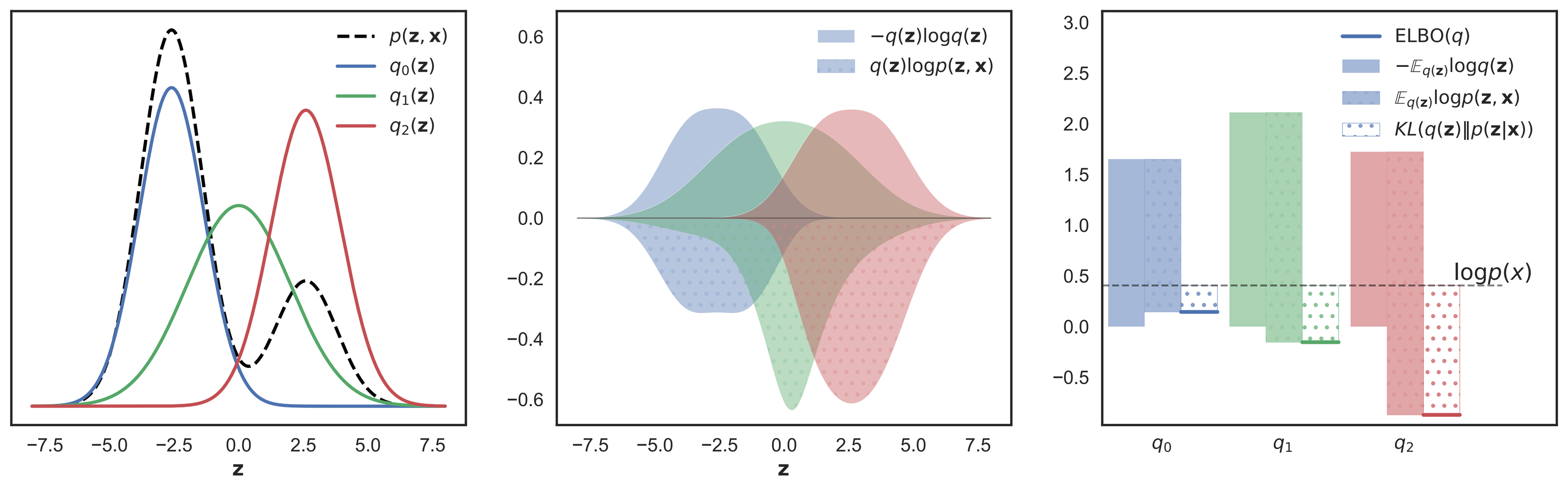
Justin Domke and Daniel Sheldon, University of Massachusetts, Amherst

- Variational autoencoders can use **importance weighting** for **better likelihood bounds**.
- But how to apply to "**pure probabilistic**" variational inference (VI)?
- We show that using importance-weighting is equivalent to **traditional VI** on **augmented distributions**. This informs test-time inference and clarifies looseness of existing bounds.
- Investigate VI on **elliptical distributions** via an "inverse CDF trick".

1. The ELBO Decomposition

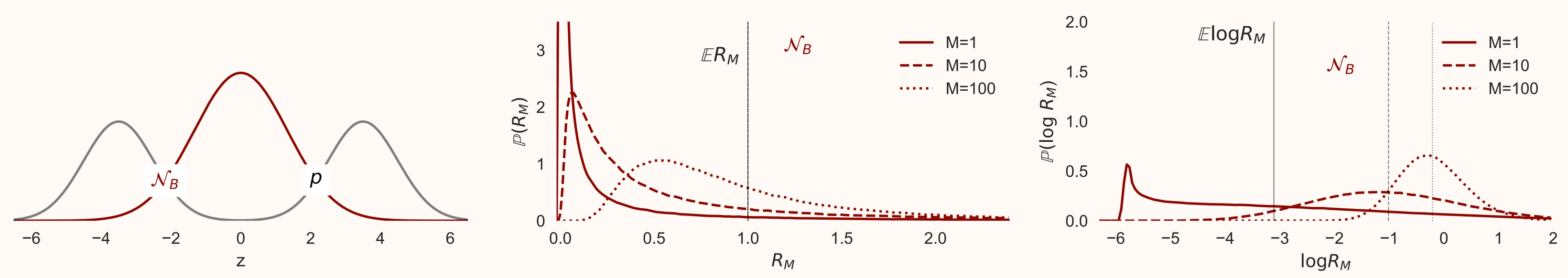
$$\log p(\boldsymbol{x}) = \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{z}\sim q} \log \frac{p(\boldsymbol{z},\boldsymbol{x})}{q(\boldsymbol{z})}}_{\mathrm{ELBO}(q\|p)} + KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x})\right)$$

- Any q gives $\mathrm{ELBO} \leq \log p(\boldsymbol{x})$.
- Looseness is KL-divergence.

