The f-Divergence Expectation Iteration Scheme

Kamélia Daudel

Télécom Paris
kamelia.daudel@telecom-paris.fr

November 22, 2019

Joint work with Randal Douc, François Portier and François Roueff

Introduction

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

$\rightarrow$ Variational Inference methods :

Goal

Approximate the posterior density $p(\cdot|\mathscr{D})$ by a variational density q_θ, where $\theta \in \mathsf{T}$:

$$\theta^\star = \mathrm{arginf}_{\theta \in \mathsf{T}} \mathcal{D}(q_\theta, p(\cdot|\mathscr{D})) \ ,$$

where $\mathcal{D}$ is a divergence.

$\hookrightarrow$ Particular case of density approximation.

Introduction

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

$\rightarrow$ Variational Inference methods :

Goal

Approximate the posterior density $p(\cdot|\mathscr{D})$ by a variational density q_θ, where $\theta \in \mathsf{T}$:

$$\theta^\star = \operatorname{arginf}_{\theta \in \mathsf{T}} \mathcal{D}(q_\theta, p(\cdot|\mathscr{D})) \; ,$$

where $\mathcal{D}$ is a divergence.

$\hookrightarrow$ Particular case of density approximation.

Introduction

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

$\rightarrow$ Variational Inference methods :

Goal

Approximate the posterior density $p(\cdot|\mathscr{D})$ by a variational density q_θ, where $\theta \in \mathsf{T}$:

$$\theta^\star = \operatorname{arginf}_{\theta \in \mathsf{T}} \mathcal{D}(q_\theta, p(\cdot|\mathscr{D})) ,$$

where $\mathcal{D}$ is a divergence.

$\hookrightarrow$ Particular case of density approximation.

Introduction

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

$\rightarrow$ Variational Inference methods :

Goal

Approximate the posterior density $p(\cdot|\mathscr{D})$ by a variational density q_θ, where $\theta \in \mathsf{T}$:

$$\theta^\star = \mathrm{arginf}_{\theta \in \mathsf{T}} \mathcal{D}(q_\theta, p(\cdot|\mathscr{D})) \ ,$$

where $\mathcal{D}$ is a divergence.

$\hookrightarrow$ Particular case of density approximation.

Our approach

Usually in Variational Inference : approximating family

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

Let us now consider a broader approximating family

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

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

Question : Can we define an iterative scheme which diminishes a given objective function at each step ?

$\rightarrow$ Yes : The f-EI(ϕ) algorithm !

Our approach

Usually in Variational Inference : approximating family

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

Let us now consider a broader approximating family

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

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

Question : Can we define an iterative scheme which diminishes a given objective function at each step ?

$\rightarrow$ Yes : The f-EI(ϕ) algorithm !

Our approach

Usually in Variational Inference : approximating family

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

Let us now consider a broader approximating family

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

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

Question : Can we define an iterative scheme which diminishes a given objective function at each step ?

$\rightarrow$ Yes : The f-EI(ϕ) algorithm !

Our approach

Usually in Variational Inference : approximating family

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

Let us now consider a broader approximating family

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

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

Question : Can we define an iterative scheme which diminishes a given objective function at each step ?

$\rightarrow$ Yes : The f-EI(ϕ) algorithm !

Outline

1. Optimisation problem
2. The f-Expectation Iteration algorithm f-EI(ϕ)
3. Application to density approximation
4. Conclusion

Outline

1. Optimisation problem
2. The f-Expectation Iteration algorithm f-EI(ϕ)
3. Application to density approximation
4. Conclusion

Objective function : the f-divergence

- $(\mathsf{Y}, \mathcal{Y}, \nu)$: measured space, where ν is a σ-finite measure on $(\mathsf{Y}, \mathcal{Y})$
- f : convex function over $(0, \infty)$ that satisfies $f(1) = 0$
- $\mathbb{P}_1$ and $\mathbb{P}_2$: two probability measures on $(\mathsf{Y}, \mathcal{Y})$ such that $\mathbb{P}_1 \preceq \nu$, $\mathbb{P}_2 \preceq \nu$ with $p_1 = \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\nu}$, $p_2 = \frac{\mathrm{d}\mathbb{P}_2}{\mathrm{d}\nu}$

Definition 1 : f-divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$

$$D_f(\mathbb{P}_1 || \mathbb{P}_2) = \int_{\mathsf{Y}} f\left(\frac{p_1(y)}{p_2(y)}\right) p_2(y) \nu(\mathrm{d}y)$$

Objective function : the f-divergence

Definition 1 : f-divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$

$$D_f(\mathbb{P}_1||\mathbb{P}_2) = \int_{\mathsf{Y}} f\left(\frac{p_1(y)}{p_2(y)}\right) p_2(y)\nu(\mathrm{d}y)$$

