Variational bounds in Variational Inference: how to choose them?

Kamélia Daudel

UNIVERSITY OF OXFORD

CoSInES-Bayes4Health VI Masterclass – 09/11/2022

Outline

Outline

Introduction

- We consider a model with joint distribution $p_\theta(x,z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$

- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density

- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta;x) := \log p_\theta(x) = \log\left(\int p_\theta(x,z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta;x) = \frac{\nabla_\theta(\int p_\theta(x,z)\mathrm{d}z)}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int p_\theta(x,z)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z}$$
$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x,z)\mathrm{d}z$

Introduction

- We consider a model with joint distribution $p_\theta(x, z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$
- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x, z)}{\int p_\theta(x, z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density
- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta; x) := \log p_\theta(x) = \log\left(\int p_\theta(x, z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta; x) = \frac{\nabla_\theta(\int p_\theta(x, z)\mathrm{d}z)}{\int p_\theta(x, z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x, z))\mathrm{d}z}{\int p_\theta(x, z)\mathrm{d}z} = \frac{\int p_\theta(x, z)\nabla_\theta(\log p_\theta(x, z))\mathrm{d}z}{\int p_\theta(x, z)\mathrm{d}z}$$

$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x, z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x, z)\mathrm{d}z$

Introduction

- We consider a model with joint distribution $p_\theta(x,z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$

- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density

- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta;x) := \log p_\theta(x) = \log\left(\int p_\theta(x,z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta;x) = \frac{\nabla_\theta(\int p_\theta(x,z)\mathrm{d}z)}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int p_\theta(x,z)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z}$$
$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x,z)\mathrm{d}z$

Introduction

- We consider a model with joint distribution $p_\theta(x,z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$

- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density

- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta;x) := \log p_\theta(x) = \log\left(\int p_\theta(x,z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta;x) = \frac{\nabla_\theta(\int p_\theta(x,z)\mathrm{d}z)}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int p_\theta(x,z)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z}$$
$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x,z)\mathrm{d}z$

Introduction

- We consider a model with joint distribution $p_\theta(x, z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$

- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x, z)}{\int p_\theta(x, z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density

- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta; x) := \log p_\theta(x) = \log\left(\int p_\theta(x, z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta; x) = \frac{\nabla_\theta(\int p_\theta(x, z)\mathrm{d}z)}{\int p_\theta(x, z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x, z))\mathrm{d}z}{\int p_\theta(x, z)\mathrm{d}z} = \frac{\int p_\theta(x, z)\nabla_\theta(\log p_\theta(x, z))\mathrm{d}z}{\int p_\theta(x, z)\mathrm{d}z}$$
$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x, z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x, z)\mathrm{d}z$

Introduction

- We consider a model with joint distribution $p_\theta(x,z)$ parameterized by θ, where x is an observation and z is a latent variable valued in $\mathbb{R}^d$
- Posterior density of the latent variable z given the observation x

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)\mathrm{d}z}$$

- What we would like : compute / sample from the posterior density
- Key example : maximize the **marginal log likelihood** w.r.t. θ

$$\ell(\theta;x) := \log p_\theta(x) = \log\left(\int p_\theta(x,z)\mathrm{d}z\right)$$

$$\nabla_\theta \ell(\theta;x) = \frac{\nabla_\theta(\int p_\theta(x,z)\mathrm{d}z)}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int \nabla_\theta(p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z} = \frac{\int p_\theta(x,z)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z}{\int p_\theta(x,z)\mathrm{d}z}$$
$$= \int p_\theta(z|x)\nabla_\theta(\log p_\theta(x,z))\mathrm{d}z$$

- Problem : for many important models, we can only evaluate $p_\theta(z|x)$ up to the marginal likelihood $\int p_\theta(x,z)\mathrm{d}z$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.
- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \ : \ \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log \left(w_{\theta,\phi}(z;x) \right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x) || p_\theta(\cdot|x))$ where

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.
- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \ : \ \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log\left(w_{\theta,\phi}(z;x)\right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ where

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.
- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \ : \ \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log \left(w_{\theta,\phi}(z;x) \right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ where

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.

- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \ : \ \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log \left(w_{\theta,\phi}(z; x) \right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z; x) = \frac{p_\theta(x, z)}{q_\phi(z|x)}$$

$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ where

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.
- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \ : \ \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log \left(w_{\theta,\phi}(z;x) \right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ where

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

Variational bounds

- Variational bounds are surrogate objective functions to the marginal log likelihood that are more amenable to optimization.

- They involve a variational family of probability densities $\mathcal{Q}$

$$\text{e.g.} \quad \mathcal{Q} = \left\{ z \mapsto q_\phi(z|x) \;:\; \phi \in \mathbb{R}^L \right\}$$

- Example : **Evidence Lower BOund (ELBO)**

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log\left(w_{\theta,\phi}(z;x)\right) \mathrm{d}z \quad \text{where} \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) \text{ where}$$

$$D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \int_{\mathsf{Y}} q_\phi(z|x) \log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right) \mathrm{d}z \quad \text{(Exclusive KL)}$$

so that $\mathrm{ELBO}(\theta, \phi; x) \leq \ell(\theta; x)$

"Traditional Variational Inference" : θ is constant, the goal is to minimize the exclusive KL divergence $\Leftrightarrow$ maximizing the ELBO

Optimisation w.r.t. (θ, ϕ): Variational Auto-Encoder (VAE) framework

Training with the ELBO

1. Unbiased Monte Carlo (MC) estimator of the ELBO

$$\begin{aligned}\mathrm{ELBO}(\theta,\phi;x) &= \int q_\phi(z|x)\log\left(w_{\theta,\phi}(z;x)\right)\mathrm{d}z \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\log\left(w_{\theta,\phi}(z_i;x)\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\ldots N\end{aligned}$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$
3. Reparameterized gradient of the ELBO:

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi;x) = \int q(\varepsilon)\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon$$

4. Unbiased SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi;x) \approx \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\ldots N$$

Training with the ELBO

1. Unbiased Monte Carlo (MC) estimator of the ELBO

$$\begin{aligned}\mathrm{ELBO}(\theta, \phi; x) &= \int q_\phi(z|x) \log\left(w_{\theta,\phi}(z;x)\right) \mathrm{d}z \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \log\left(w_{\theta,\phi}(z_i;x)\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \ldots N\end{aligned}$$

2. Reparameterization trick $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$
3. Reparameterized gradient of the ELBO:

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi; x) = \int q(\varepsilon) \nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x)\right) \mathrm{d}\varepsilon$$

4. Unbiased SGD w.r.t. (θ, ϕ)

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi; x) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x)\right), \quad \varepsilon_i \sim q, \quad i = 1 \ldots N$$

Training with the ELBO

1. Unbiased Monte Carlo (MC) estimator of the ELBO

$$\begin{aligned}\mathrm{ELBO}(\theta, \phi; x) &= \int q_\phi(z|x) \log\left(w_{\theta,\phi}(z; x)\right) \mathrm{d}z \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \log\left(w_{\theta,\phi}(z_i; x)\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \dots N\end{aligned}$$

2. Reparameterization trick $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$
3. Reparameterized gradient of the ELBO:

$$\nabla_{\theta,\phi} \mathrm{ELBO}(\phi; x) = \int q(\varepsilon) \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x)\right) \mathrm{d}\varepsilon$$

4. Unbiased SGD w.r.t. (θ, ϕ)

$$\nabla_{\theta,\phi} \mathrm{ELBO}(\phi; x) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x)\right), \quad \varepsilon_i \sim q, \quad i = 1 \dots N$$

Training with the ELBO

1. Unbiased Monte Carlo (MC) estimator of the ELBO

$$\begin{aligned}\mathrm{ELBO}(\theta,\phi;x) &= \int q_\phi(z|x)\log\left(w_{\theta,\phi}(z;x)\right)\mathrm{d}z \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\log\left(w_{\theta,\phi}(z_i;x)\right), \quad z_i \sim q_\phi(\cdot|x), \quad i=1\ldots N\end{aligned}$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$
3. Reparameterized gradient of the ELBO:

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi;x) = \int q(\varepsilon)\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon$$

4. Unbiased SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{ELBO}(\phi;x) \approx \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i=1\ldots N$$

Question

$$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

Question Can we change the regularization term?

Question

$$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

Question Can we change the regularization term?

Question

$$\mathrm{ELBO}(\theta, \phi; x) = \ell(\theta; x) - D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

Question Can we change the regularization term?

Outline

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha-1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$

Proof Set $f(\alpha) = \int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

Then, $f(1) = 1$ and $f'(\alpha) = \int q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

$$\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \lim_{\alpha \to 1} \frac{\log f(\alpha) - \log f(1)}{\alpha - 1} = \frac{f'(1)}{f(1)} = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha - 1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$

Proof Set $f(\alpha) = \int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

Then, $f(1) = 1$ and $f'(\alpha) = \int q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

$$\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \lim_{\alpha \to 1} \frac{\log f(\alpha) - \log f(1)}{\alpha - 1} = \frac{f'(1)}{f(1)} = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha - 1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$

Proof Set $f(\alpha) = \int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

Then, $f(1) = 1$ and $f'(\alpha) = \int q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

$$\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \lim_{\alpha \to 1} \frac{\log f(\alpha) - \log f(1)}{\alpha - 1} = \frac{f'(1)}{f(1)} = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha - 1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$

Proof Set $f(\alpha) = \int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

Then, $f(1) = 1$ and $f'(\alpha) = \int q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

$$\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \lim_{\alpha \to 1} \frac{\log f(\alpha) - \log f(1)}{\alpha - 1} = \frac{f'(1)}{f(1)} = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha - 1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$

  Proof Set $f(\alpha) = \int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

Then, $f(1) = 1$ and $f'(\alpha) = \int q_\phi(z|x) \log \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z$

$$\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \lim_{\alpha \to 1} \frac{\log f(\alpha) - \log f(1)}{\alpha - 1} = \frac{f'(1)}{f(1)} = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha-1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$
- $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) \leq \ell(\theta; x)$, $\mathrm{VR}^{(0)}(\theta, \phi; x) = \ell(\theta; x)$

$\to$ The VR bound generalizes the ELBO, interpolates between $\ell(\theta; x)$ and the ELBO

The Variational Rényi (VR) bound

Variational Rényi (VR) bound (Li and Turner, NeurIPS 2016): for all $\alpha > 0$ and $\neq 1$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right), \quad w_{\theta,\phi}(z;x) = \frac{p_\theta(z,x)}{q_\phi(z|x)}$$
$$= \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

where $D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$ is **Rényi's α-divergence**: for all $\alpha > 0$ and $\neq 1$

$$D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = \frac{1}{\alpha-1} \log \left(\int q_\phi(z|x) \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right)^{\alpha-1} \mathrm{d}z \right)$$

- We have that $\lim_{\alpha \to 1} D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x)) = D^{(KL)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$
- $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) \leq \ell(\theta; x)$, $\mathrm{VR}^{(0)}(\theta, \phi; x) = \ell(\theta; x)$

$\to$ The VR bound generalizes the ELBO, interpolates between $\ell(\theta; x)$ and the ELBO

Impact of α

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

- Question How does the regularization term behave?

Impact of α

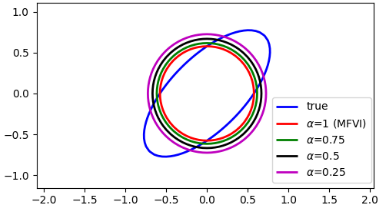
$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

- Question How does the regularization term behave?

