Infinite-dimensional α-divergence minimisation for Variational Inference

Kamélia Daudel

University of Oxford
kamelia.daudel@stats.ox.ac.uk

Séminaire MIA-Paris-Saclay
11/04/2021

Joint work with Randal Douc and François Portier

Outline

Outline

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 $p(\mathscr{D})$.

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 $p(\mathscr{D})$.

Variational Inference in a nutshell

→ Variational Inference : inference is seen as an optimisation problem.

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

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

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

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

→ Typically, D : exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: parametric family (e.g. Mean-field)

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \end{cases}$$

Variational Inference in a nutshell

→ Variational Inference : inference is seen as an optimisation problem.

1. Posit a *simpler* variational family $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density
$$\inf_{q \in \mathcal{Q}} D(\mathbb{Q}||\mathbb{P}_{|\mathscr{D}}) \ ,$$
where D is a measure of dissimilarity between the variational distribution $\mathbb{Q}$ and the posterior distribution $\mathbb{P}_{|\mathscr{D}}$

→ Typically, D : exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: parametric family (e.g. Mean-field)

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \end{cases}$$

Variational Inference in a nutshell

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

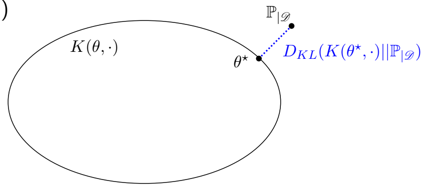
1. Posit a *simpler* variational family $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density

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

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

$\rightarrow$ Typically, D : exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: parametric family (e.g. Mean-field)

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \end{cases}$$



Variational Inference in a nutshell

→ Variational Inference : inference is seen as an optimisation problem.

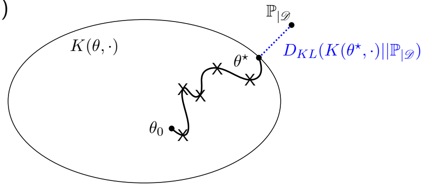
1. Posit a *simpler* variational family $\mathcal{Q}$, where $q \in \mathcal{Q}$.
2. Fit q to obtain the best approximation to the posterior density

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

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

→ Typically, D : exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: parametric family (e.g. Mean-field)

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k(\theta, y) \ : \ \theta \in \mathsf{T}\} \end{cases}$$



Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$

- Can we select alternative/more general D?
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
- Can we have theoretical guaranties?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\to$ D : exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$

- Can we select alternative/more general D?
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
- Can we have theoretical guaranties?
  $\to$ Ensure a systematic decrease in the α-divergence in our algorithms

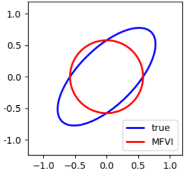
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
- Can we have theoretical guarantees?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

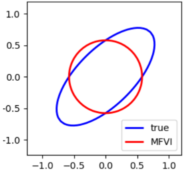
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
- Can we have theoretical guarantees?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

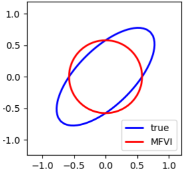
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
- Can we have theoretical guarantees?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

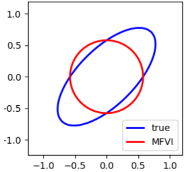
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
  $\rightarrow D$ is the α-divergence
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?

- Can we have theoretical guaranties?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

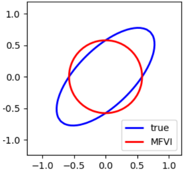
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\dfrac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
  $\rightarrow D$ is the α-divergence
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
  $\rightarrow$ Put a prior on the variational parameter θ
- Can we have theoretical guarantees?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

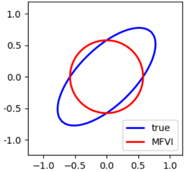
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\to D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1)k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
  $\to D$ is the α-divergence
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
  $\to$ Put a prior on the variational parameter θ
- Can we have theoretical guaranties?
  $\to$ Ensure a systematic decrease in the α-divergence in our algorithms

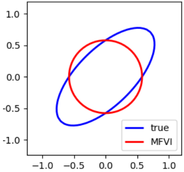
Core question in Variational Inference

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

Question : How to choose D and $\mathcal{Q}$?

