Quantized Variational Inference
Optimal Quantization for ELBO maximization

Amir DIB
ENS Paris-Saclay, Centre Borelli, Université Paris-Saclay, CNRS.

MOTIVATION

Quantized Variational Inference is a new algorithm for Evidence Lower Bound maximization. Optimal **Voronoi Tesselation** produces **variance free** gradients for ELBO optimization at the cost of introducing asymptotically decaying bias. Using the **Quantized Variational Inference framework leads to fast convergence** for both score function and the reparametrized gradient estimator at a comparable computational cost.

OVERVIEW

Let y be the data, z a latent variables and $p(y,z)$ the model. The goal of the Bayesian statistician is to find the best latent variable that fits the data, hence the likelihood $p(z\,|\,y)$. We can approximate the true posterior by maximizing the **Evidence Lower Bound** $\mathscr{L}(\lambda)$ though the variational distribution q_λ thanks to the following decomposition

$$\log p(y) = \underbrace{\mathop{\mathbb{E}}_{z\sim q_\lambda}\left[\log \frac{p(z,y)}{q_\lambda(z)}\right]}_{\mathscr{L}(\lambda)} + \mathrm{KL}\left(q_\lambda(z)\|p(z\,|\,y)\right).$$

Given a sample $\left(X_1^\lambda,\dots,X_N^\lambda\right)$ of size N, typical MCVI procedure consists of a Gradient descent at each step k

$$\lambda_{k+1} = \lambda_k - \alpha_k \underbrace{\frac{1}{N}\sum_{i=1}^{N} \nabla_\lambda f\left(X_i^{\lambda_k}\right)}_{\hat{g}^N_{MC}}.$$

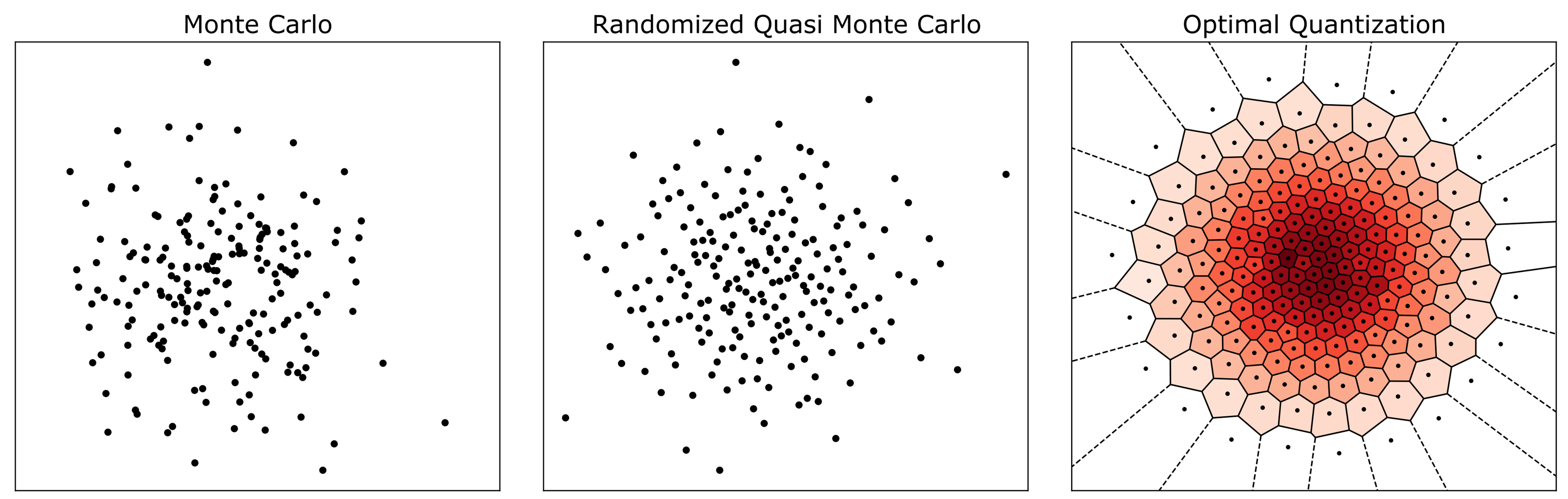
The speed of convergence crucially depends on the expected norm of the gradient and thus on the gradient variance thanks to the bias-variance decomposition

$$\mathbb{E}\,|\,g\,|^2_{\ell_2} = \operatorname{tr}\mathbb{V}g + |\,\mathbb{E}g\,|^2_{\ell_2}.$$

Our approach consists of considering alternative sampling instead of the traditional Monte Carlo. Precisely, we consider the **Optimal Quantizer** or **Voronoï Tesselation** of q_λ at level N, $X^{\Gamma_N,\lambda}$. We thus obtain a **deterministic gradient** allowing for larger step in the optimization procedure by using the following gradient descent scheme

$$\lambda_{k+1} = \lambda_k - \alpha_k \underbrace{\nabla_\lambda \sum_{i=1}^{N} \omega_i^k f\left(X_i^{\Gamma_N,\lambda_k}\right)}_{\hat{g}^N_{OQ}}.$$

OPTIMAL QUANTIZER



Different type of sampling for a bi-variate standard gaussian distribution.

