Multimodal optimization: a variational approach

Tâm Le Minh[1], Jacopo Iollo[1], Julyan Arbel[1], Thomas Möllenhoff[2], Mohammad Emtiyaz Khan[2], Florence Forbes[1]

[1]Univ. Grenoble Alpes, Inria, France [2]RIKEN-AIP, Japan

Problem statement

Fitness function

- $\ell\colon \mathbb{R}^d \to \mathbb{R} \rightsquigarrow$ can be **non-convex** and **multimodal**.

Goals

- locate **multiple local and global maxima** in one run,
- identify the **"widest" maxima**.

Approach

Our optimization algorithm is inspired by the **Bayesian learning rule** [2], using:

- a **variational formulation** of the problem and a relevant **variational family**,
- a **procedure** based on **natural gradients**.

In addition, we use an **annealed objective function**.

Annealed variational objective

Variational formulation

$$q^* = \arg\max_{q \in \mathcal{P}(\mathbb{R}^d)} \mathbb{E}_q[\ell(\boldsymbol{\xi})].$$

The solutions are of the form $q^* = \sum_{i=1}^L c_i \delta_{\boldsymbol{\xi}_i^*}$, where

- $(\boldsymbol{\xi}_i^*)_{1\le i\le L}$ are the global maxima of ℓ,
- $(c_i)_{1\le i\le L}$ are weights in $[0,1]$ such that $\sum_{\ell=1}^L c_i = 1$.

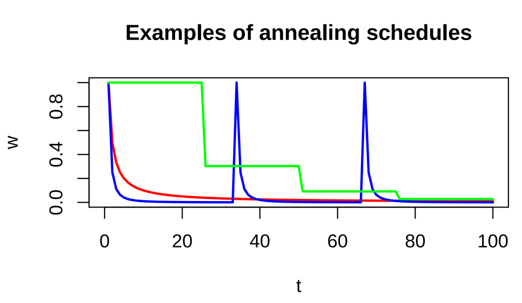
Entropy penalty

$$q^{*,\omega} = \arg\max_{q \in \mathcal{P}(\mathbb{R}^d)} \mathbb{E}_q[\underbrace{\ell(\boldsymbol{\xi}) - \omega \log q(\boldsymbol{\xi})}_{f_\omega(\boldsymbol{\xi})}], \text{ where } \omega > 0.$$

Convergence result

$$q^{*,\omega} \underset{\omega\to 0}{\longrightarrow} q^* = \sum_{i=1}^L c_i^* \delta_{\boldsymbol{\xi}_i^*},$$

where for all $1 \le i \le L$, $c_i^* \propto \det(-\nabla^2 \ell(\boldsymbol{\xi}_i^*))^{-1/2}$.



Annealing schedule

For optimization, set $(\omega_t)_{t\ge 1}$ with $\omega_t \underset{t\to 0}{\longrightarrow} 0$.

Variational family: Gaussian mixtures

Search restriction to Gaussian mixtures with K components

$$q_{\boldsymbol{\Lambda}}(\boldsymbol{\xi}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1}).$$

Parameterization

$$\boldsymbol{\Lambda} = (\log(\pi_1/\pi_K), \dots, \log(\pi_{K-1}/\pi_K), \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K), \text{ where } \boldsymbol{\lambda}_k = (\boldsymbol{S}_k \boldsymbol{m}_k, -\boldsymbol{S}_k/2).$$

Targeted result

- the means $(\boldsymbol{m}_k)_{1\le k\le K}$ converge to different modes of ℓ,
- the covariance matrices $(\boldsymbol{S}_k^{-1})_{1\le k\le K}$ shrink to 0,
- the weights $(\pi_k)_{1\le k\le K}$ give information on the curvature at the modes.

Effect of entropy penalty

$\omega > 0$ induces an (intra- and) inter-component repulsion term.

Optimization: Natural gradient ascent

Natural gradient ascent: update rule

$$\boldsymbol{\Lambda}_{t+1} = \boldsymbol{\Lambda}_t + \rho_t \boldsymbol{F}(\boldsymbol{\Lambda}_t)^{-1} \nabla_{\boldsymbol{\Lambda}} \underbrace{\mathbb{E}_{q_{\boldsymbol{\Lambda}_t}}[f_{\omega_t}(\boldsymbol{\xi}; \boldsymbol{\Lambda}_t)]}_{\mathcal{L}_{\omega_t}(\boldsymbol{\Lambda}_t)}.$$

- $\boldsymbol{F}(\boldsymbol{\Lambda}_t)$ is the **Fisher information matrix**.
- The natural gradient gives the steepest direction in the Riemannian manifold (parameter space) [1].
- Convergence is quick, but computation of $\boldsymbol{F}(\boldsymbol{\Lambda}_t)^{-1}$ is usually involving.

