The f-Expectation Iteration Algorithm

K. Daudel[1], R. Douc[2], F. Portier[1], F. Roueff[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France

[2] SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France

Motivation

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.
$\to$ Variational Inference methods see this as an optimisation problem over the variational family $\{y \mapsto q_\theta(y) \ : \ \theta \in \mathsf{T}\}$.

Let us now consider a broader approximating family

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

where M is a 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 ? $\to$ Yes : the f-EI(ϕ) Algorithm !

The f-EI(ϕ) Algorithm

General Optimisation Problem: f convex over $(0, \infty)$, $f(1) = 0$

$$\operatorname{arginf}_{\mu \in \mathsf{M}} \Psi^{(f)}(\mu) \quad \text{where} \quad \Psi^{(f)}(\mu) = \int_{\mathsf{Y}} f\left(\frac{\mu q(y)}{p(y)}\right) p(y)\nu(\mathrm{d}y) .$$

$\to$ the mapping $\mu \mapsto \Psi^{(f)}(\mu)$ is convex.
$\to$ includes f-Divergence posterior density approximation.

Let $\phi \in \mathbb{R}^\star$, $\mu \in \mathrm{M}_1(\mathsf{T})$ and let the sequence $(\mu_n)_{n \in \mathbb{N}}$ be defined by

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

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

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

(A1) For all $(\theta, y) \in \mathsf{T} \times \mathsf{Y}$, $q(\theta, y) > 0$, $p(y) > 0$ and $\int_{\mathsf{Y}} p(y)\nu(dy) < \infty$.
(A2) $f \ : \ (0, \infty) \to \mathbb{R}$ is monotonous, strictly convex and continuously differentiable, and $f(1) = 0$.

Theoretical Results

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

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

(A3) T is a compact metric space, for all $y \in \mathsf{Y}$, $\theta \mapsto q(\theta, y)$ is continuous + uniform boundedness of $\Psi^{(f)}$ and b_μ with respect to μ and θ.

Theorem 2. Assume (A1), (A3) and let (f, ϕ) belong to Table 1. Further assume that there exists $\mu, \bar{\mu} \in \mathrm{M}_1(\mathsf{T})$ such that $\mu_n \Rightarrow \bar{\mu}$ as $n \to \infty$. Then $\bar{\mu}$ is a fixed point of $\mathcal{I}^\phi$ and

$$\Psi^{(f)}(\bar{\mu}) = \inf_{\zeta \in \mathrm{M}_{1,\mu}(\mathsf{T})} \Psi^{(f)}(\zeta).$$

Let $Y_1, ..., Y_K \overset{iid}{\sim} \mu q$ and define $b_{\mu,K}(\theta) = \frac{1}{K}\sum_{k=1}^K \frac{q(\theta, Y_k)}{\mu q(Y_k)} f'\left(\frac{\mu q(Y_k)}{p(Y_k)}\right)$.

Theorem 3. Assume (A1). Let (f, ϕ) belong to Table 1. Let $\mu \in \mathrm{M}_1(\mathsf{T})$ be such that $\int_{\mathsf{T}} \mu(\mathrm{d}\theta) \mathbb{E}_{\mu q}[\{\frac{q(\theta, Y_1)}{\mu q(Y_1)} |f'\left(\frac{\mu q(Y_1)}{p(Y_1)}\right)|\}^\phi] < \infty$ and $\Psi^{(f)}(\mu) < \infty$. Then, $\mathbb{P}-$a.s.

$$\lim_{K \to \infty} \left\| \mathcal{I}_K^\phi(\mu) - \mathcal{I}^\phi(\mu) \right\|_{TV} = 0 ,$$

where $\mathcal{I}_K^\phi(\mu)(\mathrm{d}\theta) = \frac{\mu(\mathrm{d}\theta) \cdot |b_{\mu,K}(\theta)|^\phi}{\mu(|b_{\mu,K}|^\phi)}$.