Impact of α

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x) - D^{(\alpha)}(q_\phi(\cdot|x)||p_\theta(\cdot|x))$$

- Question How does the regularization term behave?
- Example : $D^{(\alpha)}(q||p)$ with $p(z) = \mathcal{N}(z; [0,0], [[3,-2],[-2,3]])$ and $\mathcal{Q} = \{q : z \mapsto \mathcal{N}(z_1; \mu_1, \sigma_1^2)\ \mathcal{N}(z_2; \mu_2, \sigma_2^2) : \mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0\}$



Adapted from (Li and Turner, NeurIPS 2016)

Training with the VR bound (Li and Turner, NeurIPS 2016)

1. MC estimator of the VR bound

$$\begin{aligned}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)\\ &\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i\sim q_\phi(\cdot|x), \quad i=1\ldots N\end{aligned}$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x)\sim q_\phi(\cdot|x)$ where $\varepsilon\sim q$
3. Reparameterized gradient of the VR bound :

$$\begin{aligned}\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \nabla_{\theta,\phi}\left[\frac{1}{1-\alpha}\log\left(\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon\right)\right]\\ &= \frac{1}{1-\alpha}\frac{\int q(\varepsilon)\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}\\ &= \frac{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{-\alpha}\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}\\ &= \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}\end{aligned}$$

4. SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) \approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right),\ \varepsilon_i\sim q, i=1\ldots N$$

Training with the VR bound (Li and Turner, NeurIPS 2016)

1. MC estimator of the VR bound

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\ldots N$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$

3. Reparameterized gradient of the VR bound :

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \nabla_{\theta,\phi}\left[\frac{1}{1-\alpha}\log\left(\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon\right)\right]$$
$$= \frac{1}{1-\alpha}\frac{\int q(\varepsilon)\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{-\alpha}\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$

4. SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) \approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \ \varepsilon_i \sim q, i = 1\ldots N$$

Training with the VR bound (Li and Turner, NeurIPS 2016)

1. MC estimator of the VR bound

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\ldots N$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$

3. Reparameterized gradient of the VR bound :

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \nabla_{\theta,\phi}\left[\frac{1}{1-\alpha}\log\left(\int q(\varepsilon)w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon\right)\right]$$
$$= \frac{1}{1-\alpha}\frac{\int q(\varepsilon)\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\right]\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon)w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{-\alpha}\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right]\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$

4. SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) \approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right),\ \varepsilon_i \sim q, i = 1\ldots N$$

Training with the VR bound (Li and Turner, NeurIPS 2016)

1. MC estimator of the VR bound

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \ldots N$$

2. Reparameterization trick $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$
3. Reparameterized gradient of the VR bound :

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \nabla_{\theta,\phi}\left[\frac{1}{1-\alpha}\log\left(\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon\right)\right]$$
$$= \frac{1}{1-\alpha}\frac{\int q(\varepsilon)\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{-\alpha}\nabla_{\theta,\phi}\left[w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right]\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$= \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$

4. SGD w.r.t. (θ,ϕ)

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) \approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \ \varepsilon_i \sim q, i = 1 \ldots N$$

Training with the VR bound (Li and Turner, NeurIPS 2016)

❶ MC estimator of the VR bound

$$\begin{aligned}
\mathrm{VR}^{(\alpha)}(\theta, \phi; x) &= \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z; x)^{1-\alpha} \mathrm{d}z \right) \\
&\approx \frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{i=1}^{N} w_{\theta,\phi}(z_i; x)^{1-\alpha} \right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \dots N
\end{aligned}$$

❷ Reparameterization trick $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$

❸ Reparameterized gradient of the VR bound :

$$\begin{aligned}
\nabla_{\theta,\phi} \mathrm{VR}^{(\alpha)}(\theta, \phi; x) &= \nabla_{\theta,\phi} \left[\frac{1}{1-\alpha} \log \left(\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha} \mathrm{d}\varepsilon \right) \right] \\
&= \frac{1}{1-\alpha} \frac{\int q(\varepsilon) \nabla_{\theta,\phi} \left[w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha} \right] \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha} \mathrm{d}\varepsilon} \\
&= \frac{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{-\alpha} \nabla_{\theta,\phi} \left[w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right] \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha} \mathrm{d}\varepsilon} \\
&= \frac{\int q(\varepsilon) w_{\theta,\phi}(z; x)^{1-\alpha} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right) \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z; x)^{1-\alpha} \mathrm{d}\varepsilon}
\end{aligned}$$

❹ SGD w.r.t. (θ, ϕ)

$$\nabla_{\theta,\phi} \mathrm{VR}^{(\alpha)}(\theta, \phi; x) \approx \sum_{i=1}^{N} \frac{w_{\theta,\phi}(z_i; x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha}} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x) \right), \ \varepsilon_i \sim q, i = 1 \dots N$$

Some important comments

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\dots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\dots N$$

→ Sanity check : $\nabla_{\theta,\phi}\mathrm{VR}^{(1)}(\theta,\phi;x) = \nabla_{\theta,\phi}\mathrm{ELBO}(\theta,\phi;x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Some important comments

$$
\begin{aligned}
\mathrm{VR}^{(\alpha)}(\theta, \phi; x) &= \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right) \\
&\approx \frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha} \right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \ldots N
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\theta,\phi} \mathrm{VR}^{(\alpha)}(\theta, \phi; x) &= \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right) \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}\varepsilon} \\
&\approx \sum_{i=1}^{N} \frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x) \right), \quad \varepsilon_i \sim q, \quad i = 1 \ldots N
\end{aligned}
$$

→ Sanity check : $\nabla_{\theta,\phi} \mathrm{VR}^{(1)}(\theta, \phi; x) = \nabla_{\theta,\phi} \mathrm{ELBO}(\theta, \phi; x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Some important comments

$$\begin{aligned}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)\\ &\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^N w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i\sim q_\phi(\cdot|x), \quad i=1\ldots N\end{aligned}$$

$$\begin{aligned}\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}\\ &\approx \sum_{i=1}^N \frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^N w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i\sim q, \quad i=1\ldots N\end{aligned}$$

→ Sanity check : $\nabla_{\theta,\phi}\mathrm{VR}^{(1)}(\theta,\phi;x) = \nabla_{\theta,\phi}\mathrm{ELBO}(\theta,\phi;x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Some important comments

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}z \right)$$
$$\approx \frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha} \right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \ldots N$$

$$\nabla_{\theta,\phi} \mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right) \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha} \mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N} \frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x) \right), \quad \varepsilon_i \sim q, \quad i = 1 \ldots N$$

→ Sanity check : $\nabla_{\theta,\phi} \mathrm{VR}^{(1)}(\theta, \phi; x) = \nabla_{\theta,\phi} \mathrm{ELBO}(\theta, \phi; x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Some important comments

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\dots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\dots N$$

→ Sanity check : $\nabla_{\theta,\phi}\mathrm{VR}^{(1)}(\theta,\phi;x) = \nabla_{\theta,\phi}\mathrm{ELBO}(\theta,\phi;x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Some important comments

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$

$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\dots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$

$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\dots N$$

→ Sanity check : $\nabla_{\theta,\phi}\mathrm{VR}^{(1)}(\theta,\phi;x) = \nabla_{\theta,\phi}\mathrm{ELBO}(\theta,\phi;x)$

→ Training with $\alpha < 1$ lead to positive empirical results

→ However,

1. The VR bound can only be estimated using biased MC estimators
2. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha = 1$)

Problem 1

- Li and Turner (Theorem 2, NeurIPS 2016) looked into the properties of the biased approximation of the VR bound

$$\begin{aligned}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\ w_{\theta,\phi}(z;x)^{1-\alpha}\,\mathrm{d}z\right)\\ &\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right), \quad z_j \sim q_\phi(z|x), \quad j = 1\ldots N\end{aligned}$$

- More precisely, they investigated the expectation of the biased MC approximation, i.e.

$$\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i|x)\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right)\mathrm{d}z_{1:N}$$

1. For all $\alpha \le 1$ and all $N \in \mathbb{N}^\star$

$$\mathrm{ELBO}(\theta,\phi;x) \le \ell_N^{(\alpha)}(\theta,\phi;x) \le \ell_{N+1}^{(\alpha)}(\theta,\phi;x) \le \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$$

2. $\ell_N^{(\alpha)}(\theta,\phi;x) \to \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ as $N \to \infty$

Problem 1

- Li and Turner (Theorem 2, NeurIPS 2016) looked into the properties of the biased approximation of the VR bound

$$\begin{aligned}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}z\right) \\ &\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right), \quad z_j \sim q_\phi(z|x), \quad j = 1\dots N\end{aligned}$$

- More precisely, they investigated the expectation of the biased MC approximation, i.e.

$$\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i|x)\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right)\mathrm{d}z_{1:N}$$

1. For all $\alpha \leq 1$ and all $N \in \mathbb{N}^\star$

$$\mathrm{ELBO}(\theta,\phi;x) \leq \ell_N^{(\alpha)}(\theta,\phi;x) \leq \ell_{N+1}^{(\alpha)}(\theta,\phi;x) \leq \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$$

2. $\ell_N^{(\alpha)}(\theta,\phi;x) \to \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ as $N \to \infty$

Problem 1

- Li and Turner (Theorem 2, NeurIPS 2016) looked into the properties of the biased approximation of the VR bound

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right), \quad z_j \sim q_\phi(z|x), \quad j = 1\ldots N$$

- More precisely, they investigated the expectation of the biased MC approximation, i.e.

$$\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i|x)\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right)\mathrm{d}z_{1:N}$$

1. For all $\alpha \leq 1$ and all $N \in \mathbb{N}^\star$

$$\mathrm{ELBO}(\theta,\phi;x) \leq \ell_N^{(\alpha)}(\theta,\phi;x) \leq \ell_{N+1}^{(\alpha)}(\theta,\phi;x) \leq \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$$

2. $\ell_N^{(\alpha)}(\theta,\phi;x) \to \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ as $N \to \infty$

Problem 1

- Li and Turner (Theorem 2, NeurIPS 2016) looked into the properties of the biased approximation of the VR bound

$$
\begin{aligned}
\mathrm{VR}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}z\right) \\
&\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right), \quad z_j \sim q_\phi(z|x), \quad j = 1\ldots N
\end{aligned}
$$

- More precisely, they investigated the expectation of the biased MC approximation, i.e.

$$
\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i|x)\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}\right)\mathrm{d}z_{1:N}
$$

1. For all $\alpha \leq 1$ and all $N \in \mathbb{N}^\star$

$$
\mathrm{ELBO}(\theta,\phi;x) \leq \ell_N^{(\alpha)}(\theta,\phi;x) \leq \ell_{N+1}^{(\alpha)}(\theta,\phi;x) \leq \mathrm{VR}^{(\alpha)}(\theta,\phi;x)
$$

2. $\ell_N^{(\alpha)}(\theta,\phi;x) \to \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ as $N \to \infty$

At this stage

- VR bound : interesting generalization of the ELBO

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\dots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\dots N$$

- Two problems :
  1. The MC estimation of the VR bound is biased
  2. The SGD with the VR bound uses biased estimators

At this stage

- VR bound : interesting generalization of the ELBO

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\dots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\dots N$$

- Two problems :

1. The MC estimation of the VR bound is biased

2. The SGD with the VR bound uses biased estimators

At this stage

- VR bound : interesting generalization of the ELBO

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\ldots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\ldots$$

- Two problems :
  1. The MC estimation of the VR bound is biased
  2. The SGD with the VR bound uses biased estimators

At this stage

- VR bound : interesting generalization of the ELBO

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}z\right)$$
$$\approx \frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{i=1}^{N} w_{\theta,\phi}(z_i;x)^{1-\alpha}\right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1\ldots N$$

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\right)\mathrm{d}\varepsilon}{\int q(\varepsilon)w_{\theta,\phi}(z;x)^{1-\alpha}\mathrm{d}\varepsilon}$$
$$\approx \sum_{i=1}^{N}\frac{w_{\theta,\phi}(z_i;x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha}}\nabla_{\theta,\phi}\left(\log w_{\theta,\phi}(f(\varepsilon_i,\phi;x);x)\right), \quad \varepsilon_i \sim q, \quad i = 1\ldots N$$

- Two problems :

  1. The MC estimation of the VR bound is biased

  2. The SGD with the VR bound uses biased estimators

At this stage

- VR bound : interesting generalization of the ELBO

$$
\begin{aligned}
\mathrm{VR}^{(\alpha)}(\theta, \phi; x) &= \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) w_{\theta,\phi}(z; x)^{1-\alpha} \mathrm{d}z \right) \\
&\approx \frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{i=1}^{N} w_{\theta,\phi}(z_i; x)^{1-\alpha} \right), \quad z_i \sim q_\phi(\cdot|x), \quad i = 1 \dots N
\end{aligned}
$$

$$
\nabla_{\theta,\phi} \mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) w_{\theta,\phi}(z; x)^{1-\alpha} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right) \mathrm{d}\varepsilon}{\int q(\varepsilon) w_{\theta,\phi}(z; x)^{1-\alpha} \mathrm{d}\varepsilon}
$$

$$
\approx \sum_{i=1}^{N} \frac{w_{\theta,\phi}(z_i; x)^{1-\alpha}}{\sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha}} \nabla_{\theta,\phi} \left(\log w_{\theta,\phi}(f(\varepsilon_i, \phi; x); x) \right), \quad \varepsilon_i \sim q, \quad i = 1 \dots N
$$

- Two problems :

  1. The MC estimation of the VR bound is biased

     $\rightarrow$ Some control of the approximation error via

$$
\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}
$$

  2. The SGD with the VR bound uses biased estimators

An idea

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

An idea

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

💡 Could this expectation be seen as a variational bound?

