Infinite-dimensional α-divergence minimisation for Variational Inference

Kamélia Daudel

Télécom Paris, Institut Polytechnique de Paris
kamelia.daudel@telecom-paris.fr

Monte Plateau
07/12/2020

Joint work with Randal Douc and François Portier

Introduction

Goal : build an iterative scheme

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \;, \qquad n \in \mathbb{N}^\star \;,$$

which extends traditional Variational Inference methods:

- it allows to move from one measure to another measure (Infinite-dimensional),
- is is such that one iteration leads to a systematic decrease of a certain criterion (α-divergence).

Introduction

Goal : build an iterative scheme

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \;, \qquad n \in \mathbb{N}^\star \;,$$

which extends traditional Variational Inference methods:

- it allows to move from one measure to another measure (Infinite-dimensional),
- is is such that one iteration leads to a systematic decrease of a certain criterion (α-divergence).

Introduction

Goal : build an iterative scheme

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \in \mathbb{N}^\star ,$$

which extends traditional Variational Inference methods:

- it allows to move from one measure to another measure (Infinite-dimensional),
- is is such that one iteration leads to a systematic decrease of a certain criterion (α-divergence).

Introduction

Goal : build an iterative scheme

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \in \mathbb{N}^\star ,$$

which extends traditional Variational Inference methods:

- it allows to move from one measure to another measure (Infinite-dimensional),
- is is such that one iteration leads to a systematic decrease of a certain criterion (α-divergence).

Outline

Outline

Variational Inference in a nutshell

- Bayesian statistics : compute / sample from the posterior density of the latent variables y given the data $\mathscr{D}$

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

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

$\rightarrow$ Variational Inference : inference is seen as an optimisation problem.

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

$$q^\star = \operatorname{arginf}_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}) \ ,$$

where D is the a divergence (e.g the Kullback-Leibler).

Variational Inference in a nutshell

- Bayesian statistics : compute / sample from the posterior density of the latent variables y given the data $\mathscr{D}$
$$p(y|\mathscr{D}) = \frac{p(\mathscr{D}, y)}{p(\mathscr{D})} \ .$$
Problem : for many important models, we can only evaluate $p(y|\mathscr{D})$ up to the constant $p(\mathscr{D})$.

$\rightarrow$ Variational Inference : inference is seen as an optimisation problem.

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

❷ Fit q to obtain the best approximation to the posterior density
$$q^\star = \mathrm{arginf}_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}) \ ,$$
where D is the a divergence (e.g the Kullback-Leibler).

Variational Inference within the α-divergence family (1)

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

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

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

where

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

Variational Inference within the α-divergence family (1)

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

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

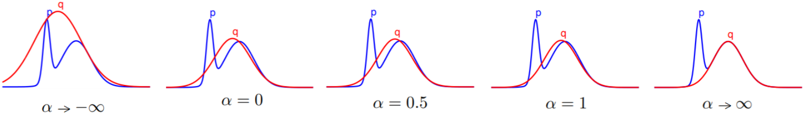
$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathscr{D})}\right) p(y|\mathscr{D})\nu(\mathrm{d}y) \ ,$$

where

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

1. A flexible family of divergences...

Figure: The Gaussian q which minimizes α-divergence to p (a mixture of two Gaussian), for varying α



[Adapted from T. Minka (2005) Divergence Measures and Message Passing. Technical Report MSR-TR-2005-173]

Variational Inference within the α-divergence family (2)

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

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

where

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

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

$\rightarrow$ We can get rid of $p(\mathscr{D})$ in the optimisation !

$$\begin{aligned} q^\star &= \operatorname{arginf}_{q\in\mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}) \\ &= \operatorname{arginf}_{q\in\mathcal{Q}} \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \quad \text{with } p(y) = p(y,\mathscr{D}) . \end{aligned}$$

Variational Inference within the α-divergence family (2)

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

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

where

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

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

$\to$ We can get rid of $p(\mathscr{D})$ in the optimisation !

