Challenges and Opportunities in Scalable Alpha-divergence Variational Inference: Application to IWAEs

Kamélia Daudel

CRiSM seminar – 19/10/2022

Joint work with Joe Benton, Arnaud Doucet and Yuyang Shi

Outline

Outline

Introduction

- Setting :
  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \text{argmax}_\theta \ \ell(\theta; x)$$

  (more generally $\theta^\star = \text{argmax}_\theta \ \sum_{i=1}^T \ell(\theta; x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :
  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathrm{argmax}_\theta\ \ell(\theta;x)$$

(more generally $\theta^\star = \mathrm{argmax}_\theta\ \sum_{i=1}^T \ell(\theta;x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :
  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathrm{argmax}_\theta \ \ell(\theta;x)$$

(more generally $\theta^\star = \mathrm{argmax}_\theta \ \sum_{i=1}^T \ell(\theta;x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :
  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathrm{argmax}_\theta\ \ell(\theta;x)$$

(more generally $\theta^\star = \mathrm{argmax}_\theta\ \sum_{i=1}^T \ell(\theta;x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :

  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathsf{argmax}_\theta \; \ell(\theta; x)$$

(more generally $\theta^\star = \mathsf{argmax}_\theta \; \sum_{i=1}^T \ell(\theta; x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :

  1. 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 $\mathbb{R}^d$
  2. In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathsf{argmax}_\theta \; \ell(\theta;x)$$

(more generally $\theta^\star = \mathsf{argmax}_\theta \; \sum_{i=1}^T \ell(\theta;x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Introduction

- Setting :

  ❶ 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 $\mathbb{R}^d$

  ❷ In that case, the **marginal log-likelihood** of x is given by

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

- Goal : find θ which best describes the observation x

$$\theta^\star = \mathsf{argmax}_\theta\ \ell(\theta; x)$$

(more generally $\theta^\star = \mathsf{argmax}_\theta\ \sum_{i=1}^T \ell(\theta; x_i)$)

- Problem : finding the optimal θ via maximum likelihood estimation is in general an intractable optimization problem

Variational Inference (VI)

- Idea : construct variational bounds, i.e. surrogate objective functions to the marginal log-likelihood that are more amenable to optimization.

- Common examples :

→ **Evidence Lower BOund (ELBO)** : rely on a variational probability density $q_\phi(z|x)$ parameterized by ϕ

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

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

→ **Importance Weighted Auto-Encoder (IWAE) bound** (Burda et al., 2016)

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

with $\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \leq \ell(\theta; x)$ and the unbiased Monte Carlo estimate

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

Variational Inference (VI)

- Idea : construct variational bounds, i.e. surrogate objective functions to the marginal log-likelihood that are more amenable to optimization.
- Common examples :

→ **Evidence Lower BOund (ELBO)** : rely on a variational probability density $q_\phi(z|x)$ parameterized by ϕ

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

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

→ **Importance Weighted Auto-Encoder (IWAE) bound** (Burda et al., 2016)

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

with $\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \leq \ell(\theta; x)$ and the unbiased Monte Carlo estimate

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

Variational Inference (VI)

- Idea : construct variational bounds, i.e. surrogate objective functions to the marginal log-likelihood that are more amenable to optimization.

- Common examples :

→ **Evidence Lower BOund (ELBO)** : rely on a variational probability density $q_\phi(z|x)$ parameterized by ϕ

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

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

→ **Importance Weighted Auto-Encoder (IWAE) bound** (Burda et al., 2016)

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

with $\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \leq \ell(\theta; x)$ and the unbiased Monte Carlo estimate

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

Variational Inference (VI)

- Idea : construct variational bounds, i.e. surrogate objective functions to the marginal log-likelihood that are more amenable to optimization.
- Common examples :

→ **Evidence Lower BOund (ELBO)** : rely on a variational probability density $q_\phi(z|x)$ parameterized by ϕ

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

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

→ **Importance Weighted Auto-Encoder (IWAE) bound** (Burda et al., 2016)

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

with $\ell_N^{(\mathrm{IWAE})}(\theta,\phi;x) \leq \ell(\theta;x)$ and the unbiased Monte Carlo estimate

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

Variational Inference (VI)

- Idea : construct variational bounds, i.e. surrogate objective functions to the marginal log-likelihood that are more amenable to optimization.