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

An idea

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

💡 Could this expectation be seen as a variational bound?

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

Outline

The VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi} \ell_N^{(\alpha)}(\theta, \phi; x)$$

$$= \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^N \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \dots N$$

The VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi} \ell_N^{(\alpha)}(\theta, \phi; x)$$

$$= \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \ldots N$$

The VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi} \ell_N^{(\alpha)}(\theta, \phi; x)$$

$$= \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \ldots N$$

The VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi} \ell_N^{(\alpha)}(\theta, \phi; x)$$

$$= \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \dots N$$

The VR-IWAE bound

For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j;x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi}\ell_N^{(\alpha)}(\theta,\phi;x)$$
$$= \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k;x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x) \right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^N \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k;x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \quad \varepsilon_j \sim q, \quad j = 1 \dots N$$

The VR-IWAE bound

For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(z_j;x)^{1-\alpha}\right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\nabla_{\theta,\phi}\ell_N^{(\alpha)}(\theta,\phi;x)$$

$$= \int\int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k;x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x)\right) \mathrm{d}\varepsilon_{1:N}.$$

$$\approx \sum_{j=1}^N \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k;x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \quad \varepsilon_j \sim q, \quad j = 1 \ldots N$$

The VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound is a **lower bound** on the marginal log likelihood that

1. Can be estimated using unbiased MC estimators
2. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

$$\begin{aligned} &\nabla_{\theta,\phi} \ell_N^{(\alpha)}(\theta, \phi; x) \\ &= \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}. \\ &\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \dots N \end{aligned}$$

The VR-IWAE bound provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the literature

Special cases of the VR-IWAE bound

For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha} \int\int \prod_{i=1}^N q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(z_j;x)^{1-\alpha}\right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta,\phi;x) = \mathrm{ELBO}(\theta,\phi;x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta,\phi;x) = \int\int \prod_{i=1}^N q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(z_j;x)\right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

Special cases of the VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta, \phi; x) = \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x) \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

Special cases of the VR-IWAE bound

For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j;x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta,\phi;x) = \mathrm{ELBO}(\theta,\phi;x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta,\phi;x) = \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j;x) \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

Special cases of the VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta, \phi; x) = \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x) \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

Special cases of the VR-IWAE bound

For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(z_j;x)^{1-\alpha}\right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha\to 1} \ell_N^{(\alpha)}(\theta,\phi;x) = \mathrm{ELBO}(\theta,\phi;x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta,\phi;x) = \int \int \prod_{i=1}^N q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(z_j;x)\right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

Special cases of the VR-IWAE bound

For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

- The case $\alpha \to 1$

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x)$$

- The case $\alpha = 0$

$$\ell_N^{(0)}(\theta, \phi; x) = \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x) \right) \mathrm{d}z_{1:N}$$

The VR-IWAE bound recovers the **Importance Weighted Auto-encoder (IWAE) bound** (Burda et al., ICLR 2016) when $\alpha = 0$

$\rightarrow$ Extension of the ELBO also leading to positive empirical results

💡 The VR-IWAE bound interpolates between the IWAE bound and the ELBO

It is the **theoretically-sound** extension of the IWAE bound originating from the VR bound methodology

At this stage

→ The VR-IWAE bound provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the literature

→ It is the **theoretically-sound** extension of the IWAE bound originating from the VR bound methodology, interpolates between the IWAE bound and the ELBO

Questions?

At this stage

→ The VR-IWAE bound provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the literature

→ It is the **theoretically-sound** extension of the IWAE bound originating from the VR bound methodology, interpolates between the IWAE bound and the ELBO

Questions?

At this stage

$\rightarrow$ The VR-IWAE bound provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the literature

$\rightarrow$ It is the **theoretically-sound** extension of the IWAE bound originating from the VR bound methodology, interpolates between the IWAE bound and the ELBO

Questions?

At this stage

→ The VR-IWAE bound provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the literature

→ It is the **theoretically-sound** extension of the IWAE bound originating from the VR bound methodology, interpolates between the IWAE bound and the ELBO

Questions?

→ Question Can we understand the behavior of the VR-IWAE bound as a function of $\alpha \in [0, 1)$ better?

Outline

Quantity of interest

Variational gap

For all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$$

$$= \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

where $\overline{w}_{\theta,\phi}(z_1; x), \ldots, \overline{w}_{\theta,\phi}(z_N; x)$ are the **relative weights** : for all $z \in \mathbb{R}^d$,

$$\overline{w}_{\theta,\phi}(z; x) := \frac{w_{\theta,\phi}(z; x)}{\mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z; x))} = \frac{w_{\theta,\phi}(z; x)}{p_\theta(x)} = \frac{p_\theta(z|x)}{q_\phi(z|x)},$$

NB : we will drop the dependency in x in $\overline{w}_{\theta,\phi}(z; x)$ for convenience

Quantity of interest

Variational gap

For all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$$

$$= \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}$$

where $\overline{w}_{\theta,\phi}(z_1; x), \ldots, \overline{w}_{\theta,\phi}(z_N; x)$ are the **relative weights** : for all $z \in \mathbb{R}^d$,

$$\overline{w}_{\theta,\phi}(z; x) := \frac{w_{\theta,\phi}(z; x)}{\mathbb{E}_{Z \sim q_\phi}\left(w_{\theta,\phi}(Z; x)\right)} = \frac{w_{\theta,\phi}(z; x)}{p_\theta(x)} = \frac{p_\theta(z|x)}{q_\phi(z|x)},$$

NB : we will drop the dependency in x in $\overline{w}_{\theta,\phi}(z; x)$ for convenience

Quantity of interest

Variational gap

For all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) := \ell_N^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$$

$$= \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j;x)^{1-\alpha}\right) \mathrm{d}z_{1:N}$$

where $\overline{w}_{\theta,\phi}(z_1;x),\ldots,\overline{w}_{\theta,\phi}(z_N;x)$ are the **relative weights** : for all $z \in \mathbb{R}^d$,

$$\overline{w}_{\theta,\phi}(z;x) := \frac{w_{\theta,\phi}(z;x)}{\mathbb{E}_{Z\sim q_\phi}\left(w_{\theta,\phi}(Z;x)\right)} = \frac{w_{\theta,\phi}(z;x)}{p_\theta(x)} = \frac{p_\theta(z|x)}{q_\phi(z|x)},$$

NB : we will drop the dependency in x in $\overline{w}_{\theta,\phi}(z;x)$ for convenience

Part I

N goes to infinity and d is fixed in the variational gap

N goes to infinity and d is fixed in the variational gap

→ Maddison et al. (NeurIPS 2017) followed by Domke and Sheldon (NeurIPS 2018) looked into the variational gap for the IWAE bound ($\alpha = 0$)

Informally, Domke and Sheldon (Theorem 3, NeurIPS 2018) states that

$$\Delta_N^{(0)}(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

where γ_0 is the variance of the relative weights, i.e.

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}(Z))$$

→ Comments :

- N is very beneficial to reduce $\Delta_N^{(0)}(\theta, \phi; x)$ (goes to 0 at a fast $1/N$ rate)
- Question What about $\Delta_N^{(\alpha)}(\theta, \phi; x)$, $\alpha \in [0, 1)$?

N goes to infinity and d is fixed in the variational gap

→ Maddison et al. (NeurIPS 2017) followed by Domke and Sheldon (NeurIPS 2018) looked into the variational gap for the IWAE bound ($\alpha = 0$)

Informally, Domke and Sheldon (Theorem 3, NeurIPS 2018) states that

$$\Delta_N^{(0)}(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

where γ_0 is the variance of the relative weights, i.e.

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}(Z))$$

→ Comments :

- N is very beneficial to reduce $\Delta_N^{(0)}(\theta, \phi; x)$ (goes to 0 at a fast $1/N$ rate)
- Question What about $\Delta_N^{(\alpha)}(\theta, \phi; x)$, $\alpha \in [0, 1)$?

N goes to infinity and d is fixed in the variational gap

→ Maddison et al. (NeurIPS 2017) followed by Domke and Sheldon (NeurIPS 2018) looked into the variational gap for the IWAE bound ($\alpha = 0$)

Informally, Domke and Sheldon (Theorem 3, NeurIPS 2018) states that

$$\Delta_N^{(0)}(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

where γ_0 is the variance of the relative weights, i.e.

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}(Z))$$

→ Comments :

- N is very beneficial to reduce $\Delta_N^{(0)}(\theta, \phi; x)$ (goes to 0 at a fast $1/N$ rate)
- Question What about $\Delta_N^{(\alpha)}(\theta, \phi; x)$, $\alpha \in [0, 1)$?

N goes to infinity and d is fixed in the variational gap

→ Maddison et al. (NeurIPS 2017) followed by Domke and Sheldon (NeurIPS 2018) looked into the variational gap for the IWAE bound ($\alpha = 0$)

Informally, Domke and Sheldon (Theorem 3, NeurIPS 2018) states that

$$\Delta_N^{(0)}(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

where γ_0 is the variance of the relative weights, i.e.

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}(Z))$$

→ Comments :

- N is very beneficial to reduce $\Delta_N^{(0)}(\theta, \phi; x)$ (goes to 0 at a fast $1/N$ rate)
- Question What about $\Delta_N^{(\alpha)}(\theta, \phi; x)$, $\alpha \in [0, 1)$?

N goes to infinity and d is fixed in the variational gap

→ Maddison et al. (NeurIPS 2017) followed by Domke and Sheldon (NeurIPS 2018) looked into the variational gap for the IWAE bound ($\alpha = 0$)

Informally, Domke and Sheldon (Theorem 3, NeurIPS 2018) states that

$$\Delta_N^{(0)}(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

where γ_0 is the variance of the relative weights, i.e.

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}(Z))$$

→ Comments :

- N is very beneficial to reduce $\Delta_N^{(0)}(\theta, \phi; x)$ (goes to 0 at a fast $1/N$ rate)
- Question What about $\Delta_N^{(\alpha)}(\theta, \phi; x)$, $\alpha \in [0, 1)$?