$$\begin{aligned} q^\star &= \mathrm{arginf}_{q\in\mathcal{Q}} D_\alpha(\mathbb{Q}||\mathbb{P}) \\ &= \mathrm{arginf}_{q\in\mathcal{Q}} \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \quad \text{with } p(y) = p(y,\mathscr{D})\ . \end{aligned}$$

Variational Inference within the α-divergence family (3)

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

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

where

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

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

$$q^\star = \operatorname{arginf}_{q \in \mathcal{Q}} \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \quad \text{with } p(y) = p(y, \mathscr{D}) \ .$$

3. ...with good convexity properties ! $\rightarrow f_\alpha$ is convex

Variational Inference within the α-divergence family (3)

α-divergence between $\mathbb{Q}$ and $\mathbb{P}$

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

where

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

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

$$q^\star = \operatorname{arginf}_{q\in\mathcal{Q}} \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \quad \text{with } p(y) = p(y,\mathscr{D}) \ .$$

3. ...with good convexity properties ! $\rightarrow$ f_α is convex

Approximating family $\mathcal{Q}$

- Usually in Variational Inference : parametric family

$$\{y \mapsto k_\theta(y) \ : \ \theta \in \mathsf{T}\} \ .$$

- What if... we consider a broader approximating family

$$\left\{y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \ : \ \mu \in \mathsf{M}\right\} \ ,$$

M : subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on $(\mathsf{T}, \mathcal{T})$?

$\rightsquigarrow$ Mixture models : $\mu = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Approximating family $\mathcal{Q}$

- Usually in Variational Inference : parametric family

$$\{y \mapsto k_\theta(y) \; : \; \theta \in \mathsf{T}\} \; .$$

- What if... we consider a broader approximating family

$$\left\{y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \; : \; \mu \in \mathsf{M}\right\} \; ,$$

M : subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on $(\mathsf{T}, \mathcal{T})$?

$\rightsquigarrow$ Mixture models : $\mu = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Approximating family $\mathcal{Q}$

- Usually in Variational Inference : parametric family

$$\{y \mapsto k_\theta(y) \; : \; \theta \in \mathsf{T}\} \; .$$

- What if... we consider a broader approximating family

$$\left\{y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \; : \; \mu \in \mathsf{M}\right\} \; ,$$

M : subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on $(\mathsf{T}, \mathcal{T})$?

$\rightsquigarrow$ Mixture models : $\mu = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Approximating family $\mathcal{Q}$

- Usually in Variational Inference : parametric family

$$\{y \mapsto k_\theta(y) \ : \ \theta \in \mathsf{T}\} \ .$$

- What if... we consider a broader approximating family

$$\left\{y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \ : \ \mu \in \mathsf{M}\right\} \ ,$$

M : subset of $\mathrm{M}_1(\mathsf{T})$, the set of probability measures on $(\mathsf{T}, \mathcal{T})$?

$\rightsquigarrow$ Mixture models : $\mu = \sum_{j=1}^J \lambda_j \delta_{\theta_j}$.

Our approach

- Let us consider the approximating family...

$$\left\{ y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \ : \ \mu \in \mathsf{M} \right\} ,$$

- and minimise the α-divergence w.r.t μ!

Optimisation problem

- $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y)$, where $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k
- p : measurable positive function on $(\mathsf{Y}, \mathcal{Y})$

$$\operatorname{arginf}_{\mu \in \mathsf{M}} \underbrace{\int_{\mathsf{Y}} f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y)}_{:= \Psi_\alpha(\mu)}$$

Our approach

- Let us consider the approximating family...

$$\left\{ y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k_\theta(y) \ : \ \mu \in \mathsf{M} \right\} ,$$

- and minimise the α-divergence w.r.t μ!