$\rightarrow D$: exclusive Kullback-Leibler (KL) divergence and $\mathcal{Q}$: Mean-field family

$$\begin{cases} D_{KL}(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} \log\left(\frac{q(y)}{p(y)}\right) q(y)\nu(\mathrm{d}y) \\ \mathcal{Q} = \{q : y \mapsto k_1(\theta_1, y_1) k_2(\theta_2, y_2) \ : \ (\theta_1, \theta_2) \in \mathsf{T}\} \end{cases}$$



- Can we select alternative/more general D?
  $\rightarrow D$ is the α-divergence
- Can we design more expressive variational families $\mathcal{Q}$ beyond traditional parametric families?
  $\rightarrow$ Put a prior on the variational parameter θ
- Can we have theoretical guaranties?
  $\rightarrow$ Ensure a systematic decrease in the α-divergence in our algorithms

Variational Inference with 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 $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu} = p$.

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

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

where

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

Variational Inference with 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 $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu} = q$, $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu} = p$.

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

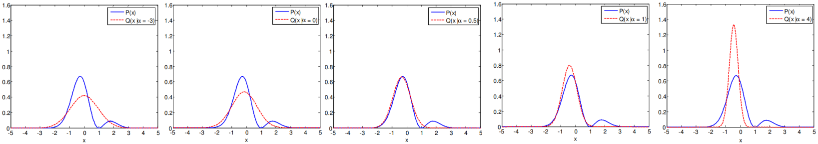
$$D_\alpha(\mathbb{Q}||\mathbb{P}) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) \ ,$$

where

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

❶ A flexible family of divergences...

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



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

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

Variational Inference with the α-divergence family

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

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

where

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

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

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

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

Black-box alpha divergence minimization. J. Hernandez-Lobato et al. (2016). ICML
Rényi divergence variational inference. Y. Li and R. E Turner (2016). NeurIPS
Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

Variational Inference with the α-divergence family

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

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

where

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

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

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

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

Black-box alpha divergence minimization. J. Hernandez-Lobato et al. (2016). ICML
Rényi divergence variational inference. Y. Li and R. E Turner (2016). NeurIPS
Variational inference via χ-upper bound minimization A. Dieng et al. (2017). NeurIPS

Variational Inference with the α-divergence family

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

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

where

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

❶ A flexible family of divergences...

❷ ...suitable for Variational Inference purposes...

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

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

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

Variational Inference with the α-divergence family

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

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

where

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

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

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

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

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

Outline

Infinite-dimensional α-divergence minimisation

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

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

Idea : Extend the traditional variational parametric family

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

by putting a prior on the variational parameter θ

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

and propose an update formula for μ that ensures a systematic decrease in the α-divergence at each step

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

NB: The mapping $\mu \mapsto \Psi_\alpha(\mu k; p)$ is convex!

Infinite-dimensional α-divergence minimisation

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

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

Idea : Extend the traditional variational parametric family

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

by putting a prior on the variational parameter θ

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

and propose an update formula for μ that ensures a systematic decrease in the α-divergence at each step

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

NB: The mapping $\mu \mapsto \Psi_\alpha(\mu k; p)$ is convex!

Infinite-dimensional α-divergence minimisation

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

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

Idea : Extend the traditional variational parametric family

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

by putting a prior on the variational parameter θ

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

and propose an update formula for μ that ensures a systematic decrease in the α-divergence at each step

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

NB: The mapping $\mu \mapsto \Psi_\alpha(\mu k; p)$ is convex!

Infinite-dimensional α-divergence minimisation

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

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

Idea : Extend the traditional variational parametric family

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

by putting a prior on the variational parameter θ

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

and propose an update formula for μ that ensures a systematic decrease in the α-divergence at each step

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

NB: The mapping $\mu \mapsto \Psi_\alpha(\mu k; p)$ is convex!

Infinite-dimensional α-divergence minimisation

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

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

Idea : Extend the traditional variational parametric family

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

by putting a prior on the variational parameter θ

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

and propose an update formula for μ that ensures a systematic decrease in the α-divergence at each step

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

NB: The mapping $\mu \mapsto \Psi_\alpha(\mu k; p)$ is convex!