- Common examples :

→ **Evidence Lower BOund (ELBO)** : rely on a variational probability density $q_\phi(z|x)$ parameterized by ϕ

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

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

→ **Importance Weighted Auto-Encoder (IWAE) bound** (Burda et al., 2016)

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

with $\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \leq \ell(\theta; x)$ and the unbiased Monte Carlo estimate

$$\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \approx \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j;x) \right), \quad z_j \sim q_\phi(\cdot|x), \quad j = 1 \dots N$$

Training procedure for the IWAE bound

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

$$\frac{\partial}{\partial \phi}\left[\int q_\phi(z|x) h(z) \mathrm{d}z\right] = \frac{\partial}{\partial \phi}\left[\int q(\varepsilon) h(f(\varepsilon, \phi; x)) \mathrm{d}\varepsilon\right] = \int q(\varepsilon) \, \frac{\partial}{\partial \phi}\left[h(f(\varepsilon, \phi; x))\right] \mathrm{d}\varepsilon$$
$$\approx \frac{\partial}{\partial \phi}\left[h(f(\varepsilon, \phi; x))\right], \quad \varepsilon \sim q$$

- **Reparameterized** gradient estimator (Burda et al., 2016)

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

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

$\rightarrow$ Unbiased SGD steps w.r.t. (θ, ϕ)

Training procedure for the IWAE bound

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

$$\frac{\partial}{\partial \phi}\left[\int q_\phi(z|x)h(z)\mathrm{d}z\right] = \frac{\partial}{\partial \phi}\left[\int q(\varepsilon)h(f(\varepsilon,\phi;x))\mathrm{d}\varepsilon\right] = \int q(\varepsilon)\ \frac{\partial}{\partial \phi}\left[h(f(\varepsilon,\phi;x))\right]\mathrm{d}\varepsilon$$
$$\approx \frac{\partial}{\partial \phi}\left[h(f(\varepsilon,\phi;x))\right], \quad \varepsilon \sim q$$

- **Reparameterized** gradient estimator (Burda et al., 2016)

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

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

$\rightarrow$ Unbiased SGD steps w.r.t. (θ,ϕ)

Training procedure for the IWAE bound

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

$$\frac{\partial}{\partial \phi}\left[\int q_\phi(z|x)h(z)\mathrm{d}z\right] = \frac{\partial}{\partial \phi}\left[\int q(\varepsilon)h(f(\varepsilon,\phi;x))\mathrm{d}\varepsilon\right] = \int q(\varepsilon)\ \frac{\partial}{\partial \phi}\left[h(f(\varepsilon,\phi;x))\right]\mathrm{d}\varepsilon$$
$$\approx \frac{\partial}{\partial \phi}\left[h(f(\varepsilon,\phi;x))\right], \quad \varepsilon \sim q$$

- **Reparameterized** gradient estimator (Burda et al., 2016)

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

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

$\rightarrow$ Unbiased SGD steps w.r.t. (θ, ϕ)

Another interesting variational bound

The Variational Rényi (VR) bound (Li and Turner, 2016) : for all $\alpha \in \mathbb{R} \setminus \{1\}$,

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

→ Lower bound on $\ell(\theta; x)$ for $\alpha > 0$ (upper for $\alpha < 0$)

→ Flexible family of variational bounds indexed by α which recovers the ELBO when $\alpha \to 1$ (also has ties to the α-divergence)

Training procedure using the **reparameterized** gradient estimator

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \; \mathrm{d}\varepsilon}{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \mathrm{d}\varepsilon}$$

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

with positive empirical results

Another interesting variational bound

The Variational Rényi (VR) bound (Li and Turner, 2016) : for all $\alpha \in \mathbb{R} \setminus \{1\}$,

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

→ Lower bound on $\ell(\theta;x)$ for $\alpha > 0$ (upper for $\alpha < 0$)

→ Flexible family of variational bounds indexed by α which recovers the ELBO when $\alpha \to 1$ (also has ties to the α-divergence)

Training procedure using the **reparameterized** gradient estimator

$$\frac{\partial}{\partial \phi}\mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon}$$
$$\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k;x)^{1-\alpha}} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \quad \varepsilon_j \sim q, \quad j = 1 \dots N$$

with positive empirical results

Another interesting variational bound

The Variational Rényi (VR) bound (Li and Turner, 2016) : for all $\alpha \in \mathbb{R} \setminus \{1\}$,

