Variational Inference
Foundations and recent advances
(Part 1)

Kamélia Daudel

University of Bristol – 09/03/2022

Outline

Outline

Bayesian inference

- Goal : model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

- Core quantity in Bayesian Inference : posterior density of the latent variables y given the data $\mathscr{D}$

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} = \frac{p(\mathscr{D}|y)p_0(y)}{p(\mathscr{D})}$$

  $p(\mathscr{D})$: normalisation constant 'marginal likelihood'

- What we would like : compute / sample from the posterior density (posterior mean, posterior predictive distribution...)

- Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

Bayesian inference

- Goal : model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

- Core quantity in Bayesian Inference : posterior density of the latent variables y given the data $\mathscr{D}$

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} = \frac{p(\mathscr{D}|y)p_0(y)}{p(\mathscr{D})}$$

$p(\mathscr{D})$: normalisation constant 'marginal likelihood'

- What we would like : compute / sample from the posterior density (posterior mean, posterior predictive distribution...)

- Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

Bayesian inference

- Goal : model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

- Core quantity in Bayesian Inference : posterior density of the latent variables y given the data $\mathscr{D}$

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} = \frac{p(\mathscr{D}|y)p_0(y)}{p(\mathscr{D})}$$

  $p(\mathscr{D})$: normalisation constant ‘marginal likelihood’

- What we would like : compute / sample from the posterior density (posterior mean, posterior predictive distribution...)

- Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

Bayesian inference

- Goal : model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

- Core quantity in Bayesian Inference : posterior density of the latent variables y given the data $\mathscr{D}$

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} = \frac{p(\mathscr{D}|y)p_0(y)}{p(\mathscr{D})}$$

  $p(\mathscr{D})$: normalisation constant 'marginal likelihood'

- What we would like : compute / sample from the posterior density (posterior mean, posterior predictive distribution...)

- Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

Bayesian inference

- Goal : model a phenomenon given some observed data while taking into account prior knowledge on the model parameters.

- Core quantity in Bayesian Inference : posterior density of the latent variables y given the data $\mathscr{D}$

$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} = \frac{p(\mathscr{D}|y)p_0(y)}{p(\mathscr{D})}$$

  $p(\mathscr{D})$: normalisation constant 'marginal likelihood'

- What we would like : compute / sample from the posterior density (posterior mean, posterior predictive distribution...)

- Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods → sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** → optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods → sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** → optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods → sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** → optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Approximate Bayesian Inference

Two broad categories of methods :

1. Monte Carlo methods $\rightarrow$ sampling methods
   - Importance Sampling (IS)
   - Markov Chain Monte Carlo (MCMC)
   - Sequential Monte Carlo (SMC) ...
2. **Variational Inference methods** $\rightarrow$ optimisation-based methods
   - Mean-field Variational Inference (MFVI)
   - Black-Box Variational Inference (BBVI)
   - Variational Auto-Encoder (VAE) ...

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

   Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

❶ Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.

❷ Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

→ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently

Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

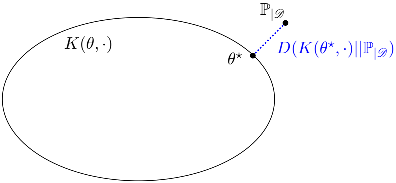
$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently



Variational Inference in a nutshell

Variational Inference methodology

1. Posit a **variational family** $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density :

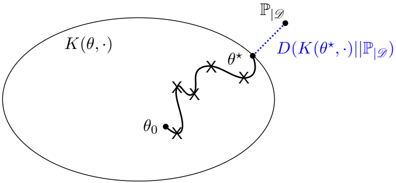
$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \tag{1}$$

Here, D is a **measure of dissimilarity** between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

$\rightarrow$ D and $\mathcal{Q}$ are key elements in the optimisation problem (1) !

What we want :

- $\mathcal{Q}$ is easy to sample from / optimise over, yet can capture the complexity inside $p(y|\mathscr{D})$ (e.g. well-chosen parametric family $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$)
- $D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ can be optimised efficiently



Outline

Mean-field Variational Inference (MFVI)

$\rightarrow D$: Kullback-Leibler (KL) divergence

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$.

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y) \ .$$

$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \geqslant 0$ and $D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = 0$ iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

$\rightarrow \mathcal{Q}$: Mean-field family

The latent variable y is made of L **independent** latent variables $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ and

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

i.e each latent variable y_ℓ is governed by its own variational probability density q_ℓ with $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$.

Mean-field Variational Inference (MFVI)

$\rightarrow D$: Kullback-Leibler (KL) divergence

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$.

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y) \ .$$

$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \geqslant 0$ and $D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = 0$ iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

$\rightarrow \mathcal{Q}$: Mean-field family

The latent variable y is made of L **independent** latent variables $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ and

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

i.e each latent variable y_ℓ is governed by its own variational probability density q_ℓ with $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$.

Mean-field Variational Inference (MFVI)

$\rightarrow D$: Kullback-Leibler (KL) divergence

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$.

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y) \ .$$

$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \geqslant 0$ and $D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = 0$ iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

$\rightarrow \mathcal{Q}$: Mean-field family

The latent variable y is made of L **independent** latent variables $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ and

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

i.e each latent variable y_ℓ is governed by its own variational probability density q_ℓ with $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$.

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y)\log\left(\frac{q(y)}{p(y,\mathscr{D})}\right)\nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y)\log\left(\frac{p(y,\mathscr{D})}{q(y)}\right)\nu(\mathrm{d}y)\ .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}\ \mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)\ .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}\ \mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)\,.$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}\ \mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)\ .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}\ \mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)\ .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}\ \mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \ .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}$ $\mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of D?

$$D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y|\mathscr{D})}\right) q(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} q(y) \log\left(\frac{q(y)}{p(y,\mathscr{D})}\right) \nu(\mathrm{d}y) + \log p(\mathscr{D})$$

$$:= -\mathsf{ELBO}(q;\mathscr{D}) + \log p(\mathscr{D})$$

Evidence Lower BOund (ELBO)

$$\mathsf{ELBO}(q;\mathscr{D}) := \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \, .$$

$\rightarrow$ We deduce :

1. $\inf_{q\in\mathcal{Q}} D_{KL}(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \sup_{q\in\mathcal{Q}}$ $\mathsf{ELBO}(q;\mathscr{D})$
2. $\mathsf{ELBO}(q;\mathscr{D}) \leqslant \log p(\mathscr{D})$ with equality iif $\mathbb{Q} = \mathbb{P}_{|\mathscr{D}}$

Why this choice of $\mathcal{Q}$?

Recall that

Mean-field assumption

The latent variable y is made of L **independent** latent variables $(y_1, \dots, y_L) \in \mathsf{Y}_1 \times \dots \times \mathsf{Y}_L$ and

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

i.e each latent variable y_ℓ is governed by its own variational probability density q_ℓ with $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$.

$\to$ Plugging this into the ELBO and keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \quad \text{(optimal rule)}$$

where $\mathbb{E}_{-\ell}$ is the expectation w.r.t q omitting the factor q_ℓ

Why this choice of $\mathcal{Q}$?

Recall that

Mean-field assumption

The latent variable y is made of L **independent** latent variables $(y_1, \ldots, y_L) \in \mathsf{Y}_1 \times \ldots \times \mathsf{Y}_L$ and

$$\mathcal{Q} = \left\{ q : y \mapsto \prod_{\ell=1}^{L} q_\ell(y_\ell) \right\}$$

i.e each latent variable y_ℓ is governed by its own variational probability density q_ℓ with $\nu(\mathrm{d}y) = \bigotimes_{\ell=1}^{L} \nu_\ell(\mathrm{d}y_\ell)$.

$\rightarrow$ Plugging this into the ELBO and keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right) \quad \text{(optimal rule)}$$

where $\mathbb{E}_{-\ell}$ is the expectation w.r.t q omitting the factor q_ℓ

CAVI algorithm

Optimal rule keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$$

CAVI algorithm

Optimal rule keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$$

Algorithm 1: *Coordinate Ascent Variational Inference (CAVI)*

Input: $(q_\ell)_{1 \leqslant \ell \leqslant L}$: initial variational factors.
Output: Return the optimised mean-field variational density q satisfying:
for all $y \in \mathsf{Y}$, $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.

while *the ELBO has not converged* **do**
  for $\ell = 1 \ldots L$ **do**
    set $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$
  end
  Compute the ELBO.
end

CAVI algorithm

Optimal rule keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$$

Algorithm 1: *Coordinate Ascent Variational Inference (CAVI)*

Input: $(q_\ell)_{1 \leqslant \ell \leqslant L}$: initial variational factors.
Output: Return the optimised mean-field variational density q satisfying:
for all $y \in \mathsf{Y}$, $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.

while *the ELBO has not converged* **do**
  for $\ell = 1 \ldots L$ **do**
    set $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$
  end
  Compute the ELBO.
end

$\rightarrow$ Convergence towards a **local** maximum of the ELBO

CAVI algorithm

Optimal rule keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$$

Algorithm 1: *Coordinate Ascent Variational Inference (CAVI)*

Input: $(q_\ell)_{1 \leqslant \ell \leqslant L}$: initial variational factors.
Output: Return the optimised mean-field variational density q satisfying: for all $y \in \mathsf{Y}$, $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.

while *the ELBO has not converged* **do**
  for $\ell = 1 \ldots L$ **do**
    set $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$
  end
  Compute the ELBO.
end

→ Convergence towards a **local** maximum of the ELBO

→ Tractable updates for **conditionally conjugate exponential** models (e.g. Bayesian mixture of Gaussians, Latent Dirichlet Allocation)

CAVI algorithm

Optimal rule keeping all factors but ℓ fixed :

$$q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$$

Algorithm 1: *Coordinate Ascent Variational Inference (CAVI)*

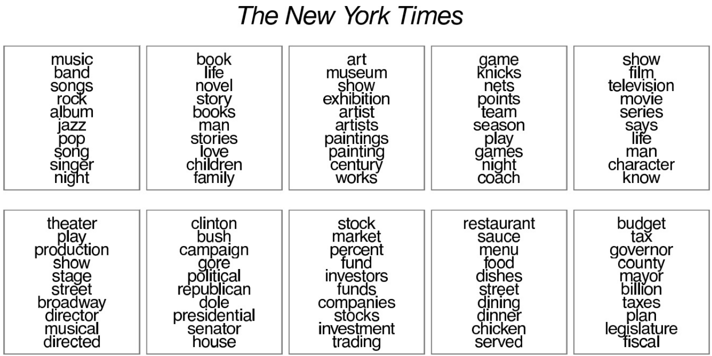
Input: $(q_\ell)_{1 \leqslant \ell \leqslant L}$: initial variational factors.
Output: Return the optimised mean-field variational density q satisfying:
for all $y \in \mathsf{Y}$, $q(y) = \prod_{\ell=1}^{L} q_\ell(y_\ell)$.

```
while the ELBO has not converged do
    for ℓ = 1 ... L do
        set q_ℓ(y_ℓ) ∝ exp(E_{-ℓ}[log p(y, 𝒟)])
    end
    Compute the ELBO.
end
```

→ Convergence towards a **local** maximum of the ELBO

→ Tractable updates for **conditionally conjugate exponential** models (e.g. Bayesian mixture of Gaussians, Latent Dirichlet Allocation)

Variational Inference: A Review for Statisticians. D. Blei et al. (2017). JASA



Data : 1.8M articles from the New York Times

Model : hierarchical Dirichlet process topic model

Taken from **Stochastic Variational Inference.** M. D. Hoffman et al. (2013). JMRL.

Toy example : Bayesian Linear Regression (BLR) - 1

- $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I 1-D class labels $(c_i)_{1 \leqslant i \leqslant I}$, I 2-D covariates $(\boldsymbol{x}_i)_{1 \leqslant i \leqslant I}$
- $y = \{y_1, y_2\} \in \mathbb{R}^2$: regression coefficients
- Model :

$$
\begin{aligned}
p(c_i|\boldsymbol{x}_i, y) &= \mathcal{N}(c_i; y^T \boldsymbol{x}_i, \sigma^2) \; , \quad 1 \leqslant i \leqslant I \\
p_0(y) &= \mathcal{N}(y; \mu_0, \Lambda_0^{-1})
\end{aligned}
$$

μ_0, Λ_0, σ : fixed hyperparameters

In that case,

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

with $\Lambda = \Lambda_0 + \sigma^{-2} \sum_{i=1}^{I} \boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\Lambda\mu = \Lambda_0 \mu_0 + \sigma^{-2} \sum_{i=1}^{I} c_i \boldsymbol{x}_i$.

Toy example : Bayesian Linear Regression (BLR) - 1

- $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I 1-D class labels $(c_i)_{1 \leqslant i \leqslant I}$, I 2-D covariates $(\boldsymbol{x}_i)_{1 \leqslant i \leqslant I}$
- $y = \{y_1, y_2\} \in \mathbb{R}^2$: regression coefficients
- Model :

$$p(c_i|\boldsymbol{x}_i, y) = \mathcal{N}(c_i; y^T \boldsymbol{x}_i, \sigma^2) \; , \quad 1 \leqslant i \leqslant I$$
$$p_0(y) = \mathcal{N}(y; \mu_0, \Lambda_0^{-1})$$

μ_0, Λ_0, σ : fixed hyperparameters

In that case,

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

with $\Lambda = \Lambda_0 + \sigma^{-2} \sum_{i=1}^{I} \boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\Lambda\mu = \Lambda_0\mu_0 + \sigma^{-2} \sum_{i=1}^{I} c_i \boldsymbol{x}_i$.

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell, k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$
  that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\to$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell, k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 2

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

- Mean-field assumption : $q(y) = q_1(y_1)q_2(y_2)$
- Optimal rules : for all $\ell = \{1, 2\}$, $q_\ell(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$ that is,

$$q_1(y_1) \propto \exp\left(\mathbb{E}_{y_2 \sim q_2}\left[\log p(y|\mathscr{D})\right]\right)$$
$$q_2(y_2) \propto \exp\left(\mathbb{E}_{y_1 \sim q_1}\left[\log p(y|\mathscr{D})\right]\right)$$

Notation : $\mu = (\mu_1 \ \mu_2)$, $\Lambda = (\Lambda_{\ell,k})_{1 \leqslant \ell,k \leqslant 2}$ with $\Lambda_{1,2} = \Lambda_{2,1}$

$$\log p(y|\mathscr{D}) = -\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(y_2 - \mu_2)\Lambda_{1,2} + (y_2 - \mu_2)^2\Lambda_{2,2}\right\} + c_{-y}$$

$\rightarrow$ Plugging this in the optimal rule,

$$q_1(y_1) \propto \exp\left(-\frac{1}{2}\left\{(y_1 - \mu_1)^2\Lambda_{1,1} + 2(y_1 - \mu_1)(\mathbb{E}_{y_2 \sim q_2}\left[y_2\right] - \mu_2)\Lambda_{1,2}\right\}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left\{y_1^2\Lambda_{1,1} - 2y_1\left[\mu_1\Lambda_{1,1} - (\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}\right]\right\}\right)$$

so that : $q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$

Toy example : Bayesian Linear Regression (BLR) - 3

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

Optimal updates :

$$q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$$
$$q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{y_1 \sim q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$$

Setting $m_1 = \mathbb{E}_{y_1 \sim q_1}[y_1]$ and $m_2 = \mathbb{E}_{y_2 \sim q_2}[y_2]$, the CAVI algorithm alternates between :

$$m_1 \leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2}$$
$$m_2 \leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2}$$

One stable fixed point : $(m_1, m_2) = (\mu_1, \mu_2)$

$\mu = (0\ 0)$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.