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n)\ , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) \,, \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y)\nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

The (α, Γ)-descent algorithm

Optimisation problem

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

- p is a nonnegative measurable function defined on $(\mathsf{Y}, \mathcal{Y})$
- M is a subset of $\mathrm{M}_1(\mathsf{T})$, the space of probability measures on T
- $K : (\theta, A) \mapsto \int_A k(\theta, y) \nu(\mathrm{d}y)$ is a Markov transition kernel defined on $\mathsf{T} \times \mathcal{Y}$ with density k

Algorithm

Let $\mu_1 \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi_\alpha(\mu_1 k) < \infty$. The sequence of probability measures $(\mu_n)_{n \geqslant 1}$ is defined iteratively by

$$\mu_{n+1} = \mathcal{I}_\alpha(\mu_n) , \qquad n \geqslant 1$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$ and all $\theta \in \mathsf{T}$,

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

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

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

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

Theorem

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

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

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

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

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

Theorem

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

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

Conditions for a monotonic decrease

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $k(\theta, y) > 0$, $p(y) \geqslant 0$ and $\int_{\mathsf{Y}} p(y)\nu(\mathrm{d}y) < \infty$.

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

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

Theorem

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

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

Examples satisfying (A2)

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

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

- Entropic Mirror Descent : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\eta \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu_n k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right]$$

$\to$ NB : η corresponds to the learning rate

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

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \left[\int_{\mathsf{Y}} k(\theta, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha - 1} \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

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

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

- Entropic Mirror Descent : $\eta \in (0, 1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \exp\left[-\eta \int_{\mathsf{Y}} k(\theta, y) \log\left(\frac{\mu_n k(y)}{p(y)}\right) \nu(\mathrm{d}y)\right]$$

$\rightarrow$ NB : η corresponds to the learning rate

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

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta) \left[\int_{\mathsf{Y}} k(\theta, y) \left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha - 1} \nu(\mathrm{d}y) + (\alpha - 1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

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

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

- Entropic Mirror Descent : $\eta \in (0,1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\exp\left[-\eta\int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu_n k(y)}{p(y)}\right)\nu(\mathrm{d}y)\right]$$

$\rightarrow$ NB : η corresponds to the learning rate

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

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\left[\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)+(\alpha-1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Examples satisfying (A2)

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

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

- Entropic Mirror Descent : $\eta \in (0,1]$, $\kappa \in \mathbb{R}$ and $\alpha = 1$

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\exp\left[-\eta\int_{\mathsf{Y}} k(\theta,y)\log\left(\frac{\mu_n k(y)}{p(y)}\right)\nu(\mathrm{d}y)\right]$$

$\to$ NB : η corresponds to the learning rate

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

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

$$\mu_{n+1}(\mathrm{d}\theta) \propto \mu_n(\mathrm{d}\theta)\left[\int_{\mathsf{Y}} k(\theta,y)\left(\frac{\mu_n k(y)}{p(y)}\right)^{\alpha-1}\nu(\mathrm{d}y)+(\alpha-1)\kappa\right]^{\frac{\eta}{1-\alpha}}$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha-1)v+1]^{\eta/(1-\alpha)}$ with $\eta \in (0,1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n\geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \quad \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\to$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\to$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n\geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \quad \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

Convergence results

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

- Entropic Mirror Descent : $\Gamma(v) = e^{-\eta v}$ with $\eta \in (0, \frac{1}{|\alpha-1||b|_{\infty,\alpha}+1})$, any α, κ
- Power Descent : $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ with $\eta \in (0, 1]$, $\alpha > 1$, $\kappa > 0$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0, 1]$, $\kappa \leqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $(\mu_n)_{n \geqslant 1}$ weakly converges to $\mu^\star$, then :

$$\mu^\star \text{ is a fixed point of } \mathcal{I}_\alpha \text{ and } \Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

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

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

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

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J,\ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

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

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

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

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J, \ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta) \cdot \Gamma(b_{\mu_n,\alpha}(\theta) + \kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha} + \kappa))}, \quad n \geqslant 1$$

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

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

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

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n} \delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n} \Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n} \Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \ldots J,\ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

The special case of finite mixture models

$$\mu_{n+1}(\mathrm{d}\theta) = \frac{\mu_n(\mathrm{d}\theta)\cdot\Gamma(b_{\mu_n,\alpha}(\theta)+\kappa)}{\mu_n(\Gamma(b_{\mu_n,\alpha}+\kappa))}, \quad n \geqslant 1$$

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

Let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \ldots, \lambda_{J,1}) \in \mathcal{S}_J$ and denote