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

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

→ Lower bound on $\ell(\theta; x)$ for $\alpha > 0$ (upper for $\alpha < 0$)

→ Flexible family of variational bounds indexed by α which recovers the ELBO when $\alpha \to 1$ (also has ties to the α-divergence)

Training procedure using the **reparameterized** gradient estimator

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \; \mathrm{d}\varepsilon}{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \mathrm{d}\varepsilon}$$

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

with positive empirical results

Another interesting variational bound

The Variational Rényi (VR) bound (Li and Turner, 2016) : for all $\alpha \in \mathbb{R} \setminus \{1\}$,

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

→ Lower bound on $\ell(\theta; x)$ for $\alpha > 0$ (upper for $\alpha < 0$)

→ Flexible family of variational bounds indexed by α which recovers the ELBO when $\alpha \to 1$ (also has ties to the α-divergence)

Training procedure using the **reparameterized** gradient estimator

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \; \mathrm{d}\varepsilon}{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \mathrm{d}\varepsilon}$$

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

with positive empirical results

Another interesting variational bound

The Variational Rényi (VR) bound (Li and Turner, 2016) : for all $\alpha \in \mathbb{R} \setminus \{1\}$,

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

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

→ Lower bound on $\ell(\theta; x)$ for $\alpha > 0$ (upper for $\alpha < 0$)

→ Flexible family of variational bounds indexed by α which recovers the ELBO when $\alpha \to 1$ (also has ties to the α-divergence)

Training procedure using the **reparameterized** gradient estimator

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon}$$

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

with positive empirical results

Recovers SGD with the IWAE bound for $\alpha = 0$ (resp. ELBO for $\alpha = 1$)

Another interesting variational bound (cont'd)

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

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

$$\frac{\partial}{\partial\phi}\mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon}$$

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

→ However

1. The VR bound can only be estimated using biased MC estimators
2. The VR bound does not recover the IWAE bound when $\alpha = 0$
3. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha \in \{0,1\}$)

Another interesting variational bound (cont'd)

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

$$\frac{\partial}{\partial\phi}\mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon}$$
$$\approx \sum_{j=1}^{N}\frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k;x)^{1-\alpha}}\frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \quad \varepsilon_j \sim q, \quad j = 1\dots N$$

→ However

1. The VR bound can only be estimated using biased MC estimators
2. The VR bound does not recover the IWAE bound when $\alpha = 0$
3. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha \in \{0,1\}$)

Another interesting variational bound (cont'd)

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

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

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \; \mathrm{d}\varepsilon}{\int q(\varepsilon) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \mathrm{d}\varepsilon}$$

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

→ However

1. The VR bound can only be estimated using biased MC estimators
2. The VR bound does not recover the IWAE bound when $\alpha = 0$
3. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha \in \{0, 1\}$)

Another interesting variational bound (cont'd)

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

$$\begin{aligned}\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) &= \frac{\int q(\varepsilon) \; w_{\theta,\phi}(z; x)^{1-\alpha} \; \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \; \mathrm{d}\varepsilon}{\int q(\varepsilon) \; w_{\theta,\phi}(z; x)^{1-\alpha} \; \mathrm{d}\varepsilon} \\ &\approx \sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)^{1-\alpha}} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \quad \varepsilon_j \sim q, \quad j = 1 \ldots N\end{aligned}$$

→ However

1. The VR bound can only be estimated using biased MC estimators
2. The VR bound does not recover the IWAE bound when $\alpha = 0$
3. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha \in \{0, 1\}$)

Another interesting variational bound (cont'd)

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

$$\frac{\partial}{\partial\phi}\mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon}$$
$$\approx \sum_{j=1}^{N}\frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k;x)^{1-\alpha}}\frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \quad \varepsilon_j \sim q, \quad j = 1\ldots N$$

→ However

1. The VR bound can only be estimated using biased MC estimators
2. The VR bound does not recover the IWAE bound when $\alpha = 0$
3. No theoretical justification as SGD with the VR bound resorts to biased estimators on top of the reparameterization trick (unless $\alpha \in \{0,1\}$)

An idea

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

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

- They investigated the expectation of the biased MC approximation, i.e.