Toy example : Bayesian Linear Regression (BLR) - 3

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

Optimal updates :

$$q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$$
$$q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{y_1 \sim q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$$

Setting $m_1 = \mathbb{E}_{y_1 \sim q_1}[y_1]$ and $m_2 = \mathbb{E}_{y_2 \sim q_2}[y_2]$, the CAVI algorithm alternates between :

$$m_1 \leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2}$$
$$m_2 \leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2}$$

One stable fixed point : $(m_1, m_2) = (\mu_1, \mu_2)$

$\mu = (0\ 0)$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.

Toy example : Bayesian Linear Regression (BLR) - 3

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

Optimal updates :

$$q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$$
$$q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{y_1 \sim q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$$

Setting $m_1 = \mathbb{E}_{y_1 \sim q_1}[y_1]$ and $m_2 = \mathbb{E}_{y_2 \sim q_2}[y_2]$, the CAVI algorithm alternates between :

$$m_1 \leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2}$$
$$m_2 \leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2}$$

One stable fixed point : $(m_1, m_2) = (\mu_1, \mu_2)$

$\mu = (0\ 0)$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.

Toy example : Bayesian Linear Regression (BLR) - 3

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

Optimal updates :

$$q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$$
$$q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{y_1 \sim q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$$

Setting $m_1 = \mathbb{E}_{y_1 \sim q_1}[y_1]$ and $m_2 = \mathbb{E}_{y_2 \sim q_2}[y_2]$, the CAVI algorithm alternates between :

$$m_1 \leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2}$$
$$m_2 \leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2}$$

One stable fixed point : $(m_1, m_2) = (\mu_1, \mu_2)$

$\mu = (0\ 0)$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.

Toy example : Bayesian Linear Regression (BLR) - 3

$$p(y|\mathscr{D}) = \mathcal{N}(y; \mu, \Lambda^{-1})$$

Optimal updates :

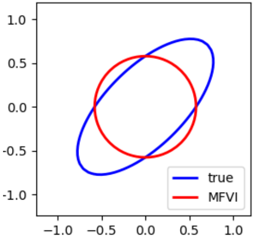
$$q_1(y_1) = \mathcal{N}(y_1; \mu_1 - \Lambda_{1,1}^{-1}(\mathbb{E}_{y_2 \sim q_2}[y_2] - \mu_2)\Lambda_{1,2}, \Lambda_{1,1}^{-1})$$
$$q_2(y_2) = \mathcal{N}(y_2; \mu_2 - \Lambda_{2,2}^{-1}(\mathbb{E}_{y_1 \sim q_1}[y_1] - \mu_1)\Lambda_{1,2}, \Lambda_{2,2}^{-1})$$

Setting $m_1 = \mathbb{E}_{y_1 \sim q_1}[y_1]$ and $m_2 = \mathbb{E}_{y_2 \sim q_2}[y_2]$, the CAVI algorithm alternates between :

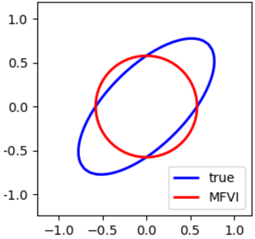
$$m_1 \leftarrow \mu_1 - \Lambda_{1,1}^{-1}(m_2 - \mu_2)\Lambda_{1,2}$$
$$m_2 \leftarrow \mu_2 - \Lambda_{2,2}^{-1}(m_1 - \mu_1)\Lambda_{1,2}$$

One stable fixed point : $(m_1, m_2) = (\mu_1, \mu_2)$

$\mu = (0\ 0)$, $\Lambda_{1,1} = \Lambda_{2,2} = 3$ and $\Lambda_{1,2} = -2$.



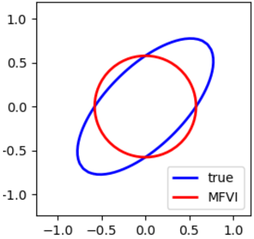
Limitations of MFVI



1. The approximative family $\mathcal{Q}$ can be too restrictive / the updates are model-specific.

2. The ELBO tends to underestimate the posterior variance.

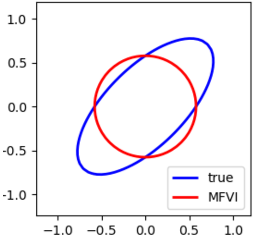
Limitations of MFVI



❶ The approximative family $\mathcal{Q}$ can be too restrictive / the updates are model-specific.

❷ The ELBO tends to underestimate the posterior variance.

Limitations of MFVI



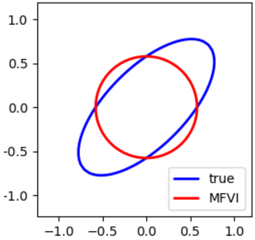
❶ The approximative family $\mathcal{Q}$ can be too restrictive / the updates are model-specific.

→ Black-box Variational Inference

Black Box Variational Inference. R. Ranganath et al. (2014). PMLR.

❷ The ELBO tends to underestimate the posterior variance.

Limitations of MFVI



❶ The approximative family $\mathcal{Q}$ can be too restrictive / the updates are model-specific.

$\rightarrow$ Black-box Variational Inference

Black Box Variational Inference. R. Ranganath et al. (2014). PMLR.

❷ The ELBO tends to underestimate the posterior variance.

$\rightarrow$ Alpha-divergence Variational Inference

Black-box alpha divergence minimization. J. Hernandez-Lobato et al. (2016). ICML

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

Outline

Black-box Variational Inference (BBVI) - 1

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n\geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n}$$

We have that :

$$\begin{aligned}
\nabla_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} k(\theta,y) \log\left(\frac{p(y,\mathscr{D})}{k(\theta,y)}\right) \nu(\mathrm{d}y) \right) \Big|_{\theta=\theta_n} \\
&= -\int_{\mathsf{Y}} \frac{\partial}{\partial\theta} \left(\frac{k(\theta,y)}{p(y,\mathscr{D})} \log\left(\frac{k(\theta,y)}{p(y,\mathscr{D})}\right)\right) \Big|_{(\theta,y)=(\theta_n,y)} p(y,\mathscr{D})\nu(\mathrm{d}y) \\
&= -\int_{\mathsf{Y}} \left(\log\left(\frac{k(\theta_n,y)}{p(y,\mathscr{D})}\right) + 1\right) \frac{\partial k(\theta,y)}{\partial\theta}\Big|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} k(\theta_n,y) \log\left(\frac{p(y,\mathscr{D})}{k(\theta_n,y)}\right) \frac{\partial \log k(\theta,y)}{\partial\theta}\Big|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \quad \text{(REINFORCE)} \\
&\qquad - \int_{\mathsf{Y}} \frac{\partial k(\theta,y)}{\partial\theta}\Big|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)
\end{aligned}$$

Black-box Variational Inference (BBVI) - 1

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

We have that :

$$\nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \nabla_\theta \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta, y)} \right) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n}$$

$$= - \int_{\mathsf{Y}} \frac{\partial}{\partial \theta} \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \log \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \right) \right) \Bigg|_{(\theta, y) = (\theta_n, y)} p(y, \mathscr{D}) \nu(\mathrm{d}y)$$

$$= - \int_{\mathsf{Y}} \left(\log \left(\frac{k(\theta_n, y)}{p(y, \mathscr{D})} \right) + 1 \right) \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} k(\theta_n, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y) \quad \text{(REINFORCE)}$$

$$- \int_{\mathsf{Y}} \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

Black-box Variational Inference (BBVI) - 1

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

We have that :

$$\nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \nabla_\theta \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta, y)} \right) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n}$$

$$= - \int_{\mathsf{Y}} \frac{\partial}{\partial \theta} \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \log \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \right) \right) \Bigg|_{(\theta, y)=(\theta_n, y)} p(y, \mathscr{D}) \nu(\mathrm{d}y)$$

$$= - \int_{\mathsf{Y}} \left(\log \left(\frac{k(\theta_n, y)}{p(y, \mathscr{D})} \right) + 1 \right) \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y)=(\theta_n, y)} \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} k(\theta_n, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y)=(\theta_n, y)} \nu(\mathrm{d}y) \quad \text{(REINFORCE)}$$

$$- \int_{\mathsf{Y}} \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y)=(\theta_n, y)} \nu(\mathrm{d}y)$$

Black-box Variational Inference (BBVI) - 1

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

We have that :

$$\nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \nabla_\theta \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta, y)} \right) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n}$$

$$= - \int_{\mathsf{Y}} \frac{\partial}{\partial \theta} \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \log \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \right) \right) \Bigg|_{(\theta, y) = (\theta_n, y)} p(y, \mathscr{D}) \nu(\mathrm{d}y)$$

$$= - \int_{\mathsf{Y}} \left(\log \left(\frac{k(\theta_n, y)}{p(y, \mathscr{D})} \right) + 1 \right) \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} k(\theta_n, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y) \quad \text{(REINFORCE)}$$

$$- \int_{\mathsf{Y}} \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

Black-box Variational Inference (BBVI) - 1

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

We have that :

$$\begin{aligned}
\nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} k(\theta, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta, y)} \right) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n} \\
&= - \int_{\mathsf{Y}} \frac{\partial}{\partial \theta} \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \log \left(\frac{k(\theta, y)}{p(y, \mathscr{D})} \right) \right) \Bigg|_{(\theta,y)=(\theta_n,y)} p(y, \mathscr{D}) \nu(\mathrm{d}y) \\
&= - \int_{\mathsf{Y}} \left(\log \left(\frac{k(\theta_n, y)}{p(y, \mathscr{D})} \right) + 1 \right) \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}} k(\theta_n, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \Bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \quad \text{(REINFORCE)} \\
&\qquad - \int_{\mathsf{Y}} \frac{\partial k(\theta, y)}{\partial \theta} \Bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)
\end{aligned}$$

Black-box Variational Inference (BBVI) - 2

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

with

$$\nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \int_{\mathsf{Y}} k(\theta_n, y) \log \left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)$$

In practice...

1. **Stochastic** Gradient Ascent using the **unbiased** estimate

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

   where $Y_1, \ldots, Y_M \ : \ M$ i.i.d. samples generated from $k(\theta_n, \cdot)$

2. Large-scale learning using **mini-batching**

$\rightarrow$ convergence towards a **local** optimum of the ELBO ($(r_n)_{n \geqslant 1}$ follows Robbins-Monro)

Black-box Variational Inference (BBVI) - 2

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

with

$$\nabla_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \int_{\mathsf{Y}} k(\theta_n, y) \log\left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta, y)=(\theta_n, y)} \nu(\mathrm{d}y)$$

In practice...

1. **Stochastic** Gradient Ascent using the **unbiased** estimate

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta, y)=(\theta_n, Y_m)}$$

   where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

2. Large-scale learning using **mini-batching**

$\rightarrow$ convergence towards a **local** optimum of the ELBO ($(r_n)_{n \geqslant 1}$ follows Robbins-Monro)

Black-box Variational Inference (BBVI) - 2

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n\geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n}$$

with

$$\nabla_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \int_{\mathsf{Y}} k(\theta_n, y) \log\left(\frac{p(y,\mathscr{D})}{k(\theta_n,y)}\right) \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)$$

In practice...

❶ **Stochastic** Gradient Ascent using the **unbiased** estimate

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n,Y_m)}\right) \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1,\ldots,Y_M$: M i.i.d. samples generated from $k(\theta_n,\cdot)$

❷ Large-scale learning using **mini-batching**

→ convergence towards a **local** optimum of the ELBO ($(r_n)_{n\geqslant 1}$ follows Robbins-Monro)

Black-box Variational Inference (BBVI) - 2

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

with

$$\nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \int_{\mathsf{Y}} k(\theta_n, y) \log\left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)}\right) \left.\frac{\partial \log k(\theta, y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)$$

In practice...

1. **Stochastic** Gradient Ascent using the **unbiased** estimate

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \left.\frac{\partial \log k(\theta, y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,Y_m)}$$

   where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

2. Large-scale learning using **mini-batching**

$\rightarrow$ convergence towards a **local** optimum of the ELBO ($(r_n)_{n \geqslant 1}$ follows Robbins-Monro)

Black-box Variational Inference (BBVI) - 2

Idea of BBVI : Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform Gradient Ascent on the ELBO with a learning policy $(r_n)_{n \geqslant 1}$

$$\theta_{n+1} = \theta_n + r_n \nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n}$$

with

$$\nabla_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \int_{\mathsf{Y}} k(\theta_n, y) \log\left(\frac{p(y, \mathscr{D})}{k(\theta_n, y)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)$$

In practice...

1. **Stochastic** Gradient Ascent using the **unbiased** estimate

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

   where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

2. Large-scale learning using **mini-batching**

$\rightarrow$ convergence towards a **local** optimum of the ELBO ($(r_n)_{n \geqslant 1}$ follows Robbins-Monro)

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta, y) = (\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta, y) = (\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\to$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n,Y_m)}\right) \frac{\partial \log k(\theta,y)}{\partial\theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1,\ldots,Y_M$: M i.i.d. samples generated from $k(\theta_n,\cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!
2. $Y_1 \sim k(\theta_n,\cdot)$ with $p(Y_1,\mathscr{D}) = 0$ makes the gradient blow up...
   $\to$ the ELBO enforces $\mathrm{supp}(k(\theta_n,\cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \;:\; \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n,Y_m)}\right) \frac{\partial \log k(\theta,y)}{\partial\theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1,\ldots,Y_M$: M i.i.d. samples generated from $k(\theta_n,\cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n,\cdot)$ with $p(Y_1,\mathscr{D}) = 0$ makes the gradient blow up...
   $\to$ the ELBO enforces $\mathrm{supp}(k(\theta_n,\cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)} \right) \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!
2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta,y)=(\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...

   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n,Y_m)}\right) \frac{\partial \log k(\theta,y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n,\cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log \left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)} \right) \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta,y)=(\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta,y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...

   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n,\cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathsf{ELBO}(k(\theta,\cdot);\mathscr{D})|_{\theta=\theta_n} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(Y_m,\mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n,Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n,\cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \text{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!
2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\text{supp}(k(\theta_n, \cdot)) \subseteq \text{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Remarks on BBVI

In short, BBVI resorts to Stochastic Gradient Ascent on the ELBO

- The updates are not model-specific ("Black-box")
- $\mathcal{Q}$: parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

- Uses the unbiased estimator

$$\hat{\nabla}_\theta \mathrm{ELBO}(k(\theta, \cdot); \mathscr{D})|_{\theta=\theta_n} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{p(Y_m, \mathscr{D})}{k(\theta_n, Y_m)}\right) \frac{\partial \log k(\theta, y)}{\partial \theta}\bigg|_{(\theta,y)=(\theta_n, Y_m)}$$

where $Y_1, \ldots, Y_M$: M i.i.d. samples generated from $k(\theta_n, \cdot)$

1. The variance of the gradient estimators is an issue :
   - Rao-blackwellisation
   - Reparameterisation (used in VAEs)
   - Control variates
   - Quasi-Monte Carlo methods
   - ... This is an active area of research!

2. $Y_1 \sim k(\theta_n, \cdot)$ with $p(Y_1, \mathscr{D}) = 0$ makes the gradient blow up...
   $\rightarrow$ the ELBO enforces $\mathrm{supp}(k(\theta_n, \cdot)) \subseteq \mathrm{supp}(p(\cdot|\mathscr{D}))$ "zero-forcing"

Outline

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.