Optimisation problem

- $\mu k(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y)$, where $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density k
- p : measurable positive function on $(\mathsf{Y}, \mathcal{Y})$

$$\boxed{\operatorname{arginf}_{\mu \in \mathsf{M}} \underbrace{\int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)}_{:= \Psi_\alpha(\mu)}}$$

Outline

The (α, Γ)-descent

Optimisation problem

$$\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi_\alpha(\mu) \quad \text{with} \quad \Psi_\alpha(\mu) := \int_{\mathsf{Y}} f_\alpha \left(\frac{\mu k(y)}{p(y)} \right) p(y) \nu(\mathrm{d}y)$$

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi_\alpha(\mu_1) < \infty$. We define the sequence of probability measures $(\mu_n)_{n \in \mathbb{N}^\star}$ iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \in \mathbb{N}^\star \,. \tag{1}$$

Algorithm 1: *Exact (α, Γ)-descent one-step transition*

1. Expectation step : $b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha \left(\frac{\mu k(y)}{p(y)} \right) \nu(\mathrm{d}y)$
2. Iteration step : $\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \dfrac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$

Monotonicity

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ 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 1

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then, the two following assertions hold.

1. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$.
2. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Monotonicity

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ 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 1

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathrm{T})$ be such that $\Psi_\alpha(\mu) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then, the two following assertions hold.

1. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$.
2. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Monotonicity

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ 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 1

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then, the two following assertions hold.

1. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$.
2. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Monotonicity

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ 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 1

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then, the two following assertions hold.

1. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$. ← proof later !
2. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ 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

$$[(\alpha - 1)(v - \kappa) + 1] (\log \Gamma)'(v) + 1 \geqslant 0 \ .$$

1. Entropic MD : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v} \ .$$

2. Power descent : $\eta \in (0, 1]$, $(\alpha - 1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha - 1) v + 1]^{\frac{\eta}{1-\alpha}} \ .$$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$\left[(\alpha - 1)(v - \kappa) + 1\right] (\log \Gamma)'(v) + 1 \geqslant 0 \;.$$

❶ Entropic MD : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v} \;.$$

❷ Power descent : $\eta \in (0, 1]$, $(\alpha - 1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = \left[(\alpha - 1)\, v + 1\right]^{\frac{\eta}{1-\alpha}} \;.$$

Examples satisfying (A2)

(A2) The function $\Gamma : \mathrm{Dom}_\alpha \to \mathbb{R}_{>0}$ is decreasing, continuously differentiable and satisfies the inequality

$$[(\alpha - 1)(v - \kappa) + 1]\,(\log \Gamma)'(v) + 1 \geqslant 0 \,.$$

❶ Entropic MD : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

$$\Gamma(v) = e^{-\eta v} \,.$$

❷ Power descent : $\eta \in (0, 1]$, $(\alpha - 1)\kappa \geqslant 0$ and $\alpha \neq 1$

$$\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}} \,.$$

Limiting behavior

Table 1: Examples of allowed (Γ, κ) in the (α, Γ)-descent

Divergence considered	Possible choice of (Γ, κ)	
Forward KL $(\alpha = 1)$	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1)$	any κ
α-*divergence* with $\alpha \in \mathbb{R} \setminus \{1\}$	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{\lvert\alpha-1\rvert \lvert b\rvert_{\infty,\alpha}+1})$	any κ
	$\alpha > 1$, $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa > 0$
	$\alpha < 1$, $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa \leqslant 0$

→ Convergence towards the optimum value at a $O(1/N)$ rate

→ Convergence towards the optimum value

Limiting behavior

Table 1: Examples of allowed (Γ, κ) in the (α, Γ)-descent

Divergence considered	Possible choice of (Γ, κ)	
Forward KL $(\alpha = 1)$	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, 1)$	any κ
α-*divergence* with $\alpha \in \mathbb{R} \setminus \{1\}$	$\Gamma(v) = e^{-\eta v}$, $\eta \in (0, \frac{1}{\lvert\alpha-1\rvert \lvert b\rvert_{\infty,\alpha}+1})$	any κ
	$\alpha > 1$, $\Gamma(v) = [(\alpha-1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa > 0$
	$\alpha < 1$, $\Gamma(v) = [(\alpha-1)\, v + 1]^{\frac{\eta}{1-\alpha}}$, $\eta \in (0, 1]$	$\kappa \leqslant 0$

$\rightarrow$ Convergence towards the optimum value at a $O(1/N)$ rate

$\rightarrow$ Convergence towards the optimum value

Mixture models and (α, Γ)-descent

$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \;:\; \forall j \in \{1, ..., J\}\,,\; \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$.
Let $\theta_1, ..., \theta_J \in \mathsf{T}$ be fixed and denote