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

e.g. they showed that it is non-decreasing with N when $\alpha \leq 1$.

- Question 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

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

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

- They investigated the expectation of the biased MC approximation, i.e.

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

e.g. they showed that it is non-decreasing with N when $\alpha \leq 1$.

- Question 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

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

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

- They investigated the expectation of the biased MC approximation, i.e.

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

  e.g. they showed that it is non-decreasing with N when $\alpha \leq 1$.
- Question 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

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

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

- They investigated the expectation of the biased MC approximation, i.e.

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

  e.g. they showed that it is non-decreasing with N when $\alpha \leq 1$.

- Question 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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

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. Recovers the IWAE objective function when $\alpha = 0$
3. Leads to the same SGD procedure as the VR bound in the reparameterized case, but this time using unbiased estimators

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

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

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

The VR-IWAE bound (cont'd)

→ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

→ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

→ Other notable advantages of the VR-IWAE bound :

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

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

→ Motivates the use of $\alpha \in [0, 1)$ in practice

The VR-IWAE bound (cont'd)

→ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

→ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

→ Other notable advantages of the VR-IWAE bound :

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

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

→ Motivates the use of $\alpha \in [0, 1)$ in practice

The VR-IWAE bound (cont'd)

→ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

→ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

→ Other notable advantages of the VR-IWAE bound :

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

→ Motivates the use of $\alpha \in [0, 1)$ in practice

The VR-IWAE bound (cont'd)

→ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

→ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

→ Other notable advantages of the VR-IWAE bound :

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

→ Motivates the use of $\alpha \in [0, 1)$ in practice

The VR-IWAE bound (cont'd)

→ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

→ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

→ Other notable advantages of the VR-IWAE bound :

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

→ Motivates the use of $\alpha \in [0, 1)$ in practice

The VR-IWAE bound (cont'd)

$\rightarrow$ The VR-IWAE bound is the **theoretically-sound** extension of the IWAE bound originating from the α-divergence VI methodology

$\rightarrow$ It provides **theoretical guarantees** behind various VR-bound gradient-based schemes previously proposed in the α-divergence VI community

$\rightarrow$ Other notable advantages of the VR-IWAE bound :

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

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

- The doubly-reparameterized gradient estimators of the IWAE generalize to the VR-IWAE bound

$\rightarrow$ Motivates the use of $\alpha \in [0, 1)$ in practice

Outline

Outline

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** 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) $\frac{\log N}{d^{1/3}} \to 0$

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** studies

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

   💡 This analysis will be tailored for **low to medium dimensions** settings

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

Overview

→ Quantity of interest : variational gap

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

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

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

→ Two **complentary** studies

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

   💡 This analysis will be 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 will be tailored for **high-dimensional** settings

Outline

N goes to infinity and d is fixed

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

Informally, Domke and Sheldon (2018, Theorem 3) 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.

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

→ 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

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

Informally, Domke and Sheldon (2018, Theorem 3) 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.

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

→ 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

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

Informally, Domke and Sheldon (2018, Theorem 3) 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.

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

→ 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

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

Informally, Domke and Sheldon (2018, Theorem 3) 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.

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

→ 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

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

Informally, Domke and Sheldon (2018, Theorem 3) 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.

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

$\rightarrow$ 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

Theorem

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

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

$\rightarrow$ 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 $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [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, motivates $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

$\rightarrow$ 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 $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [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, motivates $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

$\rightarrow$ 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 $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [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, motivates $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

$\rightarrow$ 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 $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [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, motivates $\alpha \in (0, 1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

$\rightarrow$ 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 $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [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, motivates $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

→ 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 $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ [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, motivates $\alpha \in (0, 1)$
- one of them controls γ_α^2

Main result

Theorem

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

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

→ 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 $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ [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, motivates $\alpha \in (0,1)$
- one of them controls γ_α^2

Main result (cont'd)

Theorem

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

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

→ To the best of our knowledge, first result shedding light on how α may play a role in the the success of α-divergence VI.

→ Question Can we find some limitations to this approach?

Main result (cont'd)

Theorem

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

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

$\rightarrow$ To the best of our knowledge, first result shedding light on how α may play a role in the the success of α-divergence VI.