$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \ ,$$

where

$$f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], \quad \alpha \in \mathbb{R} \setminus \{0, 1\}$$

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \ ,$$

where

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}\,, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.

$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

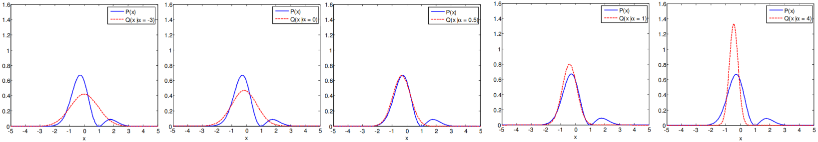
$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) ,$$

where

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

1. A flexible family of divergences...

Figure: In red, the Gaussian which minimises $D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$ for a varying α



$\alpha = -3$ $\alpha = 0$ $\alpha = 0.5$ $\alpha = 1$ $\alpha = 4$

Adapted from V. Cevher's lecture notes (2008) https://www.ece.rice.edu/~vc3/elec633/AlphaDivergence.pdf

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

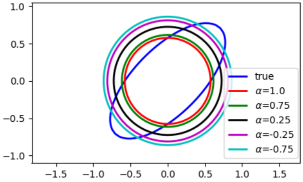
$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \;,$$

where

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}\,, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

1. A flexible family of divergences...

Figure: Optimal mean-field approximation for a varying α (BLR revisited)



Adapted from **Rényi divergence variational inference.** Y. Li and R. E Turner. (2016). NeurIPS

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D}) \nu(\mathrm{d}y) ,$$

where

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\} , \\ u \log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

1. A flexible family of divergences...
2. ...suitable for Variational Inference purposes...

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \inf_{q \in \mathcal{Q}} \Psi_\alpha(q; p)$$

with $\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ and $p = p(\cdot, \mathscr{D})$

3. ...with good convexity properties : f_α is convex!

The alpha-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D})$

Alpha-divergence between $\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$

$$D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) ,$$

where

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

1. A flexible family of divergences...
2. ...suitable for Variational Inference purposes...

$$\inf_{q \in \mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \Leftrightarrow \inf_{q \in \mathcal{Q}} \Psi_\alpha(q; p)$$

with $\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ and $p = p(\cdot, \mathscr{D})$

3. ...with good convexity properties : f_α is convex!

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta,\cdot);p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f'_\alpha(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}\nabla_\theta \Psi_\alpha(k(\theta,\cdot);p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha\left(\frac{k(\theta,y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)\right)\Big|_{\theta=\theta_n} \\ &= \dots \\ &= \int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\ &\approx \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_n,Y_m)^{\alpha-1}p(Y_m)^{1-\alpha}}{\alpha-1} \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,Y_m)}\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n,\cdot)$ as a sampler + Mini-batching + Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta, \cdot); p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f_\alpha'(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha \left(\frac{k(\theta, y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n} \\
&= \ldots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_n, Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n, \cdot)$ as a sampler + Mini-batching + Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta, \cdot); p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f'_\alpha(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha\left(\frac{k(\theta, y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \right) \Big|_{\theta=\theta_n} \\
&= \ldots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_n, Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n, \cdot)$ as a sampler
+ Mini-batching + Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta, \cdot); p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f'_\alpha(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha \left(\frac{k(\theta, y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n} \\
&= \ldots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_n, Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n, \cdot)$ as a sampler
$+$ Mini-batching $+$ Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta,\cdot);p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f'_\alpha(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta,\cdot);p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha\left(\frac{k(\theta,y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)\right)\Bigg|_{\theta=\theta_n} \\
&= \ldots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_n,Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha-1} \left.\frac{\partial \log k(\theta,y)}{\partial\theta}\right|_{(\theta,y)=(\theta_n,Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n,\cdot)$ as a sampler
+ Mini-batching + Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta, \cdot); p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f'_\alpha(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha \left(\frac{k(\theta, y)}{p(y)} \right) p(y) \nu(\mathrm{d}y) \right) \Bigg|_{\theta=\theta_n} \\
&= \dots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_n, Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n, \cdot)$ as a sampler
$+$ Mini-batching $+$ Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta,\cdot);p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f_\alpha'(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta,\cdot);p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha\left(\frac{k(\theta,y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)\right)\Bigg|_{\theta=\theta_n} \\
&= \dots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\alpha-1} \ \frac{\partial \log k(\theta,y)}{\partial\theta}\Bigg|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_n,Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha-1} \ \frac{\partial \log k(\theta,y)}{\partial\theta}\Bigg|_{(\theta,y)=(\theta_n,Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n,\cdot)$ as a sampler
$+$ Mini-batching $+$ Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A first approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Descent** on $\Psi_\alpha(k(\theta, \cdot); p)$

We have that : for all $\alpha \in \mathbb{R} \setminus \{1\}$, $f_\alpha'(u) = \frac{1}{\alpha-1}\left[u^{\alpha-1} - 1\right]$ and

$$\begin{aligned}
\nabla_\theta \Psi_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} &= \nabla_\theta \left(\int_{\mathsf{Y}} f_\alpha\left(\frac{k(\theta, y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \right)\Bigg|_{\theta=\theta_n} \\
&= \ldots \\
&= \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\alpha - 1} \left.\frac{\partial \log k(\theta, y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_n, Y_m)^{\alpha-1} p(Y_m)^{1-\alpha}}{\alpha - 1} \left.\frac{\partial \log k(\theta, y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,Y_m)}
\end{aligned}$$

In practice : **Stochastic** Gradient Descent using $k(\theta_n, \cdot)$ as a sampler
$+$ Mini-batching $+$ Reparameterisation.

Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta,\cdot);p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta,y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

$\rightarrow$ derived from Rényi's α-divergence, linked to the α-divergence.

$$\begin{aligned} \nabla_\theta \mathcal{L}_\alpha(k(\theta,\cdot);p)|_{\theta=\theta_n} &= \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n,y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y) \\ &\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,Y_m)} \end{aligned}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n,\cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

→ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

$\rightarrow$ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler
$+$ Mini-batching $+$ Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

$\rightarrow$ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y)=(\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y)=(\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

→ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler

+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

→ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta,\cdot);p) = \frac{1}{1-\alpha} \log\left(\int_{\mathsf{Y}} k(\theta,y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)\right)$$

→ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta,\cdot);p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n,y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n,y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left.\frac{\partial \log k(\theta,y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left.\frac{\partial \log k(\theta,y)}{\partial \theta}\right|_{(\theta,y)=(\theta_n,Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n,\cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

A second approach

Consider a **parametric** family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

and perform **Gradient Ascent** on the VR bound : for all $\alpha \in \mathbb{R} \setminus \{1\}$

$$\mathcal{L}_\alpha(k(\theta, \cdot); p) = \frac{1}{1-\alpha} \log \left(\int_{\mathsf{Y}} k(\theta, y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y) \right)$$

→ derived from Rényi's α-divergence, linked to the α-divergence.

$$\nabla_\theta \mathcal{L}_\alpha(k(\theta, \cdot); p)|_{\theta=\theta_n} = \frac{\alpha}{1-\alpha} \int_{\mathsf{Y}} \frac{k(\theta_n, y)^\alpha p(y)^{1-\alpha}}{\int_{\mathsf{Y}} k(\theta_n, y')^\alpha p(y')^{1-\alpha} \nu(\mathrm{d}y')} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, y)} \nu(\mathrm{d}y)$$

$$\approx \frac{\alpha}{1-\alpha} \sum_{m=1}^{M} \frac{w_{n,m}}{\sum_{m'=1}^{M} w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_n, Y_m)}$$

with $w_{n,m} = k(\theta_n, Y_m)^\alpha p(Y_m)^{1-\alpha}$

In practice : **Stochastic** Gradient Ascent using $k(\theta_n, \cdot)$ as a sampler
+ Mini-batching + Reparameterisation

Rényi divergence variational inference. Y. Li and R. E Turner. (2016). NeurIPS

Alpha-divergence Variational Inference : summary

Alpha-Divergence approach	Rényi's Alpha-Divergence approach
$\inf_{\theta \in \mathsf{T}} \Psi_\alpha(k(\theta, \cdot); p)$	$\sup_{\theta \in \mathsf{T}} \mathcal{L}_\alpha(k(\theta, \cdot); p)$
$\frac{1}{\alpha-1} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\|_{(\theta,y)=(\theta_n,Y_m)}$	$\frac{\alpha}{1-\alpha} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\|_{(\theta,y)=(\theta_n,Y_m)}$

Alpha : $\quad \theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,Y_m)}$

Rényi's Alpha : $\quad \theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n,Y_m)}$

Alpha-divergence Variational Inference : summary

Alpha-Divergence approach	Rényi's Alpha-Divergence approach
$\inf_{\theta \in \mathsf{T}} \Psi_\alpha(k(\theta, \cdot); p)$	$\inf_{\theta \in \mathsf{T}} -\alpha^{-1} \mathcal{L}_\alpha(k(\theta, \cdot); p)$
$\frac{1}{\alpha-1} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\rvert_{(\theta,y)=(\theta_n,Y_m)}$	$\frac{1}{\alpha-1} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\rvert_{(\theta,y)=(\theta_n,Y_m)}$

Alpha : $\theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right\rvert_{(\theta,y)=(\theta_n,Y_m)}$

Rényi's Alpha : $\theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right\rvert_{(\theta,y)=(\theta_n,Y_m)}$

Alpha-divergence Variational Inference : summary

Alpha-Divergence approach	Rényi's Alpha-Divergence approach
$\inf_{\theta \in \mathsf{T}} \Psi_\alpha(k(\theta, \cdot); p)$	$\inf_{\theta \in \mathsf{T}} -\alpha^{-1} \mathcal{L}_\alpha(k(\theta, \cdot); p)$
$\frac{1}{\alpha-1} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\|_{(\theta,y)=(\theta_n, Y_m)}$	$\frac{1}{\alpha-1} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta,y)}{\partial \theta} \right\|_{(\theta,y)=(\theta_n, Y_m)}$

Alpha : $\quad \theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \frac{1}{M} \sum_{m=1}^M w_{n,m} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n, Y_m)}$

Rényi's Alpha : $\quad \theta_{n+1} = \theta_n + \frac{r_n}{1-\alpha} \sum_{m=1}^M \frac{w_{n,m}}{\sum_{m'=1}^M w_{n,m'}} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_n, Y_m)}$

Outline

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

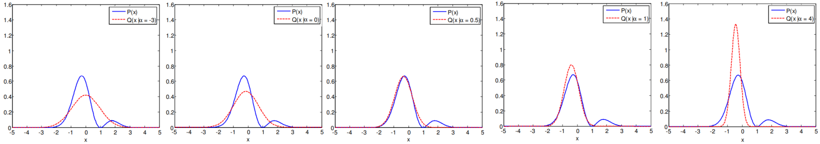
Conclusion of Part 1

Variational Inference : optimisation-based methods for Bayesian Inference

Core questions in Variational Inference :
- choice of the **variational family** $\mathcal{Q}$
- choice of the **measure of dissimilarity** D

1. MFVI : mean-field family, model-specific updates using the ELBO
2. SVI : scales MFVI to large datasets
3. BBVI : parametric family, Stochastic Gradient Ascent on the ELBO
4. **Alpha-divergence Variational Inference** : parametric family, extends BBVI to more general objective functions derived from the Alpha-divergence

Food for thoughts

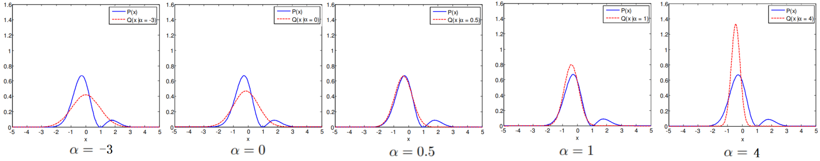


$\alpha = -3$ $\alpha = 0$ $\alpha = 0.5$ $\alpha = 1$ $\alpha = 4$

Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?

Some answers in Part 2 and 3!

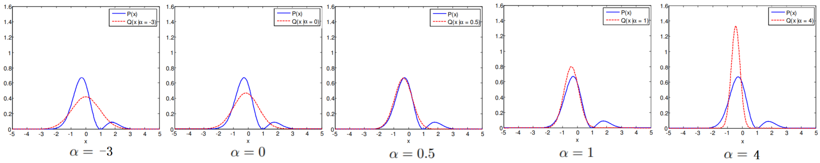
Food for thoughts



Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?

Some answers in Part 2 and 3!

Food for thoughts



Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?

Some answers in Part 2 and 3!

Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell) q_{-\ell}(y_{-\ell})$ and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y, \mathscr{D})}{q_\ell(y_\ell) q_{-\ell}(y_{-\ell})}\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y, \mathscr{D}) \nu_{-\ell}(\mathrm{d}y_{-\ell})\right) \nu_\ell(\mathrm{d}y_\ell)$$
$$- \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell) q_{-\ell}(y_{-\ell})\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(q_\ell(y_\ell)\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}$$

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}\,.$$

Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell) q_{-\ell}(y_{-\ell})$ and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y, \mathscr{D})}{q_\ell(y_\ell) q_{-\ell}(y_{-\ell})}\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y, \mathscr{D}) \nu_{-\ell}(\mathrm{d}y_{-\ell})\right) \nu_\ell(\mathrm{d}y_\ell)$$

$$- \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell) q_{-\ell}(y_{-\ell})\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(q_\ell(y_\ell)\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}$$

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}\,.$$

Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell)q_{-\ell}(y_{-\ell})$ and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$
\begin{aligned}
\mathsf{ELBO}(q;\mathscr{D}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y,\mathscr{D})}{q(y)}\right)\nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}_\ell}\int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell)q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y,\mathscr{D})}{q_\ell(y_\ell)q_{-\ell}(y_{-\ell})}\right)\nu_{-\ell}(\mathrm{d}y_{-\ell})\nu_\ell(\mathrm{d}y_\ell) \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell)\left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y,\mathscr{D})\nu_{-\ell}(\mathrm{d}y_{-\ell})\right)\nu_\ell(\mathrm{d}y_\ell) \\
&\quad - \int_{\mathsf{Y}_\ell}\int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell)q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell)q_{-\ell}(y_{-\ell})\right)\nu_{-\ell}(\mathrm{d}y_{-\ell})\nu_\ell(\mathrm{d}y_\ell) \\
&:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell)\mathbb{E}_{-\ell}\left[\log p(y,\mathscr{D})\right]\nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell)\log\left(q_\ell(y_\ell)\right)\nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}
\end{aligned}
$$

$$
\mathsf{ELBO}(q;\mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell)\log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y,\mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right)\nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}\,.
$$

Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell) q_{-\ell}(y_{-\ell})$ and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y, \mathscr{D})}{q_\ell(y_\ell) q_{-\ell}(y_{-\ell})}\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y, \mathscr{D}) \nu_{-\ell}(\mathrm{d}y_{-\ell})\right) \nu_\ell(\mathrm{d}y_\ell)$$
$$- \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell) q_{-\ell}(y_{-\ell})\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(q_\ell(y_\ell)\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}$$