$\rightarrow$ a flexible family of divergences

$f(u)$	Corresponding divergence
$u \log(u)$	$D_{KL}(\mathbb{P}_1\|\|\mathbb{P}_2) = \int_{\mathsf{Y}} \log\left(\frac{p_1(y)}{p_2(y)}\right) p_1(y)\nu(\mathrm{d}y)$
$-\log(u)$	$D_{rKL}(\mathbb{P}_1\|\|\mathbb{P}_2) = \int_{\mathsf{Y}} -\log\left(\frac{p_1(y)}{p_2(y)}\right) p_2(y)\nu(\mathrm{d}y)$
$\dfrac{1}{\alpha(\alpha-1)}[u^\alpha - 1]$	$D_A^{(\alpha)}(\mathbb{P}_1\|\|\mathbb{P}_2) = \frac{1}{\alpha(\alpha-1)}\left[\int_{\mathsf{Y}} \left(\frac{p_1(y)}{p_2(y)}\right)^\alpha p_2(y)\nu(\mathrm{d}y) - 1\right]$

Table 1: Special cases in the f-divergence family

Optimisation problem

- $(\mathsf{T}, \mathcal{T})$: measurable space
- p : measurable positive function on $(\mathsf{Y}, \mathcal{Y})$
- $Q : (\theta, A) \mapsto \int_A q(\theta, y)\nu(\mathrm{d}y)$: Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density q

$\forall \mu \in \mathrm{M}_1(\mathsf{T})$, $\forall y \in \mathsf{Y}$, $\mu q(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) q(\theta, y)$

General optimisation problem

$$\boxed{\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi^{(f)}(\mu)}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\Psi^{(f)}(\mu) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ The mapping $\mu \mapsto \Psi^{(f)}(\mu)$ is convex.

Optimisation problem

- $(\mathsf{T}, \mathcal{T})$: measurable space
- p : measurable positive function on $(\mathsf{Y}, \mathcal{Y})$
- $Q : (\theta, A) \mapsto \int_A q(\theta, y)\nu(\mathrm{d}y)$: Markov transition kernel on $\mathsf{T} \times \mathcal{Y}$ with kernel density q

$\forall \mu \in \mathrm{M}_1(\mathsf{T})$, $\forall y \in \mathsf{Y}$, $\mu q(y) = \int_{\mathsf{T}} \mu(\mathrm{d}\theta) q(\theta, y)$

General optimisation problem

$$\boxed{\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi^{(f)}(\mu)}$$

where for all $\mu \in \mathrm{M}_1(\mathsf{T})$, $\Psi^{(f)}(\mu) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y)$.

$\rightarrow$ The mapping $\mu \mapsto \Psi^{(f)}(\mu)$ is convex.

Outline

1. Optimisation problem
2. The f-Expectation Iteration algorithm f-EI(ϕ)
3. Application to density approximation
4. Conclusion

The f-Expectation Iteration algorithm f-EI(ϕ)

Let $\phi \in \mathbb{R}^\star$, $\mu \in \mathrm{M}_1(\mathsf{T})$ such that $\Psi^{(f)}(\mu) < \infty$. We define the sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ iteratively by

$$\begin{cases} \mu_0 = \mu, \\ \mu_{n+1} = \mathcal{I}^\phi(\mu_n) \;, \qquad n \in \mathbb{N} \;. \end{cases} \tag{1}$$

Algorithm 1: *Exact f-EI(ϕ) transition*

1. Expectation step : $b_\mu(\theta) = \int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\mu q(y)}{p(y)}\right) \nu(\mathrm{d}y)$
2. Iteration step : $\mathcal{I}^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}$

When is the f-EI(ϕ) algorithm well-defined ?

$$b_\mu(\theta) = \int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\mu q(y)}{p(y)}\right) \nu(\mathrm{d}y)$$
$$\mathcal{I}^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}$$

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

(A2) $f : (0, \infty) \to \mathbb{R}$ is monotonous, strictly convex and continuously differentiable, and $f(1) = 0$.

$\to$ Under (A1) and (A2), b_μ is well-defined and $|b_\mu| \in (0, \infty]$.

$\to$ The iteration $\mu \mapsto \mathcal{I}^\phi(\mu)$ is well-defined if moreover we have

$$0 < \mu(|b_\mu|^\phi) < \infty \, . \tag{2}$$

When is the f-EI(ϕ) algorithm well-defined ?

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

$$\mathcal{I}^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}$$

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

(A2) $f : (0, \infty) \to \mathbb{R}$ is monotonous, strictly convex and continuously differentiable, and $f(1) = 0$.

→ Under (A1) and (A2), b_μ is well-defined and $|b_\mu| \in (0, \infty]$.

→ The iteration $\mu \mapsto \mathcal{I}^\phi(\mu)$ is well-defined if moreover we have

$$0 < \mu(|b_\mu|^\phi) < \infty \,. \tag{2}$$

When is the f-EI(ϕ) algorithm well-defined ?

$$b_\mu(\theta) = \int_{\mathsf{Y}} q(\theta, y) f'\left(\frac{\mu q(y)}{p(y)}\right)\nu(\mathrm{d}y)$$
$$\mathcal{I}^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}$$

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

(A2) $f : (0, \infty) \to \mathbb{R}$ is monotonous, strictly convex and continuously differentiable, and $f(1) = 0$.

$\to$ Under (A1) and (A2), b_μ is well-defined and $|b_\mu| \in (0, \infty]$.