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{with} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \;.$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\begin{cases} \boldsymbol{\lambda}_1 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)} \end{cases} . \tag{2}$$

- In practice, we will use

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

  with $Y_{1,n}, ..., Y_{M,n}$ drawn independently from $\mu_n k$.
- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)

Mixture models and (α, Γ)-descent

$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\}\,,\ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$.
Let $\theta_1, ..., \theta_J \in \mathsf{T}$ be fixed and denote

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{with} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\begin{cases} \boldsymbol{\lambda}_1 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)} \end{cases} . \tag{2}$$

- In practice, we will use

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^M \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

  with $Y_{1,n}, ..., Y_{M,n}$ drawn independently from $\mu_n k$.
- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)

Mixture models and (α, Γ)-descent

$S_J = \left\{ \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J) \in \mathbb{R}^J \ : \ \forall j \in \{1, ..., J\} \, , \ \lambda_j \geqslant 0 \text{ and } \sum_{j=1}^J \lambda_j = 1 \right\}$.

Let $\theta_1, ..., \theta_J \in \mathsf{T}$ be fixed and denote

$$\mu_{\boldsymbol{\lambda}} = \sum_{j=1}^J \lambda_j \delta_{\theta_j} \quad \text{with} \quad \boldsymbol{\lambda} \in \mathcal{S}_J \ .$$

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\begin{cases} \boldsymbol{\lambda}_1 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)} \end{cases} . \tag{2}$$

- In practice, we will use

$$\hat{b}_{\mu_n,\alpha,M}(\theta_j) = \frac{1}{M} \sum_{m=1}^{M} \frac{k(\theta_j, Y_{m,n})}{\mu_n k(Y_{m,n})} f'_\alpha \left(\frac{\mu_n k(Y_{m,n})}{p(Y_{m,n})} \right),$$

  with $Y_{1,n}, ..., Y_{M,n}$ drawn independently from $\mu_n k$.
- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)

Outline

Numerical Experiments

- Framework

Kernel: Gaussian transition kernel k_h with bandwidth h.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, (\theta_j)_{1 \leqslant j \leqslant J} \in \mathsf{T}^J \right\} .$$

At time t,

1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Stochastic Mixture (α, Γ)-descent.
2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$, with $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

- Toy example

$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$, $s = 2$

- Bayesian Logistic Regression

Covertype dataset ($581,012$ data points and 54 features)

Numerical Experiments

- Framework

Kernel: Gaussian transition kernel k_h with bandwidth h.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, (\theta_j)_{1 \leqslant j \leqslant J} \in \mathsf{T}^J \right\} .$$

At time t,

1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Stochastic Mixture (α, Γ)-descent.
2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$, with $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

- Toy example

$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$, $s = 2$

- Bayesian Logistic Regression

Covertype dataset ($581,012$ data points and 54 features)

Numerical Experiments

- Framework

Kernel: Gaussian transition kernel k_h with bandwidth h.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, (\theta_j)_{1 \leqslant j \leqslant J} \in \mathsf{T}^J \right\} .$$

At time t,

1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Stochastic Mixture (α, Γ)-descent.
2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$, with $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

- Toy example

$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$, $s = 2$

- Bayesian Logistic Regression

Covertype dataset ($581,012$ data points and 54 features)

Numerical Experiments

- Framework

Kernel: Gaussian transition kernel k_h with bandwidth h.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, (\theta_j)_{1 \leqslant j \leqslant J} \in \mathsf{T}^J \right\} .$$

At time t,

1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Stochastic Mixture (α, Γ)-descent.
2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$, with $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

- Toy example

$p(y) = Z \times [0.5\mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5\mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$, $s = 2$

- Bayesian Logistic Regression

Covertype dataset ($581,012$ data points and 54 features)

Numerical Experiments

- Framework

Kernel: Gaussian transition kernel k_h with bandwidth h.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_h(y) = \sum_{j=1}^{J} \lambda_j k_h(y - \theta_j) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, (\theta_j)_{1 \leqslant j \leqslant J} \in \mathsf{T}^J \right\} .$$