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

Then, $\mu_n = \underbrace{\mathcal{I}_\alpha \circ \cdots \circ \mathcal{I}_\alpha}_{n \text{ times}}(\mu_{\boldsymbol{\lambda}_1})$ is of the form $\mu_n = \sum_{j=1}^J \lambda_{j,n}\delta_{\theta_j}$ with

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\ldots J,\ n \geqslant 1$$

NB : $\mu_n k(y) = \sum_{j=1}^J \lambda_{j,n} k(\theta_j, y)$

$$\mathcal{Q} = \left\{ q : y \mapsto \mu_{\boldsymbol{\lambda}} k(y) = \sum_{j=1}^J \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J \right\}$$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^{J}\lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing
e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$
Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log \Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2 \log J} |b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1, \cdot), \dots K(\theta_J, \cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n \geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1 \dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j)+\kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i)+\kappa)}, \quad j = 1\dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta\in\mathsf{T},\mu\in\mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow$ $O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n}k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot),\dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta\in\mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Convergence results for finite mixture models

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^J \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J,\ n \geqslant 1$$

Assume (A1) and that $|b|_{\infty,\alpha} = \sup_{\theta \in \mathsf{T}, \mu \in \mathrm{M}_1(\mathsf{T})} |b_{\mu,\alpha}(\theta)| < \infty$

$\rightarrow O(1/N)$ convergence rates when Γ is L-smooth and $-\log\Gamma$ is concave increasing

e.g. Entropic Mirror Descent: when $\alpha = 1$, we have for all $\eta \in (0,1)$

$$\Psi_\alpha(\mu_{\boldsymbol{\lambda}_n} k) - \Psi_\alpha(\mu^\star k) \leqslant \frac{\log J}{\eta N} + \frac{\sqrt{2\log J}|b|_{\infty,1}}{(1-\eta)N}$$

$\rightarrow$ The case $\alpha < 1$ for the Power Descent is trickier... $\eta \in (0,1]$, $\kappa \geqslant 0$

Under additionnal assumptions on Ψ_α and $b_{\mu,\alpha}$, if $\{K(\theta_1,\cdot), \dots K(\theta_J,\cdot)\}$ are linearly independent, then :

- $(\lambda_n)_{n\geqslant 1}$ converges to some $\boldsymbol{\lambda}^\star$
- $\mu^\star = \mu_{\boldsymbol{\lambda}^\star}$ is a fixed point of $\mathcal{I}_\alpha$ and $\Psi_\alpha(\mu^\star k) = \inf_{\zeta \in \mathrm{M}_{1,\mu_1}(\mathsf{T})} \Psi_\alpha(\zeta k)$

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\rightarrow$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

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

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\rightarrow$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\rightarrow$ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

→ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

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

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

→ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

→ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

$\rightarrow$ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

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

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

$\rightarrow$ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

$\rightarrow$ Idea : combine this step with and *Exploration Step* updating Θ

Towards a practical implementation

Algorithm

Let $\Theta = (\theta_1, ..., \theta_J) \in \mathsf{T}^J$ be fixed and let $\boldsymbol{\lambda}_1 \in \mathcal{S}_J$. At time $n \geqslant 1$, define

$$\mu_{n+1}k = \sum_{j=1}^{J} \lambda_{j,n+1} k(\theta_j, \cdot)$$

where

$$\lambda_{j,n+1} = \frac{\lambda_{j,n}\Gamma(b_{\mu_n,\alpha}(\theta_j) + \kappa)}{\sum_{i=1}^{J} \lambda_{i,n}\Gamma(b_{\mu_n,\alpha}(\theta_i) + \kappa)}, \quad j = 1 \dots J$$

→ Monte Carlo approximations to estimate $b_{\mu_n,\alpha}(\theta_j)$, e.g.

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

with $Y_{1,n}, ..., Y_{M,n} \overset{\text{i.i.d}}{\sim} \mu_n k$.

→ Exploitation step not requiring any information on the distribution of $\theta_1, ..., \theta_J$

→ Idea : combine this step with and *Exploration Step* updating Θ

Outline

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

Algorithm

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

  Algorithm

  1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
  2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Numerical experiments

- Gaussian kernel with density k_h and bandwidth h, $\mathsf{T} = \mathbb{R}^d$

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

<u>Algorithm</u>

1. Exploitation step : optimise $\boldsymbol{\lambda}$ using the (α, Γ)-descent.
2. *Exploration step* : update Θ (e.g. by sampling under $\mu_{\boldsymbol{\lambda},\Theta} k_h$, $h \propto J^{-1/(4+d)}$)