$\rightarrow$ Question Can we find some limitations to this approach?

A key example

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

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

with

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

→ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

A key example

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

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

with

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

$\rightarrow$ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

A key example

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

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

with

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

→ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

A key example

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

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

with

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

$\rightarrow$ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

A key example

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

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

with

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

→ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

A key example

Log-normal distribution of the relative weights

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

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

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

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

with

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

→ Our theorem 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? $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$

Outline

$N, d \to \infty$ with either $\frac{\log N}{d} \to 0$ or $\frac{\log N}{d^{1/3}} \to 0$

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

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

$\to$ 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 $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ with either $\frac{\log N}{d} \to 0$ or $\frac{\log N}{d^{1/3}} \to 0$

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

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

$\to$ 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 $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ with either $\frac{\log N}{d} \to 0$ or $\frac{\log N}{d^{1/3}} \to 0$

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

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

$\to$ 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 $\frac{\log N}{d^{1/3}} \to 0$

$N, d \to \infty$ with either $\frac{\log N}{d} \to 0$ or $\frac{\log N}{d^{1/3}} \to 0$

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

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

$\to$ 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 $\frac{\log N}{d^{1/3}} \to 0$

Main result in the log-normal case

Theorem

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

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

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

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

→ Previously, we had

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

- 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

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

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

Main result in the log-normal case

Theorem

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

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

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

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

→ Previously, we had

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

- 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

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

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

Main result in the log-normal case

Theorem

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

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

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

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

→ Previously, we had

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

- 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

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

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

Main result in the log-normal case

Theorem

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

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

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

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

$\to$ Previously, we had

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

- 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

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

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

Main result in the log-normal case

Theorem

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

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

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

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

→ Previously, we had

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

- 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

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

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

Gaussian example

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, 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

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

with $\sigma = 1$.

- Asymptotic result 1

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

- Asymptotic result 2

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

$\rightarrow$ Weight collapse phenomenon might occur even for simple examples!

Gaussian example

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, 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

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

with $\sigma = 1$.

- Asymptotic result 1

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

- Asymptotic result 2

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

$\rightarrow$ Weight collapse phenomenon might occur even for simple examples!

Gaussian example

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, 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

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

with $\sigma = 1$.

- <u>Asymptotic result 1</u>

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

- <u>Asymptotic result 2</u>

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

$\rightarrow$ Weight collapse phenomenon might occur even for simple examples!

Gaussian example

Gaussian example

Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, 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

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

with $\sigma = 1$.

- Asymptotic result 1

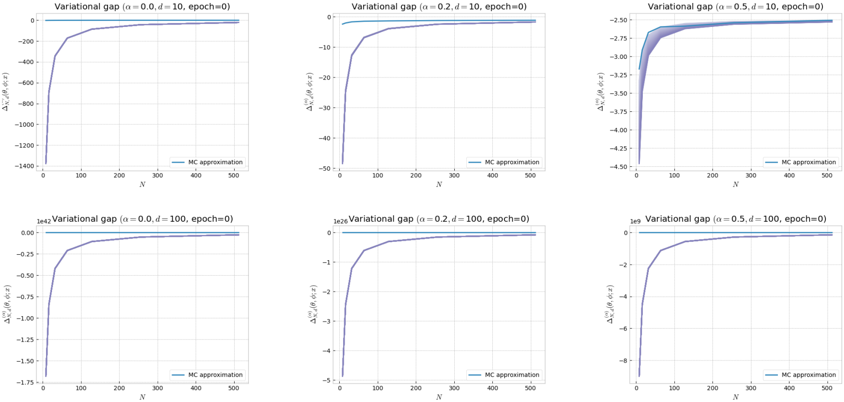
$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = -\alpha \cdot \frac{\sigma^2 d}{2} - \frac{\exp\left[(1-\alpha)^2 \sigma^2 d\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right)$$