$\to$ The iteration $\mu \mapsto \mathcal{I}^\phi(\mu)$ is well-defined if moreover we have

$$0 < \mu(|b_\mu|^\phi) < \infty \; . \tag{2}$$

Monotonicity

Divergence considered		Corresponding range
Reverse KL $f(u) = -\log(u)$		$\phi \in (0, 1]$
α-*divergence* $f(u) = \frac{1}{\alpha(\alpha-1)}(u^\alpha - 1)$	$\alpha \in (-\infty, -1]$	$\phi \in (0, -1/\alpha]$
	$\alpha \in (-1, 1) \setminus \{0\}$	$\phi \in (0, 1]$
	$\alpha \in (1, \infty)$	$\phi \in (1/(1-\alpha), 0)$

Table 2 : Allowed (f, ϕ) in the f-EI(ϕ) algorithm

Theorem 1

Assume that p and q are as in (A1). Let (f, ϕ) belong to any of the cases in Table 2.
Then (A2) holds. Moreover, let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\Psi^{(f)}(\mu) < \infty$. Then the sequence $(\mu_n)_{n \in \mathbb{N}}$ defined by (1) is well-defined and the sequence $(\Psi^{(f)}(\mu_n))_{n \in \mathbb{N}}$ is non-increasing.

Limiting behavior

(A3) T is a compact metric space, $\theta \mapsto q(\theta, y)$ is continuous for all $y \in$ Y, $\Psi^{(f)}$ and b_μ are uniformly bounded w.r.t μ and θ.

Theorem 2

Assume (A1), (A2) and (A3). Further assume that there exists $\mu, \bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n \in \mathbb{N}}$ defined by (1) weakly converges to $\bar{\mu}$ as $n \to \infty$. Then

1. $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$,
2. $\Psi^{(f)}(\bar{\mu}) = \inf_{\zeta \in \mathrm{M}_{1,\mu}(\mathsf{T})} \Psi^{(f)}(\zeta)$,

for f non-increasing and $\phi > 0$ or f non-decreasing and $\phi < 0$.

$\mathrm{M}_{1,\mu}(\mathsf{T})$: set of probability measures dominated by μ

$\rightarrow$ Compatible with Theorem 1 for (f, ϕ)

Limiting behavior

(A3) T is a compact metric space, $\theta \mapsto q(\theta, y)$ is continuous for all $y \in \mathsf{Y}$, $\Psi^{(f)}$ and b_μ are uniformly bounded w.r.t μ and θ.

Theorem 2

Assume (A1), (A2) and (A3). Further assume that there exists $\mu, \bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ such that the (well-defined) sequence $(\mu_n)_{n \in \mathbb{N}}$ defined by (1) weakly converges to $\bar{\mu}$ as $n \to \infty$. Then

1. $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$,
2. $\Psi^{(f)}(\bar{\mu}) = \inf_{\zeta \in \mathrm{M}_{1,\mu}(\mathsf{T})} \Psi^{(f)}(\zeta)$,

for f non-increasing and $\phi > 0$ or f non-decreasing and $\phi < 0$.

$\mathrm{M}_{1,\mu}(\mathsf{T})$: set of probability measures dominated by μ

$\to$ Compatible with Theorem 1 for (f, ϕ)

Approximate f-EI(ϕ)

Algorithm 1 typically involves an intractable integral in the Expectation step :

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

$\rightarrow$ Approximate f-EI(ϕ)

Algorithm 2: *Approximate f-EI(ϕ) transition*

1. Sampling step : Draw independently $Y_1, ..., Y_K \sim \mu q$
2. Expectation step : $b_{\mu,K}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{q(\theta, Y_k)}{\mu q(Y_k)} f'\left(\frac{\mu q(Y_k)}{p(Y_k)}\right)$
3. Iteration step : $\mathcal{I}_K^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}$

Approximate f-EI(ϕ)

Algorithm 1 typically involves an intractable integral in the Expectation step :

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

$\rightarrow$ Approximate f-EI(ϕ)

Algorithm 2: *Approximate f-EI(ϕ) transition*

1. Sampling step : Draw independently $Y_1, ..., Y_K \sim \mu q$
2. Expectation step : $b_{\mu,K}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \frac{q(\theta, Y_k)}{\mu q(Y_k)} f'\left(\frac{\mu q(Y_k)}{p(Y_k)}\right)$
3. Iteration step : $\mathcal{I}_K^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}$

Total variation convergence

Let $Y_1, Y_2, \ldots$ be i.i.d random variables with common density μq w.r.t ν, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Proposition 2

Assume (A1) and (A2). Let $\mu \in \mathrm{M}_1(\mathsf{T})$, $\phi \in \mathbb{R}^\star$ be such that $\mu(|b_\mu|) \vee \mu(|b_\mu|^\phi) < \infty$ and

$$\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}_{\mu q} \left[\left\{ \frac{q(\theta, Y_1)}{\mu q(Y_1)} \left| f'\left(\frac{\mu q(Y_1)}{p(Y_1)} \right) \right| \right\}^\phi \right] < \infty \, . \tag{3}$$

Then,

$$\lim_{K \to \infty} \left\| \mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu) \right\|_{TV} = 0 \, , \quad \mathbb{P} - \text{a.s.}$$