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}\,.$$

Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell) q_{-\ell}(y_{-\ell})$
and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y, \mathscr{D})}{q_\ell(y_\ell) q_{-\ell}(y_{-\ell})}\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y, \mathscr{D}) \nu_{-\ell}(\mathrm{d}y_{-\ell})\right) \nu_\ell(\mathrm{d}y_\ell)$$

$$- \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell) q_{-\ell}(y_{-\ell})\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell)$$

$$:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(q_\ell(y_\ell)\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}$$

$$\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}\,.$$

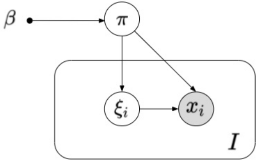
Proof of the optimal rule $q_\ell^*(y_\ell) \propto \exp\left(\mathbb{E}_{-\ell}[\log p(y, \mathscr{D})]\right)$

For notational convenience : $\mathsf{Y} = \mathsf{Y}_\ell \times \mathsf{Y}_{-\ell}$, $q(y) = q_\ell(y_\ell) q_{-\ell}(y_{-\ell})$ and $\nu(\mathrm{d}y) = \nu_\ell(\mathrm{d}y_\ell)\nu_{-\ell}(\mathrm{d}y_{-\ell})$.

$$
\begin{aligned}
\mathsf{ELBO}(q; \mathscr{D}) &= \int_{\mathsf{Y}} q(y) \log\left(\frac{p(y, \mathscr{D})}{q(y)}\right) \nu(\mathrm{d}y) \\
&= \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(\frac{p(y, \mathscr{D})}{q_\ell(y_\ell) q_{-\ell}(y_{-\ell})}\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell) \\
&= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \left(\int_{\mathsf{Y}_{-\ell}} q_{-\ell}(y_{-\ell}) \log p(y, \mathscr{D}) \nu_{-\ell}(\mathrm{d}y_{-\ell})\right) \nu_\ell(\mathrm{d}y_\ell) \\
&\qquad - \int_{\mathsf{Y}_\ell} \int_{\mathsf{Y}_{-\ell}} q_\ell(y_\ell) q_{-\ell}(y_{-\ell}) \log\left(q_\ell(y_\ell) q_{-\ell}(y_{-\ell})\right) \nu_{-\ell}(\mathrm{d}y_{-\ell}) \nu_\ell(\mathrm{d}y_\ell) \\
&:= \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right] \nu_\ell(\mathrm{d}y_\ell) - \int_{\mathsf{Y}_\ell} q_\ell(y_\ell) \log\left(q_\ell(y_\ell)\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell}
\end{aligned}
$$

$$
\mathsf{ELBO}(q; \mathscr{D}) = \int_{\mathsf{Y}} q_\ell(y_\ell) \log\left(\frac{\exp\left(\mathbb{E}_{-\ell}\left[\log p(y, \mathscr{D})\right]\right)}{q_\ell(y_\ell)}\right) \nu_\ell(\mathrm{d}y_\ell) + c_{-\ell} \ .
$$

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \dots x_I\}$, $x_1, \dots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \dots, \xi_I\}$, π : global latent variable, $\xi_1, \dots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

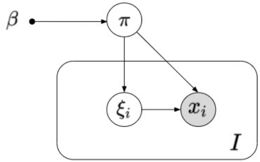
γ : global variational parameter, $\phi_1, \dots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \ldots x_I\}$, $x_1, \ldots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \ldots, \xi_I\}$, π : global latent variable, $\xi_1, \ldots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

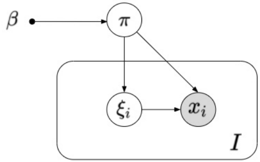
γ : global variational parameter, $\phi_1, \ldots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \ldots x_I\}$, $x_1, \ldots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \ldots, \xi_I\}$, π : global latent variable, $\xi_1, \ldots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

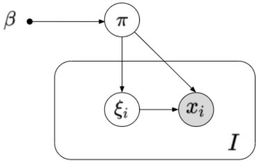
γ : global variational parameter, $\phi_1, \ldots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \ldots x_I\}$, $x_1, \ldots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \ldots, \xi_I\}$, π : global latent variable, $\xi_1, \ldots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

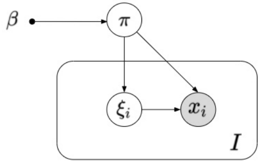
γ : global variational parameter, $\phi_1, \ldots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \ldots x_I\}$, $x_1, \ldots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \ldots, \xi_I\}$, π : global latent variable, $\xi_1, \ldots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

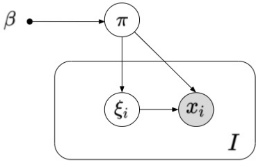
γ : global variational parameter, $\phi_1, \ldots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \ldots x_I\}$, $x_1, \ldots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \ldots, \xi_I\}$, π : global latent variable, $\xi_1, \ldots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

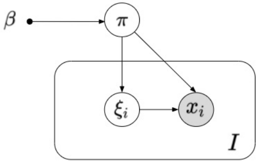
γ : global variational parameter, $\phi_1, \ldots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

CAVI for large datasets : Stochastic Variational Inference



- $\mathscr{D} = \{x_1, \dots x_I\}$, $x_1, \dots, x_I$: i.i.d. observations
- $y = \{\pi, \xi_1, \dots, \xi_I\}$, π : global latent variable, $\xi_1, \dots, \xi_I$: local latent variables (β : hyperparameter)

In that case, $p(y, \mathscr{D}) = p(\pi|\beta) \prod_{i=1}^{I} p(\xi_i|\pi) p(x_i|\xi_i, \pi)$

Mean-field approximation :

$$q(y) = q(\pi|\gamma) \prod_{i=1}^{I} q(\xi_i|\phi_i)$$

γ : global variational parameter, $\phi_1, \dots, \phi_I$: local variational parameters

Problem : I is often very large (e.g. 1.8M articles from the New York Times)

$\rightarrow$ The use of **stochastic** optimisation enabled large scale learning

Stochastic Variational Inference. M. D. Hoffman et al. (2013). JMRL.

Variational Inference
Foundations and recent advances
(Part 2)

Kamélia Daudel

UNIVERSITY OF OXFORD

University of Bristol – 09/03/2022

Reminder - 1

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D}) = \frac{p(\cdot,\mathscr{D})}{p(\mathscr{D})}$

- Variational Inference optimisation problem :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$$

where $\mathcal{Q}$ is the variational family and D is the measure of dissimilarity

- **Alpha-Divergence Variational Inference :** Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$
$$\mathcal{L}_\alpha(q;p) = \frac{1}{1-\alpha} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

Reminder - 1

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D}) = \frac{p(\cdot,\mathscr{D})}{p(\mathscr{D})}$

- Variational Inference optimisation problem :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$$

  where $\mathcal{Q}$ is the variational family and D is the measure of dissimilarity

- **Alpha-Divergence Variational Inference :** Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$
$$\mathcal{L}_\alpha(q;p) = \frac{1}{1-\alpha} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

  with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

Reminder - 1

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D}) = \frac{p(\cdot,\mathscr{D})}{p(\mathscr{D})}$

- Variational Inference optimisation problem :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}})$$

where $\mathcal{Q}$ is the variational family and D is the measure of dissimilarity

- **Alpha-Divergence Variational Inference :** Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$
$$\mathcal{L}_\alpha(q;p) = \frac{1}{1-\alpha} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

Reminder - 1

$(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$.
$\mathbb{Q}$ and $\mathbb{P}_{|\mathscr{D}}$: $\mathbb{Q} \preceq \nu$, $\mathbb{P}_{|\mathscr{D}} \preceq \nu$ with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}_{|\mathscr{D}}}{\mathrm{d}\nu} = p(\cdot|\mathscr{D}) = \frac{p(\cdot,\mathscr{D})}{p(\mathscr{D})}$

- Variational Inference optimisation problem :

$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q} || \mathbb{P}_{|\mathscr{D}})$$

where $\mathcal{Q}$ is the variational family and D is the measure of dissimilarity

- **Alpha-Divergence Variational Inference :** Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R} \setminus \{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

Reminder - 2

When $\mathcal{Q}$ is parametric,

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

we can perform Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?

Reminder - 2

When $\mathcal{Q}$ is parametric,

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

we can perform Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q; p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q; p)$)

Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?

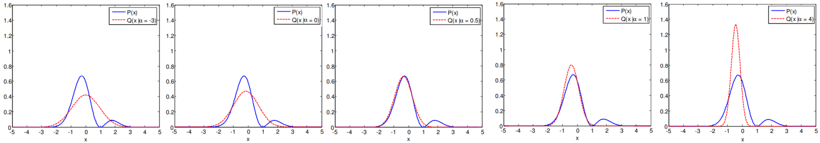
Reminder - 2

When $\mathcal{Q}$ is parametric,

$$\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

we can perform Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q; p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q; p)$)

Question : Can we further extend the approximating family $\mathcal{Q}$ in the context of Alpha-divergence Variational Inference?



$\alpha = -3$ $\alpha = 0$ $\alpha = 0.5$ $\alpha = 1$ $\alpha = 4$

Outline

1. Infinite-dimensional Alpha-divergence minimisation
2. Numerical experiments
3. Conclusion of Part 2

Outline

Infinite-dimensional Alpha-divergence minimisation

Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
K. Daudel, R. Douc and F. Portier. Ann. Statist. 49 (4) 2250 - 2270, August 2021.
https://doi.org/10.1214/20-AOS2035.

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Idea : Extend the traditional variational parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

by putting a prior on the variational parameter θ

$$\mathcal{Q} = \left\{ q : y \mapsto \mu k(y) := \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

and propose an update formula for μ that ensures a systematic decrease in $\mu \mapsto \Psi_\alpha(\mu k; p)$ at each step

$\to$ Finite Mixture Models : $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$

Infinite-dimensional Alpha-divergence minimisation

Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
K. Daudel, R. Douc and F. Portier. Ann. Statist. 49 (4) 2250 - 2270, August 2021.
https://doi.org/10.1214/20-AOS2035.

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Idea : Extend the traditional variational parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

by putting a prior on the variational parameter θ

$$\mathcal{Q} = \left\{ q : y \mapsto \mu k(y) := \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

and propose an update formula for μ that ensures a systematic decrease in $\mu \mapsto \Psi_\alpha(\mu k; p)$ at each step

$\rightarrow$ Finite Mixture Models : $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$

Infinite-dimensional Alpha-divergence minimisation

Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
K. Daudel, R. Douc and F. Portier. Ann. Statist. 49 (4) 2250 - 2270, August 2021.
https://doi.org/10.1214/20-AOS2035.

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Idea : Extend the traditional variational parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

by putting a prior on the variational parameter θ

$$\mathcal{Q} = \left\{q : y \mapsto \mu k(y) := \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$$

and propose an update formula for μ that ensures a systematic decrease in $\mu \mapsto \Psi_\alpha(\mu k; p)$ at each step

$\to$ Finite Mixture Models : $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$

Infinite-dimensional Alpha-divergence minimisation

Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
K. Daudel, R. Douc and F. Portier. Ann. Statist. 49 (4) 2250 - 2270, August 2021.
https://doi.org/10.1214/20-AOS2035.

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Idea : Extend the traditional variational parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

by putting a prior on the variational parameter θ

$$\mathcal{Q} = \left\{q : y \mapsto \mu k(y) := \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$$

and propose an update formula for μ that ensures a systematic decrease in $\mu \mapsto \Psi_\alpha(\mu k; p)$ at each step

$\rightarrow$ Finite Mixture Models : $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \quad \text{with} \quad \Psi_\alpha(\mu k; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \quad \text{with} \quad \Psi_\alpha(\mu k; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f_\alpha'\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \quad \text{with} \quad \Psi_\alpha(\mu k; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)$$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; p) \quad \text{with} \quad \Psi_\alpha(\mu k; p) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y) \nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; \cancel{p}) \quad \text{with} \quad \Psi_\alpha(\mu k; \cancel{p}) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n)\,, \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right)\nu(\mathrm{d}y)$$

The (α, Γ)-descent algorithm

Optimisation problem

$$\inf_{\mu \in \mathsf{M}} \Psi_\alpha(\mu k; \cancel{p}) \quad \text{with} \quad \Psi_\alpha(\mu k; \cancel{p}) := \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} \text{ with } b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1]\,(\log \Gamma)'(v) + 1 \geqslant 0\,.$$

Theorem

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu k) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then,

1. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$
2. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Theorem

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu k) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then,

1. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$
2. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Theorem

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu k) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then,

1. $\Psi_\alpha(\mathcal{I}_\alpha(\mu) k) \leqslant \Psi_\alpha(\mu k)$
2. $\Psi_\alpha(\mathcal{I}_\alpha(\mu) k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

First idea

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + f'_\alpha\left(\frac{\zeta k(y)}{p(y)}\right)\frac{\mu k(y) - \zeta k(y)}{p(y)}$$

✗ Not the best idea!

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

First idea

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + f'_\alpha\left(\frac{\zeta k(y)}{p(y)}\right)\frac{\mu k(y) - \zeta k(y)}{p(y)}$$

✗ Not the best idea!

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

<u>First idea</u>

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + f'_\alpha\left(\frac{\zeta k(y)}{p(y)}\right)\frac{\mu k(y) - \zeta k(y)}{p(y)}$$

✗ Not the best idea!

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

First idea

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + f'_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) \frac{\mu k(y) - \zeta k(y)}{p(y)}$$

✗ Not the best idea!

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

Second idea

By convexity of f_α : for all $y \in \mathsf{Y}$

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \ .$$

$\to$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

Second idea

By convexity of f_α : for all $y \in \mathsf{Y}$

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)]\ .$$

$\rightarrow$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

<u>Second idea</u>

By convexity of f_α : for all $y \in \mathsf{Y}$

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \ .$$

$\rightarrow$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \displaystyle\int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

<u>Second idea</u>

By convexity of f_α : for all $y \in \mathsf{Y}$

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \;.$$

$\to$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage : for all $y \in \mathsf{Y}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) \frac{1}{p(y)}[1 - g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y)p(y)$, we deduce

$$\Psi_\alpha(\mu k) \geqslant \Psi_\alpha(\zeta k) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$$

Choice of A_α

We take $\quad A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage : for all $y \in \mathsf{Y}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1-g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y)p(y)$, we deduce

$$\Psi_\alpha(\mu k) \geqslant \Psi_\alpha(\zeta k) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)]$$

Choice of A_α

We take $\quad A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)]$

Proof : 1) Proving a general lower bound

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu k) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\zeta k)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu k) = \int_{\mathsf{Y}} f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage : for all $y \in \mathsf{Y}$,

$$f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \geqslant f_\alpha \left(\frac{\zeta k(y)}{p(y)} \right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta) \mu k(y)}{p(y)} \right) \frac{1}{p(y)} [1 - g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y) p(y)$, we deduce

$$\Psi_\alpha(\mu k) \geqslant \Psi_\alpha(\zeta k) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta) \mu k(y)}{p(y)} \right) [1 - g(\theta)]$$

Choice of A_α

We take $\quad A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta) \mu k(y)}{p(y)} \right) [1 - g(\theta)]$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Setting $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} = \mathcal{I}_\alpha(\mu)(\mathrm{d}\theta)$$

and thus

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu) k)$$

with $\quad A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Setting $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} = \mathcal{I}_\alpha(\mu)(\mathrm{d}\theta)$$

and thus

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu) k)$$

with $\quad A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Setting $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} = \mathcal{I}_\alpha(\mu)(\mathrm{d}\theta)$$

and thus

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu) k)$$

with $\quad A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f_\alpha' \left(\frac{g(\theta) \mu k(y)}{p(y)} \right) [1 - g(\theta)] \ .$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

Proving that $A_\alpha \geqslant 0$ $\rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f_\alpha'(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Setting $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} = \mathcal{I}_\alpha(\mu)(\mathrm{d}\theta)$$

and thus

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu) k)$$

with $\quad A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha \left(\frac{g(\theta) \mu k(y)}{p(y)} \right) [1 - g(\theta)] \ .$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha - 1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha - 1} - 1 \right] [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Setting $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$\zeta(\mathrm{d}\theta) = \mu(\mathrm{d}\theta) g(\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))} = \mathcal{I}_\alpha(\mu)(\mathrm{d}\theta)$$

and thus

$$A_\alpha \leqslant \Psi_\alpha(\mu k) - \Psi_\alpha(\mathcal{I}_\alpha(\mu) k)$$

with $\quad A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha - 1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0$ $\rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} \cancel{-1} \right] \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} \cancel{- 1} \right] \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$\begin{aligned}
A_\alpha &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} \cancel{-1} \right] \right) [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]
\end{aligned}$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0$ $\rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

In this case $f'_\alpha(u) = \frac{1}{\alpha-1}[u^{\alpha-1} - 1]$ and

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$\begin{aligned}
A_\alpha &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} \cancel{-1} \right] \right) [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} \nu(\mathrm{d}y) \right) g(\theta)^{\alpha-1} [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]
\end{aligned}$$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$A_\alpha = \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right)$$
$$= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1\right] (\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

Conclusion: $A_\alpha \geqslant 0$! □

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \rightarrow \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1 \right] (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Conclusion: $A_\alpha \geqslant 0$! □

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1}\right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1}\right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V)\right]\right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1}\right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V)\right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1\right] (\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$A_\alpha = \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right)$$

$$= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1\right] (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \to$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \geqslant 0 \; .$$

Conclusion: $A_\alpha \geqslant 0$! □

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \rightarrow \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \, .$$

Conclusion: $A_\alpha \geqslant 0$! □

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \rightarrow \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

Conclusion: $A_\alpha \geqslant 0$!

