Adaptive Monte Carlo methods for Complex Models

Kamélia Daudel[1], Randal Douc[2], François Portier[1] and François Roueff[1]

[1] Télécom Paris, [2] Télécom SudParis

What this PhD is about

▷ Bayesian Inference: **model** a phenomenon based on a set of observed **data** while incorporating **uncertainty** in the model parameters.

▷ Why is it difficult? The context of **Big data** often requires **complex models**, rendering cores quantities in Bayesian Inference **intractable** (e.g posterior density, predictive distribution).

▷ In this PhD: our goal is to build novel **scalable** Approximated Bayesian Inference algorithms at the intersection of **Monte Carlo** (MC) and **Variational Inference** (VI) methods to better approximate the **posterior density**.

Problem statement

▷ Target: Posterior density of the latent variable y given the data $\mathcal{D}$:

$$p(y|\mathcal{D}) = \frac{p(y,\mathcal{D})}{\int_{\mathsf{Y}} p(y,\mathcal{D})\nu(\mathrm{d}y)}$$

▷ Goal in VI: choose a **measure of discrepancy** D and an **approximating family** $\mathcal{Q}$; then find

$$\inf_{q \in \mathcal{Q}} D(q||p(\cdot|\mathcal{D})) \tag{1}$$

Typically, D is the **forward Kullback-Leibler** (fKL) divergence and $\mathcal{Q}$ is a **parametric** family

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

⇝ Problems: (i) posterior **variance underestimation** due to the fKL (ii) $\mathcal{Q}$ is sometimes **not large enough** to capture $p(\cdot|\mathcal{D})$ (see figure).

▷ Our approach of (1): D is the **α-divergence D_α**

$$D_\alpha(q||p(\cdot|\mathcal{D})) = \int_{\mathsf{Y}} f_\alpha\left(\frac{q(y)}{p(y|\mathcal{D})}\right) p(y|\mathcal{D})\nu(\mathrm{d}y)$$

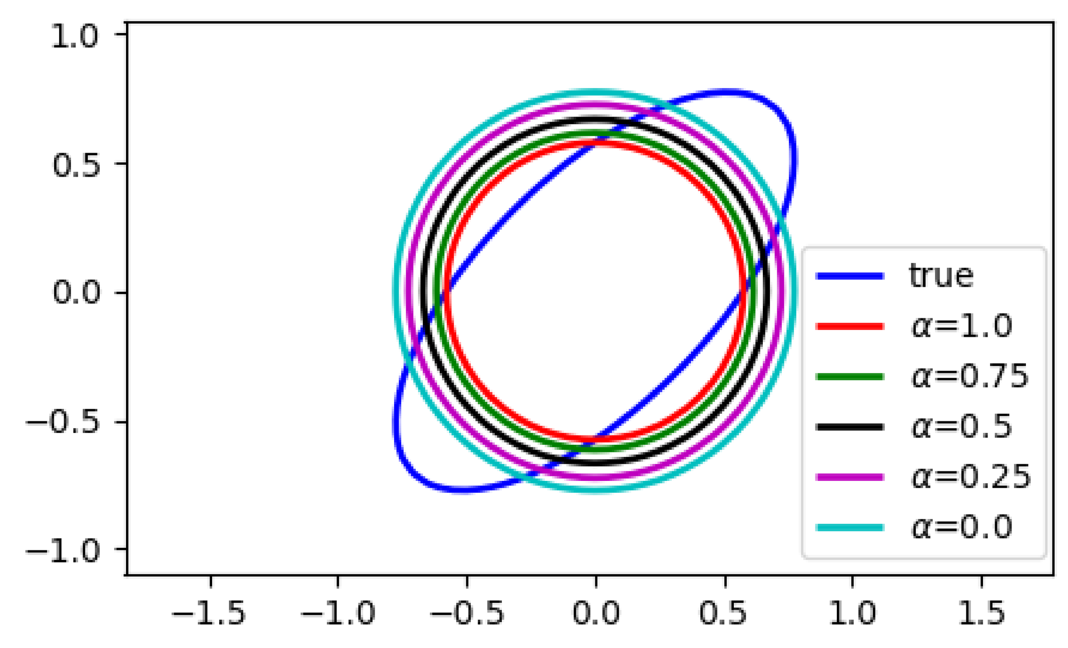
and **enrich** $\mathcal{Q}$ by considering either

$$[1] \quad \left\{q : y \mapsto \int_{\mathsf{T}} \mu(\mathrm{d}\theta)k(\theta,y) \ : \ \mu \in \mathsf{M}\right\}$$

$$[2] \quad \left\{q : y \mapsto \sum_{j=1}^{J} \lambda_j k(\theta_j, y) \ : \ \boldsymbol{\lambda} \in \mathcal{S}_J, \Theta \in \mathsf{T}^J\right\},$$

where $(\boldsymbol{\lambda}, \Theta) = (\lambda_j, \theta_j)_{1 \leqslant j \leqslant J}$, $\mathcal{S}_J$: simplex of $\mathbb{R}^J$.