Sketch of the proof (1)

- Triangular inequality :

$$\left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right| = \left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} + \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right|$$
$$\leqslant \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|}{\mu(|b_\mu|^\phi)}$$

which implies :

$$\left\|\mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu)\right\|_{TV} = \mu\left(\left|\frac{|b_{\mu,K}|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu|^\phi}{\mu(|b_\mu|^\phi)}\right|\right)$$
$$\leqslant \left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$$

- First term of the r.h.s : $\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right|$

Lemma $\lim_{K\to\infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi)\,, \quad \mathbb{P} - \text{a.s.}$

Sketch of the proof (1)

- Triangular inequality :

$$\left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right| = \left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} \quad + \quad \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right|$$

$$\leqslant \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|}{\mu(|b_\mu|^\phi)}$$

  which implies :

$$\left\|\mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu)\right\|_{TV} = \mu\left(\left|\frac{|b_{\mu,K}|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu|^\phi}{\mu(|b_\mu|^\phi)}\right|\right)$$

$$\leqslant \left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$$

- First term of the r.h.s : $\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right|$

  Lemma $\lim_{K\to\infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi)\,, \quad \mathbb{P}-\text{a.s.}$

Sketch of the proof (1)

- Triangular inequality :

$$\left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right| = \left|\frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} + \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_\mu|^\phi)} - \frac{|b_\mu(\theta)|^\phi}{\mu(|b_\mu|^\phi)}\right|$$
$$\leqslant \frac{|b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|}{\mu(|b_\mu|^\phi)}$$

  which implies :

$$\left\|\mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu)\right\|_{TV} = \mu\left(\left|\frac{|b_{\mu,K}|^\phi}{\mu(|b_{\mu,K}|^\phi)} - \frac{|b_\mu|^\phi}{\mu(|b_\mu|^\phi)}\right|\right)$$
$$\leqslant \left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right| + \frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$$

- First term of the r.h.s : $\left|1 - \frac{\mu(|b_{\mu,K}|^\phi)}{\mu(|b_\mu|^\phi)}\right|$

  Lemma $\lim_{K\to\infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi)\,, \quad \mathbb{P}-\text{a.s.}$

Sketch of the proof (2)

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$

  Generalized Dominated Convergence Theorem :

  1. For all $K \in \mathbb{N}^\star$ and for μ-almost all $\theta \in \mathsf{T}$,

     $$a_K(\theta) \leqslant b_K(\theta) \leqslant c_K(\theta) ,$$

     and the limits $\lim_{K \to \infty} a_K(\theta)$, $\lim_{K \to \infty} b_K(\theta)$, $\lim_{K \to \infty} c_K(\theta)$ exist.

  2. - $\mu|\lim_{K \to \infty} a_K| + \mu|\lim_{K \to \infty} c_K| < \infty$
     - $\mu(\lim_{K \to \infty} a_K) = \lim_{K \to \infty} \mu(a_K)$ and $\mu(\lim_{K \to \infty} c_K) = \lim_{K \to \infty} \mu(c_K)$

$$\Rightarrow \mu(\lim_{K \to \infty} b_K) = \lim_{K \to \infty} \mu(b_K)$$

Sketch of the proof (2)

$$b_K(\theta) = ||b_{\mu,K}(\theta)|^{\phi} - |b_\mu(\theta)|^{\phi}|$$

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^{\phi} - |b_\mu|^{\phi}|)}{\mu(|b_\mu|^{\phi})}$

Generalized Dominated Convergence Theorem :

❶ For all $K \in \mathbb{N}^\star$ and for μ-almost all $\theta \in \mathsf{T}$,

$$a_K(\theta) \leqslant b_K(\theta) \leqslant c_K(\theta) \ ,$$

and the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist.

❷
- $\mu|\lim_{K\to\infty} a_K| + \mu|\lim_{K\to\infty} c_K| < \infty$
- $\mu(\lim_{K\to\infty} a_K) = \lim_{K\to\infty} \mu(a_K)$ and $\mu(\lim_{K\to\infty} c_K) = \lim_{K\to\infty} \mu(c_K)$

$$\Rightarrow \mu(\lim_{K\to\infty} b_K) = \lim_{K\to\infty} \mu(b_K)$$

Sketch of the proof (2)

$$b_K(\theta) = ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|$$

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$

Generalized Dominated Convergence Theorem :

❶ For all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$0 \leqslant ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi| \leqslant |b_{\mu,K}(\theta)|^\phi + |b_\mu(\theta)|^\phi$$

and the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist.

❷
- $\mu|\lim_{K\to\infty} a_K| + \mu|\lim_{K\to\infty} c_K| < \infty$
- $\mu(\lim_{K\to\infty} a_K) = \lim_{K\to\infty} \mu(a_K)$ and $\mu(\lim_{K\to\infty} c_K) = \lim_{K\to\infty} \mu(c_K)$

$$\Rightarrow \mu(\lim_{K\to\infty} b_K) = \lim_{K\to\infty} \mu(b_K)$$

Sketch of the proof (2)

$b_K(\theta) = ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|$

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$

Generalized Dominated Convergence Theorem :

❶ For all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$0 \leqslant ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi| \leqslant |b_{\mu,K}(\theta)|^\phi + |b_\mu(\theta)|^\phi$$

$a_K(\theta)$ $c_K(\theta)$

and the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist.