Proof : 2) take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$

Proving that $A_\alpha \geqslant 0 \rightarrow$ We treat the case $\alpha \in \mathbb{R} \setminus \{1\}$.

We have obtained that

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

It's time to use that $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$!

(i) Let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$ (probability space $(\mathsf{T}, \mathcal{T}, \mu)$)

(ii) Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$.

Then,

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \quad \text{since} \quad \mathbb{E}[1 - \tilde{\Gamma}(V)] = 0 \end{aligned}$$

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1 \right] (\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

Conclusion: $A_\alpha \geqslant 0$! $\square$

Reminder : Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] \, (\log \Gamma)'(v) + 1 \geqslant 0 \; .$$

Theorem

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu k) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then,

1. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) \leqslant \Psi_\alpha(\mu k)$
2. $\Psi_\alpha(\mathcal{I}_\alpha(\mu)k) = \Psi_\alpha(\mu k)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v)+1 \geqslant 0\ .$$

- Entropic Mirror Descent : $\eta \in (0,1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\exp\left[-\eta\int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu_n k(y)}{p(y)}\right)\nu(\mathrm{d}y)\right]$$

$\to$ NB : η corresponds to the learning rate

- Power descent : $\eta \in (0,1]$, $(\alpha-1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha-1)\,v+1]^{\frac{\eta}{1-\alpha}}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\left[\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)+(\alpha-1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v)+1 \geqslant 0\ .$$

- Entropic Mirror Descent : $\eta \in (0,1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\exp\left[-\eta\int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu_n k(y)}{p(y)}\right)\nu(\mathrm{d}y)\right]$$

$\rightarrow$ NB : η corresponds to the learning rate

- Power descent : $\eta \in (0,1]$, $(\alpha-1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha-1)\,v+1]^{\frac{\eta}{1-\alpha}}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\left[\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)+(\alpha-1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \ .$$

- Entropic Mirror Descent : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\eta \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu_n k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right]$$

$\rightarrow$ NB : η corresponds to the learning rate

- Power descent : $\eta \in (0, 1]$, $(\alpha - 1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \left[\int_{\mathsf{Y}} k(\theta, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1} \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1]\,(\log \Gamma)'(v) + 1 \geqslant 0\,.$$

- Entropic Mirror Descent : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\eta \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu_n k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right]$$

$\rightarrow$ NB : η corresponds to the learning rate

- Power descent : $\eta \in (0, 1]$, $(\alpha - 1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}}$$

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \left[\int_{\mathsf{Y}} k(\theta, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha - 1} \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$
Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n\geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\cdot\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha}+\kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n\geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \quad \Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\}, \ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{where} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J, \ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

$$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\}\,,\ \lambda_j \geqslant 0 \text{ and } \textstyle\sum_{j=1}^J \lambda_j = 1 \right\}$$

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

$$\mu_{\boldsymbol{\lambda}} = \textstyle\sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{where} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J,\ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\}\, , \ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{where} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J, \ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

$$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\}\,,\ \lambda_j \geqslant 0 \text{ and } \textstyle\sum_{j=1}^J \lambda_j = 1 \right\}$$

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

$$\mu_{\boldsymbol{\lambda}} = \textstyle\sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{where} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J,\ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing
e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$
Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\ldots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1 \dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\ldots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n}k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\ldots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n}k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\ldots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n}k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\ldots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing
e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2 \log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$
Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot), \dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n \geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\to$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M}\sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha\left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})}\right),$$

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\to$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\to$ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\rightarrow$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\rightarrow$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\rightarrow$ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\rightarrow$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\rightarrow$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\rightarrow$ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\rightarrow$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^{M} \tfrac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\tfrac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\rightarrow$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\rightarrow$ Idea : combine this step with and *Exploration Step* updating Θ

Outline

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \; : \; \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

<u>Algorithm</u>

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
  $p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

<u>Algorithm</u>

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

Algorithm

❶ Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
❷ *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} .$$

<u>Algorithm</u>

❶ Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
❷ *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

- Toy example
$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

- 0.5-Mirror descent : $\Gamma(v) = e^{-\eta v}$ and $\alpha = 0.5$,
- 0.5-Power descent : $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ and $\alpha = 0.5$.

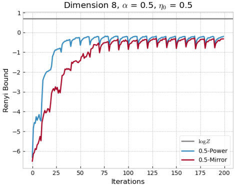
$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials

Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ and $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ and $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials

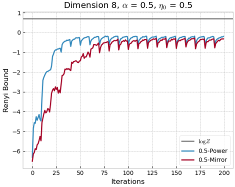


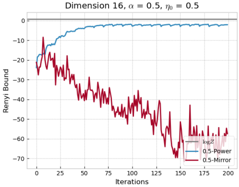
Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

- 0.5-Mirror descent : $\Gamma(v) = e^{-\eta v}$ and $\alpha = 0.5$,
- 0.5-Power descent : $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ and $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials



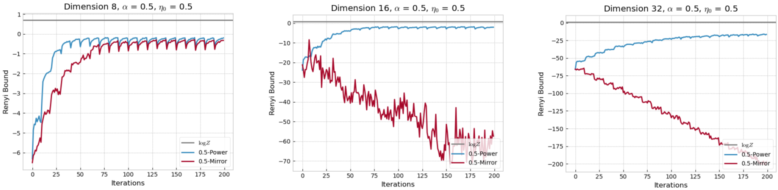


Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ and $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ and $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials



Toy example : the case $\alpha = 1$

Comparison between:

- <u>1-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

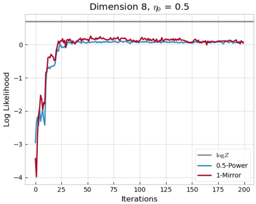
$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials

Toy example : the case $\alpha = 1$

Comparison between:

- <u>1-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials

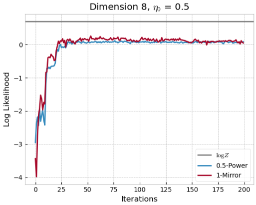


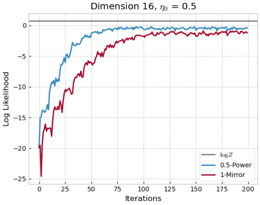
Toy example : the case $\alpha = 1$

Comparison between:

- 1-Mirror descent : $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$,
- 0.5-Power descent : $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials



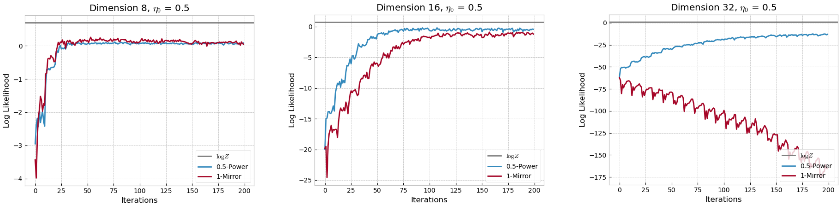


Toy example : the case $\alpha = 1$

Comparison between:

- <u>1-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials



Bayesian Logistic Regression

$\rightarrow$ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

$\rightarrow$ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

$\rightarrow$ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$\begin{aligned} p_0(\beta) &= \mathrm{Gamma}(\beta; a, b) \ , \\ p_0(w_l|\beta) &= \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L \\ p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) &= \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I \end{aligned}$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$
\begin{aligned}
p_0(\beta) &= \mathrm{Gamma}(\beta; a, b) \ , \\
p_0(w_l|\beta) &= \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L \\
p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) &= \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I
\end{aligned}
$$

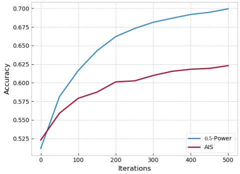
where $a = 1$ and $b = 0.01$

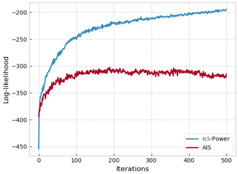
Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS





$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Outline

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \; : \; \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \; : \; \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Conclusion of Part 2

General framework for infinite-dimensional α-divergence minimisation over

$$\mathcal{Q} = \left\{ q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M} \right\}$$

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Food for thoughts - 1

- Question : What about Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$?

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

$$\text{Alpha :} \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1} \int_{\mathsf{Y}} k(\theta, y) \mu_n k(y)^{\alpha-1} p(y)^{1-\alpha} \nu(\mathrm{d}y)\right]$$

$$\text{Rényi's Alpha :} \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1} \frac{\int_{\mathsf{Y}} k(\theta, y) \mu_n k(y)^{\alpha-1} p(y)^{1-\alpha} \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^{\alpha} p(y)^{1-\alpha} \nu(\mathrm{d}y)}\right]$$

$\rightarrow$ The Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ is in fact closely-related to the Power Descent

Food for thoughts - 1

- Question : What about Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$?

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Alpha : $\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha-1}\int_{\mathsf{Y}} k(\theta, y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)\right]$

Rényi's Alpha : $\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha-1}\frac{\int_{\mathsf{Y}} k(\theta, y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^{\alpha} p(y)^{1-\alpha}\nu(\mathrm{d}y)}\right]$

$\rightarrow$ The Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ is in fact closely-related to the Power Descent

Food for thoughts - 1

- Question : What about Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$?

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

$$\text{Alpha :} \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1}\int_{\mathsf{Y}} k(\theta, y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)\right]$$

$$\text{Rényi's Alpha :} \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1}\frac{\int_{\mathsf{Y}} k(\theta, y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^{\alpha} p(y)^{1-\alpha}\nu(\mathrm{d}y)}\right]$$

→ The Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ is in fact closely-related to the Power Descent

Food for thoughts - 1

- Question : What about Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$?

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

Alpha : $\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1} \int_{\mathsf{Y}} k(\theta, y) \mu_n k(y)^{\alpha-1} p(y)^{1-\alpha} \nu(\mathrm{d}y)\right]$

Rényi's Alpha : $\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha - 1} \frac{\int_{\mathsf{Y}} k(\theta, y) \mu_n k(y)^{\alpha-1} p(y)^{1-\alpha} \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^{\alpha} p(y)^{1-\alpha} \nu(\mathrm{d}y)}\right]$

→ The Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ is in fact closely-related to the Power Descent

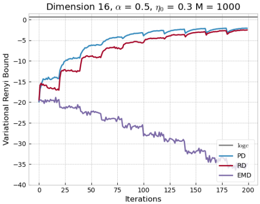
Food for thoughts - 1

- Question : What about Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$?

Mixture weights optimisation for Alpha-Divergence Variational Inference.
K. Daudel and R. Douc (2021). To appear in NeurIPS2021

$$\text{Alpha}: \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha-1}\int_{\mathsf{Y}} k(\theta,y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)\right]$$

$$\text{Rényi's Alpha}: \quad \mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\frac{\eta}{\alpha-1}\frac{\int_{\mathsf{Y}} k(\theta,y)\mu_n k(y)^{\alpha-1} p(y)^{1-\alpha}\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} \mu_n k(y)^{\alpha} p(y)^{1-\alpha}\nu(\mathrm{d}y)}\right]$$



→ The Entropic Mirror Descent applied to $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ is in fact closely-related to the Power Descent

Food for thoughts - 2

- Question : What is a good choice for the exploration step?

  Some answers in Part 3!

Food for thoughts - 2

- Question : What is a good choice for the exploration step?

  Some answers in Part 3!

Variational Inference
Foundations and recent advances
(Part 3)

Kamélia Daudel

University of Bristol – 09/03/2022

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \; : \; \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \; : \; \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \; : \; \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \;:\; \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \;:\; \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \;:\; \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \;:\; \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \;:\; \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \;:\; \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 1

Alpha-Divergence Variational Inference : Two possible objective functions

$$\Psi_\alpha(q;p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

$$-\alpha^{-1}\mathcal{L}_\alpha(q;p) = \frac{1}{\alpha(\alpha-1)} \log\left(\int_{\mathsf{Y}} q(y)^\alpha p(y)^{1-\alpha}\nu(\mathrm{d}y)\right)$$

with $p = p(\cdot, \mathscr{D})$ and

$$f_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right], & \text{if } \alpha \in \mathbb{R}\setminus\{0,1\}, \\ u\log(u) + 1 - u, & \text{if } \alpha = 1 \text{ (Exclusive KL)}, \\ -\log(u) + u - 1, & \text{if } \alpha = 0 \text{ (Inclusive KL)}. \end{cases}$$

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Reminder - 2

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)
$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Question : Can we propose valid updates for

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\} ?$$

Reminder - 2

- $\mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$

Stochastic Gradient Descent w.r.t θ on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

- $\mathcal{Q} = \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \ : \ \mu \in \mathsf{M}\right\}$

Power Descent, Entropic Mirror Descent on $\Psi_\alpha(q;p)$ (resp. $-\alpha^{-1}\mathcal{L}_\alpha(q;p)$)

$\rightarrow$ applies to $\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J\right\}$

Question : Can we propose valid updates for

$$\mathcal{Q} = \left\{q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J\right\} ?$$

Outline

Outline

Monotonic Alpha-Divergence Minimisation

Monotonic Alpha-divergence Minimisation.
K. Daudel, R. Douc and F. Roueff (2021). https://arxiv.org/abs/2103.05684

Idea : Extend the typical variational parametric family

$$\mathcal{Q} = \{y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\}$$

by considering the mixture model variational family

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda},\Theta} k(y) := \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

and propose an update formula for $(\boldsymbol{\lambda}, \Theta)$ that ensures a systematic decrease in the alpha-divergence (i.e. Ψ_α) at each step.

$\rightarrow$ Optimising w.r.t $\boldsymbol{\lambda}$ and Θ is the novelty compared to Part 2!