At time t,

1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Stochastic Mixture (α, Γ)-descent.
2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$, with $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

- Toy example

$p(y) = Z \times [0.5 \mathcal{N}(\boldsymbol{y}; -s\boldsymbol{u_d}, \boldsymbol{I_d}) + 0.5 \mathcal{N}(\boldsymbol{y}; s\boldsymbol{u_d}, \boldsymbol{I_d})]$, $Z = 2$, $s = 2$

- Bayesian Logistic Regression

Covertype dataset ($581,012$ data points and 54 features)

Toy Example : Mirror Descent vs Power Descent

We compare :

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial weights: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

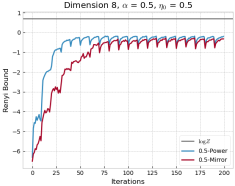
Toy Example : Mirror Descent vs Power Descent

We compare :

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial weights: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

Figure: Average Renyi-Bound for the 0.5-Power and 0.5-Mirror descent computed over 100 replicates with $\eta_0 = 0.5$.



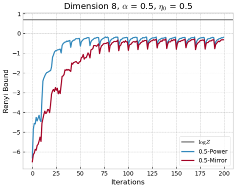
Toy Example : Mirror Descent vs Power Descent

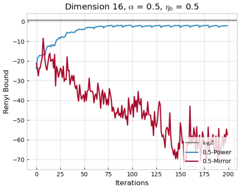
We compare :

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial weights: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

Figure: Average Renyi-Bound for the 0.5-Power and 0.5-Mirror descent computed over 100 replicates with $\eta_0 = 0.5$.





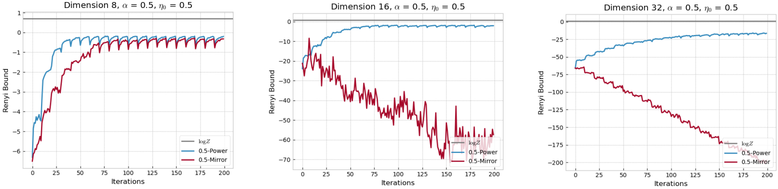
Toy Example : Mirror Descent vs Power Descent

We compare :

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial weights: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

Figure: Average Renyi-Bound for the 0.5-Power and 0.5-Mirror descent computed over 100 replicates with $\eta_0 = 0.5$.



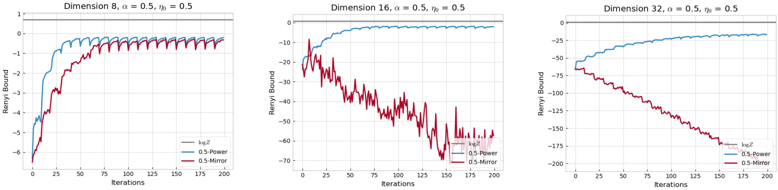
Toy Example : Mirror Descent vs Power Descent

We compare :

- <u>0.5-Mirror descent :</u> $\Gamma(v) = e^{-\eta v}$ with $\alpha = 0.5$,
- <u>0.5-Power descent :</u> $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

$J = M = 100$, initial weights: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

Figure: Average Renyi-Bound for the 0.5-Power and 0.5-Mirror descent computed over 100 replicates with $\eta_0 = 0.5$.



Mirror $\quad \lambda_{j,n} \propto \exp\left(\frac{\eta}{1-\alpha}\left((\alpha-1) b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j) + (\alpha-1)\kappa\right)\right)$

Power $\quad \lambda_{j,n} \propto \exp\left(\frac{\eta}{1-\alpha} \mathbf{log}\left((\alpha-1) b_{\mu_{\boldsymbol{\lambda}_n},\alpha}(\theta_j) + (\alpha-1)\kappa\right)\right)$.

Toy Example : $\alpha = 1$

We compare:

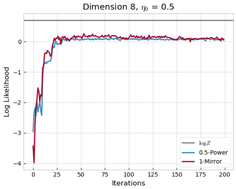
- 1-Mirror descent : $\Gamma(v) = e^{-\eta v}$ with $\alpha = 1$,
- 0.5-Power descent : $\Gamma(v) = [(\alpha - 1)\, v + 1]^{\eta/(1-\alpha)}$ with $\alpha = 0.5$.