Case of Gaussian mixtures [4]

$$\boldsymbol{S}_{k,t+1} = \boldsymbol{S}_{k,t} - \frac{2\rho_t}{\pi_{k,t}} \nabla_{\boldsymbol{S}_k^{-1}} \mathcal{L}_{\omega_t}(\boldsymbol{\Lambda}_t),$$

$$\boldsymbol{m}_{k,t+1} = \boldsymbol{m}_{k,t} + \frac{\rho_t}{\pi_{k,t}} \boldsymbol{S}_{k,t+1}^{-1} \nabla_{\boldsymbol{m}_k} \mathcal{L}_{\omega_t}(\boldsymbol{\Lambda}_t),$$

$$\log(\pi_{k,t+1}/\pi_{K,t+1}) = \log(\pi_{k,t}/\pi_{K,t}) + \rho_t \nabla_{\pi_k} \mathcal{L}_{\omega_t}(\boldsymbol{\Lambda}_t).$$

Weight gradient

$$\nabla_{\pi_k} \mathcal{L}_\omega(\boldsymbol{\Lambda}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1})}[f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})] - \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_K, \boldsymbol{S}_K^{-1})}[f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})].$$

Black-box method

$$\nabla_{\boldsymbol{m}_k} \mathcal{L}_\omega(\boldsymbol{\Lambda}) = \pi_k \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1})}[\boldsymbol{S}_k(\boldsymbol{\xi} - \boldsymbol{m}_k) f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})],$$

$$\nabla_{\boldsymbol{S}_k^{-1}} \mathcal{L}_\omega(\boldsymbol{\Lambda}) = \frac{\pi_k}{2} \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1})}[(\boldsymbol{S}_k(\boldsymbol{\xi} - \boldsymbol{m}_k)(\boldsymbol{\xi} - \boldsymbol{m}_k)^T \boldsymbol{S}_k - \boldsymbol{S}_k) f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})].$$

Bonnet and Price's theorems

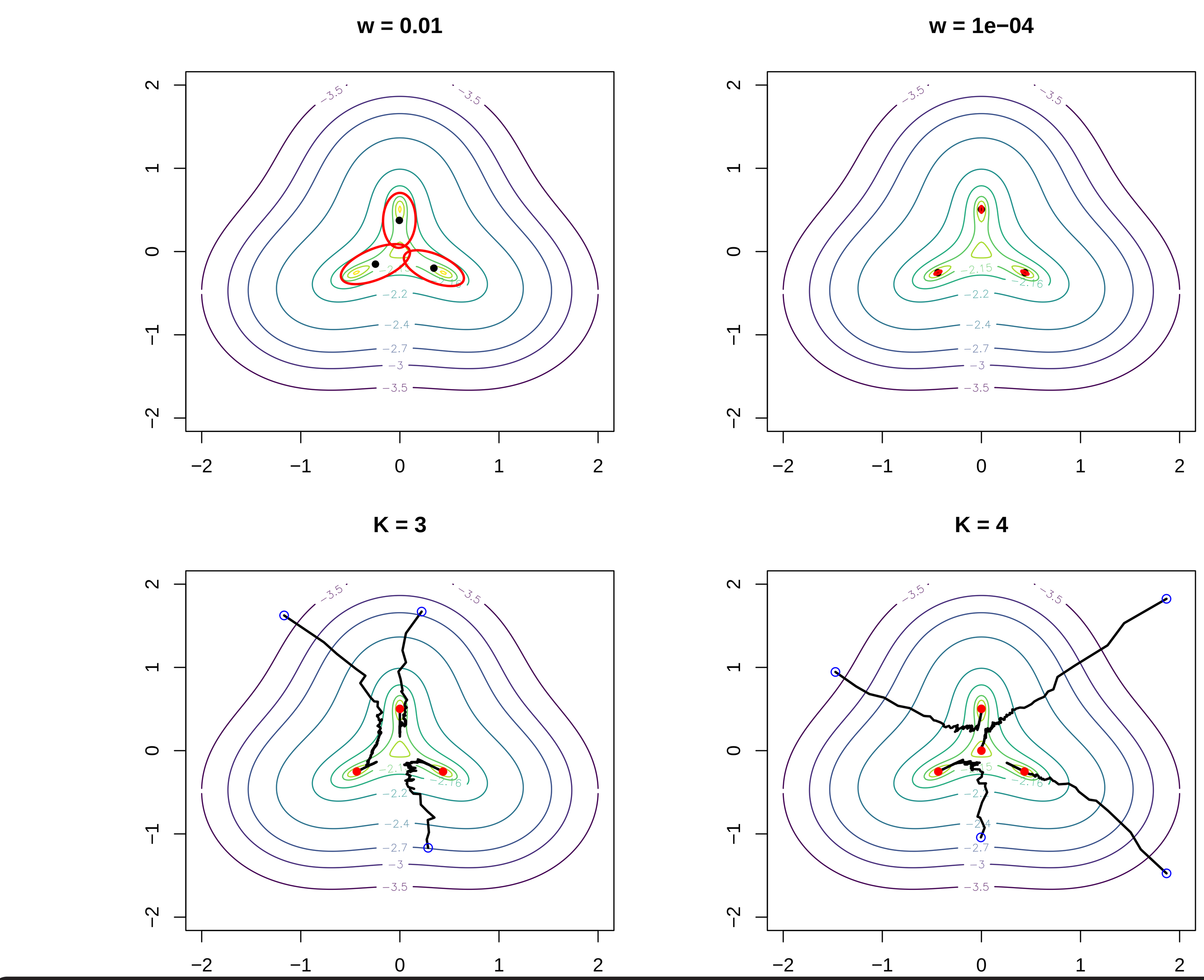
$$\nabla_{\boldsymbol{m}_k} \mathcal{L}_\omega(\boldsymbol{\Lambda}) = \pi_k \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1})}[\nabla_{\boldsymbol{\xi}} f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})],$$

