Infinite-dimensional α-divergence minimisation for Variational Inference

Kamélia Daudel

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

MCQMC2020

Joint work with Randal Douc, François Portier and François Roueff

Introduction

Goal : build an iterative scheme

- which allows to move from one measure to another measure (Infinite-dimensional)

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

- such that one iteration leads to a systematic decrease of a certain criterion (α-divergence),
- applicable to Bayesian Inference problems (Variational Inference).

Introduction

Goal : build an iterative scheme

- which allows to move from one measure to another measure (Infinite-dimensional)

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

- such that one iteration leads to a systematic decrease of a certain criterion (α-divergence),
- applicable to Bayesian Inference problems (Variational Inference).

Introduction

Goal : build an iterative scheme

- which allows to move from one measure to another measure (Infinite-dimensional)

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

- such that one iteration leads to a systematic decrease of a certain criterion (α-divergence),
- applicable to Bayesian Inference problems (Variational Inference).

Introduction

Goal : build an iterative scheme

- which allows to move from one measure to another measure (Infinite-dimensional)

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

- such that one iteration leads to a systematic decrease of a certain criterion (α-divergence),
- applicable to Bayesian Inference problems (Variational Inference).

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 the marginal likelihood $p(\mathscr{D})$ is untractable.

→ 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

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

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 the marginal likelihood $p(\mathscr{D})$ is untractable.

$\rightarrow$ 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

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

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

Variational Inference within the α-divergence family

$(\mathsf{Y}, \mathcal{Y}, \nu)$: 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

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

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

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

$\rightarrow$ f_α is convex !

Variational Inference within the α-divergence family

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

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

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

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

$\rightarrow$ f_α is convex !

Variational Inference within the α-divergence family

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

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

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

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

$\rightarrow$ f_α is convex !

Our approach

- Usually in Variational Inference : parametric family

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

- Let us 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 $(\mathsf{T}, \mathcal{T})$.
$\rightsquigarrow$ Example : Mixture models $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$.

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 $(\mathsf{Y}, \mathcal{Y})$

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

Our approach

- Usually in Variational Inference : parametric family

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

- Let us 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 $(\mathsf{T}, \mathcal{T})$.
$\rightsquigarrow$ Example : Mixture models $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$.

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 $(\mathsf{Y}, \mathcal{Y})$

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

Our approach

- Usually in Variational Inference : parametric family

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

- Let us 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 $(\mathsf{T}, \mathcal{T})$.
  $\rightsquigarrow$ Example : Mixture models $\mu = \sum_{j=1}^{J} \lambda_j \delta_{\theta_j}$.

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 $(\mathsf{Y}, \mathcal{Y})$

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

Outline

The (α, Γ)-descent

Optimisation problem

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

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*

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

❷ 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 $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\mu = \mathcal{I}_\alpha(\mu)$.

Examples satisfying (A2)

(A2) The function $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\Gamma : \Delta_\alpha \to \mathbb{R}_{>0}$ 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 $\alpha \neq 1$

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

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

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

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

✗ β-smooth assumption not satisfied in general for the α-divergence (e.g Gaussian family).

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

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

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

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

✗ β-smooth assumption not satisfied in general for the α-divergence (e.g Gaussian family).

✓ 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 $L_{\alpha,1} := \inf_{v\in\Delta_\alpha} \{[(\alpha-1)(v-\kappa)+1]\,(\log\Gamma)'(v)+1\} \times \inf_{v\in\Delta_\alpha} -\Gamma'(v)$.

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

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

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

✗ β-smooth assumption not satisfied in general for the α-divergence (e.g Gaussian family).

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

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

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

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

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

- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)
- 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 $\mu_n k$.

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

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

- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)
- 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 $\mu_n k$.

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

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

- Exploitation step which requires no information on the distribution of $\{\theta_1, ..., \theta_J\}$ (as opposed to Importance Sampling)
- 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 $\mu_n k$.

Outline

Numerical Experiments

- Framework

  Kernel: Gaussian transition kernel k_{h_t} with bandwidth h_t.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_{h_t}(y) = \sum_{j=1}^{J_t} \lambda_j k_{h_t}(y - \theta_{j,t}) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_{J_t}, (\theta_{j,t})_{1 \leqslant j \leqslant J_t} \in \mathsf{T}^{J_t} \right\} .$$

  $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

  1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Mixture (α, Γ)-descent (with Monte Carlo approximation of $b_{\mu,\alpha}$).
  2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$.

- Toy example

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

- Bayesian Logistic Regression

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

Numerical Experiments

- Framework

  Kernel: Gaussian transition kernel k_{h_t} with bandwidth h_t.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_{h_t}(y) = \sum_{j=1}^{J_t} \lambda_j k_{h_t}(y - \theta_{j,t}) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_{J_t}, (\theta_{j,t})_{1 \leqslant j \leqslant J_t} \in \mathsf{T}^{J_t} \right\} .$$

  $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

  1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Mixture (α, Γ)-descent (with Monte Carlo approximation of $b_{\mu,\alpha}$).
  2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$.