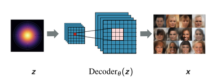


2. Importance Weighting

- For any $R > 0$ with $\mathbb{E}\,R = p(\boldsymbol{x})$: $\quad \log p(\boldsymbol{x}) = \underbrace{\mathbb{E}\log R}_{\text{bound}} + \underbrace{\mathbb{E}\log\frac{p(\boldsymbol{x})}{R}}_{\text{looseness}}$.
- Traditional VI: $R = \frac{p(\boldsymbol{z},\boldsymbol{x})}{q(\boldsymbol{z})}$, $\boldsymbol{z}\sim q$.
- Better bound: Average i.i.d. samples: $R_M = \frac{1}{M}\sum_{m=1}^{M}\frac{p(\boldsymbol{z}_m,\boldsymbol{x})}{q(\boldsymbol{z}_m)}$, $\boldsymbol{z}_m \sim q$



IWAEs: [Burda et al., 2015]

- $p(\boldsymbol{z},\boldsymbol{x})$ = 
  - Input $\boldsymbol{x}$ (dataset)
  - Maximize $\mathbb{E}\log R_M$ w.r.t. p and q
  - Use p

Importance Weighted VI (IWVI):

- $p(\boldsymbol{z},\boldsymbol{x})$ = (Some model)
  - Input $\boldsymbol{x}$ (evidence)
  - Maximize $\mathbb{E}\log R_M$ w.r.t. q
  - Use q

Good-old-fashioned VI:

$$\log p(\boldsymbol{x}) = \mathrm{ELBO}\left(q(\boldsymbol{z})\|p(\boldsymbol{z},\boldsymbol{x})\right) + KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x})\right)$$

- **Learning**: $\mathrm{ELBO} \leq \log p(\boldsymbol{x})$
- **Inference**: $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}\,t(\boldsymbol{z}) \approx \mathbb{E}_{q(\boldsymbol{z})}\,t(\boldsymbol{z})$

IWVI:

$$\log p(\boldsymbol{x}) = \mathbb{E}\log R_M + \mathbb{E}\log\frac{p(\boldsymbol{x})}{R_M}$$

- **Learning**: $\mathbb{E}\log R_M \leq \log p(\boldsymbol{x})$
- **Inference**: ???

3. Main Technical Results

Summary:

- Theorem 1: For augmented p_M / q_M, IWVI minimizes $KL\left(q_M(\boldsymbol{z}_{1:M})\|p_M(\boldsymbol{z}_{1:M}|\boldsymbol{x})\right)$.
- Theorem 2: That is exactly $\underbrace{KL\left(q_M(\boldsymbol{z}_1)\|p(\boldsymbol{z}_1|\boldsymbol{x})\right)}_{\text{what we care about}} + \underbrace{KL\left(q_M(\boldsymbol{z}_{2:M})\|q(\boldsymbol{z}_{2:M})\right)}_{\text{other stuff}}$.
- Theorem 3: When M is large that is approximately $\frac{1}{M}\frac{\mathbb{V}[R]}{2p(\boldsymbol{x})}$.

3.1 Theorem 1: IWVI is Normal VI on Augmented Distributions

Definition of $q_M(\boldsymbol{z}_{1:M})$:

- Draw $\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_M$ independently from $q(\boldsymbol{z})$.
- Choose $m \in \{1,\cdots,M\}$ with prob $\mathbb{P}(m) \propto \frac{p(\hat{\boldsymbol{z}}_m,\boldsymbol{x})}{q(\hat{\boldsymbol{z}}_m)}$.
- Set $\boldsymbol{z}_1 = \hat{\boldsymbol{z}}_m$ and $\boldsymbol{z}_{2:M} = \hat{\boldsymbol{z}}_{-m}$

(Self-normalized importance sampling for $\hat{\boldsymbol{z}}_m$; also keep and relabel unselected $\hat{\boldsymbol{z}}_i$)

Definition of $p_M(\boldsymbol{z}_{1:M})$:

- $p_M(\boldsymbol{z}_{1:M},\boldsymbol{x}) = p(\boldsymbol{z}_1,\boldsymbol{x})q(\boldsymbol{z}_2)\cdots q(\boldsymbol{z}_M)$.