VI according to α: an illustration



Mean-Field approximation with varying values of α for a toy Bayesian Linear Regression model

First paper [1]

▷ Proposed algorithm: the (α, Γ)-descent

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

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

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

Given an initial μ_1, $(\mu_n)_{n\geqslant 1}$ is defined by

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

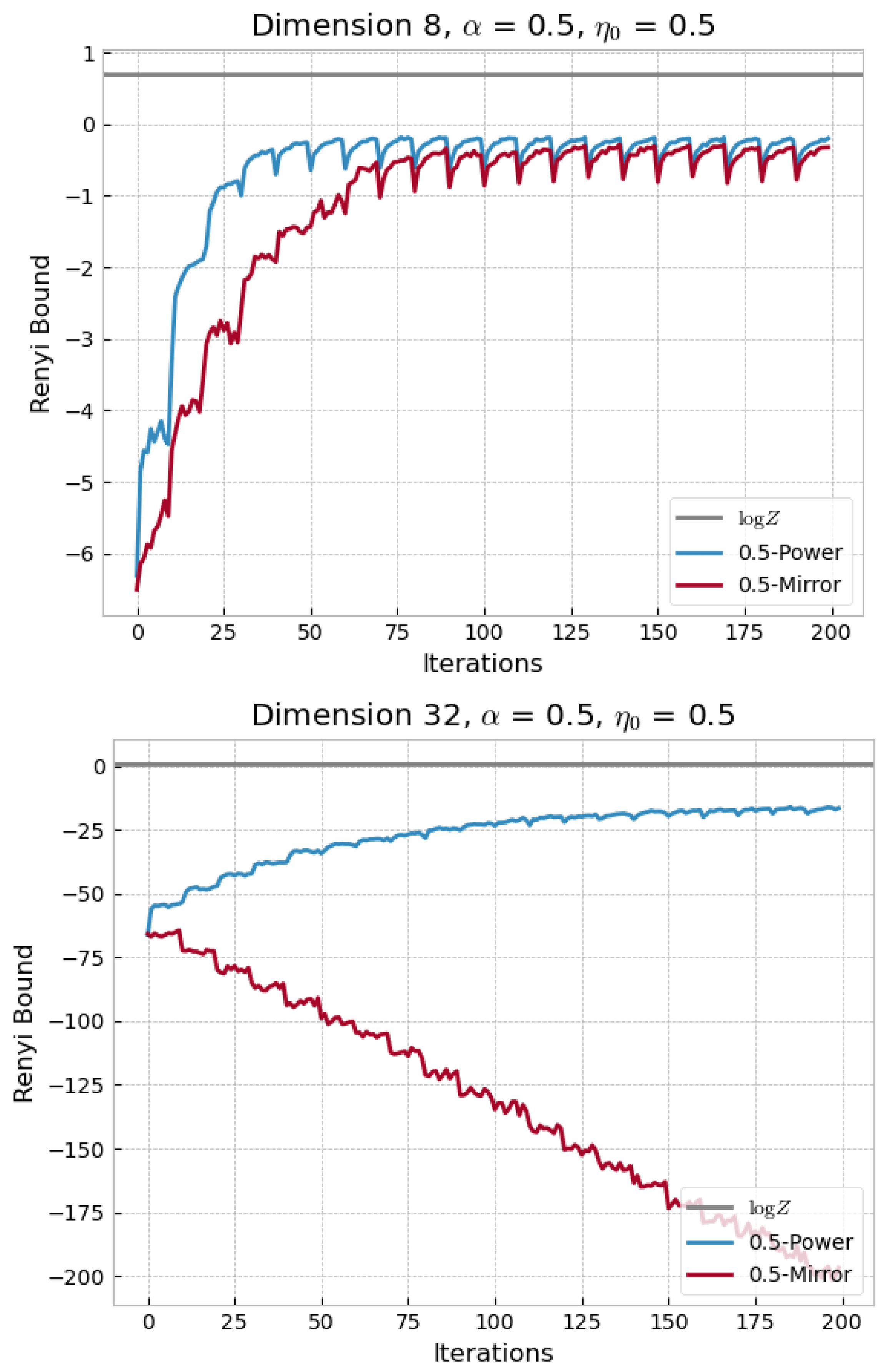
Here, f_α is the **convex** function defined by

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

▶ Findings:

1. Sufficient conditions for a **systematic decrease** in the α-divergence; convergence results/rates.
2. Recovers the **Mirror Descent** for $\Gamma(v) = e^{-\eta v}$.
3. Novel algorithm: **Power Descent** with $\Gamma(v) = [(\alpha-1)v+1]^{\eta/1-\alpha}$.
4. Applicable to **Mixture weights optimisation** for any kernel K using **MC methods**.
5. **Empirical benefit** of using the Power descent (see figure on the right).

