Mixture weights optimisation for Alpha-Divergence Variational Inference

Kamélia Daudel (Télécom Paris and University of Oxford) and Randal Douc (Télécom SudParis)

Introduction

▷ Variational Inference with the **exclusive KL** and a **parametric** variational family of the form

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

has some known **limitations** : (i) the exclusive KL leads to **posterior variance underestimation** and has difficulty capturing **multimodality** (ii) $\mathcal{Q}$ can be **too restrictive** to capture complex posterior densities.

▷ Idea: Consider the **α-divergence**, let $\Theta = (\theta_1, \ldots, \theta_J) \in \mathsf{T}^J$, $\mathcal{S}_J$ be the simplex of dimension $J > 1$ and

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

Why is that a good idea? (i) The α-divergence **recovers** the exclusive KL when $\alpha \to 1$ and permits to bypass the issues of the exclusive KL **when $\alpha < 1$** (ii) optimising the α-divergence w.r.t the mixture weights $\boldsymbol{\lambda}$ **expands** the traditional parametric variational family and enables to **select** the mixture components **according to their overall importance** in the set of component parameters.

How to do it? The **Power Descent** algorithm from [1] carries out the mixture weights optimisation **regardless of how** Θ is obtained. This **gradient-based** procedure (it notably involves a learning rate η) is defined **for all** $\alpha \neq 1$ and it **outperforms** the Entropic Mirror Descent when $\alpha < 1$ as d increases.

▷ **Problems** :

1. The convergence result for the Power Descent in [1] **assumes the existence of the limit when $\alpha < 1$**.
2. The Power Descent is defined for $\alpha \neq 1$.
3. **No convergence rate** is available for the Power Descent when $\alpha < 1$.

Contributions

We make the three following contributions:

1. We prove the **full convergence** of the Power Descent **towards the optimal mixture weights when $\alpha < 1$** [Theorem 2].

2. We investigate the **extension** to the case $\alpha = 1$ and show that we obtain an Entropic Mirror Descent performing exclusive KL minimisation [Proposition 1].

3. We introduce the **Rényi Descent**, an algorithm **closely-related** to the Power Descent that converges at an **$O(1/N)$ rate when $\alpha < 1$** [Theorem 3].

The Rényi Descent :

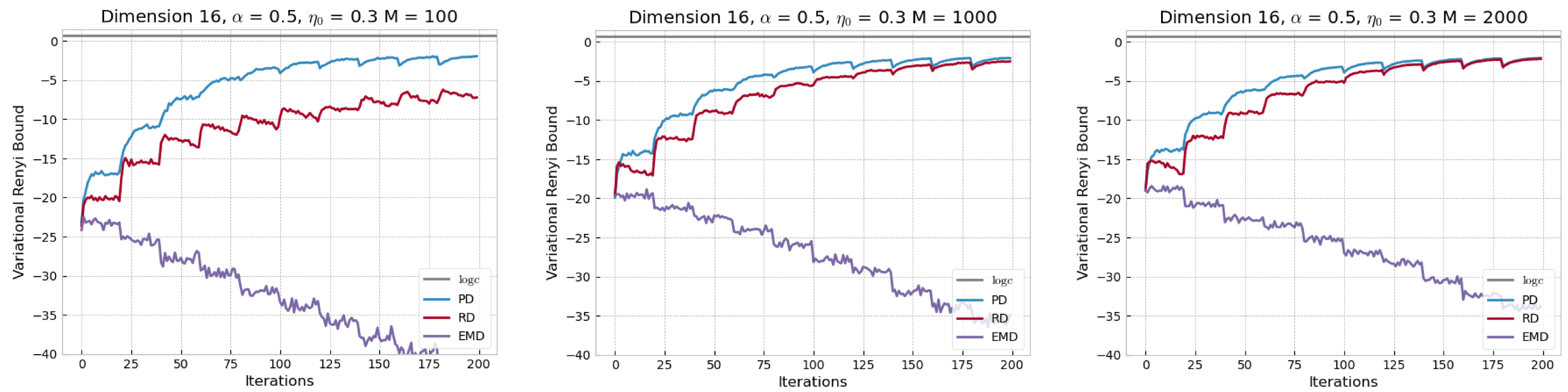
- shares the same **first-order approximation** as the Power Descent,
- can be linked to **Entropic Mirror Descent** steps applied to the **Variational Rényi (VR)** Bound from [2] (hence the name!),
- differs from the Entropic Mirror Descent considered in [1] as it uses **adaptive learning rates**. This shows that a deeper connection runs between Power Descent and Entropic Mirror Descent **beyond the one identified in [1]**.