Density Approximation

We can access an unnormalized version p^* of the probability density $\tilde{p}$

$$\tilde{p}(y) = \frac{p^*(y)}{Z} , \quad \text{where} \quad Z := \int_{\mathsf{Y}} p^*(y)\nu(\mathrm{d}y) .$$

Lemma 1. Assume (A1). For the α-divergence, optimising $D_f(\mu Q||\tilde{\mathbb{P}})$ (with respect to μ) is equivalent to optimising $\Psi^{(f)}(\mu; p)$ with $p = p^*$. Furthermore, for all $\alpha_+ \in (0, 1) \cup (1, +\infty)$ and all $\alpha_- < 0$, we have

$$\forall \mu \in \mathrm{M}_1(\mathsf{T}), \quad \xi^{(\alpha^+)}(\mu q) \leqslant Z \leqslant \xi^{(\alpha^-)}(\mu q) ,$$

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

Mixture α-Approximate f-EI(ϕ)

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

Iteration step : Set

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

$$\boldsymbol{\lambda} \leftarrow \frac{1}{b_{\boldsymbol{\lambda}}} \boldsymbol{B_\lambda}$$

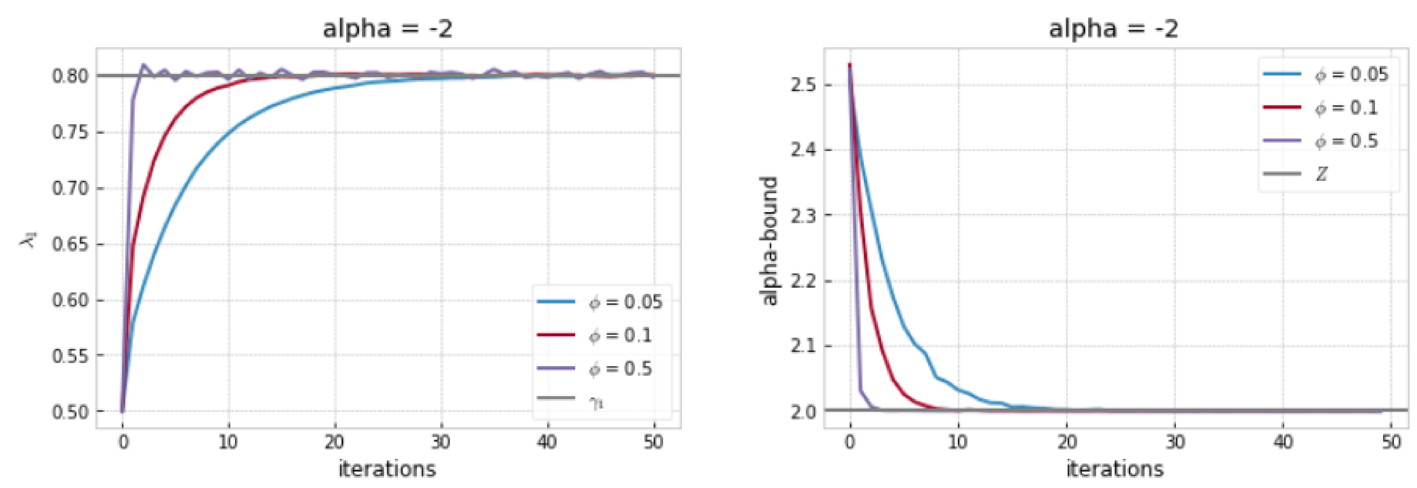
end

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

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

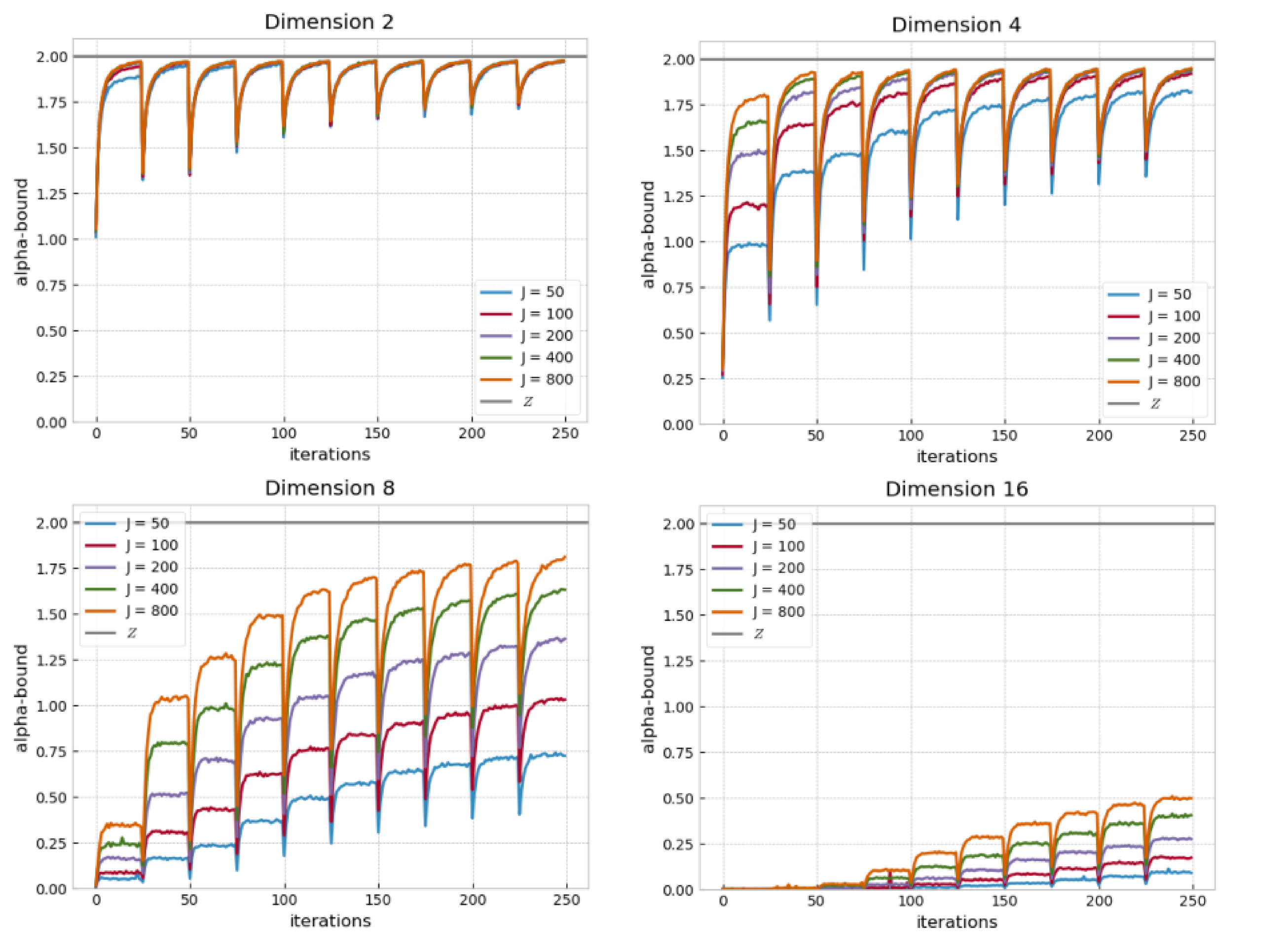
Numerical Experiments

Impact of ϕ (with fixed parameter set $\Theta = \{-2, 2\}$)



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

Impact of d and J (fully adaptive algorithm)



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

References

[1] R. Douc et al. Convergence of adaptive mixtures of importance sampling schemes. AOS 2007

[2] JM Hernández-Lobato et al. Black-box α-divergence Minimization. ICML 2015.

[3] Y. Li and RE. Turner. Rényi Divergence Variational Inference. NIPS 2016.