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0, 1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

→ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

→ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0, 1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha}/\mathbb{E}_{Z\sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1}\mathbb{V}_{Z\sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0, 1)$, denote $\overline{w}^{(\alpha)}_{\theta,\phi}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}^{(\alpha)}_{\theta,\phi}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0, 1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\rightarrow$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\rightarrow$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result when $N \to \infty$ and d is fixed

Theorem 1

Let $\alpha \in [0,1)$, denote $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$ and $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi(\cdot|x)}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$. Then, under "*some conditions*", we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

$\to$ Two main terms :

1. A term going to zero at a fast $1/N$ rate that depends on γ_α^2
2. An error term $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [decreases away from 0 as α increases]

The hyperparameter α balances between these two terms meaning that a proper tuning of α may be beneficial in practice

$\to$ "*some conditions*"

- generalize the conditions from Domke and Sheldon (2018)
- do not get more restrictive as α increases, motivating $\alpha \in (0,1)$
- one of them controls γ_α^2

💡 To the best of our knowledge, first result shedding light on how α may play a role in Rényi's α-divergence VI

Example

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1 - \alpha}.$$

→ Sanity check : $\mathbb{E}(\overline{w}_{\theta,\phi}) = \mathbb{E}(\exp(-\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_1)) = 1$

→ **Gaussian example** Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, with $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ is the d-dimensional vector whose coordinates are all equal to 1. Then $\sigma = 1$.

Example

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1-\alpha}.$$

$\rightarrow$ Sanity check : $\mathbb{E}(\overline{w}_{\theta,\phi}) = \mathbb{E}(\exp(-\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_1)) = 1$

$\rightarrow$ **Gaussian example** Set $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$, with $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ is the d-dimensional vector whose coordinates are all equal to 1. Then $\sigma = 1$.

Example

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1 - \alpha}.$$

$\rightarrow$ Sanity check : $\mathbb{E}(\overline{w}_{\theta,\phi}) = \mathbb{E}(\exp(-\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_1)) = 1$

$\rightarrow$ **Gaussian example** Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, with $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ is the d-dimensional vector whose coordinates are all equal to 1. Then $\sigma = 1$.

Gaussian example and Theorem 1 empirically

- $\Delta_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j; x)^{1-\alpha} \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \dots N$$

- Theorem 1 is represented through functions of the form:

$$c \mapsto -\frac{\alpha d}{2} - \frac{\exp\left[(1-\alpha)^2 d\right] - 1}{2(1-\alpha)N} + \frac{c}{N}$$

Gaussian example and Theorem 1 empirically

- $\Delta_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j; x)^{1-\alpha} \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \ldots N$$

- Theorem 1 is represented through functions of the form:

$$c \mapsto -\frac{\alpha d}{2} - \frac{\exp\left[(1-\alpha)^2 d\right] - 1}{2(1-\alpha)N} + \frac{c}{N}$$

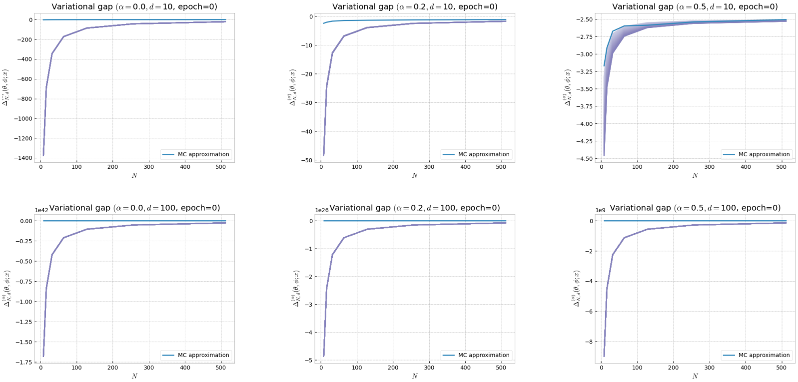
Gaussian example and Theorem 1 empirically

- $\Delta_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j; x)^{1-\alpha} \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \ldots N$$

- Theorem 1 is represented through functions of the form:

$$c \mapsto -\frac{\alpha d}{2} - \frac{\exp\left[(1-\alpha)^2 d\right] - 1}{2(1-\alpha)N} + \frac{c}{N}$$



Example 1 revisited

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1-\alpha}.$$

→ Theorem 1 may not capture what is happening in **high dimensions**
i.e. we may never use N large enough in high-dimensional settings for the asymptotic regime to kick in

→ Question Analysis as both d and N go to infinity? $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$

Example 1 revisited

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1-\alpha}.$$

→ Theorem 1 may not capture what is happening in **high dimensions**
i.e. we may never use N large enough in high-dimensional settings for the asymptotic regime to kick in

→ Question Analysis as both d and N go to infinity? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

Example 1 revisited

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1 - \alpha}.$$

$\rightarrow$ Theorem 1 may not capture what is happening in **high dimensions**
i.e. we may never use N large enough in high-dimensional settings for the asymptotic regime to kick in

$\rightarrow$ Question Analysis as both d and N go to infinity? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

Example 1 revisited

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1 - \alpha}.$$

→ Theorem 1 may not capture what is happening in **high dimensions**
i.e. we may never use N large enough in high-dimensional settings for the asymptotic regime to kick in

→ Question Analysis as both d and N go to infinity? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

Example 1 revisited

Example 1 : Log-normal distribution of the relative weights

Let $\sigma > 0$, $S_1, \ldots, S_N$ be **i.i.d. normal r.v** and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \ldots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

Then, for all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1-\alpha}.$$

$\rightarrow$ Theorem 1 may not capture what is happening in **high dimensions**
i.e. we may never use N large enough in high-dimensional settings for the asymptotic regime to kick in

$\rightarrow$ Question Analysis as both d and N go to infinity? $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$

Part II

N and d go to infinity in the variational gap

$N, d \to \infty$ in the variational gap

$\to$ Key intuition : it is typically possible to approximate the distribution of the relative weights by a **log-normal distribution** of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \dots N.$$

$\to$ Theoretical study in two steps :

1. Log-normal case : $d, N \to \infty$ with $\frac{\log N}{d} \to 0$
2. Approximate log-normal case : $d, N \to \infty$ with $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ in the variational gap

$\to$ Key intuition : it is typically possible to approximate the distribution of the relative weights by a **log-normal distribution** of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \ldots N.$$

$\to$ Theoretical study in two steps :

1. Log-normal case : $d, N \to \infty$ with $\frac{\log N}{d} \to 0$
2. Approximate log-normal case : $d, N \to \infty$ with $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ in the variational gap

$\rightarrow$ Key intuition : it is typically possible to approximate the distribution of the relative weights by a **log-normal distribution** of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \dots N.$$

$\rightarrow$ Theoretical study in two steps :

1. Log-normal case : $d, N \to \infty$ with $\frac{\log N}{d} \to 0$
2. Approximate log-normal case : $d, N \to \infty$ with $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ in the variational gap

$\to$ Key intuition : it is typically possible to approximate the distribution of the relative weights by a **log-normal distribution** of the form

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \dots N.$$

$\to$ Theoretical study in two steps :

1. Log-normal case : $d, N \to \infty$ with $\frac{\log N}{d} \to 0$
2. Approximate log-normal case : $d, N \to \infty$ with $\frac{\log N}{d^{1/3}} \to 0$

Part II.1

Log-normal case

Main result in the log-normal case

Theorem 2

Let $S_1, \ldots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

$\rightarrow$ Informally

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

Main result in the log-normal case

Theorem 2

Let $S_1, \ldots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

$\rightarrow$ Informally

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

Main result in the log-normal case

Theorem 2

Let $S_1, \ldots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0, 1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Informally

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

→ Comparison with Theorem 1

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) = -\alpha \cdot \frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- While increasing N decreases the variational gap for N large enough, it does so by a factor which is negligible before the term $-d\sigma^2/2$
- This time, the term $-d\sigma^2/2$ **does not** depend on α

Main result in the log-normal case

Theorem 2

Let $S_1, \ldots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \ldots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0, 1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

$\rightarrow$ Informally

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

$\rightarrow$ Comparison with Theorem 1

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) = -\alpha \cdot \frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- While increasing N decreases the variational gap for N large enough, it does so by a factor which is negligible before the term $-d\sigma^2/2$
- This time, the term $-d\sigma^2/2$ **does not** depend on α

Main result in the log-normal case

Theorem 2

Let $S_1, \dots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \dots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0, 1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Informally

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

→ Comparison with Theorem 1

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = -\alpha \cdot \frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- While increasing N decreases the variational gap for N large enough, it does so by a factor which is negligible before the term $-d\sigma^2/2$
- This time, the term $-d\sigma^2/2$ **does not** depend on α

Main result in the log-normal case

Theorem 2

Let $S_1, \dots, S_N$ be **i.i.d. normal random variables**. Further assume that

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \dots N.$$

with $\sigma > 0$. Then, for all $\alpha \in [0, 1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta^{(\alpha)}_{N,d}(\theta, \phi; x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Informally

$$\Delta^{(\alpha)}_{N,d}(\theta, \phi; x) \approx -\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right)$$

→ Weight collapse phenomenon : for all $\alpha \in [0, 1)$,

$$\Delta^{(\alpha)}_{N,d}(\theta, \phi; x) \approx \mathrm{ELBO}(\theta, \phi; x) - \ell(\theta; x), \quad \text{as } N, d \to \infty \text{ with } \tfrac{\log N}{d} \to 0.$$

Gaussian example revisited

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$, with $\theta = 0\cdot\boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ is the d-dimensional vector whose coordinates are all equal to 1. Then

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1\ldots N$$

with $\sigma = 1$.

- Theorem 1

$$\Delta^{(\alpha)}_{N,d}(\theta,\phi;x) = -\alpha\cdot\frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right]-1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Theorem 2

$$\lim_{\substack{N,d\to\infty \\ \log N/d\to 0}} \Delta^{(\alpha)}_{N,d}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

Gaussian example revisited

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, with $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ is the d-dimensional vector whose coordinates are all equal to 1. Then

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \ldots N$$

with $\sigma = 1$.

- Theorem 1

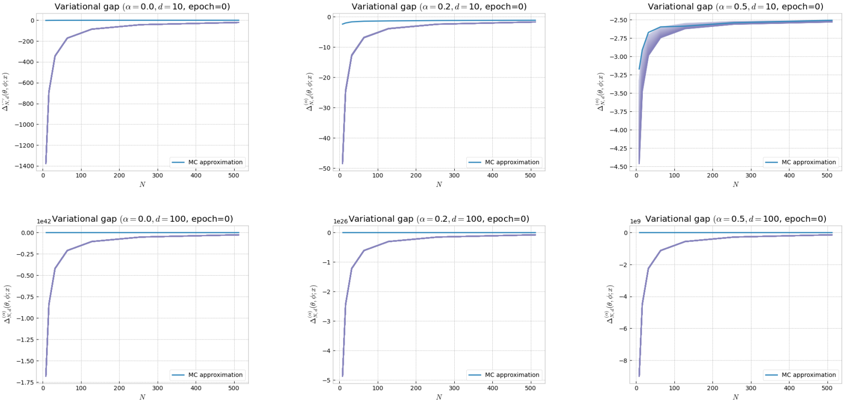
$$\Delta^{(\alpha)}_{N,d}(\theta, \phi; x) = -\alpha \cdot \frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Theorem 2

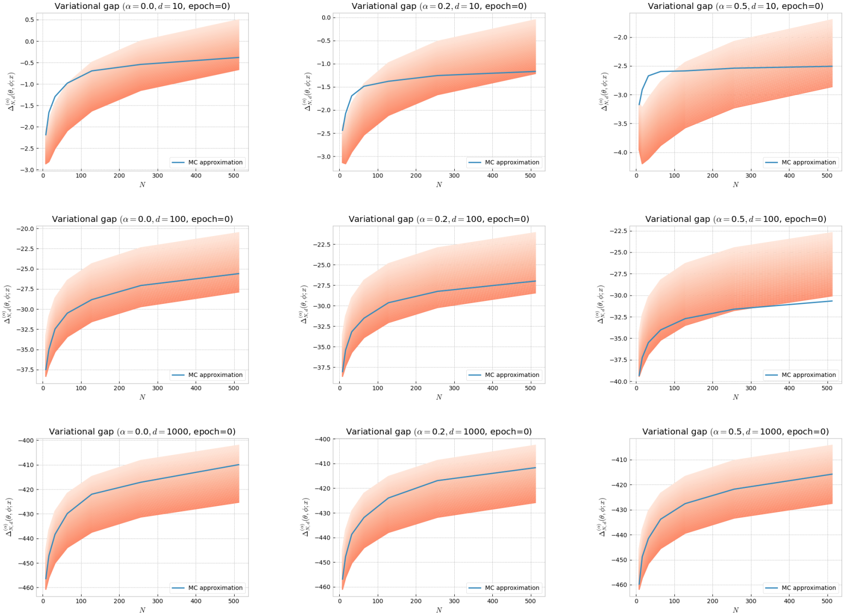
$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta^{(\alpha)}_{N,d}(\theta, \phi; x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d \log N}}\right)\right) = 0.$$

Weight collapse phenomenon might occur even for simple examples!