- Toy example

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

- Bayesian Logistic Regression

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

Numerical Experiments

- Framework

  Kernel: Gaussian transition kernel k_{h_t} with bandwidth h_t.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_{h_t}(y) = \sum_{j=1}^{J_t} \lambda_j k_{h_t}(y - \theta_{j,t}) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_{J_t}, (\theta_{j,t})_{1 \leqslant j \leqslant J_t} \in \mathsf{T}^{J_t} \right\} .$$

  $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

  1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Mixture (α, Γ)-descent (with Monte Carlo approximation of $b_{\mu,\alpha}$).
  2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$.

- Toy example

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

- Bayesian Logistic Regression

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

Numerical Experiments

- Framework

  Kernel: Gaussian transition kernel k_{h_t} with bandwidth h_t.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_{h_t}(y) = \sum_{j=1}^{J_t} \lambda_j k_{h_t}(y - \theta_{j,t}) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_{J_t}, (\theta_{j,t})_{1 \leqslant j \leqslant J_t} \in \mathsf{T}^{J_t} \right\} .$$

  $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

  1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Mixture (α, Γ)-descent (with Monte Carlo approximation of $b_{\mu,\alpha}$).
  2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$.

- Toy example

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

- Bayesian Logistic Regression

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

Numerical Experiments

- Framework

  Kernel: Gaussian transition kernel k_{h_t} with bandwidth h_t.

$$\left\{ y \mapsto \mu_{\boldsymbol{\lambda}} k_{h_t}(y) = \sum_{j=1}^{J_t} \lambda_j k_{h_t}(y - \theta_{j,t}) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_{J_t}, (\theta_{j,t})_{1 \leqslant j \leqslant J_t} \in \mathsf{T}^{J_t} \right\} .$$

  $h_t \propto J_t^{-1/(4+d)}$, where d is the dimension of the latent space.

  1. Exploitation step Optimise $\boldsymbol{\lambda}$ using the Mixture (α, Γ)-descent (with Monte Carlo approximation of $b_{\mu,\alpha}$).
  2. Exploration step Sample $(\theta_{j,t+1})_{1 \leqslant j \leqslant J_{t+1}}$ according to $\mu_{\boldsymbol{\lambda}} k_{h_t}$.

- Toy example

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

- Bayesian Logistic Regression

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

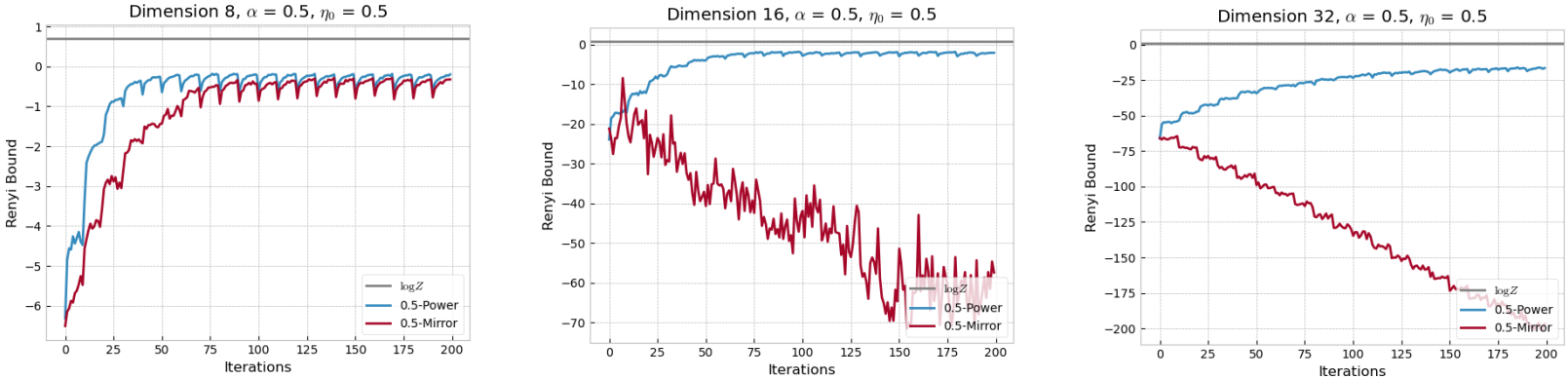
Toy Example: Mirror Descent vs Power Descent

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: $[1/J, ..., 1/J]$, $N = 10$, $T = 20$.