(One sample from p and $M-1$ "dummy" samples from q.)

Previously known [Bachman and Precup, 2015, Cremer et al., 2017, Naesseth et al., 2018, Le et al., 2018]: $\log p(\boldsymbol{x}) \geq \mathrm{ELBO}(q_M(\boldsymbol{z}_1)\|p_M(\boldsymbol{z}_1,\boldsymbol{x})) \geq \mathbb{E}\log R_M$.

Our Result: $\log p(\boldsymbol{x}) = \mathbb{E}\log R_M + KL\left(q_M(\boldsymbol{z}_{1:M})\|p_M(\boldsymbol{z}_{1:M}|\boldsymbol{x})\right)$.

Thus, approximate test integrals as

$$\mathop{\mathbb{E}}_{p(\boldsymbol{z}|\boldsymbol{x})} t(\boldsymbol{z}) = \mathop{\mathbb{E}}_{p_M(\boldsymbol{z}_1|\boldsymbol{x})} t(\boldsymbol{z}_1) \approx \mathop{\mathbb{E}}_{q_M(\boldsymbol{z}_1)} t(\boldsymbol{z}_1).$$

3.2 Theorem 2: IWVI is Tightening an Upper Bound

Result:

$$\underbrace{KL\left(q_M(\boldsymbol{z}_{1:M})\|p_M(\boldsymbol{z}_{1:M}|\boldsymbol{x})\right)}_{\text{what IWVI minimizes}} = \underbrace{KL\left(q_M(\boldsymbol{z}_1)\|p(\boldsymbol{z}_1|\boldsymbol{x})\right)}_{\text{what we care about}} + \underbrace{KL\left(q_M(\boldsymbol{z}_{2:M})\|q(\boldsymbol{z}_{2:M})\right)}_{\text{other stuff}}$$

Proof: KL chain rule + definition of p_M.

If you will use normalized importance sampling, IWVI truly optimizes a bound.

3.3 Theorem 3: Asymptotics

Result: If $\mathbb{E}\left|R - p(\boldsymbol{x})\right|^{2+\alpha} < \infty$ for $\alpha > 0$ and $\limsup_{M\to\infty}\mathbb{E}\frac{1}{R_M} < \infty$,

$$\lim_{M\to\infty} M\left(\log p(\boldsymbol{x}) - \mathbb{E}\log R_M\right) = \frac{\mathbb{V}[R_1]}{2p(\boldsymbol{x})}.$$

(Maddison et al. [2017] showed for $\alpha = 4$)

Non-Proof: CLT + 2nd-order delta-method:

$$M\left(\log p(\boldsymbol{x}) - \mathbb{E}\log R_M\right) \xrightarrow{d} \frac{\mathbb{V}[R_1]}{2p(\boldsymbol{x})}\chi_1^2$$

Problem: $X_M \xrightarrow{d} X$ does not imply $\mathbb{E}[X_M] \to \mathbb{E}[X]$.

Proof: Long. Broadly follows Maddison et al. [2017] to bound higher terms in a Taylor expansion. Biggest technical innovation is using Marcinkiewicz–Zygmund inequality to bound sample moments from true moments.

4. Elliptical Distributions

- For $M = 1$, IWVI minimizes KL.
  - This is **mode finding**.
- For M large, IWVI minimizes $\mathbb{V}[R]$ (Equiv. χ^2 divergence).
  - This is **mode spanning**.

Suggests we want **different tail behavior** as M changes.