Gaussian example and Theorem 1 empirically



Gaussian example and Theorem 2 empirically



Part II.2

Approximate log-normal case

Assumptions

Let $S_1, \ldots, S_N$ be such that :

$$S_i = \frac{1}{\sigma\sqrt{d}} \sum_{j=1}^{d} \xi_{i,j}, \quad i = 1 \ldots N$$

We will work under (A1) :

(A1) For all $i = 1 \ldots N$,

1. $\xi_{i,1}, \ldots, \xi_{i,d}$ are i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure and satisfy $\mathbb{E}(\xi_{i,1}) = 0$ and $\mathbb{V}(\xi_{i,1}) = \sigma^2 < \infty$.
2. There exists $K > 0$ such that:

$$|\mathbb{E}(\xi_{i,1}^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Approximate log-normal weights

$$\begin{aligned} \log \overline{w}_{\theta,\phi}(z_i) &= -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N \\ &= -da - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N \end{aligned}$$

with $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$

Assumptions

Let $S_1, \ldots, S_N$ be such that :

$$S_i = \frac{1}{\sigma\sqrt{d}} \sum_{j=1}^{d} \xi_{i,j}, \quad i = 1 \ldots N$$

We will work under (A1) :

(A1) For all $i = 1 \ldots N$,

1. $\xi_{i,1}, \ldots, \xi_{i,d}$ are i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure and satisfy $\mathbb{E}(\xi_{i,1}) = 0$ and $\mathbb{V}(\xi_{i,1}) = \sigma^2 < \infty$.
2. There exists $K > 0$ such that:

$$|\mathbb{E}(\xi_{i,1}^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Approximate log-normal weights

$$\log \overline{w}_{\theta,\phi}(z_i) = -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N$$
$$= -da - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N$$

with $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$

Assumptions

Let $S_1, \dots, S_N$ be such that :

$$S_i = \frac{1}{\sigma\sqrt{d}} \sum_{j=1}^{d} \xi_{i,j}, \quad i = 1 \dots N$$

We will work under (A1) :

(A1) For all $i = 1 \dots N$,

1. $\xi_{i,1}, \dots, \xi_{i,d}$ are i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure and satisfy $\mathbb{E}(\xi_{i,1}) = 0$ and $\mathbb{V}(\xi_{i,1}) = \sigma^2 < \infty$.
2. There exists $K > 0$ such that:

$$|\mathbb{E}(\xi_{i,1}^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Approximate log-normal weights

$$\begin{aligned} \log \overline{w}_{\theta,\phi}(z_i) &= -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S_1)) - \sigma\sqrt{d} S_i, \quad i = 1 \dots N \\ &= -da - \sigma\sqrt{d} S_i, \quad i = 1 \dots N \end{aligned}$$

with $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$

Main result in the approximate log-normal case

Theorem 3

Assume (A1) and that

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d}S_i, \quad i = 1 \dots N.$$

Then, $a > 0$ and for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta^{(\alpha)}_{N,d}(\theta,\phi;x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Weight collapse phenomenon : for all $\alpha \in [0,1)$,

$$\Delta^{(\alpha)}_{N,d}(\theta,\phi;x) \approx \mathrm{ELBO}(\theta,\phi;x) - \ell(\theta;x), \quad \text{as } N,d\to\infty \text{ with } \frac{\log N}{d^{1/3}} \to 0.$$

The condition that N should grow at least exponentially with d has been replaced by the less restrictive yet still stringent condition that N should grow at least exponentially with $d^{1/3}$.

→ NB : no dependency in α left in the asymptotic regime

Main result in the approximate log-normal case

Theorem 3

Assume (A1) and that

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N.$$

Then, $a > 0$ and for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta^{(\alpha)}_{N,d}(\theta,\phi;x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Weight collapse phenomenon : for all $\alpha \in [0,1)$,

$$\Delta^{(\alpha)}_{N,d}(\theta,\phi;x) \approx \mathrm{ELBO}(\theta,\phi;x) - \ell(\theta;x), \quad \text{as } N,d\to\infty \text{ with } \frac{\log N}{d^{1/3}} \to 0.$$

The condition that N should grow at least exponentially with d has been replaced by the less restrictive yet still stringent condition that N should grow at least exponentially with $d^{1/3}$.

→ NB : no dependency in α left in the asymptotic regime

Main result in the approximate log-normal case

Theorem 3

Assume (A1) and that

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d}S_i, \quad i = 1 \dots N.$$

Then, $a > 0$ and for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

→ Weight collapse phenomenon : for all $\alpha \in [0,1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx \mathrm{ELBO}(\theta,\phi;x) - \ell(\theta;x), \quad \text{as } N,d\to\infty \text{ with } \frac{\log N}{d^{1/3}} \to 0.$$

The condition that N should grow at least exponentially with d has been replaced by the less restrictive yet still stringent condition that N should grow at least exponentially with $d^{1/3}$.

→ NB : no dependency in α left in the asymptotic regime

Linear Gaussian example

Linear Gaussian example (Rainforth et al., ICML 2018)

Set $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ with $\theta \in \mathbb{R}^d$, and $q_\phi(z|x) = \mathcal{N}(z; Ax + b, 2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\phi = (\tilde{a}, b) \in \mathbb{R}^d \times \mathbb{R}^d$.
Then, we can write

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N.$$

with $\sigma^2 = \frac{1}{18} + \frac{8}{3}\lambda^2$ and $a = \lambda^2 + \frac{1}{6} + \frac{1}{2}\log(3/4)$, where $\lambda = \frac{\left\| \frac{\theta+x}{2} - Ax - b \right\|}{\sqrt{d}}$

$\rightarrow$ (A1) holds if we set $(\theta, \phi) = (\theta^\star, \phi^\star)$!
$[\theta^\star = T^{-1}\sum_{t=1}^T x_t,\ \phi^\star = (a^\star, b^\star)$ with $a^\star = \frac{1}{2}\boldsymbol{u}_d,\ b^\star = \frac{\theta^\star}{2}]$

- Theorem 1

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = \frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] - \frac{(4-\alpha)^d(15-6\alpha)^{-\frac{d}{2}} - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Theorem 3

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d}\log N}\right)\right) = 0$$

Linear Gaussian example

Linear Gaussian example (Rainforth et al., ICML 2018)

Set $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ with $\theta \in \mathbb{R}^d$, and $q_\phi(z|x) = \mathcal{N}(z; Ax + b, 2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\phi = (\tilde{a}, b) \in \mathbb{R}^d \times \mathbb{R}^d$.
Then, we can write

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d} S_i, \quad i = 1 \dots N.$$

with $\sigma^2 = \frac{1}{18} + \frac{8}{3}\lambda^2$ and $a = \lambda^2 + \frac{1}{6} + \frac{1}{2}\log(3/4)$, where $\lambda = \frac{\left\| \frac{\theta + x}{2} - Ax - b \right\|}{\sqrt{d}}$

$\rightarrow$ (A1) holds if we set $(\theta, \phi) = (\theta^\star, \phi^\star)$!
$[\theta^\star = T^{-1} \sum_{t=1}^T x_t,\ \phi^\star = (a^\star, b^\star)$ with $a^\star = \frac{1}{2}\boldsymbol{u}_d,\ b^\star = \frac{\theta^\star}{2}]$

- Theorem 1

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = \frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] - \frac{(4-\alpha)^d(15-6\alpha)^{-\frac{d}{2}} - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Theorem 3

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d \log N}}\right)\right) = 0$$

Linear Gaussian example

Linear Gaussian example (Rainforth et al., ICML 2018)

Set $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ with $\theta \in \mathbb{R}^d$, and $q_\phi(z|x) = \mathcal{N}(z; Ax + b, 2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\phi = (\tilde{a}, b) \in \mathbb{R}^d \times \mathbb{R}^d$. Then, we can write

$$\log \overline{w}_{\theta,\phi}(z_i) = -da - \sigma\sqrt{d} S_i, \quad i = 1 \dots N.$$

with $\sigma^2 = \frac{1}{18} + \frac{8}{3}\lambda^2$ and $a = \lambda^2 + \frac{1}{6} + \frac{1}{2}\log(3/4)$, where $\lambda = \frac{\left\|\frac{\theta+x}{2} - Ax - b\right\|}{\sqrt{d}}$

$\to$ (A1) holds if we set $(\theta, \phi) = (\theta^\star, \phi^\star)$!
$[\theta^\star = T^{-1}\sum_{t=1}^{T} x_t,\ \phi^\star = (a^\star, b^\star)$ with $a^\star = \frac{1}{2}\boldsymbol{u}_d,\ b^\star = \frac{\theta^\star}{2}]$

- Theorem 1

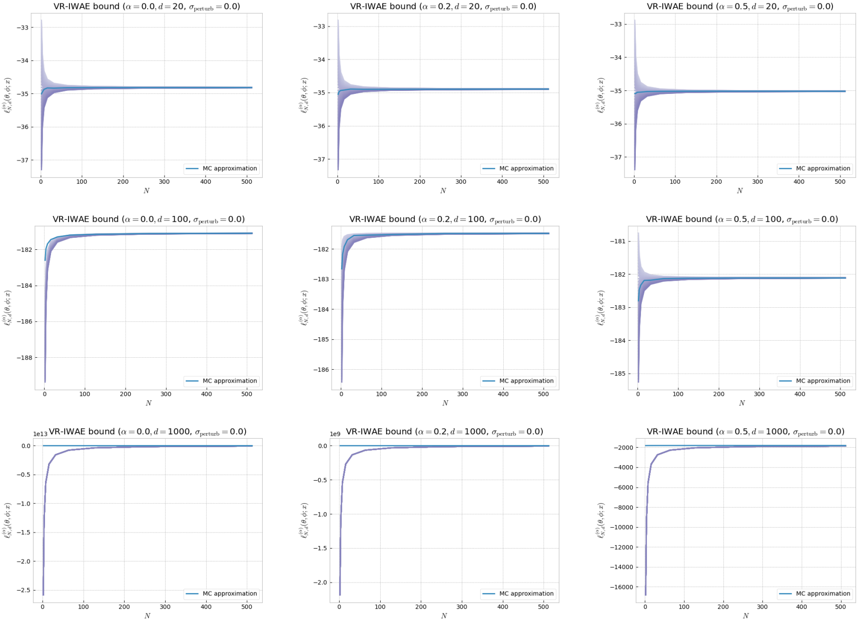
$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = \frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] - \frac{(4-\alpha)^d(15-6\alpha)^{-\frac{d}{2}} - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Theorem 3

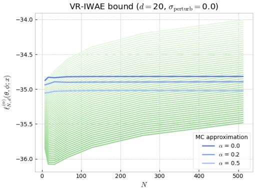
$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0$$

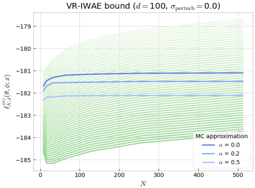
The choice of the variational approximation $q_\phi(\cdot|x)$ matters a lot!