NB : Our work contributes towards **deriving convergence results of variational objective functions**, with the particularity that we focus on mixture weights updates in the optimisation procedures, which are carried out **for general choices of kernel k**.

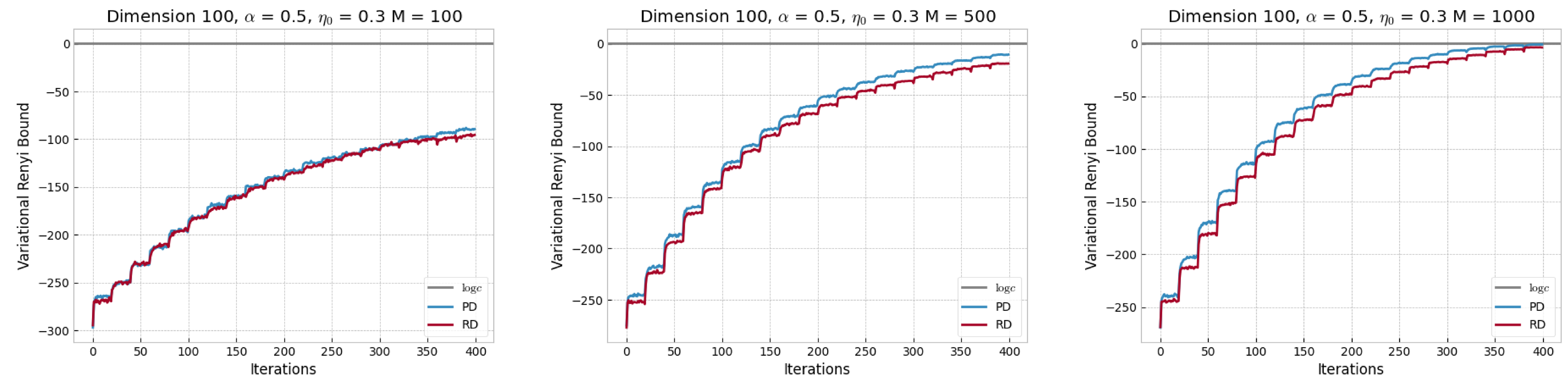
Numerical experiments

The Power Descent and the Rényi Descent are **gradient-based** algorithms. In practice, the gradients are approximated by using **Monte Carlo methods**. Since these algorithms act on the mixture weights $\boldsymbol{\lambda}$ only, they are paired up with an **Exploration step** that **updates the components parameters Θ** in our numerical experiments.

▷ In dimension $d = 16$ with an increasing number of Monte Carlo samples M...



▷ In dimension $d = 100$ with an increasing number of Monte Carlo samples M...



These figures permit us to **illustrate the newly-found proximity** between the Power Descent (PD) and the Rényi Descent (RD), as opposed to the Entropic Mirror Descent (EMD) considered in [1].

Discussion

In their stochastic versions, the Power Descent applies the function $\Gamma(v) = [(\alpha - 1)v + 1]^{\eta/(1-\alpha)}$ to an **unbiased** estimator of the gradient, while the Rényi Descent applies the function $\Gamma(v) = e^{-\eta v}$ to a **biased** estimator of the gradient.

Finding which approach is most suitable between biased and unbiased α-divergence minimisation is still an active area of research in Variational Inference. Our work sheds light on links between unbiased and biased α-divergence methods **beyond the framework of stochastic gradient descent**, as both the *unbiased* Power Descent and the *biased* Rényi Descent share the same first-order approximation.

References

[1] Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250 – 2270, 2021.

[2] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1073–1081. Curran Associates, Inc., 2016.