$$\nabla_{\boldsymbol{S}_k^{-1}} \mathcal{L}_\omega(\boldsymbol{\Lambda}) = \frac{\pi_k}{2} \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1})}[\nabla_{\boldsymbol{\xi}}^2 f_\omega(\boldsymbol{\xi}; \boldsymbol{\Lambda})].$$

Simulations

Example 1: Gaussian mixture with 3 components, 3 global + 1 local modes

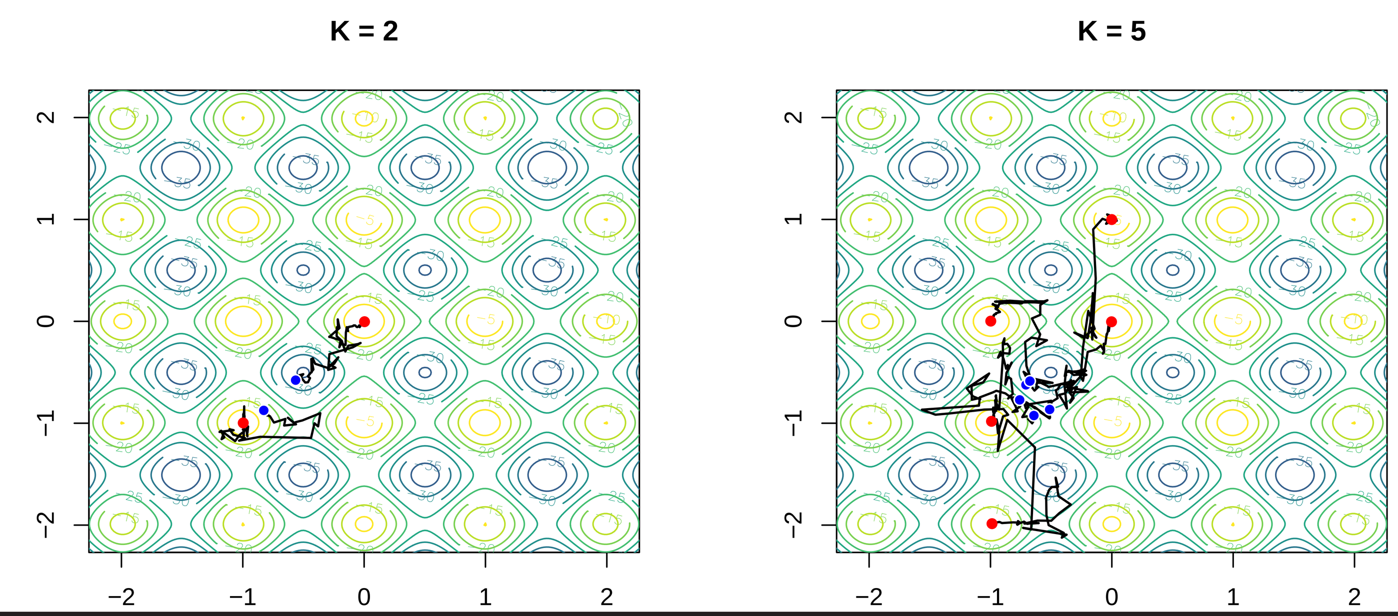
Effect of the entropy penalty: approached solutions $q_{\boldsymbol{\Lambda}^{*,\omega}}$ for a fixed $\omega > 0$, $K = 3$



▶ The entropy term helps to **prevent the means from converging to similar modes**.