Conditions for a monotonic decrease

Optimisation problem

$$\inf_{\boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J} \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta} k; p) \quad \text{with} \quad \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta} k; p) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_{\boldsymbol{\lambda},\Theta} k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

Theorem

Assume (A1). Let $\alpha \in [0, 1)$, $J \in \mathbb{N}^\star$. Then, choosing $(\boldsymbol{\lambda}_n, \Theta_n)_{n \geqslant 1}$ so that: $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_1,\Theta_1} k) < \infty$ and $\forall n \geqslant 1$,

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_{\boldsymbol{\lambda}_n,\Theta_n} k(y)}{p(y)}\right)^{\alpha-1}$, yields a systematic decrease in Ψ_α at each step.

Conditions for a monotonic decrease

Optimisation problem

$$\inf_{\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J} \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k\cancel{;p}) \quad \text{with} \quad \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k\cancel{;p}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_{\boldsymbol{\lambda},\Theta}k(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)$$

(A1) For all $(\theta,y)\in\mathsf{T}\times\mathsf{Y}$, $k(\theta,y)>0$, $p(y)\geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y)<\infty$.

Theorem

Assume (A1). Let $\alpha\in[0,1)$, $J\in\mathbb{N}^\star$. Then, choosing $(\boldsymbol{\lambda}_n,\Theta_n)_{n\geqslant 1}$ so that: $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_1,\Theta_1}k)<\infty$ and $\forall n\geqslant 1$,

$$\int_{\mathsf{Y}}\sum_{j=1}^J \lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y)\log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right)\nu(\mathrm{d}y)\geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}}\sum_{j=1}^J \lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y)\log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right)\nu(\mathrm{d}y)\geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y)=k(\theta_{j,n},y)\left(\frac{\mu_{\boldsymbol{\lambda}_n,\Theta_n}k(y)}{p(y)}\right)^{\alpha-1}$, yields a systematic decrease in Ψ_α at each step.

Conditions for a monotonic decrease

Optimisation problem

$$\inf_{\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J}\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k;\cancel{p}) \quad \text{with} \quad \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k;\cancel{p}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_{\boldsymbol{\lambda},\Theta}k(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)$$

(A1) For all $(\theta,y)\in\mathsf{T}\times\mathsf{Y}$, $k(\theta,y)>0$, $p(y)\geqslant 0$ and $\int_{\mathsf{Y}}p(y)\nu(\mathrm{d}y)<\infty$.

Theorem

Assume (A1). Let $\alpha\in[0,1)$, $J\in\mathbb{N}^\star$. Then, choosing $(\boldsymbol{\lambda}_n,\Theta_n)_{n\geqslant 1}$ so that: $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_1,\Theta_1}k)<\infty$ and $\forall n\geqslant 1$,

$$\int_{\mathsf{Y}}\sum_{j=1}^J\lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y)\log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right)\nu(\mathrm{d}y)\geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}}\sum_{j=1}^J\lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y)\log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right)\nu(\mathrm{d}y)\geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y)=k(\theta_{j,n},y)\left(\frac{\mu_{\boldsymbol{\lambda}_n,\Theta_n}k(y)}{p(y)}\right)^{\alpha-1}$, yields a systematic decrease in Ψ_α at each step.

Conditions for a monotonic decrease

Optimisation problem

$$\inf_{\boldsymbol{\lambda}\in\mathcal{S}_J,\Theta\in\mathsf{T}^J} \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k\cancel{;p}) \quad \text{with} \quad \Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta}k\cancel{;p}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu_{\boldsymbol{\lambda},\Theta}k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$$

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

Theorem

Assume (A1). Let $\alpha \in [0, 1)$, $J \in \mathbb{N}^\star$. Then, choosing $(\boldsymbol{\lambda}_n, \Theta_n)_{n\geqslant 1}$ so that: $\Psi_\alpha(\mu_{\boldsymbol{\lambda}_1,\Theta_1}k) < \infty$ and $\forall n \geqslant 1$,

$$\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^J \lambda_{j,n}\varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y)\left(\frac{\mu_{\boldsymbol{\lambda}_n,\Theta_n}k(y)}{p(y)}\right)^{\alpha-1}$, yields a systematic decrease in Ψ_α at each step.

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \displaystyle\int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

→ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_\mathsf{Y} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_\mathsf{Y} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_\mathsf{Y} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_\mathsf{Y} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_\mathsf{Y} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_\mathsf{Y} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

→ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right) p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right) p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 1

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)$.

$\to$ **Case** $\alpha = 0$ **:** $f_0(u) = -\log(u) + u - 1$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}}\left(-\log\left(\frac{q(y)}{p(y)}\right) + \frac{q(y)}{p(y)} - 1\right)p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}}\left(-\log\left(\frac{q(y)}{p(y)}\right) + \left(\frac{q(y)}{p(y)} - \frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right)p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}}\left(-\log\left(\frac{q(y)}{q'(y)}\right) - \log\left(\frac{q'(y)}{p(y)}\right) + \frac{q'(y)}{p(y)} - 1\right)p(y)\nu(\mathrm{d}y)$$

so that

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} -\log\left(\frac{q(y)}{q'(y)}\right)p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \displaystyle\int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha + \left(\frac{q'(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha + \left(\frac{q'(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha + \left(\frac{q'(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha + \left(\frac{q'(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right] p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q(y)}{p(y)} - 1\right)\right]p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right]p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha + \left(\frac{q'(y)}{p(y)}\right)^\alpha - 1 - \alpha\left(\frac{q'(y)}{p(y)} - 1\right)\right]p(y)\nu(\mathrm{d}y)$$

$$= \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right]p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \displaystyle\int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0, 1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0, 1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

$$= \int_{\mathsf{Y}} \left(\frac{q'(y)}{p(y)}\right)^\alpha \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{q'(y)}\right)^\alpha - 1\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \displaystyle\int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

$$= \int_{\mathsf{Y}} \left(\frac{q'(y)}{p(y)}\right)^\alpha \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{q(y)}{q'(y)}\right)^\alpha - 1\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Since $\log(u) \leqslant u - 1$ for all $u > 0$ and $\alpha \in (0,1)$,

$$\frac{u-1}{\alpha(\alpha-1)} \leqslant \frac{\log(u)}{\alpha(\alpha-1)}$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

$$= \int_{\mathsf{Y}} \left(\frac{q'(y)}{p(y)}\right)^\alpha \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{q'(y)}\right)^\alpha - 1\right] p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Since $\log(u) \leqslant u-1$ for all $u > 0$ and $\alpha \in (0,1)$,

$$\frac{u^{\alpha}-1}{\alpha(\alpha-1)} \leqslant \frac{\log(u^{\alpha})}{\alpha(\alpha-1)}$$

Proof : 1) Proving a general lower bound - 2

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha}\log\left(\frac{q(y)}{q'(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

By definition $\Psi_\alpha(q) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right)p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ **Case** $\alpha \in (0,1)$ **:** $f_\alpha(u) = \frac{1}{\alpha(\alpha-1)}\left[u^\alpha - 1 - \alpha(u-1)\right]$

$$\Psi_\alpha(q) = \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{p(y)}\right)^\alpha - \left(\frac{q'(y)}{p(y)}\right)^\alpha\right]p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

$$= \int_{\mathsf{Y}} \left(\frac{q'(y)}{p(y)}\right)^\alpha \frac{1}{\alpha(\alpha-1)}\left[\left(\frac{q(y)}{q'(y)}\right)^\alpha - 1\right]p(y)\nu(\mathrm{d}y) + \Psi_\alpha(q')$$

Since $\log(u) \leqslant u-1$ for all $u > 0$ and $\alpha \in (0,1)$,

$$\frac{u^{\alpha}-1}{\alpha(\alpha-1)} \leqslant \frac{\log(u^{\alpha})}{\alpha(\alpha-1)} = \frac{\log(u)}{\alpha-1}$$

Proof : 2) Derive (Weights) and (Components)

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n, \Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

Proof : 2) Derive (Weights) and (Components)

Let $q, q' \in \mathcal{Q}$ and assume that $\Psi_\alpha(q') < \infty$. For all $\alpha \in [0, 1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} q'(y)^\alpha p(y)^{1-\alpha} \log\left(\frac{q(y)}{q'(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(q') - \Psi_\alpha(q)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n, \Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

Proof : 2) Derive (Weights) and (Components)

Assume that $\Psi_\alpha(\mu_n k) < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu_{n+1}k)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n,\Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

Proof : 2) Derive (Weights) and (Components)

Assume that $\Psi_\alpha(\mu_n k) < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha} \int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu_{n+1}k)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n, \Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

💡 $u \mapsto \frac{1}{1-\alpha} \log(u)$ is concave

Proof : 2) Derive (Weights) and (Components)

Assume that $\Psi_\alpha(\mu_n k) < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right)\nu(\mathrm{d}y) \leqslant \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu_{n+1}k)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n,\Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

$u \mapsto \frac{1}{1-\alpha}\log(u)$ is concave

Jensen's inequality: for all $y \in \mathsf{Y}$ and all $n \geqslant 1$,

$$\begin{aligned}\frac{1}{1-\alpha}\log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) &= \frac{1}{1-\alpha}\log\left(\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \frac{\lambda_{j,n+1}k(\theta_{j,n+1},y)}{\lambda_{j,n}k(\theta_{j,n},y)}\right) \\ &\geqslant \frac{1}{1-\alpha}\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \log\left(\frac{\lambda_{j,n+1}k(\theta_{j,n+1})}{\lambda_{j,n}k(\theta_{j,n},y)}\right)\end{aligned}$$

Proof : 2) Derive (Weights) and (Components)

Assume that $\Psi_\alpha(\mu_n k) < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu_{n+1} k)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n,\Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

$u \mapsto \frac{1}{1-\alpha}\log(u)$ is concave

Jensen's inequality: for all $y \in \mathsf{Y}$ and all $n \geqslant 1$,

$$\frac{1}{1-\alpha}\log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) = \frac{1}{1-\alpha}\log\left(\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \frac{\lambda_{j,n+1}k(\theta_{j,n+1},y)}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

$$\geqslant \frac{1}{1-\alpha}\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \log\left(\frac{\lambda_{j,n+1}k(\theta_{j,n+1})}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

that is :

$$\frac{1}{1-\alpha}\log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) \geqslant \frac{1}{1-\alpha}\sum_{j=1}^J \lambda_{j,n}\frac{k(\theta_{j,n},y)}{\mu_n k(y)} \log\left(\frac{\lambda_{j,n+1}k(\theta_{j,n+1})}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

Proof : 2) Derive (Weights) and (Components)

Assume that $\Psi_\alpha(\mu_n k) < \infty$. For all $\alpha \in [0,1)$, it holds that

$$\frac{1}{1-\alpha}\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) \nu(\mathrm{d}y) \leqslant \Psi_\alpha(\mu_n k) - \Psi_\alpha(\mu_{n+1}k)$$

Notation : $\mu_n k(y) := \mu_{\boldsymbol{\lambda}_n, \Theta_n} k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_{j,n}, y)$, for all $n \geqslant 1$ and all $y \in \mathsf{Y}$

$u \mapsto \frac{1}{1-\alpha}\log(u)$ is concave

Jensen's inequality: for all $y \in \mathsf{Y}$ and all $n \geqslant 1$,

$$\frac{1}{1-\alpha}\log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) = \frac{1}{1-\alpha}\log\left(\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \frac{\lambda_{j,n+1}k(\theta_{j,n+1},y)}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

$$\geqslant \frac{1}{1-\alpha}\sum_{j=1}^J \frac{\lambda_{j,n}k(\theta_{j,n},y)}{\sum_{\ell=1}^J \lambda_{\ell,n}k(\theta_{\ell,n},y)} \log\left(\frac{\lambda_{j,n+1}k(\theta_{j,n+1})}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

that is :

$$\frac{1}{1-\alpha}\log\left(\frac{\mu_{n+1}k(y)}{\mu_n k(y)}\right) \geqslant \frac{1}{1-\alpha}\sum_{j=1}^J \lambda_{j,n}\frac{k(\theta_{j,n},y)}{\mu_n k(y)} \log\left(\frac{\lambda_{j,n+1}k(\theta_{j,n+1})}{\lambda_{j,n}k(\theta_{j,n},y)}\right)$$

To finish the proof :

(i) multiply by $(\mu_n k(y))^\alpha p(y)^{1-\alpha}$ on both sides ($\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n},y)\left(\frac{p(y)}{\mu_n k(y)}\right)^{1-\alpha}$)

(ii) integrate with respect to $\nu(\mathrm{d}y)$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

   $\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\boldsymbol{\lambda}_{n+1} = \operatorname{argmax}_{\boldsymbol{\lambda} \in \mathcal{S}_J} \sum_{j=1}^{J} \left[\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)\right] \log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

   $\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\boldsymbol{\lambda}_{n+1} = \mathrm{argmax}_{\boldsymbol{\lambda} \in \mathcal{S}_J} \sum_{j=1}^{J} \left[\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)\right] \log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha - 1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

   $\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\boldsymbol{\lambda}_{n+1} = \operatorname{argmax}_{\boldsymbol{\lambda} \in \mathcal{S}_J} \sum_{j=1}^{J} \left[\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)\right] \log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \tag{Weights}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \tag{Components}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

$\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\boldsymbol{\lambda}_{n+1} = \mathrm{argmax}_{\boldsymbol{\lambda} \in \mathcal{S}_J} \sum_{j=1}^{J} \left[\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)\right] \log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \tag{Weights}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \tag{Components}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

   $\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\boldsymbol{\lambda}_{n+1} = \operatorname{argmax}_{\boldsymbol{\lambda} \in \mathcal{S}_J} \sum_{j=1}^{J} \left[\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)\right] \log\left(\frac{\lambda_j}{\lambda_{j,n}}\right)$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n},y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

   $\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y)}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y)}, \quad j = 1 \ldots J$$

Commenting the conditions for a monotonic decrease

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Weights)}$$

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

where $\varphi_{j,n}^{(\alpha)}(y) = k(\theta_{j,n}, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}$

1. (Weights) and (Components) permit separate/simultaneous updates
2. They are satified for $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n$ and $\Theta_{n+1} = \Theta_n$ respectively
3. The dependency is **simpler** in (Weights)

$\rightarrow$ (Weights) holds for $\boldsymbol{\lambda}_{n+1}$ such that

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}, \quad j = 1 \ldots J$$

where $\eta_n \in (0, 1]$ and κ is such that $(\alpha - 1)\kappa \geqslant 0$

Understanding the mixture weights update

(Weights) and (Components) hold for $\boldsymbol{\lambda}_{n+1}$ and Θ_{n+1} such that:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}, \quad j = 1 \ldots J$$

$$\Theta_{n+1} = \Theta_n$$

where $\eta_n \in (0,1]$ and κ is such that $(\alpha-1)\kappa \geqslant 0$

→ We recover the Power Descent algorithm from Part 2

Core insights :

1. The mixture weights update is gradient-based, η_n plays the role of a learning rate
2. We can improve on the Power Descent by proposing simultaneous updates for Θ with convergence guarantees!

Understanding the mixture weights update

(Weights) and (Components) hold for $\boldsymbol{\lambda}_{n+1}$ and Θ_{n+1} such that:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}, \quad j = 1 \ldots J$$

$$\Theta_{n+1} = \Theta_n$$

where $\eta_n \in (0,1]$ and κ is such that $(\alpha-1)\kappa \geqslant 0$

$\rightarrow$ We recover the Power Descent algorithm from Part 2

Core insights :

1. The mixture weights update is gradient-based, η_n plays the role of a learning rate
2. We can improve on the Power Descent by proposing simultaneous updates for Θ with convergence guarantees!