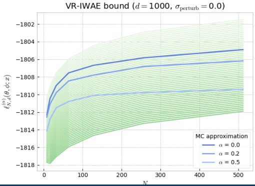
Linear Gaussian example and Theorem 1 empirically



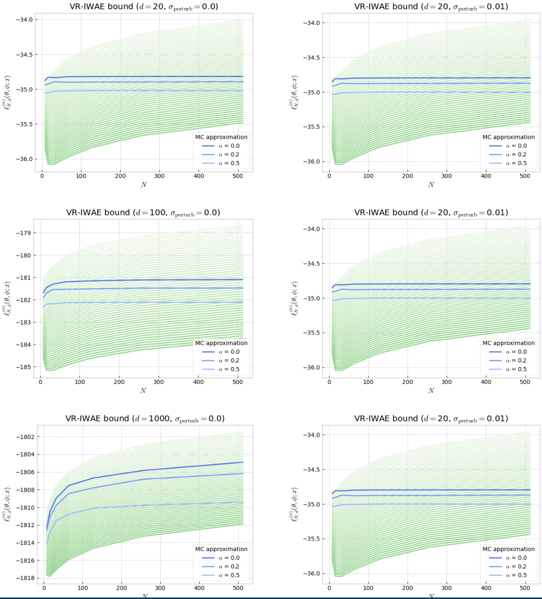
Linear Gaussian example and Theorem 3 empirically



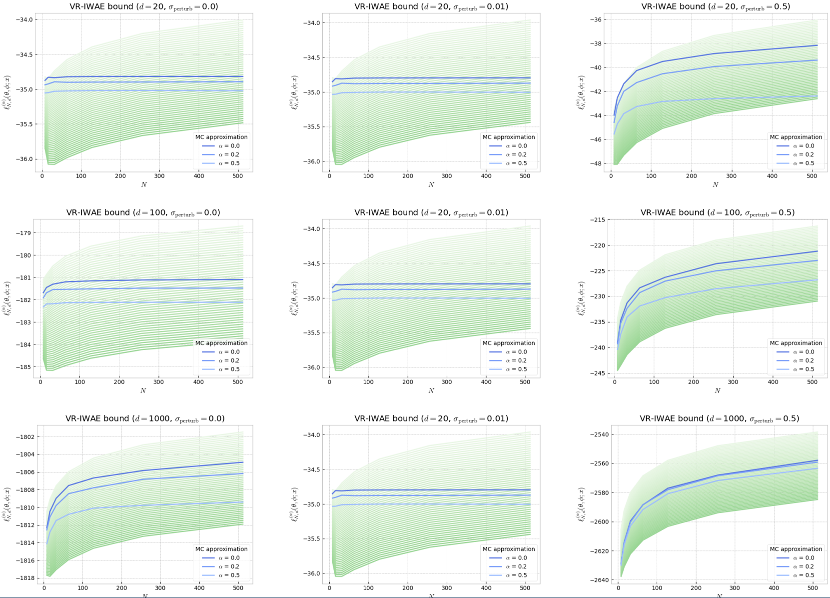




Linear Gaussian example and Theorem 3 empirically



Linear Gaussian example and Theorem 3 empirically



At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

$\rightarrow$ Two **complementary** studies

1. When $N \to \infty$ and the dimension of the latent space d is fixed

2. When $N, d \to \infty$ with (i) $\frac{\log N}{d} \to 0$ or (ii) $\frac{\log N}{d^{1/3}} \to 0$

$\rightarrow$ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

$\rightarrow$ Two **complementary** studies

1. When $N \rightarrow \infty$ and the dimension of the latent space d is fixed
2. When $N, d \rightarrow \infty$ with (i) $\frac{\log N}{d} \rightarrow 0$ or (ii) $\frac{\log N}{d^{1/3}} \rightarrow 0$

$\rightarrow$ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

$\rightarrow$ Two **complementary** studies

❶ When $N \to \infty$ and the dimension of the latent space d is fixed

❷ When $N, d \to \infty$ with (i) $\frac{\log N}{d} \to 0$ or (ii) $\frac{\log N}{d^{1/3}} \to 0$

$\rightarrow$ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

$\rightarrow$ Two **complementary** studies

1. When $N \rightarrow \infty$ and the dimension of the latent space d is fixed

2. When $N, d \rightarrow \infty$ with (i) $\frac{\log N}{d} \rightarrow 0$ or (ii) $\frac{\log N}{d^{1/3}} \rightarrow 0$

$\rightarrow$ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

→ Two **complementary** studies

1. When $N \to \infty$ and the dimension of the latent space d is fixed

   💡 This analysis is tailored for **low to medium dimensions** settings

2. When $N, d \to \infty$ with (i) $\frac{\log N}{d} \to 0$ or (ii) $\frac{\log N}{d^{1/3}} \to 0$

→ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

→ Two **complementary** studies

1. When $N \to \infty$ and the dimension of the latent space d is fixed

   💡 This analysis is tailored for **low to medium dimensions** settings

2. When $N, d \to \infty$ with (i) $\frac{\log N}{d} \to 0$ or (ii) $\frac{\log N}{d^{1/3}} \to 0$

   💡 This analysis is tailored for **high-dimensional** settings

→ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

At this stage

Quantity of interest : **variational gap**

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x), \quad \alpha \in [0, 1)$$

$\rightarrow$ Two **complementary** studies

1. When $N \to \infty$ and the dimension of the latent space d is fixed

   💡 This analysis is tailored for **low to medium dimensions** settings

2. When $N, d \to \infty$ with (i) $\frac{\log N}{d} \to 0$ or (ii) $\frac{\log N}{d^{1/3}} \to 0$

   💡 This analysis is tailored for **high-dimensional** settings

$\rightarrow$ Question Can we apply what we have learnt to a scenario where the posterior density is known up to a constant?

From theory to practice

- $\ell_N^{(\alpha)}(\theta,\phi;x)$ is estimated using the unbiased MC estimator

$$\frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j)^{1-\alpha}\right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1\ldots N$$

- Theorem 1

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

- Theorem 3 *Assuming that the weights are approximately log-normal*

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0$$

From theory to practice

- $\ell_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j)^{1-\alpha} \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \ldots N$$

- Theorem 1

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

- Theorem 3 *Assuming that the weights are approximately log-normal*

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da \left(1 - \frac{\sigma}{a} \sqrt{\frac{2 \log N}{d}} + O\left(\frac{\log \log N}{\sqrt{d \log N}} \right) \right) = 0$$

From theory to practice

- $\ell_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator
$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j)^{1-\alpha} \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \dots N$$
- Theorem 1
$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$
becomes
$$\ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$
- Theorem 3 *Assuming that the weights are approximately log-normal*
$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0$$

From theory to practice

- $\ell_N^{(\alpha)}(\theta, \phi; x)$ is estimated using the unbiased MC estimator
$$\frac{1}{1-\alpha} \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j)^{1-\alpha}\right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \dots N$$
- Theorem 1
$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$
becomes
$$\ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$
- Theorem 3 *Assuming that the weights are approximately log-normal*
$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0$$
becomes
$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \ell_{N,d}^{(\alpha)}(\theta, \phi; x) - \left[\mathrm{ELBO}(\theta, \phi; x) + \sqrt{d}\sigma\sqrt{2\log N} + O\left(\frac{\sqrt{d}\log\log N}{\sqrt{\log N}}\right)\right] = 0$$

Outline

VAE on MNIST dataset

- More details about this framework in the afternoon lecture!
- Here, we only want to look at
  1. the behavior of the relative weights
  2. the behavior of the VR-IWAE bound

VAE on MNIST dataset

- More details about this framework in the afternoon lecture!
- Here, we only want to look at
  1. the behavior of the relative weights
  2. the behavior of the VR-IWAE bound

VAE on MNIST dataset

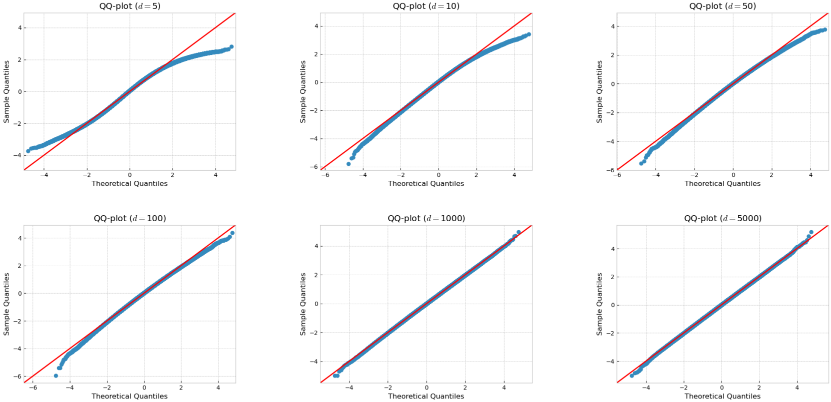
- More details about this framework in the afternoon lecture!
- Here, we only want to look at
  1. the behavior of the relative weights
  2. the behavior of the VR-IWAE bound

VAE on MNIST dataset

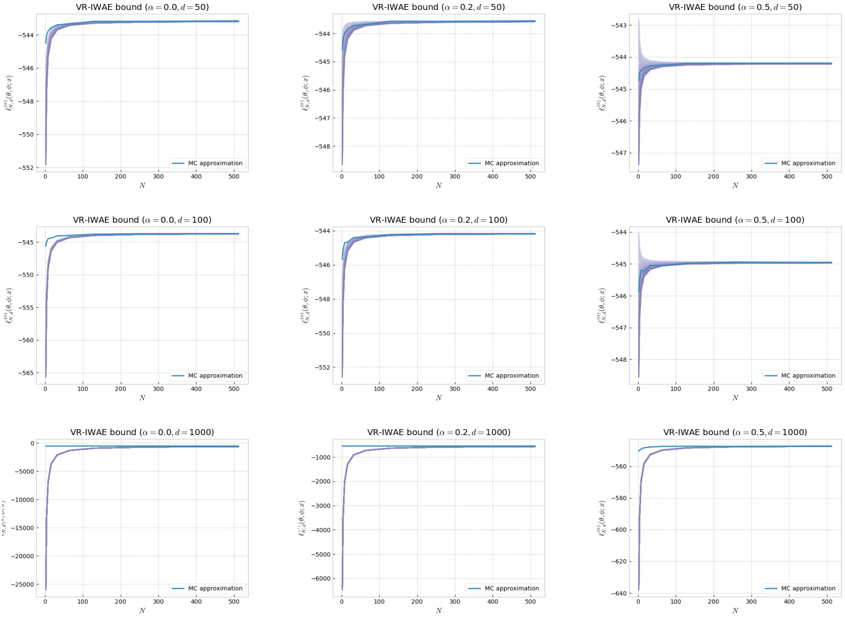
- More details about this framework in the afternoon lecture!
- Here, we only want to look at
  1. the behavior of the relative weights
  2. the behavior of the VR-IWAE bound

VAE on MNIST dataset

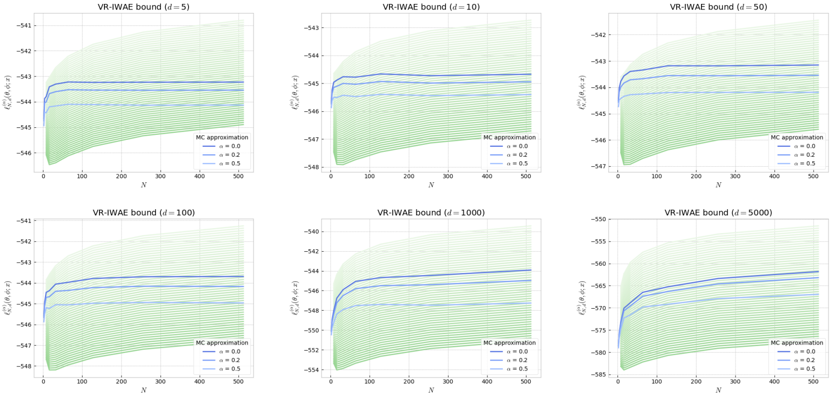
- More details about this framework in the afternoon lecture!
- Here, we only want to look at
  1. the behavior of the relative weights
  2. the behavior of the VR-IWAE bound



VAE on MNIST dataset and Theorem 1



VAE on MNIST dataset and Theorem 3



At this stage

→ Two complementary analyses of the VR-IWAE bound that we verified on a **real-world scenario**

1. Theorem 1 is tailored for low to medium dimensions settings
2. Theorem 3 is tailored for high-dimensional settings

Questions?