Toy Example : $\alpha = 1$

We compare:

- <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 $\alpha = 0.5$.

Figure: Plotted is the average Log-likelihood for 0.5-Power and 1-Mirror descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.

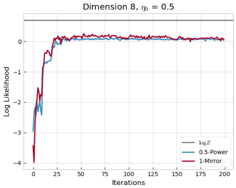


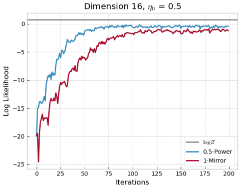
Toy Example : $\alpha = 1$

We compare:

- 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 $\alpha = 0.5$.

Figure: Plotted is the average Log-likelihood for 0.5-Power and 1-Mirror descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



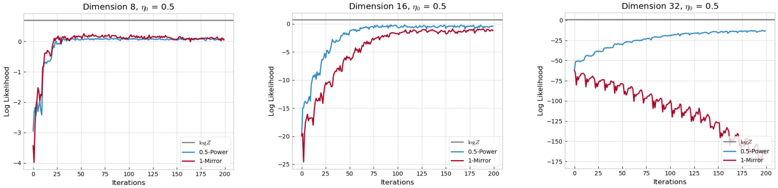


Toy Example : $\alpha = 1$

We compare:

- <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 $\alpha = 0.5$.

Figure: Plotted is the average Log-likelihood for 0.5-Power and 1-Mirror descent in dimension $d = \{8, 16, 32\}$ computed over 100 replicates with $\eta_0 = 0.5$.



Outline

Take-away message

The (α, Γ)-descent

- performs an update of probability measures
  - sufficient conditions on (α, Γ) leading to a systematic decrease
  - includes Entropic Mirror Descent
  - convergence to an optimum and $O(1/N)$ convergence rates,
- can be applied to density approximation
  - handles the case of Mixture Models for any kernel K
  - requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$
  - empirical benefit of using the Power descent.

[Kamélia Daudel, Randal Douc and François Portier (2020). Infinite-dimensional gradient-based descent for alpha-divergence minimisation. To be published in the Annals of Statistics. https://arxiv.org/abs/2005.10618]

Take-away message

The (α, Γ)-descent

- performs an update of probability measures
  - sufficient conditions on (α, Γ) leading to a systematic decrease
  - includes Entropic Mirror Descent
  - convergence to an optimum and $O(1/N)$ convergence rates,
- can be applied to density approximation
  - handles the case of Mixture Models for any kernel K
  - requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$
  - empirical benefit of using the Power descent.

[Kamélia Daudel, Randal Douc and François Portier (2020). Infinite-dimensional gradient-based descent for alpha-divergence minimisation. To be published in the Annals of Statistics. https://arxiv.org/abs/2005.10618]

Take-away message

The (α, Γ)-descent

- performs an update of probability measures
  - sufficient conditions on (α, Γ) leading to a systematic decrease
  - includes Entropic Mirror Descent
  - convergence to an optimum and $O(1/N)$ convergence rates,
- can be applied to density approximation
  - handles the case of Mixture Models for any kernel K
  - requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$
  - empirical benefit of using the Power descent.

[Kamélia Daudel, Randal Douc and François Portier (2020). Infinite-dimensional gradient-based descent for alpha-divergence minimisation. To be published in the Annals of Statistics. https://arxiv.org/abs/2005.10618]

Outline

The result we want to prove

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) > 0$ 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 1

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu) < \infty$ and $\mu(\Gamma(b_{\mu,\alpha} + \kappa)) < \infty$. Then, the two following assertions hold.

1. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \leqslant \Psi_\alpha(\mu)$.
2. We have $\Psi_\alpha \circ \mathcal{I}_\alpha(\mu) = \Psi_\alpha(\mu)$ if and only if $\mu = \mathcal{I}_\alpha(\mu)$.

Recall that :

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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$$

Step 1 : Proving a general lower bound (1)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ 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!