❷
- $\mu|\lim_{K\to\infty} a_K| + \mu|\lim_{K\to\infty} c_K| < \infty$
- $\mu(\lim_{K\to\infty} a_K) = \lim_{K\to\infty} \mu(a_K)$ and $\mu(\lim_{K\to\infty} c_K) = \lim_{K\to\infty} \mu(c_K)$

$$\Rightarrow \mu(\lim_{K\to\infty} b_K) = \lim_{K\to\infty} \mu(b_K)$$

Sketch of the proof (2)

$b_K(\theta) = ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|$

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$

Generalized Dominated Convergence Theorem :

❶ For all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,

$$\underbrace{0}_{a_K(\theta)} \leqslant ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi| \leqslant \underbrace{|b_{\mu,K}(\theta)|^\phi + |b_\mu(\theta)|^\phi}_{c_K(\theta)}$$

and the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist.

❷ $$\lim_{K\to\infty} \mu\left[|b_{\mu,K}|^\phi + |b_\mu|^\phi\right] = \mu\left[\lim_{K\to\infty}\left(|b_{\mu,K}|^\phi + |b_\mu|^\phi\right)\right] < \infty$$

(*since* $\mu(|b_\mu|^\phi) < \infty$ *and* $\lim_{K\to\infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi)\,, \quad \mathbb{P}-\mathrm{a.s.}$)

$$\Rightarrow \mu(\lim_{K\to\infty} b_K) = \lim_{K\to\infty} \mu(b_K)$$

Sketch of the proof (2)

$b_K(\theta) = ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi|$

- Second term of the r.h.s : $\frac{\mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|)}{\mu(|b_\mu|^\phi)}$

Generalized Dominated Convergence Theorem :

1. For all $K \in \mathbb{N}^\star$, for all $\theta \in \mathsf{T}$,
$$\underbrace{0}_{a_K(\theta)} \leqslant ||b_{\mu,K}(\theta)|^\phi - |b_\mu(\theta)|^\phi| \leqslant \underbrace{|b_{\mu,K}(\theta)|^\phi + |b_\mu(\theta)|^\phi}_{c_K(\theta)}$$
and the limits $\lim_{K\to\infty} a_K(\theta)$, $\lim_{K\to\infty} b_K(\theta)$, $\lim_{K\to\infty} c_K(\theta)$ exist.

2. $$\lim_{K\to\infty} \mu\left[|b_{\mu,K}|^\phi + |b_\mu|^\phi\right] = \mu\left[\lim_{K\to\infty}\left(|b_{\mu,K}|^\phi + |b_\mu|^\phi\right)\right] < \infty$$
(*since* $\mu(|b_\mu|^\phi) < \infty$ *and* $\lim_{K\to\infty} \mu(|b_{\mu,K}|^\phi) = \mu(|b_\mu|^\phi)\,, \quad \mathbb{P} - \text{a.s.}$)

$$\Rightarrow \lim_{K\to\infty} \mu(||b_{\mu,K}|^\phi - |b_\mu|^\phi|) = \mu(\lim_{K\to\infty} ||b_{\mu,K}|^\phi - |b_\mu|^\phi|) = 0\,, \quad \mathbb{P} - \text{a.s.}$$

Outline

1. Optimisation problem
2. The f-Expectation Iteration algorithm f-EI(ϕ)
3. Application to density approximation
4. Conclusion

f-EI(ϕ) applied to density approximation

Let $\tilde{p}$ be a probability density function on $(\mathsf{Y}, \mathcal{Y})$ and assume that we only have access to an unnormalized version p^* of the density $\tilde{p}$, that is for all $y \in \mathsf{Y}$,

$$\tilde{p}(y) = \frac{p^*(y)}{Z} \ , \tag{4}$$

where $Z := \int_{\mathsf{Y}} p^*(y)\nu(\mathrm{d}y)$ is called the *normalizing constant* or *partition function*.

$\rightarrow$ Posterior density approximation : $\tilde{p} = p(\cdot|\mathscr{D})$, $p^* = p(\mathscr{D}, \cdot)$ and $Z = p(\mathscr{D})$.

Reformulation of the optimisation problem

- $\tilde{\mathbb{P}}$: probability measure on $(\mathsf{Y}, \mathcal{Y})$ with density $\tilde{p}$ with respect to ν
- for all $\mu \in \mathrm{M}_1(\mathsf{T})$, μQ : probability measure on $(\mathsf{Y}, \mathcal{Y})$ with density μq with respect to ν

Lemma 3

Assume (A1). Then, for both the reverse Kullback-Leibler and the α-divergence, optimising the objective

$$D_f(\mu Q || \tilde{\mathbb{P}})$$

(with respect to μ) is equivalent to optimising the objective

$$\Psi^{(f)}(\mu; p) \text{ with } p = p^*.$$

Particular case of the α-divergence

$\rightarrow$ α-bound : $\tilde{q} \mapsto \xi^{(\alpha)}(\tilde{q}) := \left[\int_{\mathsf{Y}} \left(\frac{\tilde{q}(y)}{p^*(y)}\right)^{\alpha} p^*(y)\nu(\mathrm{d}y)\right]^{\frac{1}{1-\alpha}}$