At this stage

→ Two complementary analyses of the VR-IWAE bound that we verified on a **real-world scenario**

1. Theorem 1 is tailored for low to medium dimensions settings
2. Theorem 3 is tailored for high-dimensional settings

Questions?

At this stage

→ Two complementary analyses of the VR-IWAE bound that we verified on a **real-world scenario**

1. Theorem 1 is tailored for low to medium dimensions settings
2. Theorem 3 is tailored for high-dimensional settings

Questions?

At this stage

→ Two complementary analyses of the VR-IWAE bound that we verified on a **real-world scenario**

1. Theorem 1 is tailored for low to medium dimensions settings

2. Theorem 3 is tailored for high-dimensional settings

Questions?

At this stage

→ Two complementary analyses of the VR-IWAE bound that we verified on a **real-world scenario**

1. Theorem 1 is tailored for low to medium dimensions settings
2. Theorem 3 is tailored for high-dimensional settings

Questions?

Question Can we say something about the gradient of the VR-IWAE bound as a function of $\alpha \in [0, 1)$?

Outline

Study of the gradient(s) of the VR-IWAE bound

Quantities of interest

- MC estimates of the reparameterized gradients of the VR-IWAE bound

$$\delta_N^{(\alpha)}(\phi_\ell) = \frac{\partial}{\partial \phi_\ell} \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(f(\varepsilon_j, \phi; x))^{1-\alpha} \right), \quad \ell = 1 \ldots L$$

$$\delta_N^{(\alpha)}(\theta_{\ell'}) = \frac{\partial}{\partial \theta_{\ell'}} \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(f(\varepsilon_j, \phi; x))^{1-\alpha} \right), \quad \ell' = 1 \ldots L'$$

  with $\phi = (\phi_1, \ldots, \phi_L)$, $\theta = (\theta_1, \ldots, \theta_{L'})$

- Signal-to-Noise Ratio

  Letting $X = (X_1, \ldots, X_L)$ be a random vector of dimension L,

$$\mathrm{SNR}[X] = \left(\frac{|\mathbb{E}(X_1)|}{\sqrt{\mathbb{V}(X_1)}}, \ldots, \frac{|\mathbb{E}(X_L)|}{\sqrt{\mathbb{V}(X_L)}} \right).$$

Study of the gradient(s) of the VR-IWAE bound

Quantities of interest

- MC estimates of the reparameterized gradients of the VR-IWAE bound

$$\delta_N^{(\alpha)}(\phi_\ell) = \frac{\partial}{\partial \phi_\ell} \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_j,\phi;x))^{1-\alpha}\right), \quad \ell = 1\dots L$$

$$\delta_N^{(\alpha)}(\theta_{\ell'}) = \frac{\partial}{\partial \theta_{\ell'}} \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_j,\phi;x))^{1-\alpha}\right), \quad \ell' = 1\dots L'$$

  with $\phi = (\phi_1, \dots, \phi_L)$, $\theta = (\theta_1, \dots, \theta_{L'})$

- Signal-to-Noise Ratio

  Letting $X = (X_1, \dots, X_L)$ be a random vector of dimension L,

$$\mathrm{SNR}[X] = \left(\frac{|\mathbb{E}(X_1)|}{\sqrt{\mathbb{V}(X_1)}}, \dots, \frac{|\mathbb{E}(X_L)|}{\sqrt{\mathbb{V}(X_L)}}\right).$$

SNR analysis in the reparameterized case

Theorem 4

Let $\alpha \in [0,1)$. Define $\tilde{w}_j = w_{\theta,\phi}(f(\varepsilon_j, \phi; x))$ and $\hat{Z}_{N,\alpha} = N^{-1} \sum_{j=1}^{N} \tilde{w}_j^{1-\alpha}$. Assume that the eighth moments of $\tilde{w}_1^{1-\alpha}$, $\partial \tilde{w}_1^{1-\alpha} / \partial \phi_\ell$ and $\partial \tilde{w}_1^{1-\alpha} / \partial \theta_{\ell'}$ are finite. Furthermore, assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{N,\alpha})^4) < \infty$. Lastly, assume that

$$\begin{aligned} &\partial \mathbb{V}(\tilde{w}_1^{1-\alpha}) / \partial \phi_\ell > 0, \quad \text{if } \alpha = 0 \\ &\partial \mathbb{E}(\tilde{w}_1^{1-\alpha}) / \partial \phi_\ell \neq 0, \quad \text{if } \alpha \in (0,1) \end{aligned}$$

and that $\partial \mathbb{E}(\tilde{w}_1^{1-\alpha}) / \partial \theta_{\ell'} \neq 0$. Then,

$$\begin{aligned} \mathrm{SNR}[\delta_N^{(\alpha)}(\phi_\ell)] &= \begin{cases} \Theta(\sqrt{1/N}) & \text{if } \alpha = 0, \\ \Theta(\sqrt{N}) & \text{if } \alpha \in (0,1) \end{cases} \\ \mathrm{SNR}[\delta_N^{(\alpha)}(\theta_{\ell'})] &= \Theta(\sqrt{N}). \end{aligned}$$

→ The IWAE case was already known from Rainforth et al. (ICML 2018)

→ Motivates $\alpha \in (0,1)$

SNR analysis in the reparameterized case

Theorem 4

Let $\alpha \in [0,1)$. Define $\tilde{w}_j = w_{\theta,\phi}(f(\varepsilon_j, \phi; x))$ and $\hat{Z}_{N,\alpha} = N^{-1} \sum_{j=1}^{N} \tilde{w}_j^{1-\alpha}$. Assume that the eighth moments of $\tilde{w}_1^{1-\alpha}$, $\partial \tilde{w}_1^{1-\alpha}/\partial \phi_\ell$ and $\partial \tilde{w}_1^{1-\alpha}/\partial \theta_{\ell'}$ are finite. Furthermore, assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{N,\alpha})^4) < \infty$. Lastly, assume that

$$\partial \mathbb{V}(\tilde{w}_1^{1-\alpha})/\partial \phi_\ell > 0, \quad \text{if } \alpha = 0$$
$$\partial \mathbb{E}(\tilde{w}_1^{1-\alpha})/\partial \phi_\ell \neq 0, \quad \text{if } \alpha \in (0,1)$$

and that $\partial \mathbb{E}(\tilde{w}_1^{1-\alpha})/\partial \theta_{\ell'} \neq 0$. Then,

$$\mathrm{SNR}[\delta_N^{(\alpha)}(\phi_\ell)] = \begin{cases} \Theta(\sqrt{1/N}) & \text{if } \alpha = 0, \\ \Theta(\sqrt{N}) & \text{if } \alpha \in (0,1) \end{cases}$$
$$\mathrm{SNR}[\delta_N^{(\alpha)}(\theta_{\ell'})] = \Theta(\sqrt{N}).$$

→ The IWAE case was already known from Rainforth et al. (ICML 2018)

→ Motivates $\alpha \in (0,1)$

SNR analysis in the reparameterized case

Theorem 4

Let $\alpha \in [0, 1)$. Define $\tilde{w}_j = w_{\theta,\phi}(f(\varepsilon_j, \phi; x))$ and $\hat{Z}_{N,\alpha} = N^{-1} \sum_{j=1}^{N} \tilde{w}_j^{1-\alpha}$. Assume that the eighth moments of $\tilde{w}_1^{1-\alpha}$, $\partial \tilde{w}_1^{1-\alpha} / \partial \phi_\ell$ and $\partial \tilde{w}_1^{1-\alpha} / \partial \theta_{\ell'}$ are finite. Furthermore, assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{N,\alpha})^4) < \infty$. Lastly, assume that

$$\begin{aligned} &\partial \mathbb{V}(\tilde{w}_1^{1-\alpha}) / \partial \phi_\ell > 0, \quad \text{if } \alpha = 0 \\ &\partial \mathbb{E}(\tilde{w}_1^{1-\alpha}) / \partial \phi_\ell \neq 0, \quad \text{if } \alpha \in (0, 1) \end{aligned}$$

and that $\partial \mathbb{E}(\tilde{w}_1^{1-\alpha}) / \partial \theta_{\ell'} \neq 0$. Then,

$$\begin{aligned} &\mathrm{SNR}[\delta_N^{(\alpha)}(\phi_\ell)] = \begin{cases} \Theta(\sqrt{1/N}) & \text{if } \alpha = 0, \\ \Theta(\sqrt{N}) & \text{if } \alpha \in (0, 1) \end{cases} \\ &\mathrm{SNR}[\delta_N^{(\alpha)}(\theta_{\ell'})] = \Theta(\sqrt{N}). \end{aligned}$$

→ The IWAE case was already known from Rainforth et al. (ICML 2018)

→ Motivates $\alpha \in (0, 1)$

Doubly-reparameterized gradients

→ Introduced in Tucker (ICLR 2019) for the IWAE bound

Theorem 5

For all $\alpha \in [0, 1]$,

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x))|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}$$

with $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \ldots J$ and

$$h_j(\alpha) = \alpha \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k)^{1-\alpha}} + (1-\alpha) \left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k)^{1-\alpha}} \right)^2 .$$

An unbiased estimator of $\partial \ell_N^{(\alpha)}(\theta, \phi; x)/\partial \phi$ is then given by

$$\sum_{j=1}^{N} h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \ldots J$.

Doubly-reparameterized gradients

→ Introduced in Tucker (ICLR 2019) for the IWAE bound

Theorem 5

For all $\alpha \in [0,1]$,

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x))|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}$$

with $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \dots J$ and

$$h_j(\alpha) = \alpha \, \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} + (1-\alpha) \, \left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \right)^2 .$$

An unbiased estimator of $\partial \ell_N^{(\alpha)}(\theta, \phi; x)/\partial \phi$ is then given by

$$\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi}$$

where $\varepsilon_1, \dots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \dots J$.