Understanding the mixture weights update

(Weights) and (Components) hold for $\boldsymbol{\lambda}_{n+1}$ and Θ_{n+1} such that:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}\ , \quad j = 1 \ldots J$$

$$\Theta_{n+1} = \Theta_n$$

where $\eta_n \in (0,1]$ and κ is such that $(\alpha-1)\kappa \geqslant 0$

$\rightarrow$ We recover the Power Descent algorithm from Part 2

Core insights :

1. The mixture weights update is gradient-based, η_n plays the role of a learning rate
2. We can improve on the Power Descent by proposing simultaneous updates for Θ with convergence guarantees!

Understanding the mixture weights update

(Weights) and (Components) hold for $\boldsymbol{\lambda}_{n+1}$ and Θ_{n+1} such that:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}\,, \quad j = 1 \ldots J$$

$$\Theta_{n+1} = \Theta_n$$

where $\eta_n \in (0,1]$ and κ is such that $(\alpha-1)\kappa \geqslant 0$

→ We recover the Power Descent algorithm from Part 2

Core insights :

1. The mixture weights update is gradient-based, η_n plays the role of a learning rate
2. We can improve on the Power Descent by proposing simultaneous updates for Θ with convergence guarantees!

Understanding the mixture weights update

(Weights) and (Components) hold for $\boldsymbol{\lambda}_{n+1}$ and Θ_{n+1} such that:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n}\left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha-1)\kappa\right]^{\eta_n}}\,, \quad j = 1 \ldots J$$

$$\Theta_{n+1} = \Theta_n$$

where $\eta_n \in (0,1]$ and κ is such that $(\alpha-1)\kappa \geqslant 0$

$\rightarrow$ We recover the Power Descent algorithm from Part 2

Core insights :

1. The mixture weights update is gradient-based, η_n plays the role of a learning rate
2. We can improve on the Power Descent by proposing simultaneous updates for Θ with convergence guarantees!

Towards simultaneous updates

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

- Maximisation approach : for all $j = 1 \dots J$,

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y)$$

- Gradient-based approach : for all $j = 1 \dots J$, $\gamma_{j,n} \in (0, 1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \ .$$

Towards simultaneous updates

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

- Maximisation approach : for all $j = 1 \dots J$,

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y)$$

- Gradient-based approach : for all $j = 1 \dots J$, $\gamma_{j,n} \in (0,1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta = \theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \; .$$

Towards simultaneous updates

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

- Maximisation approach : for all $j = 1 \ldots J$,

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y)$$

- Gradient-based approach : for all $j = 1 \ldots J$, $\gamma_{j,n} \in (0, 1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \; .$$

Two questions at this stage

1. Can we derive practical updates from the maximisation / gradient-based approaches?

2. Do those approaches relate to the existing litterature?

Two questions at this stage

1. Can we derive practical updates from the maximisation / gradient-based approaches?

2. Do those approaches relate to the existing litterature?

Outline

Maximisation approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \dots J$,

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y)$$

Maximisation approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \dots J$,

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y)$$

Maximisation approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \dots J$, $b_{j,n} \geqslant 0$ and

$$\theta_{j,n+1} = \mathrm{argmax}_{\theta \in \mathsf{T}} \int_{\mathsf{Y}} \left[\varphi_{j,n}^{(\alpha)}(y) + b_{j,n} k(\theta_{j,n}, y)\right] \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y)$$

$\rightarrow$ We have added a regularisation term!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)\ y\ \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \ y \ \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) (y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)\ y\ \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) (y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)(y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

where $\tilde{\Sigma}_{j,n} = \Sigma_{j,n} + (m_{j,n+1} - m_{j,n})(m_{j,n+1} - m_{j,n})^T$ and $\gamma_{j,n}$ depends on $b_{j,n}$.

$\rightarrow$ Considering all possible values of $b_{j,n}$, we have $\gamma_{j,n} \in (0, 1]$

Interpretation : tradeoff between

- an update close to $\theta_{j,n} = (m_{j,n}, \Sigma_{j,n})$ $[\gamma_{j,n} \to 0]$
- an update that chooses the Gaussian with the same mean and covariance matrix as $\check{\varphi}_{j,n}^{(\alpha)}$ $[\gamma_{j,n} = 1]$

Why does it matter? In practice, Monte Carlo approximations!

Maximisation approach for GMMs : related work

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) (y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

Consider the case $\alpha = 0$, $\gamma_{j,n} = 1$, $\eta_n = 1$, $\kappa = 0$, set $t_{j,n} = \frac{\lambda_{j,n} k(\theta_{j,n}, \cdot)}{\mu_{\boldsymbol{\lambda}_n, \Theta_n} k}$ and $\tilde{p} = p / \int p \mathrm{d}\nu$

$\rightarrow$ The M-PMC algorithm a.k.a 'Integrated EM' for GMMs

Adaptive importance sampling in general mixture classes. O. Cappé, R. Douc, A. Guillin, J-M Marin and C. P Robert (2008). Statistics and Computing, 18(4):447–459

Core insight : We have generalised an integrated EM algorithm for mixture models optimisation!

Maximisation approach for GMMs : related work

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = (1 - \gamma_{j,n}) \tilde{\Sigma}_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) (y - m_{j,n+1})(y - m_{j,n+1})^T \nu(\mathrm{d}y)$$

Consider the case $\alpha = 0$, $\gamma_{j,n} = 1$, $\eta_n = 1$, $\kappa = 0$, set $t_{j,n} = \frac{\lambda_{j,n} k(\theta_{j,n}, \cdot)}{\mu_{\boldsymbol{\lambda}_n, \Theta_n} k}$ and $\tilde{p} = p / \int p \mathrm{d}\nu$

$\rightarrow$ The M-PMC algorithm a.k.a 'Integrated EM' for GMMs

Adaptive importance sampling in general mixture classes. O. Cappé, R. Douc, A. Guillin, J-M Marin and C. P Robert (2008). Statistics and Computing, 18(4):447–459

Core insight : We have generalised an integrated EM algorithm for mixture models optimisation!

Maximisation approach for GMMs : related work

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\breve{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ \nu(\mathrm{d}y)$$

$$m_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ y\ \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ (y - m_{j,n+1})(y - m_{j,n+1})^T\ \nu(\mathrm{d}y)$$

Consider the case $\alpha = 0$, $\gamma_{j,n} = 1$, $\eta_n = 1$, $\kappa = 0$, set $t_{j,n} = \frac{\lambda_{j,n} k(\theta_{j,n}, \cdot)}{\mu_{\boldsymbol{\lambda}_n, \Theta_n} k}$ and $\tilde{p} = p / \int p \mathrm{d}\nu$

$\rightarrow$ The M-PMC algorithm a.k.a 'Integrated EM' for GMMs

Adaptive importance sampling in general mixture classes. O. Cappé, R. Douc, A. Guillin, J-M Marin and C. P Robert (2008). Statistics and Computing, 18(4):447–459

Core insight : We have generalised an integrated EM algorithm for mixture models optimisation!

Maximisation approach for GMMs : related work

Set $k(\theta, y) = \mathcal{N}(y; m, \Sigma)$ with $\theta = (m, \Sigma)$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$.

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ \nu(\mathrm{d}y)$$

$$m_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ y\ \nu(\mathrm{d}y)$$

$$\Sigma_{j,n+1} = \int_{\mathsf{Y}} t_{j,n}(y)\ \tilde{p}(y)\ (y - m_{j,n+1})(y - m_{j,n+1})^T\ \nu(\mathrm{d}y)$$

Consider the case $\alpha = 0$, $\gamma_{j,n} = 1$, $\eta_n = 1$, $\kappa = 0$, set $t_{j,n} = \frac{\lambda_{j,n} k(\theta_{j,n}, \cdot)}{\mu_{\boldsymbol{\lambda}_n, \Theta_n} k}$ and $\tilde{p} = p / \int p \mathrm{d}\nu$

$\rightarrow$ The M-PMC algorithm a.k.a 'Integrated EM' for GMMs

Adaptive importance sampling in general mixture classes. O. Cappé, R. Douc, A. Guillin, J-M Marin and C. P Robert (2008). Statistics and Computing, 18(4):447–459

Core insight : We have generalised an integrated EM algorithm for mixture models optimisation!

Outline

Gradient-based approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \ldots J$, $\gamma_{j,n} \in (0,1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \; .$$

We have that

$$\nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \frac{\partial \log k(\theta, y)}{\partial \theta} \bigg|_{(\theta,y)=(\theta_{j,n},y)} \nu(\mathrm{d}y)$$

$\rightarrow$ There might be links with Gradient Descent steps...

Gradient-based approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log\left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \dots J$, $\gamma_{j,n} \in (0,1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log\left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)}\right) \nu(\mathrm{d}y) \ .$$

We have that

$$\nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \left.\frac{\partial \log k(\theta, y)}{\partial \theta}\right|_{(\theta,y)=(\theta_{j,n},y)} \nu(\mathrm{d}y)$$

$\rightarrow$ There might be links with Gradient Descent steps...

Gradient-based approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \ldots J$, $\gamma_{j,n} \in (0, 1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \ .$$

We have that

$$\nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta, y) = (\theta_{j,n}, y)} \nu(\mathrm{d}y)$$

$\rightarrow$ There might be links with Gradient Descent steps...

Gradient-based approach

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \varphi_{j,n}^{(\alpha)}(y) \log \left(\frac{k(\theta_{j,n+1}, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \geqslant 0 \qquad \text{(Components)}$$

For all $j = 1 \dots J$, $\gamma_{j,n} \in (0, 1]$

$$\theta_{j,n+1} = \theta_{j,n} - \frac{\gamma_{j,n}}{\beta_{j,n}} \nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}}$$

where $g_{j,n}$ is assumed to be $\beta_{j,n}$-smooth on $\mathsf{T} = \mathbb{R}^d$ with

$$g_{j,n}(\theta) = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \log \left(\frac{k(\theta, y)}{k(\theta_{j,n}, y)} \right) \nu(\mathrm{d}y) \ .$$

We have that

$$\nabla g_{j,n}(\theta)|_{\theta=\theta_{j,n}} = \int_{\mathsf{Y}} \frac{\varphi_{j,n}^{(\alpha)}(y)}{\alpha - 1} \left. \frac{\partial \log k(\theta, y)}{\partial \theta} \right|_{(\theta,y)=(\theta_{j,n},y)} \nu(\mathrm{d}y)$$

$\rightarrow$ There might be links with Gradient Descent steps...

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \ y \ \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

$\rightarrow$ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1} \mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

→ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1} \mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \; y \; \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

$\rightarrow$ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1} \mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$

For all $j = 1 \ldots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y) \ y \ \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

$\rightarrow$ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1} \mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}_{j,n}^{(\alpha)} = \varphi_{j,n}^{(\alpha)} / \int \varphi_{j,n}^{(\alpha)} \mathrm{d}\nu$

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi_{\ell,n}^{(\alpha)}(y)\nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}_{j,n}^{(\alpha)}(y)\ y\ \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

$\rightarrow$ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi_{j,n}^{(\alpha)}(y)\nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

Gradient-based approach for GMMs

Set $k(\theta, y) = \mathcal{N}(y; m, \sigma^2 \boldsymbol{I}_d)$ with $\theta = m$, fixed $\sigma > 0$ and $\check{\varphi}^{(\alpha)}_{j,n} = \varphi^{(\alpha)}_{j,n} / \int \varphi^{(\alpha)}_{j,n} \mathrm{d}\nu$

For all $j = 1 \dots J$,

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\int_{\mathsf{Y}} \varphi^{(\alpha)}_{j,n}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\int_{\mathsf{Y}} \varphi^{(\alpha)}_{\ell,n}(y) \nu(\mathrm{d}y) + (\alpha - 1)\kappa \right]^{\eta_n}}$$

$$m_{j,n+1} = (1 - \gamma_{j,n}) m_{j,n} + \gamma_{j,n} \int_{\mathsf{Y}} \check{\varphi}^{(\alpha)}_{j,n}(y) \; y \; \nu(\mathrm{d}y)$$

where $\gamma_{j,n} \in (0, 1]$

→ Interpretation :

1. Maximisation and gradient-based approach coincide when $\Sigma = \sigma^2 \boldsymbol{I}_d$ with σ fixed
2. We recognise Gradient Descent steps w.r.t Θ on $-\alpha^{-1} \mathcal{L}_\alpha(\mu k; p)$ by setting

$$\gamma_{j,n} = \gamma'_{j,n} \frac{\lambda_{j,n} \int_{\mathsf{Y}} \varphi^{(\alpha)}_{j,n}(y) \nu(\mathrm{d}y)}{\int_{\mathsf{Y}} (\mu_n k(y))^\alpha p(y)^{1-\alpha} \nu(\mathrm{d}y)} \quad \text{with} \quad \gamma'_{j,n} \in (0, 1]$$

Compatibility between Gradient Descent steps w.r.t Θ and mixture weights updates (and even covariance matrices updates)!

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n\geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0,1)$ (prev. $\alpha = 0$) $\eta_n \in (0,1]$ and $(\alpha-1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1,\ \kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n\geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0,1)$ (prev. $\alpha = 0$) $\eta_n \in (0,1]$ and $(\alpha-1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1,\ \kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n\geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0,1)$ (prev. $\alpha = 0$) $\eta_n \in (0,1]$ and $(\alpha-1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1,\ \kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n\geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0,1)$ (prev. $\alpha = 0$) $\eta_n \in (0,1]$ and $(\alpha-1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1,\ \kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n\geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n\geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0,1)$ (prev. $\alpha = 0$) $\eta_n \in (0,1]$ and $(\alpha-1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1,\ \kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

At this stage

We expressed conditions on $\boldsymbol{\lambda}$ and Θ ensuring a systematic decrease in $\Psi_\alpha(\mu_{\boldsymbol{\lambda},\Theta})$:

1. Updates on $\boldsymbol{\lambda}$ linked to the gradient-based Power Descent
2. Updates on Θ :
   - Maximisation approach : generalises an Integrated EM
   - Gradient-based approach : links with Gradient Descent algorithms

	Improvements of our framework
Gradient Descent w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$	Simultaneous optimisation w.r.t $(\boldsymbol{\lambda}_n)_{n \geqslant 1}$ $\lambda_{j,n}$ needs not to be as a factor in the means updates Covariance matrices update formulas
Power Descent	Simultaneous optimisation w.r.t $(\Theta_n)_{n \geqslant 1}$ Convergence towards a local optimum of the full algorithm
M-PMC algorithm	$\alpha \in [0, 1)$ (prev. $\alpha = 0$) $\eta_n \in (0, 1]$ and $(\alpha - 1)\kappa_n \geqslant 0$ (prev. $\eta_n = 1$, $\kappa_n = 0$) $b_{j,n} \geqslant 0$ (prev. $b_{j,n} = 0$)

Outline

Monte Carlo approximations

Algorithm 1: Gaussian Mixture Models optimisation

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1 \leqslant m \leqslant M}$ from the proposal q_n.
2. For all $j = 1 \ldots J$, set:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n} \left[\sum_{m=1}^M \hat{\varphi}_{\ell,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n\right]^{\eta_n}}$$

$$(RGD) \quad m_{j,n+1} = m_{j,n} + \gamma_n \frac{\lambda_{j,n} \sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^J \sum_{m=1}^M \lambda_{j,n} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

$$(MG) \quad m_{j,n+1} = (1 - \gamma_n)\, m_{j,n} + \gamma_n \frac{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

→ Here $\hat{\varphi}_{j,n}^{(\alpha)}(y) = \frac{\varphi_{j,n}^{(\alpha)}(y)}{q_n(y)}$, $\gamma_{j,n} := \gamma_n \in (0, 1]$

→ 2 possible algorithms :

- RGD : updates derived from Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$
- MG : maximisation approach without $\lambda_{j,n}$ as a factor

→ 2 possible samplers : $q_n = \mu_{\boldsymbol{\lambda}_n, \Theta_n}$ (IS-n) and $q_n = J^{-1} \sum_{j=1}^J k(\theta_{j,n}, \cdot)$ (IS-unif).