Then,

$$\Psi^{(f)}(\mu; p) = \frac{1}{\alpha(\alpha-1)}\left(\xi^{(\alpha)}(\mu q)^{1-\alpha} - Z\right) \quad \text{with} \quad p = p^* .$$

Lemma 4

Assume (A1). Let $\mu \in \mathrm{M}_1(\mathsf{T})$. Then, for all $\alpha_+ \in (0,1) \cup (1, +\infty)$ and all $\alpha_- < 0$, we have

$$\xi^{(\alpha^+)}(\mu q) \leqslant Z \leqslant \xi^{(\alpha^-)}(\mu q) . \tag{5}$$

$\rightarrow$ We can observe the convergence / monotonicity and obtain a bound on the normalising constant Z.

Particular case of the α-divergence

$\rightarrow$ α-bound : $\tilde{q} \mapsto \xi^{(\alpha)}(\tilde{q}) := \left[\int_{\mathsf{Y}} \left(\frac{\tilde{q}(y)}{p^*(y)} \right)^{\alpha} p^*(y)\nu(\mathrm{d}y) \right]^{\frac{1}{1-\alpha}}$

Then,

$$\Psi^{(f)}(\mu; p) = \frac{1}{\alpha(\alpha-1)} \left(\xi^{(\alpha)}(\mu q)^{1-\alpha} - Z \right) \quad \text{with} \quad p = p^* \, .$$

Lemma 4

Assume (A1). Let $\mu \in \mathrm{M}_1(\mathrm{T})$. Then, for all $\alpha_+ \in (0,1) \cup (1, +\infty)$ and all $\alpha_- < 0$, we have

$$\xi^{(\alpha^+)}(\mu q) \leqslant Z \leqslant \xi^{(\alpha^-)}(\mu q) \, . \tag{5}$$

$\rightarrow$ We can observe the convergence / monotonicity and obtain a bound on the normalising constant Z.

Particular case of the α-divergence

$\rightarrow$ α-bound : $\tilde{q} \mapsto \xi^{(\alpha)}(\tilde{q}) := \left[\int_{\mathsf{Y}} \left(\frac{\tilde{q}(y)}{p^*(y)}\right)^{\alpha} p^*(y)\nu(\mathrm{d}y)\right]^{\frac{1}{1-\alpha}}$

Then,

$$\Psi^{(f)}(\mu; p) = \frac{1}{\alpha(\alpha-1)}\left(\xi^{(\alpha)}(\mu q)^{1-\alpha} - Z\right) \quad \text{with} \quad p = p^* \; .$$

Lemma 4

Assume (A1). Let $\mu \in \mathrm{M}_1(\mathsf{T})$. Then, for all $\alpha_+ \in (0,1) \cup (1, +\infty)$ and all $\alpha_- < 0$, we have

$$\xi^{(\alpha^+)}(\mu q) \leqslant Z \leqslant \xi^{(\alpha^-)}(\mu q) \; . \tag{5}$$

$\rightarrow$ We can observe the convergence / monotonicity and obtain a bound on the normalising constant Z.

μ_0 is a weighted sum of Dirac measures

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

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

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

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

with

$$\begin{cases} \boldsymbol{\lambda}_0 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \lambda_{j,n} \frac{|b_{\mu_n,K}(\theta_j)|^\phi}{\mu_n(|b_{\mu_n,K}|^\phi)} \ , \quad n \in \mathbb{N},\ j \in \{1, \dots, J\} \end{cases}$$

μ_0 is a weighted sum of Dirac measures

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

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

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

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

with

$$\begin{cases} \boldsymbol{\lambda}_0 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \lambda_{j,n} \frac{|b_{\mu_n,K}(\theta_j)|^\phi}{\mu_n(|b_{\mu_n,K}|^\phi)} \, , \quad n \in \mathbb{N}, \ j \in \{1, \ldots, J\} \end{cases}$$

μ_0 is a weighted sum of Dirac measures

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

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

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

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

with

$$\begin{cases} \boldsymbol{\lambda}_0 = \boldsymbol{\lambda} \\ \lambda_{j,n+1} = \lambda_{j,n} \frac{|b_{\mu_n,K}(\theta_j)|^\phi}{\mu_n(|b_{\mu_n,K}|^\phi)}\ , \quad n \in \mathbb{N},\ j \in \{1, \dots, J\} \end{cases}$$

Mixing the two

Algorithm 3: Mixture α-Approximate f-EI(ϕ)

Input: p^*: unnormalized version of the density $\tilde{p}$, Q: Markov transition kernel, K: number of samples, $\Theta_J = \{\theta_1, ..., \theta_J\} \subset \mathsf{T}$: parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\frac{1}{J}, ..., \frac{1}{J}]$.

while *the α-bound has not converged* **do**

Sampling step : Draw independently K samples $Y_1, ..., Y_K$ from $\mu_{\boldsymbol{\lambda}} q$.