Example 2: Rastrigin function

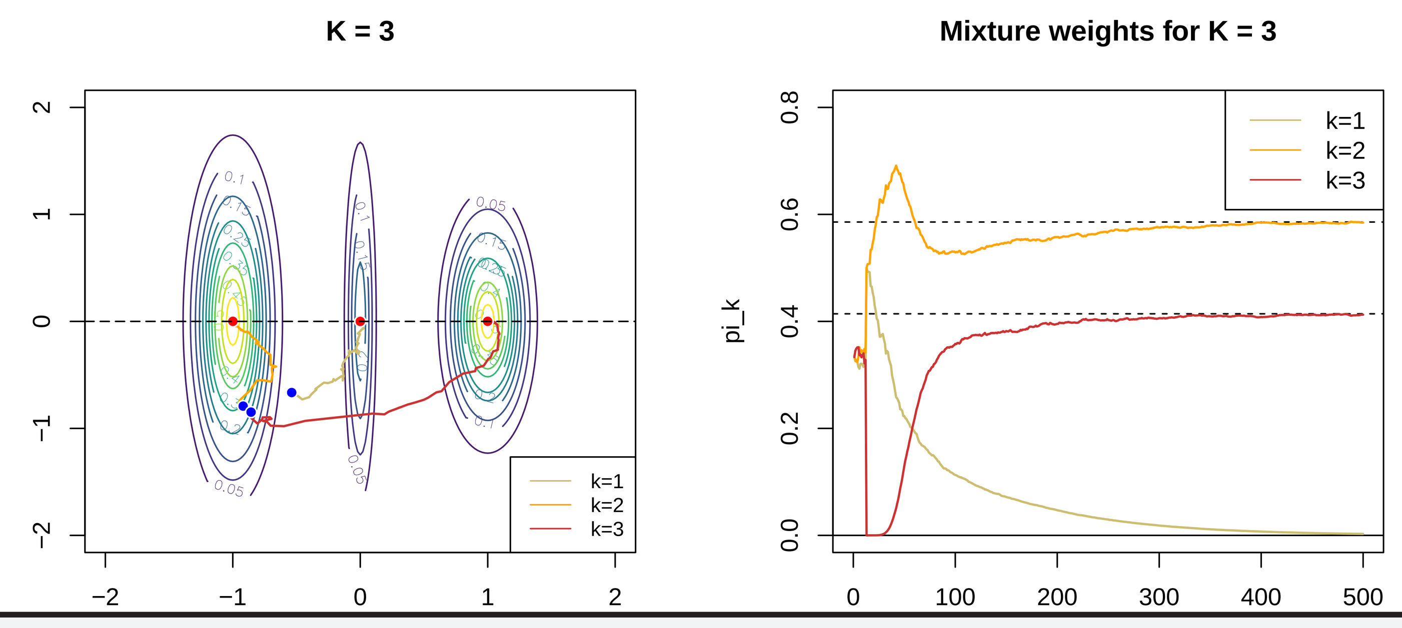
Paths of the means $(\boldsymbol{m}_k)_{1\le k\le K}$ under annealing schedule $\omega_t = \omega_0/t^{1.5}$



▶ The optimization problem is non-convex, therefore **local modes can be found**.

Example 3: Gaussian mixture with 3 components, 2 global + 1 local modes

Paths of the means $(\boldsymbol{m}_k)_{1\le k\le K}$ under annealing schedule $\omega_t = \omega_0/t$



▶ The **weights are proportional to the determinant of the Hessian** matrix of ℓ at the "highest" modes found.

Future work

- Elements from **evolutionary algorithms** [5] can be incorporated to find the global maxima more easily.
  $\rightsquigarrow$ However, this means local maxima are less likely to be detected.
- **Application** to posterior mode identification in **Bayesian inverse problems** [3].

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neur. Comp.*, 1998.
[2] M. E. Khan and H. Rue. The Bayesian learning rule. *JMLR*, 2023.
[3] T. Le Minh, J. Arbel, T. Möllenhoff, M. E. Khan, and F. Forbes. Natural variational annealing for multimodal optimization. In preparation.
[4] W. Lin, M. E. Khan, and M. Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *ICML*, 2019.
[5] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *JMLR*, 2017.