Monte Carlo approximations

Algorithm 1: Gaussian Mixture Models optimisation

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1 \leqslant m \leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n} \left[\sum_{m=1}^M \hat{\varphi}_{\ell,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n\right]^{\eta_n}}$$

$$(RGD) \quad m_{j,n+1} = m_{j,n} + \gamma_n \frac{\lambda_{j,n} \sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^J \sum_{m=1}^M \lambda_{j,n} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

$$(MG) \quad m_{j,n+1} = (1 - \gamma_n)\, m_{j,n} + \gamma_n \frac{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

→ Here $\hat{\varphi}_{j,n}^{(\alpha)}(y) = \frac{\varphi_{j,n}^{(\alpha)}(y)}{q_n(y)}$, $\gamma_{j,n} := \gamma_n \in (0, 1]$

→ 2 possible algorithms :

- RGD : updates derived from Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$
- MG : maximisation approach without $\lambda_{j,n}$ as a factor

→ 2 possible samplers : $q_n = \mu_{\boldsymbol{\lambda}_n, \Theta_n}$ (IS-n) and $q_n = J^{-1} \sum_{j=1}^J k(\theta_{j,n}, \cdot)$ (IS-unif).

Monte Carlo approximations

Algorithm 1: Gaussian Mixture Models optimisation

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1\leqslant m\leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\sum_{m=1}^{M}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) + (\alpha-1)\kappa_n\right]^{\eta_n}}{\sum_{\ell=1}^{J}\lambda_{\ell,n}\left[\sum_{m=1}^{M}\hat{\varphi}_{\ell,n}^{(\alpha)}(Y_{m,n}) + (\alpha-1)\kappa_n\right]^{\eta_n}}$$

$$(RGD) \quad m_{j,n+1} = m_{j,n} + \gamma_n \frac{\lambda_{j,n}\sum_{m=1}^{M}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})\cdot(Y_{m,n}-\theta_{j,n})}{\sum_{j=1}^{J}\sum_{m=1}^{M}\lambda_{j,n}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

$$(MG) \quad m_{j,n+1} = (1-\gamma_n)\, m_{j,n} + \gamma_n \frac{\sum_{m=1}^{M}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})\cdot Y_{m,n}}{\sum_{m=1}^{M}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

→ Here $\hat{\varphi}_{j,n}^{(\alpha)}(y) = \frac{\varphi_{j,n}^{(\alpha)}(y)}{q_n(y)}$, $\gamma_{j,n} := \gamma_n \in (0,1]$

→ 2 possible algorithms :

- RGD : updates derived from Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$
- MG : maximisation approach without $\lambda_{j,n}$ as a factor

→ 2 possible samplers : $q_n = \mu_{\boldsymbol{\lambda}_n,\Theta_n}$ (IS-n) and $q_n = J^{-1}\sum_{j=1}^{J} k(\theta_{j,n},\cdot)$ (IS-unif).

Monte Carlo approximations

Algorithm 1: Gaussian Mixture Models optimisation

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1\leqslant m\leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\left[\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) + (\alpha-1)\kappa_n\right]^{\eta_n}}{\sum_{\ell=1}^J \lambda_{\ell,n}\left[\sum_{m=1}^M \hat{\varphi}_{\ell,n}^{(\alpha)}(Y_{m,n}) + (\alpha-1)\kappa_n\right]^{\eta_n}}$$

$$(RGD) \quad m_{j,n+1} = m_{j,n} + \gamma_n \frac{\lambda_{j,n}\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^J \sum_{m=1}^M \lambda_{j,n}\hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

$$(MG) \quad m_{j,n+1} = (1-\gamma_n)\, m_{j,n} + \gamma_n \frac{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^M \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

→ Here $\hat{\varphi}_{j,n}^{(\alpha)}(y) = \frac{\varphi_{j,n}^{(\alpha)}(y)}{q_n(y)}$, $\gamma_{j,n} := \gamma_n \in (0,1]$

→ 2 possible algorithms :

- RGD : updates derived from Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$
- MG : maximisation approach without $\lambda_{j,n}$ as a factor

→ 2 possible samplers : $q_n = \mu_{\boldsymbol{\lambda}_n,\Theta_n}$ (IS-n) and $q_n = J^{-1}\sum_{j=1}^J k(\theta_{j,n},\cdot)$ (IS-unif).

Monte Carlo approximations

Algorithm 1: Gaussian Mixture Models optimisation

At iteration n,

1. Draw independently M samples $(Y_{m,n})_{1 \leqslant m \leqslant M}$ from the proposal q_n.
2. For all $j = 1 \dots J$, set:

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \left[\sum_{m=1}^{M} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n \right]^{\eta_n}}{\sum_{\ell=1}^{J} \lambda_{\ell,n} \left[\sum_{m=1}^{M} \hat{\varphi}_{\ell,n}^{(\alpha)}(Y_{m,n}) + (\alpha - 1)\kappa_n \right]^{\eta_n}}$$

$$(RGD) \quad m_{j,n+1} = m_{j,n} + \gamma_n \frac{\lambda_{j,n} \sum_{m=1}^{M} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot (Y_{m,n} - \theta_{j,n})}{\sum_{j=1}^{J} \sum_{m=1}^{M} \lambda_{j,n} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

$$(MG) \quad m_{j,n+1} = (1 - \gamma_n)\, m_{j,n} + \gamma_n \frac{\sum_{m=1}^{M} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n}) \cdot Y_{m,n}}{\sum_{m=1}^{M} \hat{\varphi}_{j,n}^{(\alpha)}(Y_{m,n})}$$

→ Here $\hat{\varphi}_{j,n}^{(\alpha)}(y) = \frac{\varphi_{j,n}^{(\alpha)}(y)}{q_n(y)}$, $\gamma_{j,n} := \gamma_n \in (0, 1]$

→ 2 possible algorithms :

- RGD : updates derived from Gradient Descent steps w.r.t Θ on $-\alpha^{-1}\mathcal{L}_\alpha(\mu k; p)$
- MG : maximisation approach without $\lambda_{j,n}$ as a factor

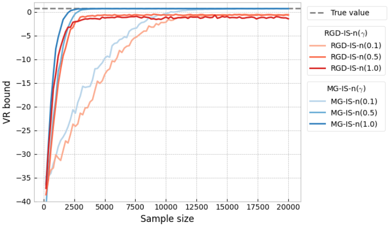
→ 2 possible samplers : $q_n = \mu_{\boldsymbol{\lambda}_n, \Theta_n}$ (IS-n) and $q_n = J^{-1} \sum_{j=1}^{J} k(\theta_{j,n}, \cdot)$ (IS-unif).

Comparing RGD to MG (fixed $\boldsymbol{\lambda}$)

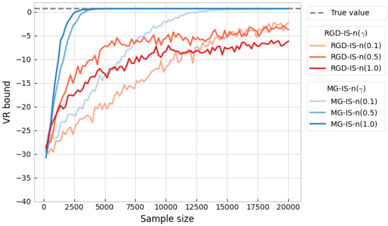
$$\text{Target :} \quad p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$$

- MC estimate of the VR Bound averaged over 30 trials for RGD and MG. [Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\kappa_n = 0$, $\eta_n = 0$. and $q_n = \mu_n k$.]

$J = 10$



$J = 50$



- LogMSE averaged over 30 trials for RGD and MG.

	$J = 10$			$J = 50$		
	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
RGD-IS-n(γ)	-0.081	-0.076	-0.218	-1.640	-1.673	-1.560
MG-IS-n(γ)	**-3.702**	**-1.875**	**-2.711**	**-2.760**	**-2.771**	**-2.788**

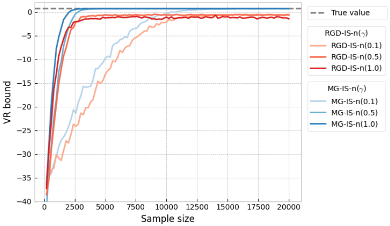
Comparing RGD to MG (fixed $\boldsymbol{\lambda}$)

$$\text{Target}: \quad p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$$

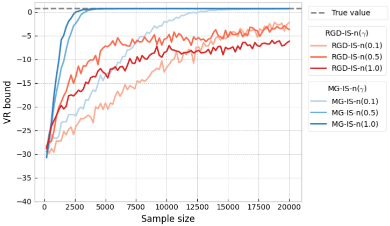
- MC estimate of the VR Bound averaged over 30 trials for RGD and MG.

[Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\kappa_n = 0$, $\eta_n = 0$. and $q_n = \mu_n k$.]

$J = 10$



$J = 50$



- LogMSE averaged over 30 trials for RGD and MG.

	$J = 10$			$J = 50$		
	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
RGD-IS-n(γ)	-0.081	-0.076	-0.218	-1.640	-1.673	-1.560
MG-IS-n(γ)	**-3.702**	**-1.875**	**-2.711**	**-2.760**	**-2.771**	**-2.788**

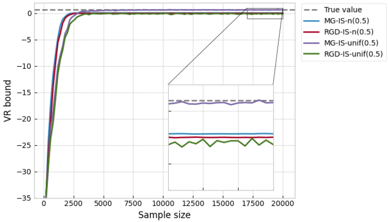
Comparing RGD to MG (varying $\boldsymbol{\lambda}$)

Target : $\quad p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$

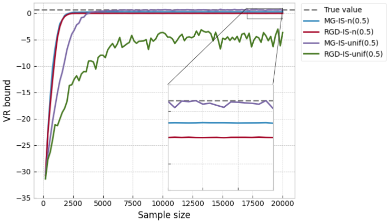
- MC estimate of the VR Bound averaged over 30 trials for RGD and MG.

[Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\eta = 0.1$, $\kappa_n = 0$.]

$J = 10$



$J = 50$



- LogMSE averaged over 30 trials for RGD and MG.

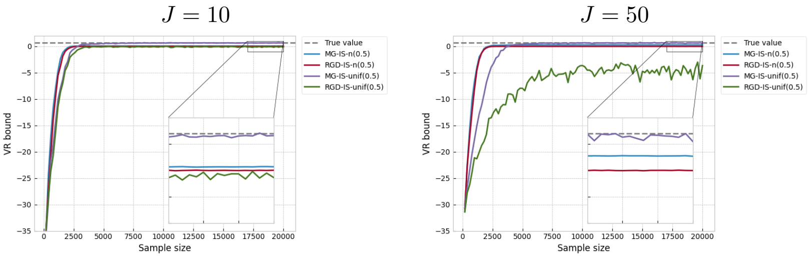
	$J = 10$			$J = 50$		
	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
RGD-IS-n(γ)	0.372	0.510	0.384	-0.616	-0.713	-0.778
MG-IS-n(γ)	1.104	1.074	0.387	1.135	-0.077	-0.060
RGD-IS-unif(γ)	0.359	0.469	0.458	-0.688	-0.670	-0.583
MG-IS-unif(γ)	**-0.200**	**-0.229**	**-0.515**	**-1.500**	**-1.462**	**-1.246**

Comparing RGD to MG (varying $\boldsymbol{\lambda}$)

Target : $\quad p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$

- MC estimate of the VR Bound averaged over 30 trials for RGD and MG.

[Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\eta = 0.1$, $\kappa_n = 0$.]



- LogMSE averaged over 30 trials for RGD and MG.

	$J=10$			$J=50$		
	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$
RGD-IS-n(γ)	0.372	0.510	0.384	-0.616	-0.713	-0.778
MG-IS-n(γ)	1.104	1.074	0.387	1.135	-0.077	-0.060
RGD-IS-unif(γ)	0.359	0.469	0.458	-0.688	-0.670	-0.583
MG-IS-unif(γ)	**-0.200**	**-0.229**	**-0.515**	**-1.500**	**-1.462**	**-1.246**

Comparing RGD to MG (varying $\boldsymbol{\lambda}$) - 2

Target : $p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$

- LogMSE averaged over 30 trials for RGD and MG.

[Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\gamma = 0.5$, $\kappa_n = 0$.]

		$J = 10$			$J = 50$	
	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.5$	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.5$
RGD-IS-n(γ)	0.045	0.510	1.299	-1.355	-0.713	0.924
MG-IS-n(γ)	0.087	1.074	1.343	-1.205	-0.077	1.329
RGD-IS-unif(γ)	-0.018	0.469	1.328	-1.385	-0.670	0.928
MG-IS-unif(γ)	**-1.244**	**-0.229**	**1.100**	**-2.524**	**-1.462**	**0.309**

Comparing RGD to MG (varying $\boldsymbol{\lambda}$) - 2

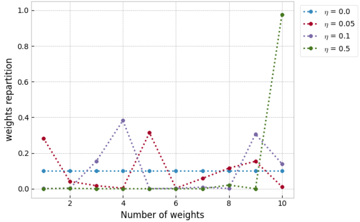
Target : $p(y) = 2 \times [0.5\mathcal{N}(\boldsymbol{y}; -2\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; 2\boldsymbol{u_d}, \boldsymbol{I_d})]$

- LogMSE averaged over 30 trials for RGD and MG.

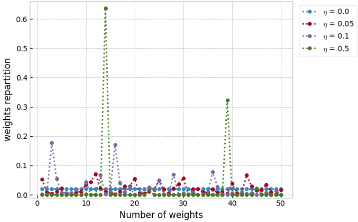
[Here, $\alpha = 0.2$, $d = 16$, $M = 200$, $\gamma = 0.5$, $\kappa_n = 0$.]

	$J = 10$			$J = 50$		
	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.5$	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.5$
RGD-IS-n(γ)	0.045	0.510	1.299	-1.355	-0.713	0.924
MG-IS-n(γ)	0.087	1.074	1.343	-1.205	-0.077	1.329
RGD-IS-unif(γ)	-0.018	0.469	1.328	-1.385	-0.670	0.928
MG-IS-unif(γ)	**-1.244**	**-0.229**	**1.100**	**-2.524**	**-1.462**	**0.309**

$J = 10$



$J = 50$



Outline

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Conclusion of Part 3

Novel framework for monotonic alpha-divergence minimisation

- applicable to mixture models optimisation
- mixture weights and mixture components parameters can be updated simultaneously
- links with an Integrated EM algorithm and with gradient-based approaches
- empirical benefits of our general framework

Some perspectives

- Additionnal convergence results
- Hyperparameters tuning
- ML applications...

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!

Overall conclusion

- **Part 1.** *General introduction to Variational Inference :*
  MFVI, BBVI, Alpha-divergence VI.

- **Part 2.** *Infinite-dimensional Alpha-divergence minimisation :*
  Mixture weights optimisation to **increase expressiveness**
  Links with the Entropic Mirror Descent algorithm

- **Part 3.** *Monotonic Alpha-divergence minimisation :*
  Mixture models optimisation with **convergence guarantees**
  Links with an Integrated EM algorithm and gradient-based approaches.

There is still a lot to do in Variational Inference regarding :

1. the expressiveness of the variational family
2. the choice of the measure of dissimilarity
3. the theory of Variational Inference
4. the interface between Variational Inference and Monte Carlo methods
5. and so much more!