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

- Bayesian Logistic Regression Covertype dataset ($581,012$ data points and 54 features)

Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

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

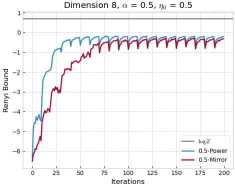
$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials

Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials

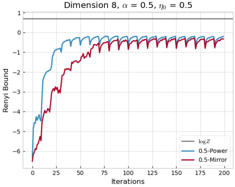


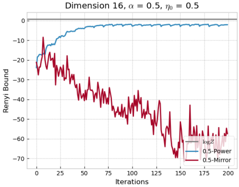
Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials



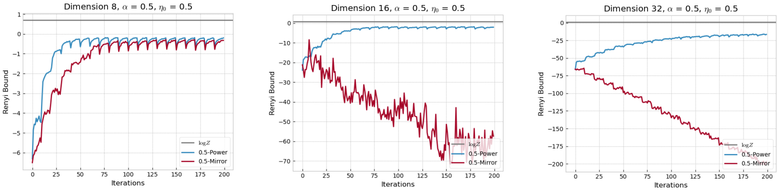


Toy example : Entropic Mirror Descent vs Power Descent

Comparison between

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : VR-Bound averaged over 100 trials



Toy example : the case $\alpha = 1$

Comparison between:

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

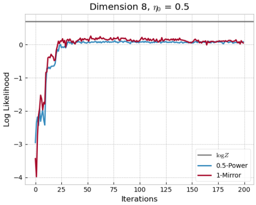
$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials

Toy example : the case $\alpha = 1$

Comparison between:

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials

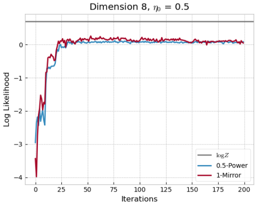


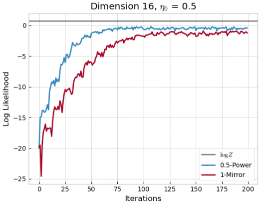
Toy example : the case $\alpha = 1$

Comparison between:

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials



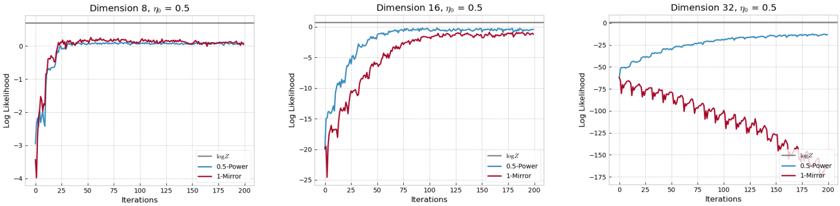


Toy example : the case $\alpha = 1$

Comparison between:

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

$J = M = 100$, initial mixture weights : $[1/J, ..., 1/J]$, $N = 10$, $T = 20$
$\eta_n = \eta_0/\sqrt{n}$, $\eta_0 = 0.5$, cv criterion : llh averaged over 100 trials



Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \; ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \; , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \; , \quad 1 \leqslant i \leqslant I$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I$$

where $a = 1$ and $b = 0.01$

Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS

$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Bayesian Logistic Regression

→ $\mathscr{D} = \{\boldsymbol{c}, \boldsymbol{x}\}$: I binary class labels, $c_i \in \{-1, 1\}$, L covariates for each datapoint, $\boldsymbol{x}_i \in \mathbb{R}^L$

→ Model : L regression coefficients $w_l \in \mathbb{R}$, precision parameter $\beta \in \mathbb{R}^+$

$$p_0(\beta) = \mathrm{Gamma}(\beta; a, b) \ ,$$
$$p_0(w_l|\beta) = \mathcal{N}(w_l; 0, \beta^{-1}) \ , \quad 1 \leqslant l \leqslant L$$
$$p(c_i = 1|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} \ , \quad 1 \leqslant i \leqslant I$$

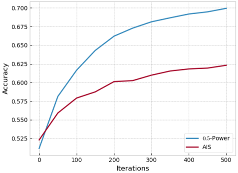
where $a = 1$ and $b = 0.01$

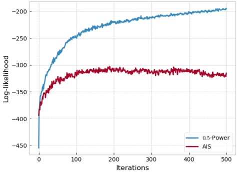
Nonparametric variational inference S. Gershman, M. Hoffman, and D. Blei (2012). ICML