Step 1 : Proving a general lower bound (1)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ 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!

Step 1 : Proving a general lower bound (1)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ 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!

Step 1 : Proving a general lower bound (1)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ 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!

Step 1 : Proving a general lower bound (2)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

<u>Second idea</u>

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)]\ .$$

$\rightarrow$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Step 1 : Proving a general lower bound (2)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

<u>Second idea</u>

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \; .$$

$\rightarrow$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Step 1 : Proving a general lower bound (2)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

Second idea

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \ .$$

$\to$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Step 1 : Proving a general lower bound (2)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

Second idea

By convexity of f_α,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{\mu k(y)}{p(y)}[1 - g(\theta)] \ .$$

$\to$ Next, we integrate w.r.t to $\frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)}$,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant \int_{\mathsf{T}} \frac{\mu(\mathrm{d}\theta)k(\theta,y)}{\mu k(y)} f_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

$$\geqslant f_\alpha\left(\frac{\int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)g(\theta)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1 - g(\theta)]$$

Step 1 : Proving a general lower bound (3)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)\frac{1}{p(y)}[1-g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y)p(y)$, we deduce

$$\Psi_\alpha(\mu) \geqslant \Psi_\alpha(\zeta) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)]$$

Choice of A_α

We take $A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y)f_\alpha'\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1-g(\theta)]$

Step 1 : Proving a general lower bound (3)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) \frac{1}{p(y)}[1 - g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y)p(y)$, we deduce

$$\Psi_\alpha(\mu) \geqslant \Psi_\alpha(\zeta) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1 - g(\theta)]$$

Choice of A_α

We take $\quad A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)[1 - g(\theta)]$

Step 1 : Proving a general lower bound (3)

Let $\mu, \zeta \in \mathrm{M}_1(\mathsf{T})$ s.t $\zeta \preceq \mu$ and $\Psi_\alpha(\mu) < \infty$. Denote by g the density of ζ w.r.t μ.

We want to find A_α such that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha(\zeta)$$

and equality holds iif $\zeta = \mu$.

By definition $\Psi_\alpha(\mu) = \int_{\mathsf{Y}} f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$ with f_α convex.

$\rightarrow$ At this stage,

$$f_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \geqslant f_\alpha\left(\frac{\zeta k(y)}{p(y)}\right) + \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) \frac{1}{p(y)}[1 - g(\theta)]$$

Now integrating w.r.t to $\nu(\mathrm{d}y)p(y)$, we deduce

$$\Psi_\alpha(\mu) \geqslant \Psi_\alpha(\zeta) + \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$$

Choice of A_α

We take $\quad A_\alpha := \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (1)

Recall that :

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$

$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$$

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu)$$

where $A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (1)

Recall that :
$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$
$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$$

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that
$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu)$$
where $A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$
$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (1)

Recall that :
$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) f'_\alpha\left(\frac{\mu k(y)}{p(y)}\right) \nu(\mathrm{d}y)$$
$$\mathcal{I}_\alpha(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot \Gamma(b_{\mu,\alpha}(\theta) + \kappa)}{\mu(\Gamma(b_{\mu,\alpha} + \kappa))}$$

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that
$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu)$$
where $A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) f'_\alpha\left(\frac{g(\theta)\mu k(y)}{p(y)}\right) [1 - g(\theta)]$.

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] \nu(\mathrm{d}y)$$
$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1}\left[\left(\frac{g(\theta)\mu k(y)}{p(y)}\right)^{\alpha-1} - 1\right] [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (2)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta,y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta,y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta,y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} g(\theta)^{\alpha-1} - 1 \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (2)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove 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

$$b_{\mu,\alpha}(\theta) = \int_\mathsf{Y} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_\mathsf{Y} \nu(\mathrm{d}y) \int_\mathsf{T} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_\mathsf{T} \mu(\mathrm{d}\theta) \left(\int_\mathsf{Y} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} g(\theta)^{\alpha-1} - 1 \right) [1 - g(\theta)]$$

$$= \int_\mathsf{T} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (2)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

→ 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta) \mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} g(\theta)^{\alpha-1} - 1 \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (2)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu)\ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