- Asymptotic result 2

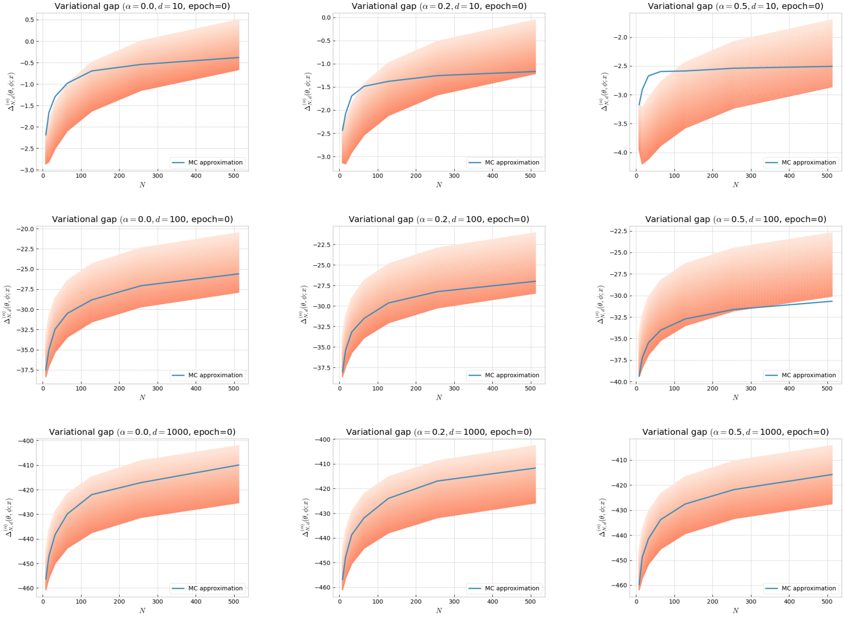
$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d \log N}}\right)\right) = 0.$$

$\rightarrow$ Weight collapse phenomenon might occur even for simple examples!

Empirical verification



Empirical verification (cont'd)



Main result in the approximate log-normal case

(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:

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

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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

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

→ NB : $-da = -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S_1))$

Main result in the approximate log-normal case

(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:

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

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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

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

$\rightarrow$ NB : $-da = -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S_1))$

Main result in the approximate log-normal case

(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:

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

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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

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

$\rightarrow$ NB : $-da = -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S_1))$

Main result in the approximate log-normal case

(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:

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

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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

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

$\rightarrow$ NB : $-da = -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S_1))$

Main result in the approximate log-normal case (cont'd)

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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 sub-exponentially with $d^{1/3}$.

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

Main result in the approximate log-normal case (cont'd)

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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 sub-exponentially with $d^{1/3}$.

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

Main result in the approximate log-normal case (cont'd)

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

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

Theorem

Assume (A1). Set $a := \log \mathbb{E}(\exp(-\xi_{1,1}))$ and further assume 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 $\alpha \in [0, 1)$,

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

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 sub-exponentially with $d^{1/3}$.

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

Linear Gaussian example (Rainforth et al., 2018)

Linear Gaussian example (Rainforth et al., 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 $\lambda = \frac{\left\| \frac{\theta + x}{2} - Ax - b \right\|}{\sqrt{d}}$

→ Set $(\theta, \phi) = (\theta^\star, \phi^\star)$ $[\theta^\star = T^{-1} \sum_{t=1}^T x_t,\ \phi^\star = (a^\star, b^\star)$ with $a^\star = \frac{1}{2}\boldsymbol{u}_d,\ b^\star = \frac{\theta^\star}{2}]$

- Asymptotic result 1

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

- Asymptotic result 2

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

→ The choice of the variational approximation q_ϕ matters a lot!

Linear Gaussian example (Rainforth et al., 2018)

Linear Gaussian example (Rainforth et al., 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 $\lambda = \frac{\left\| \frac{\theta+x}{2} - Ax - b \right\|}{\sqrt{d}}$

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

- Asymptotic result 1

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

- Asymptotic result 2

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

$\rightarrow$ The choice of the variational approximation q_ϕ matters a lot!

Linear Gaussian example (Rainforth et al., 2018)

Linear Gaussian example (Rainforth et al., 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 $\lambda = \frac{\left\| \frac{\theta+x}{2} - Ax - b \right\|}{\sqrt{d}}$

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

- Asymptotic result 1

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

- Asymptotic result 2

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

$\rightarrow$ The choice of the variational approximation q_ϕ matters a lot!