→ Quantity of interest : $p(y|\mathscr{D})$ with $y = [\boldsymbol{w}, \log \beta]$

Comparison between

- 0.5-Power descent
- Typical AIS





$N = 1$, $T = 500$, $J_0 = M_0 = 20$, $J_{t+1} = M_{t+1} = J_t + 1$
initial mixture weights : $[1/J_t, ..., 1/J_t]$, $\eta_n = \eta_0/\sqrt{n}$ with $\eta_0 = 0.05$

Outline

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \; : \; \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \;:\; \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Summary

General framework for infinite-dimensional α-divergence minimisation over

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

- recovers the Entropic Mirror Descent algorithm
- novel Power Descent algorithm
- conditions for a systematic decrease + convergence results
- applicable to mixture models :

$$\mathcal{Q} = \left\{ q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J \right\}$$

$\rightarrow$ Exploitation - Exploration algorithm

1. Update for Θ not specified (e.g. your favorite update for Θ)
2. Empirical advantages of using the Power Descent algorithm

Extensions

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

   $\rightarrow$ Extension of the Power Descent to the case $\alpha = 1$

   $\rightarrow$ Full proof of convergence for finite mixture models ($\alpha < 1$)

   $\rightarrow$ Closely-related algorithm : Rényi Descent

2. **Monotonic Alpha-divergence Minimisation.**
   K. Daudel, R. Douc and F. Roueff (2021). https://arxiv.org/abs/2103.05684

   $\rightarrow$ Conditions for a simultaneous optimisation w.r.t $\boldsymbol{\lambda}$ and Θ
   (that preserve the monotonic decrease!)

   $\rightarrow$ Simple updates on Θ for speficic kernels k (e.g Gaussian, Student's)

   $\rightarrow$ Links with Gradient Descent schemes and an Integrated EM algorithm
   with empirical benefits

Extensions

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

   $\rightarrow$ Extension of the Power Descent to the case $\alpha = 1$

   $\rightarrow$ Full proof of convergence for finite mixture models ($\alpha < 1$)

   $\rightarrow$ Closely-related algorithm : Rényi Descent

2. **Monotonic Alpha-divergence Minimisation.**
   K. Daudel, R. Douc and F. Roueff (2021). https://arxiv.org/abs/2103.05684

   $\rightarrow$ Conditions for a simultaneous optimisation w.r.t $\boldsymbol{\lambda}$ and Θ (that preserve the monotonic decrease!)

   $\rightarrow$ Simple updates on Θ for speficic kernels k (e.g Gaussian, Student's)

   $\rightarrow$ Links with Gradient Descent schemes and an Integrated EM algorithm with empirical benefits

Extensions

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

$\rightarrow$ Extension of the Power Descent to the case $\alpha = 1$

$\rightarrow$ Full proof of convergence for finite mixture models ($\alpha < 1$)

$\rightarrow$ Closely-related algorithm : Rényi Descent

❷ **Monotonic Alpha-divergence Minimisation.**
K. Daudel, R. Douc and F. Roueff (2021). https://arxiv.org/abs/2103.05684

$\rightarrow$ Conditions for a simultaneous optimisation w.r.t $\boldsymbol{\lambda}$ and Θ
(that preserve the monotonic decrease!)

$\rightarrow$ Simple updates on Θ for speficic kernels k (e.g Gaussian, Student's)

$\rightarrow$ Links with Gradient Descent schemes and an Integrated EM algorithm with empirical benefits

Extensions

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

   $\rightarrow$ Extension of the Power Descent to the case $\alpha = 1$

   $\rightarrow$ Full proof of convergence for finite mixture models ($\alpha < 1$)

   $\rightarrow$ Closely-related algorithm : Rényi Descent

2. **Monotonic Alpha-divergence Minimisation.**
   K. Daudel, R. Douc and F. Roueff (2021). https://arxiv.org/abs/2103.05684

   $\rightarrow$ Conditions for a simultaneous optimisation w.r.t $\boldsymbol{\lambda}$ and Θ
   (that preserve the monotonic decrease!)

   $\rightarrow$ Simple updates on Θ for speficic kernels k (e.g Gaussian, Student's)

   $\rightarrow$ Links with Gradient Descent schemes and an Integrated EM algorithm
   with empirical benefits

Thank you for your attention!

kamelia.daudel@stats.ox.ac.uk