At this stage

- Setting $\alpha > 0$ instead of $\alpha = 0$ (IWAE bound) can improve on the SNR for the reparameterized estimated gradients of the VR-IWAE bound

$$\mathrm{SNR}_{\phi_\ell} = \begin{cases} \Theta(\sqrt{1/N}) & \text{if } \alpha = 0 \text{ (Rainforth et al., ICML 2018)}, \\ \Theta(\sqrt{N}) & \text{if } \alpha \in (0,1) \end{cases}$$

$$\mathrm{SNR}_{\theta_{\ell'}} = \Theta(\sqrt{N})$$

- The doubly-reparameterized gradient estimators of the IWAE (Tucker et al. ICLR 2019) generalize to the VR-IWAE bound

At this stage

- Setting $\alpha > 0$ instead of $\alpha = 0$ (IWAE bound) can improve on the SNR for the reparameterized estimated gradients of the VR-IWAE bound

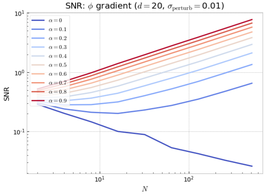
$$\mathrm{SNR}_{\phi_\ell} = \begin{cases} \Theta(\sqrt{1/N}) & \text{if } \alpha = 0 \text{ (Rainforth et al., ICML 2018)}, \\ \Theta(\sqrt{N}) & \text{if } \alpha \in (0,1) \end{cases}$$

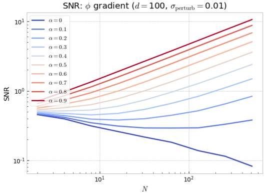
$$\mathrm{SNR}_{\theta_{\ell'}} = \Theta(\sqrt{N})$$

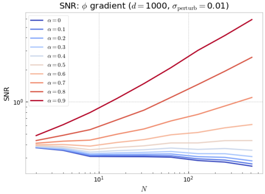
- The doubly-reparameterized gradient estimators of the IWAE (Tucker et al. ICLR 2019) generalize to the VR-IWAE bound

SNR analysis for the Linear Gaussian example : ϕ

Reparameterized

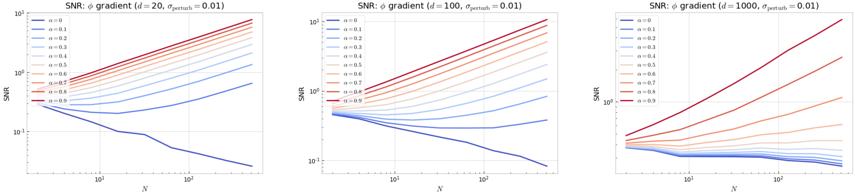




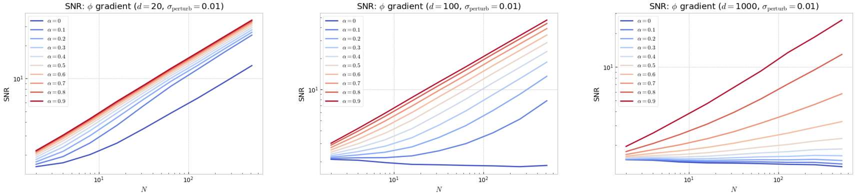


SNR analysis for the Linear Gaussian example : ϕ

Reparameterized

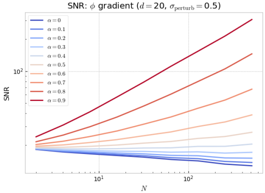


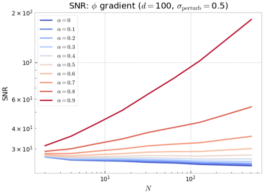
Doubly-reparameterized

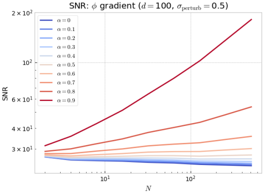


SNR analysis for the Linear Gaussian example : ϕ (cont'd)

Reparameterized

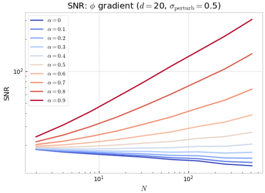


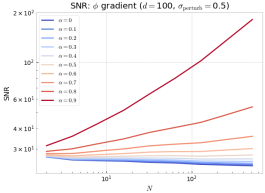


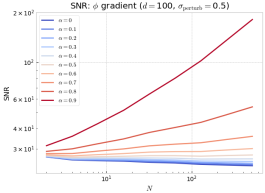


SNR analysis for the Linear Gaussian example : ϕ (cont'd)

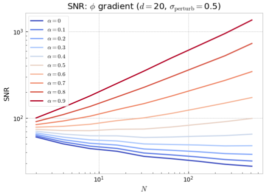
Reparameterized

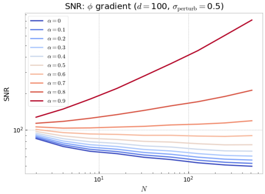


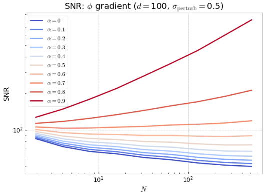




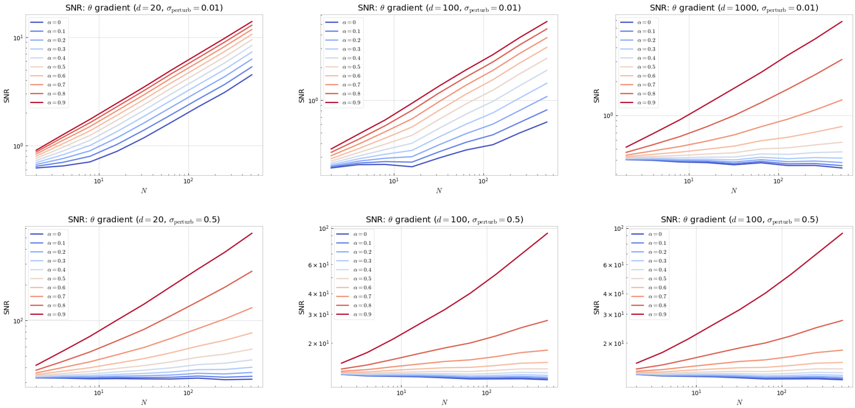
Doubly-reparameterized





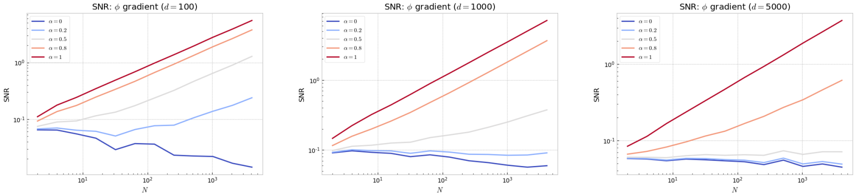


SNR analysis for the Linear Gaussian example : θ



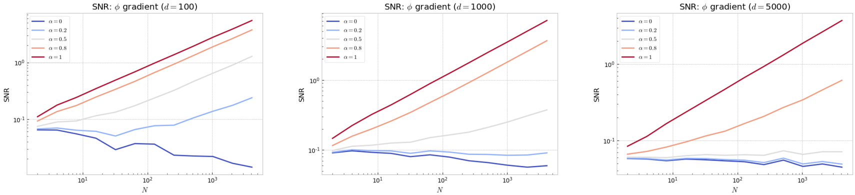
SNR analysis for VAE with MNIST : ϕ

Reparameterized

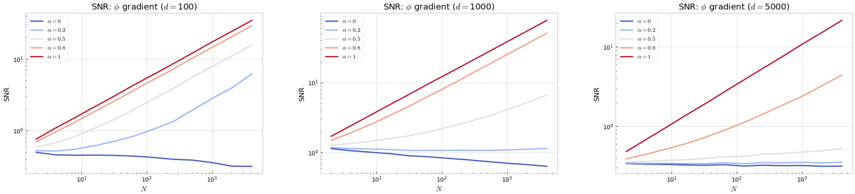


SNR analysis for VAE with MNIST : ϕ

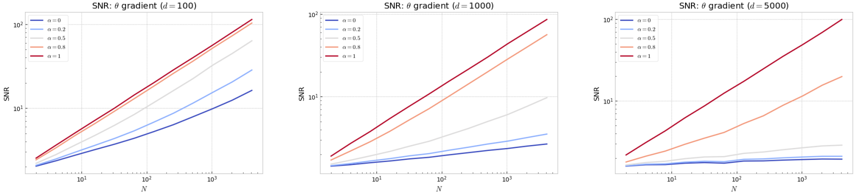
Reparameterized



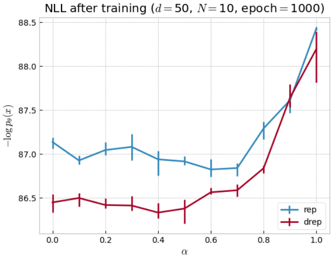
Doubly-reparameterized

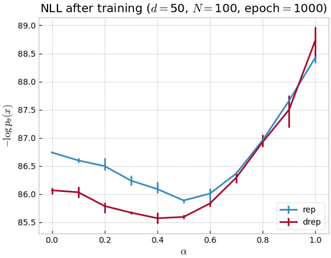


SNR analysis for VAE with MNIST : θ



Final plots


NLL after training ($d = 50$, $N = 10$, epoch = 1000)
$-\log p_\theta(x)$
88.5
88.0
87.5
87.0
86.5
0.0
0.2
0.4
0.6
0.8
1.0
α
rep
drep


NLL after training ($d = 50$, $N = 100$, epoch = 1000)
$-\log p_\theta(x)$
89.0
88.5
88.0
87.5
87.0
86.5
86.0
85.5
0.0
0.2
0.4
0.6
0.8
1.0
α
rep
drep

Outline

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature

2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound

3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)

4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature

2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound

3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)

4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
    - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
    - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
    - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
    - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Conclusion

Daudel, Benton, Shi and Doucet (2022). **Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics.**

1. We formalized and motivated the VR-IWAE bound
   - Theoretically-sound extension of the IWAE bound ($\alpha = 0$)
   - Provides theoretical guarantees behind various VR bound-based schemes proposed in the literature
2. We provided two complementary analyses of the VR-IWAE bound
   - Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
   - Encompass the case of the IWAE bound
3. We looked into the gradient(s) of the VR-IWAE bound and found desirable properties (SNR, doubly-reparameterized)
4. Empirical verification of our theoretical results

Perspectives

Some open questions:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we use the fact that the VR-IWAE bound extends the IWAE bound? (e.g. to build better gradient estimators / to enrich the variational family $\mathcal{Q}$)

Thank you for your attention !

Perspectives

Some open questions:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we use the fact that the VR-IWAE bound extends the IWAE bound? (e.g. to build better gradient estimators / to enrich the variational family $\mathcal{Q}$)

Thank you for your attention !

Perspectives

Some open questions:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we use the fact that the VR-IWAE bound extends the IWAE bound? (e.g. to build better gradient estimators / to enrich the variational family $\mathcal{Q}$)

Thank you for your attention !

Perspectives

Some open questions:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we use the fact that the VR-IWAE bound extends the IWAE bound? (e.g. to build better gradient estimators / to enrich the variational family $\mathcal{Q}$)

Thank you for your attention !

Perspectives

Some open questions:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we use the fact that the VR-IWAE bound extends the IWAE bound? (e.g. to build better gradient estimators / to enrich the variational family $\mathcal{Q}$)

Thank you for your attention !

References

- Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In 4th International Conference on Learning Representations (ICLR), 2016.
- Kamélia Daudel and Randal Douc. Mixture weights optimisation for alpha-divergence variational inference. In Advances in Neural Information Processing Systems, 2021.
- Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. The Annals of Statistics, 49(4):2250 - 2270, 2021a. doi: 10.1214/20-AOS2035.
- Kamélia Daudel, Randal Douc, and François Roueff. Monotonic alpha-divergence minimisation for variational inference. Arxiv preprint 2022.
- Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In Advances in Neural Information Processing Systems, 2018.
- Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. 3rd Symposium on Advances in Approximate Bayesian Inference, 2020.
- Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In Proceedings of the 38th International Conference on Machine Learning, 2021.

References (cont'd)

- Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In International Conference on Machine Learning, 2016.
- Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In Advances in Neural Information Processing Systems, 2016.
- Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In Advances in Neural Information Processing Systems, 2017.
- Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, 2018.
- George Tucker, Dieterich Lawson, Shixiang Shane Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. In Proceedings of the 7th International Conference on Learning Representations, 2019.