Linear Gaussian example (Rainforth et al., 2018)

Linear Gaussian example (Rainforth et al., 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 $\lambda = \frac{\left\|\frac{\theta+x}{2} - Ax - b\right\|}{\sqrt{d}}$

→ Set $(\theta, \phi) = (\theta^\star, \phi^\star)$ $[\theta^\star = T^{-1}\sum_{t=1}^T x_t$, $\phi^\star = (a^\star, b^\star)$ with $a^\star = \frac{1}{2}\boldsymbol{u}_d,\ b^\star = \frac{\theta^\star}{2}]$

- Asymptotic result 1

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

- Asymptotic result 2

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

→ The choice of the variational approximation q_ϕ matters a lot!

Linear Gaussian example (Rainforth et al., 2018)

Linear Gaussian example (Rainforth et al., 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 $\lambda = \frac{\left\| \frac{\theta + x}{2} - Ax - b \right\|}{\sqrt{d}}$

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

- Asymptotic result 1

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

- Asymptotic result 2

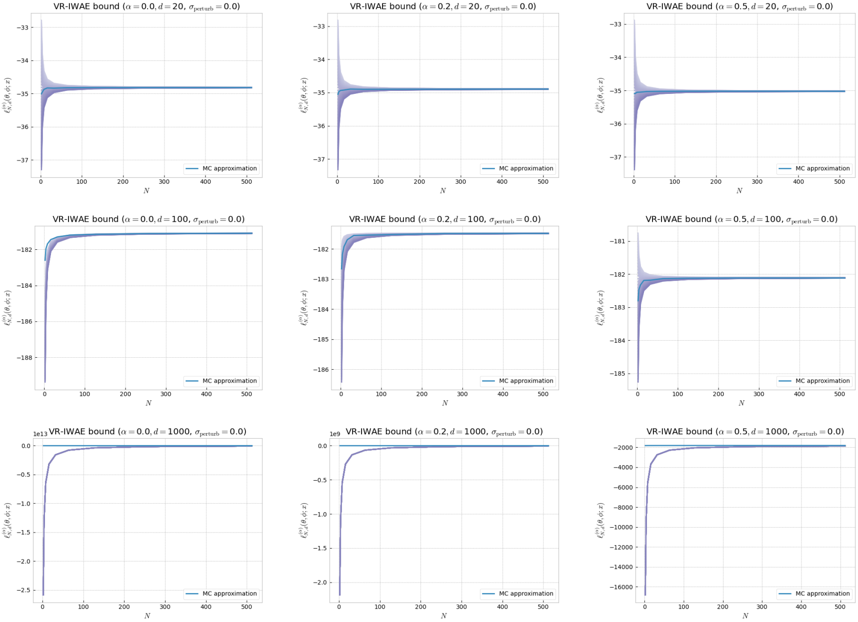
$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

$\to$ The choice of the variational approximation q_ϕ matters a lot!