Numerical experiments for [1]



Comparison between the Power Descent and the Mirror Descent as the dimension grows

Second paper [2]

Algorithm 1 optimises $\boldsymbol{\lambda}$ by decreasing the α-divergence at each step, while keeping Θ **fixed**.

What about Θ?

▶ Findings:

1. Sufficient conditions for a systematic decrease of the α-divergence:

$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log\left(\frac{\lambda_{j,n+1}}{\lambda_{j,n}}\right)\nu(\mathrm{d}y) \leqslant 0$$

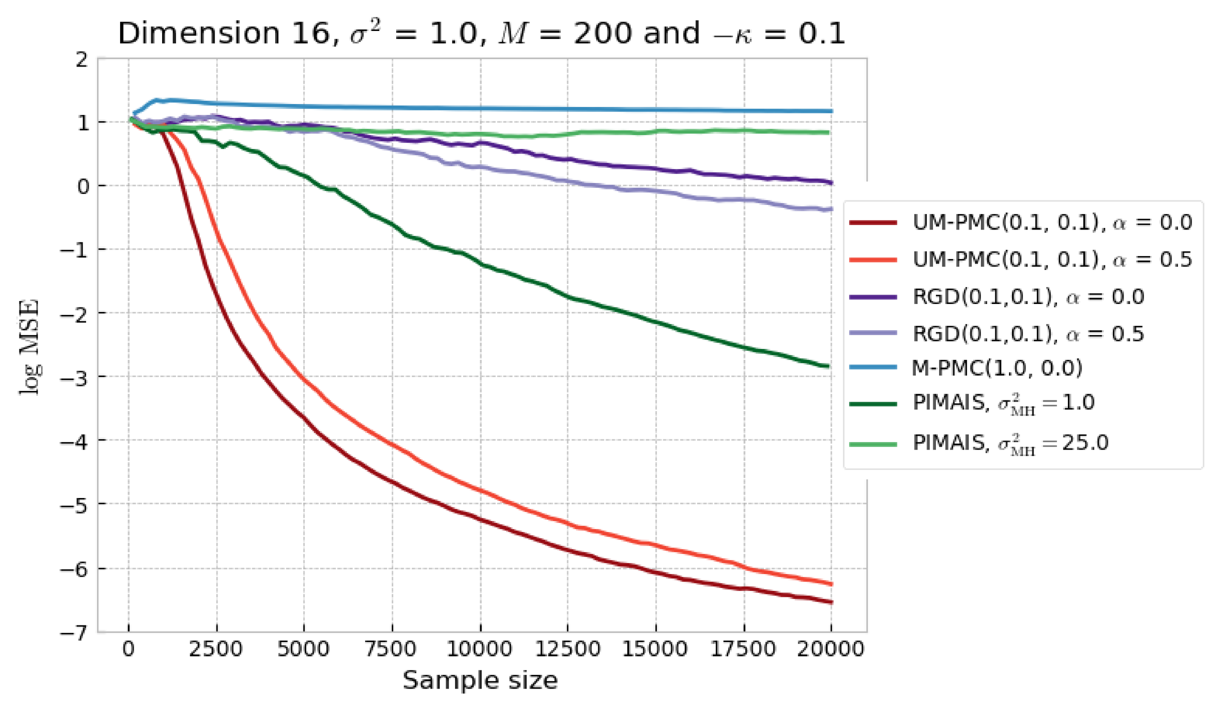
$$\int_{\mathsf{Y}} \sum_{j=1}^{J} \lambda_{j,n} \frac{\gamma_{j,\alpha}^n(y)}{\alpha - 1} \log\left(\frac{k(\theta_{j,n+1},y)}{k(\theta_{j,n},y)}\right)\nu(\mathrm{d}y) \leqslant 0\ .$$

$\boldsymbol{\lambda}$ and Θ are updated simultaneously!!!

Second paper [2] continued

2. Valid updates based on the **Power Descent** for the mixture weights $\boldsymbol{\lambda}$.
3. As for Θ: **explicit** updates when k is Gaussian with **Gradient Descent** (GD) as a special case.
4. Recovers the **M-PMC** algorithm when $\alpha = 0$ (Integrated EM).
5. Applicable to **Gaussian Mixture Models** using **MC methods**.
6. **Empirical benefits**: outperforms the M-PMC algorithm and GD-based algorithms (see figure below).

Numerical experiments for [2]



Comparison between our approach (UM-PMC) and existing methods in the literature.

Conclusion & Perspectives

▶ Novel framework for **mixture models** optimisation with **theoretical guarantees** and **numerical advantages**.

▶ **Future work**: additional convergence rates, variance reduction methods, alternative divergence...

References

[1] K. Daudel, R. Douc, and F. Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *To appear in the Annals of Statistics*, 2020.

[2] K. Daudel, R. Douc, and F. Roueff. Monotonic alpha- divergence minimisation. *Submitted*, 2021.