Expectation step : Compute $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1 \leqslant j \leqslant J}$ where

$$a_j = \frac{1}{K} \sum_{k=1}^{K} q(\theta_j, Y_k) \mu_{\boldsymbol{\lambda}} q(Y_k)^{\alpha-2} p^*(Y_k)^{1-\alpha}$$

and deduce $\boldsymbol{B}_{\boldsymbol{\lambda}} = (\lambda_j a_j^{\phi})_{1 \leqslant j \leqslant J}$, $b_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j^{\phi}$ and $c_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j$.

Iteration step : Set

$$\xi_K^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) \leftarrow c_{\boldsymbol{\lambda}}^{1/(1-\alpha)}$$

$$\boldsymbol{\lambda} \leftarrow \frac{1}{b_{\boldsymbol{\lambda}}} \boldsymbol{B}_{\boldsymbol{\lambda}}$$

end

Mixing the two

Algorithm 3: Mixture α-Approximate f-EI(ϕ)

Input: p^*: unnormalized version of the density $\tilde{p}$, Q: Markov transition kernel, K: number of samples, $\Theta_J = \{\theta_1, ..., \theta_J\} \subset \mathsf{T}$: parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\frac{1}{J}, ..., \frac{1}{J}]$.

while *the α-bound has not converged* **do**

Sampling step : Draw independently K samples $Y_1, ..., Y_K$ from $\mu_{\boldsymbol{\lambda}} q$.

Expectation step : Compute $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1 \leqslant j \leqslant J}$ where

$$a_j = \frac{1}{K} \sum_{k=1}^{K} q(\theta_j, Y_k) \mu_{\boldsymbol{\lambda}} q(Y_k)^{\alpha-2} p^*(Y_k)^{1-\alpha}$$

and deduce $\boldsymbol{B}_{\boldsymbol{\lambda}} = (\lambda_j a_j^{\phi})_{1 \leqslant j \leqslant J}$, $b_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j^{\phi}$ and $c_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j$.

Iteration step : Set

$$\xi_K^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) \leftarrow c_{\boldsymbol{\lambda}}^{1/(1-\alpha)}$$

$$\boldsymbol{\lambda} \leftarrow \frac{1}{b_{\boldsymbol{\lambda}}} \boldsymbol{B}_{\boldsymbol{\lambda}}$$

end

Mixing the two

Algorithm 3: Mixture α-Approximate f-EI(ϕ)

Input: p^*: unnormalized version of the density $\tilde{p}$, Q: Markov transition kernel, K: number of samples, $\Theta_J = \{\theta_1, ..., \theta_J\} \subset \mathsf{T}$: parameter set.

Output: Optimised weights $\boldsymbol{\lambda}$.

Set $\boldsymbol{\lambda} = [\frac{1}{J}, ..., \frac{1}{J}]$.

while *the α-bound has not converged* **do**

- Sampling step : Draw independently K samples $Y_1, ..., Y_K$ from $\mu_{\boldsymbol{\lambda}} q$.
- Expectation step : Compute $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1 \leqslant j \leqslant J}$ where

$$a_j = \frac{1}{K} \sum_{k=1}^{K} q(\theta_j, Y_k) \mu_{\boldsymbol{\lambda}} q(Y_k)^{\alpha-2} p^*(Y_k)^{1-\alpha}$$

  and deduce $\boldsymbol{B}_{\boldsymbol{\lambda}} = (\lambda_j a_j^{\phi})_{1 \leqslant j \leqslant J}$, $b_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j^{\phi}$ and $c_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j a_j$.
- Iteration step : Set

$$\xi_K^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) \leftarrow c_{\boldsymbol{\lambda}}^{1/(1-\alpha)}$$

$$\boldsymbol{\lambda} \leftarrow \frac{1}{b_{\boldsymbol{\lambda}}} \boldsymbol{B}_{\boldsymbol{\lambda}}$$

end

Most of the computing effort

One interesting remark

- Most of the computing effort : compute $(b_{\mu_n,K}(\theta_j))_{1\leqslant j\leqslant J}$ (or equivalently $\boldsymbol{A}_{\boldsymbol{\lambda}} = (a_j)_{1\leqslant j\leqslant J}$).
- The score gradient of the function

$$\tilde{q} \mapsto \mathcal{L}_A^{(\alpha)}(\tilde{q}) := \int_{\mathsf{Y}} \frac{1}{\alpha(\alpha-1)} \left(\frac{\tilde{q}(y)}{p^*(y)}\right)^{\alpha} p^*(y)\nu(\mathrm{d}y) \ ,$$

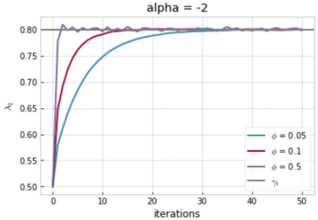
is linked to the quantities approximated in our algorithm

$$\boxed{\nabla_{\boldsymbol{\lambda}} \mathcal{L}_A^{(\alpha)}(\mu_{\boldsymbol{\lambda}} q) = (b_{\mu_{\boldsymbol{\lambda}}}(\theta_j))_{1\leqslant j\leqslant J} \ .}$$

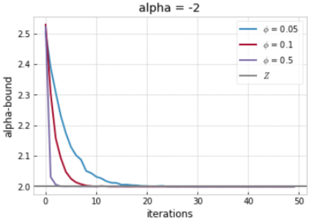
$\rightarrow$ similar to computations required in gradient-based methods involving the α-divergence or Renyi's α-divergence.