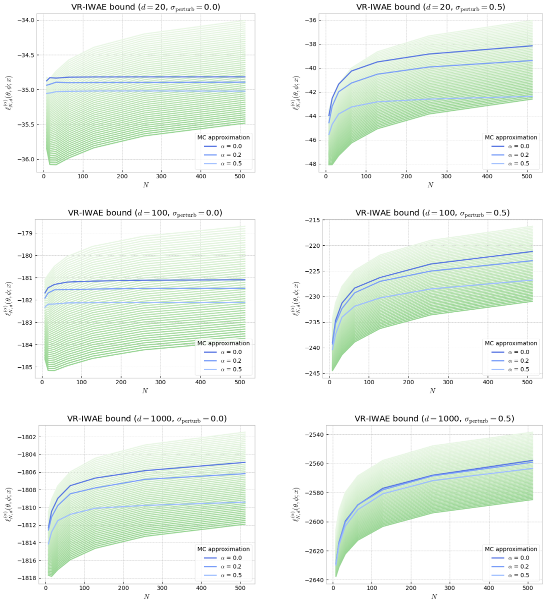
Outline

1. Introduction
2. The VR-IWAE bound
3. Theoretical study of the VR-IWAE bound
4. Numerical experiments
5. Conclusion

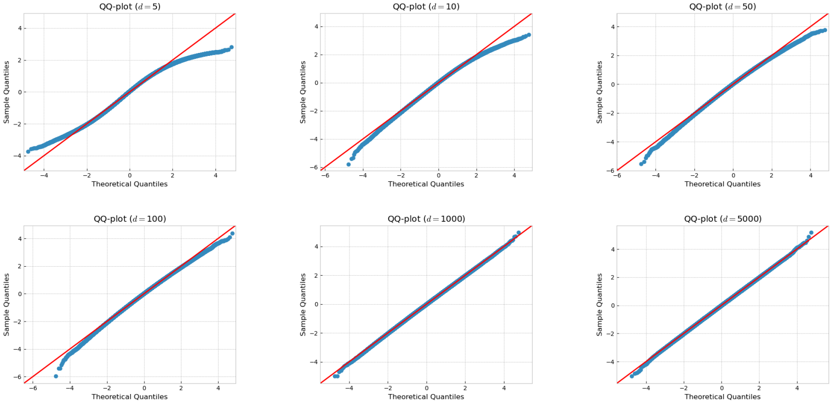
Linear Gaussian example



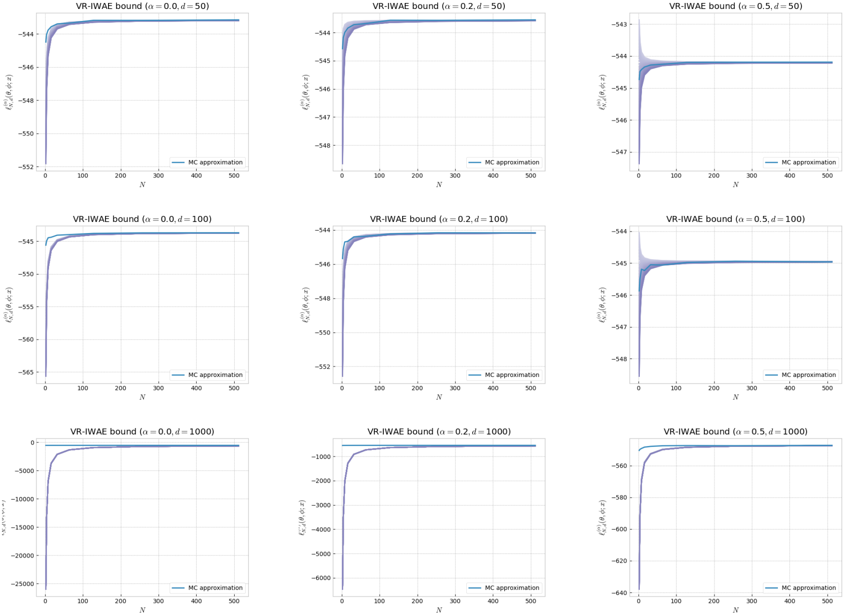
Linear Gaussian example (cont'd)



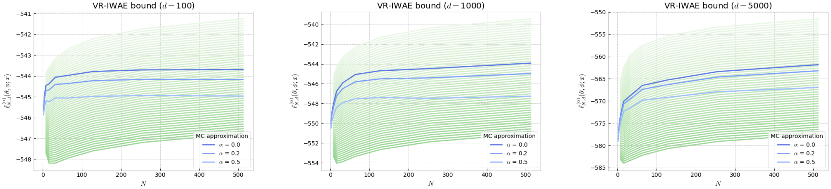
Variational auto-encoder on MNIST



Variational auto-encoder on MNIST (cont'd)



Variational auto-encoder on MNIST (cont'd - 2)



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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

→ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

→ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

Conclusion

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

❶ 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 α-Divergence VI community
- Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

❷ We provided two complementary analyses of the VR-IWAR bound

- Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
- Encompass the case of the IWAE bound

❸ Empirical verification of our theoretical results

→ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

Conclusion

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

❶ 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 α-Divergence VI community
- Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

❷ We provided two complementary analyses of the VR-IWAR bound

- Shed light on the conditions behind the success or failure of the VR-IWAE bound methodology
- Encompass the case of the IWAE bound

❸ Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

→ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?

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 α-Divergence VI community
   - Enjoys other additional desirable properties of this bound (SNR, doubly-reparameterized)

2. We provided two complementary analyses of the VR-IWAR 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. Empirical verification of our theoretical results

$\rightarrow$ Further work:

- Does the weight collapse behavior apply beyond the cases highlighted here?
- How does the weight collapse affect the gradient descent procedures?
- Can we further use the fact that the VR-IWAE bound extends the IWAE bound?