→ 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$A_\alpha = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} g(\theta)^{\alpha-1} - 1 \right) [1 - g(\theta)]$$

$$= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (2)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove 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

$$b_{\mu,\alpha}(\theta) = \int_{\mathsf{Y}} k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] \nu(\mathrm{d}y)$$

$$\begin{aligned}
A_\alpha &= \int_{\mathsf{Y}} \nu(\mathrm{d}y) \int_{\mathsf{T}} \mu(\mathrm{d}\theta) k(\theta, y) \frac{1}{\alpha - 1} \left[\left(\frac{g(\theta)\mu k(y)}{p(y)} \right)^{\alpha-1} - 1 \right] [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left(\int_{\mathsf{Y}} \nu(\mathrm{d}y) k(\theta, y) \frac{1}{\alpha - 1} \left(\frac{\mu k(y)}{p(y)} \right)^{\alpha-1} g(\theta)^{\alpha-1} - 1 \right) [1 - g(\theta)] \\
&= \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} [1 - g(\theta)]
\end{aligned}$$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that
$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \, .$$
The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,
$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! $\square$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! $\square$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \; .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! $\square$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! □

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \; .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta) \right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! $\square$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$!

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$!

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

$\rightarrow$ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! $\square$

Step 2 : take $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$ and show that $A_\alpha \geqslant 0$ (3)

For $g \propto \Gamma(b_{\mu,\alpha} + \kappa)$, we have that

$$A_\alpha \leqslant \Psi_\alpha(\mu) - \Psi_\alpha \circ \mathcal{I}_\alpha(\mu) \ .$$

The proof is complete if we prove that $A_\alpha \geqslant 0$.

→ At this stage,

$$A_\alpha = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) \left[b_{\mu,\alpha}(\theta) + \frac{1}{\alpha - 1} \right] g(\theta)^{\alpha-1} \left[1 - g(\theta)\right]$$

On the probability space $(\mathsf{T}, \mathcal{T}, \mu)$, we let V be the random variable $V(\theta) = b_{\mu,\alpha}(\theta) + \kappa$. Set $\tilde{\Gamma}(v) = \Gamma(v)/\mu(\Gamma(b_{\mu,\alpha} + \kappa))$ for all $v \in \mathrm{Dom}_\alpha$. Then, $\mathbb{E}[1 - \tilde{\Gamma}(V)] = 0$ and

$$\begin{aligned} A_\alpha &= \mathbb{E}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V) \left[1 - \tilde{\Gamma}(V) \right] \right) \\ &= \mathbb{C}\mathrm{ov}\left(\left[V - \kappa + \frac{1}{\alpha - 1} \right] \tilde{\Gamma}^{\alpha-1}(V), 1 - \tilde{\Gamma}(V) \right) \end{aligned}$$

Time to recall (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: $A_\alpha \geqslant 0$! □

Thank you !

An $O(1/N)$ convergence rate?

$\rightarrow$ Usually requires strong assumptions e.g β-smoothness assumption for the Entropic MD.

$\rightarrow$ We are able to quantify the improvement in terms of the variance of $b_{\mu,\alpha}$

$$\frac{L_{\alpha,1}}{2}\mathbb{V}\mathrm{ar}_{\mu}\left(b_{\mu,\alpha}\right) \leqslant \Psi_{\alpha}(\mu) - \Psi_{\alpha} \circ \mathcal{I}_{\alpha}(\mu) \; ,$$

where $L_{\alpha,1} := \inf_{v \in \mathrm{Dom}_{\alpha}} \left\{ \left[(\alpha-1)(v-\kappa)+1\right](\log\Gamma)'(v)+1 \right\} \times \inf_{v \in \mathrm{Dom}_{\alpha}} -\Gamma'(v)$.

$\rightsquigarrow$ Projected Gradient descent : f is β-smooth on $\mathbb{R}$

$$\forall u \in \mathbb{R}, \quad \frac{1}{\beta}\|\nabla f(u)\|^2 \leqslant f(u) - f\left(u - \frac{1}{\beta}\nabla f(u)\right) \; .$$