Numerical experiments : impact of ϕ

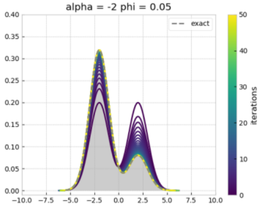
$$p^*(y) = Z \times [\gamma_1 \mathcal{N}(y; -s, 1) + \gamma_2 \mathcal{N}(y; s, 1)]$$, where $\gamma_1 = 0.8$ $\gamma_2 = 0.2$, $s = 2$ and $Z = 2$



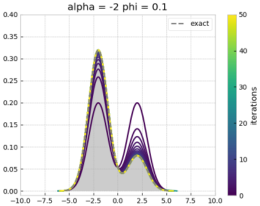
(1)



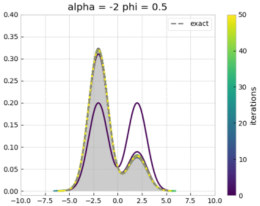
(2)



(3)



(4)



(5)

Towards an adaptive algorithm

- Algorithm 3 leaves $\{\theta_1, ..., \theta_J\}$ unchanged (Exploitation Step)

$\rightarrow$ Combine it with an Exploration step that modifies the parameter set !

Example : resampling + stochastic perturbation

Towards an adaptive algorithm

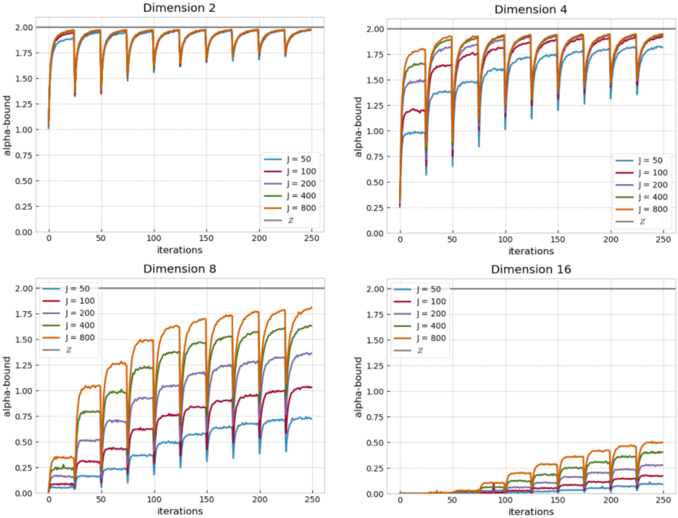
- Algorithm 3 leaves $\{\theta_1, ..., \theta_J\}$ unchanged (Exploitation Step)

$\rightarrow$ Combine it with an Exploration step that modifies the parameter set !

Example : resampling + stochastic perturbation

Numerical experiments : impact of d and J

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



Outline

1. Optimisation problem
2. The f-Expectation Iteration algorithm f-EI(ϕ)
3. Application to density approximation
4. Conclusion

Conclusion

f-EI(ϕ) algorithm : novel iterative scheme that

- performs an update of measures
  1. Sufficient conditions on (f, ϕ) leading to a systematic decrease
  2. Convergence to an optimum
  3. Approximate version of the algorithm

- can be applied to density approximation
  1. α-bound: bound on Z, which also measures the convergence
  2. the computations involved in the Mixture α-Approximate f-EI(ϕ) algorithm mostly rely on gradient-based calculations

Conclusion

f-EI(ϕ) algorithm : novel iterative scheme that

- performs an update of measures
  1. Sufficient conditions on (f, ϕ) leading to a systematic decrease
  2. Convergence to an optimum
  3. Approximate version of the algorithm

- can be applied to density approximation
  1. α-bound: bound on Z, which also measures the convergence
  2. the computations involved in the Mixture α-Approximate f-EI(ϕ) algorithm mostly rely on gradient-based calculations

Perspectives

- ϕ constant in the f-EI(ϕ) algorithm $\Rightarrow$ decaying learning rate
- Convergence rate of the f-EI(ϕ) algorithm
- Large scale learning
- Try other types of Exploration steps
- Variance reduction schemes in the approximation of b_{μ_n}
- ...

References

[1] Kamélia Daudel, Randal Douc, François Portier, François Roueff (2019). The f-Divergence Expectation Iteration Scheme. arXiv e-prints, page arXiv:1909.12239.

[2] Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert (2007). Convergence of adaptive mixtures of importance sampling schemes. arXiv e-prints, page arXiv:0708.0711.

[3] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard E. Turner (2015). Black-box α-divergence Minimization. arXiv e-prints, page arXiv:1511.03243.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians (2016). arXiv e-prints, page arXiv:1601.00670.

[5] Yingzhen Li and Richard E. Turner (2016). Rényi Divergence Variational Inference. arXiv e-prints, page arXiv:1602.02311.