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

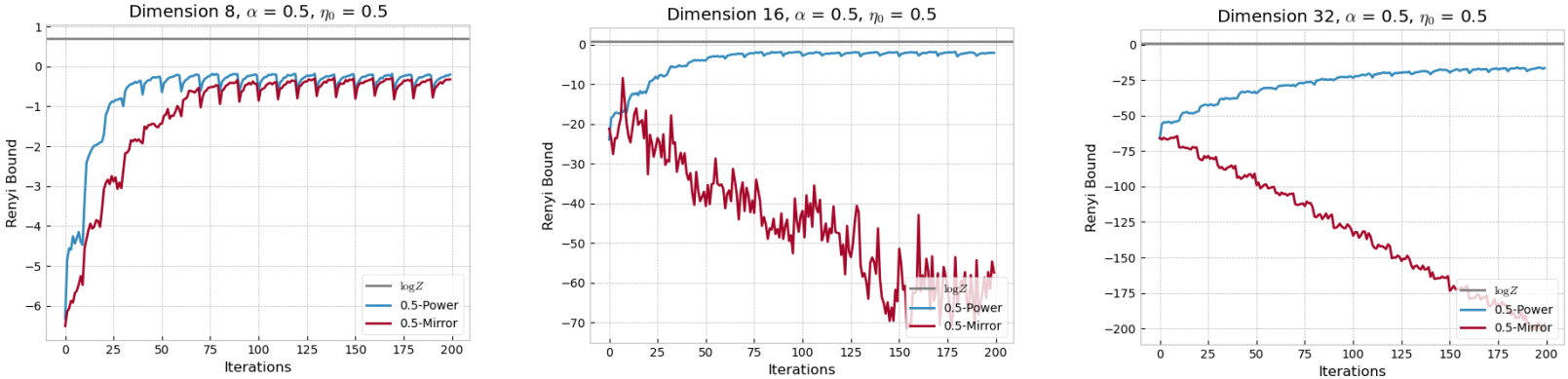
Toy Example: Mirror Descent vs Power Descent

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

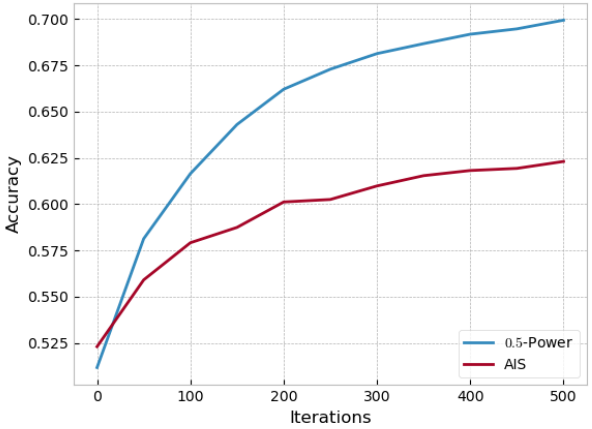
Bayesian Logistic Regression

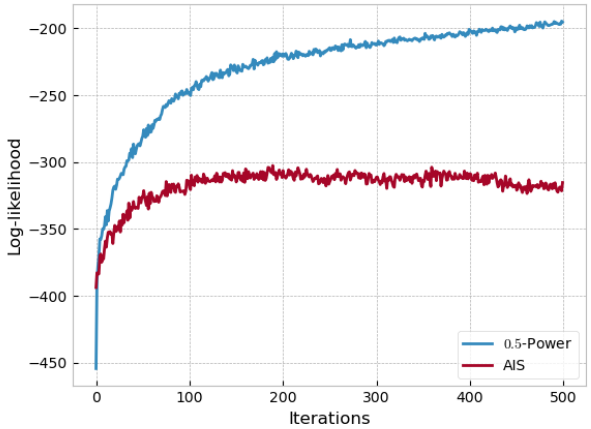
We compare :

- 0.5-Power descent
- AIS (Adaptive Importance Sampling)

$N = 1$, $T = 500$, $J_0 = M_0 = 20$
$J_{t+1} = M_{t+1} = J_t + 1$, initial weights: $[1/J_t, ..., 1/J_t]$.

Figure: Average Accuracy and Log-likelihood computed over 100 replicates for the 0.5-Power descent ($\eta_0 = 0.05$) and the AIS algorithm.





Outline

Conclusion

The (α, Γ)-descent

- performs an update of 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.

[1] Kamélia Daudel, Randal Douc and François Portier (2020).
Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
https://hal.telecom-paris.fr/hal-02614605.

[2] Kamélia Daudel, Randal Douc, François Portier and François Roueff (2019).
The f-Divergence Expectation Iteration Scheme. https://arxiv.org/abs/1909.12239.

Conclusion

The (α, Γ)-descent

- performs an update of 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.

[1] Kamélia Daudel, Randal Douc and François Portier (2020).
Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
https://hal.telecom-paris.fr/hal-02614605.
[2] Kamélia Daudel, Randal Douc, François Portier and François Roueff (2019).
The f-Divergence Expectation Iteration Scheme. https://arxiv.org/abs/1909.12239.

Conclusion

The (α, Γ)-descent

- performs an update of 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.

[1] Kamélia Daudel, Randal Douc and François Portier (2020).
Infinite-dimensional gradient-based descent for alpha-divergence minimisation.
https://hal.telecom-paris.fr/hal-02614605.

[2] Kamélia Daudel, Randal Douc, François Portier and François Roueff (2019).
The f-Divergence Expectation Iteration Scheme. https://arxiv.org/abs/1909.12239.