Given some spherically symmetric distribution g_ν, an "Elliptical" distribution is

$$q\left(\boldsymbol{z}|\mu,\Sigma,\nu\right) = \frac{1}{|\Sigma|^{1/2}}g_\nu\left(\left(\boldsymbol{z}-\boldsymbol{\mu}\right)^\top\Sigma^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\right).$$

- Fit $\boldsymbol{\mu}$, Σ as "Normal". (Reparameterization trick with $g_\nu(\boldsymbol{\epsilon})$ as base density).
- Fit ν by backpropagating through inverse CDF of $\|\boldsymbol{\epsilon}\|$, $\boldsymbol{\epsilon}\sim g_\nu$.
- No inverse-CDF? Sample $\boldsymbol{\epsilon}\sim g_\nu$, then "pretend": (Same idea at this conference: *Implicit Reparameterization Gradients*, Figurnov et al.)

$$\nabla_\nu F_\nu^{-1}(u) = -\frac{\nabla_\nu F_\nu(\|\boldsymbol{\epsilon}\|)}{\nabla_r F_\nu(\|\boldsymbol{\epsilon}\|)},\quad u = F_\nu(\|\boldsymbol{\epsilon}\|).$$

5. Experiments

Variational Families:

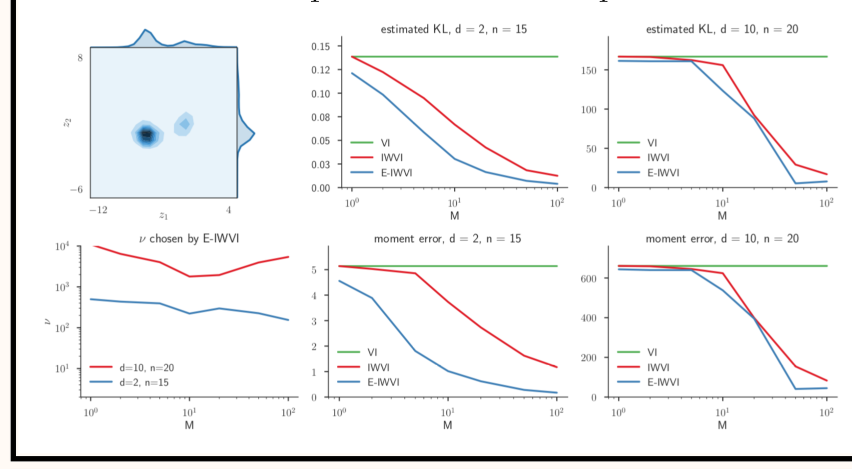
- IWVI : Gaussians
- E-IWVI: Student-T with ν deg. of freedom.

Error metrics:

- $KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z})\right)$
- $\mathbb{C}[p(\boldsymbol{z})]$ vs. $\mathbb{C}[q_M(\boldsymbol{z}_1)]$

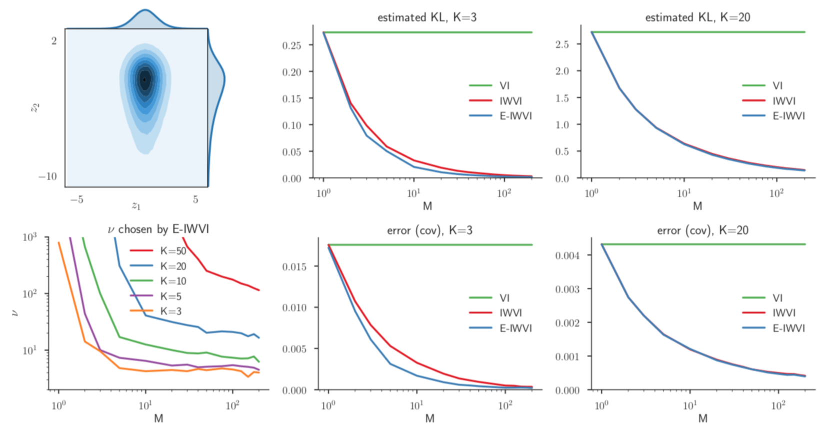
Clutter model: [Minka, 2001]

- $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}, 0, 100I)$ - hidden location
- $\boldsymbol{x} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)$ noisy observations
  - $p(\boldsymbol{x}_i|\boldsymbol{z}) = \frac{1}{4}\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{z}_i, I) + \frac{3}{4}\mathcal{N}(\boldsymbol{x}_i|0, 10I)$



Random Dirichlets:

- Sample $\alpha_1,\cdots,\alpha_K \sim \mathrm{Gamma}(10)$
- $p(\boldsymbol{z})$ is density of $\mathrm{StickBreak}(\theta)$, $\theta \sim \mathrm{Dirichlet}(\theta|\boldsymbol{\alpha})$



Logistic Regression (Cauchy Prior)