Take $\Gamma_N = \{x_1,\dots,x_N\} \subset \mathbb{R}$, the $L^p_{\mathbb{R}^d}$ quantiser Is defined as the probability measure on the convex subset of probability measure on Γ_N that minimizes the Wassertein distance

$$\inf_{\Gamma\in\mathbb{R}^d} \int_{\mathbb{R}^d} \min_{1\le k\le K} \left|\xi - x_k\right|^p \mu(d\xi).$$

Any expectation can be computed with the cubature formula

$$\mathbb{E}f(X^{\Gamma_N}) = \sum_{i=1}^{N} \mathbb{P}\left(X^{\Gamma_N} = x_i\right) f\left(x_i\right).$$

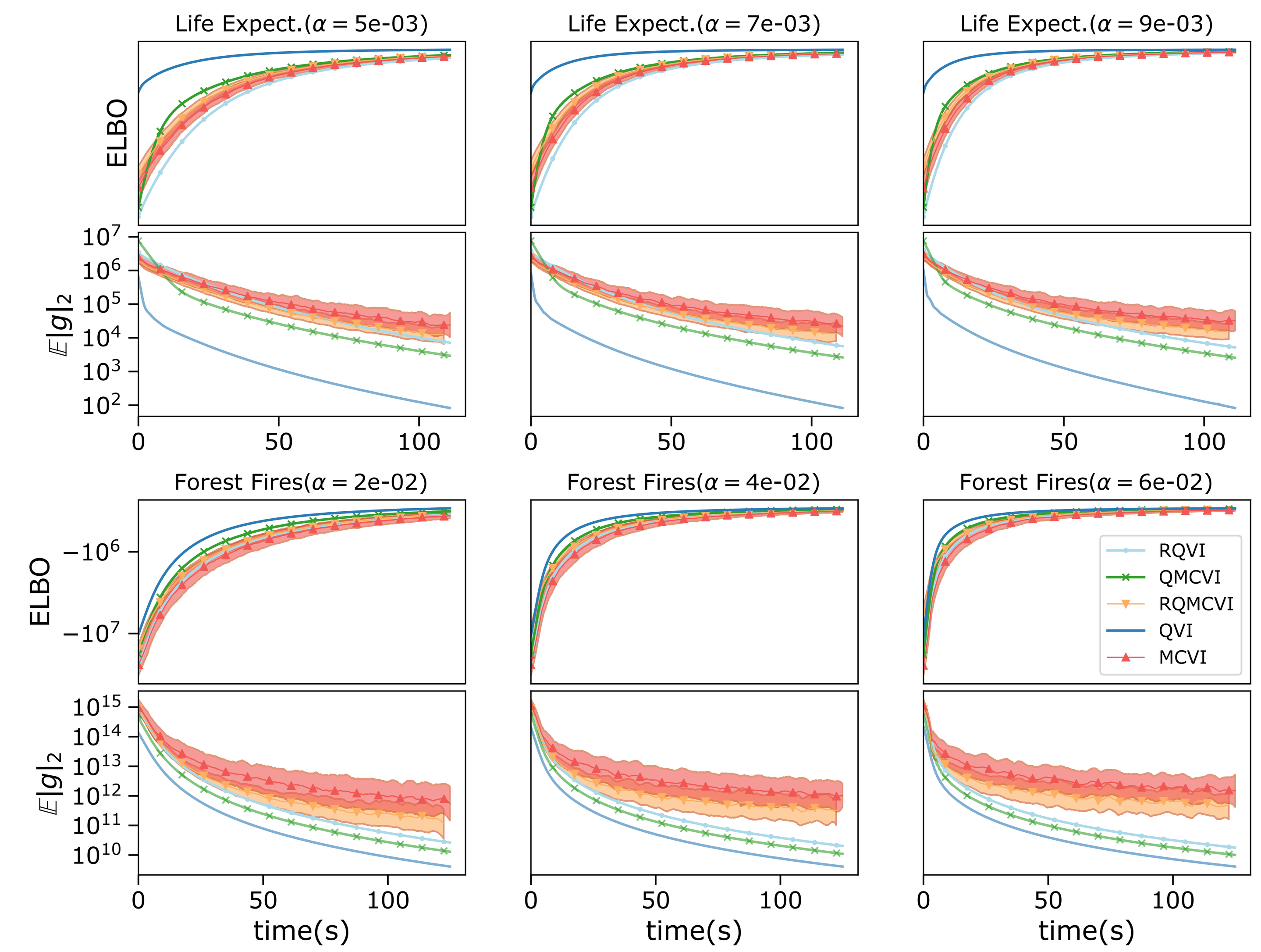
Denoting $\widehat{\mathscr{L}}^N_{OQ}(\lambda)$ the ELBO obtained by sampling with the **OQ** and $\mathscr{L}(\lambda)$ the true ELBO, we can obtain a bound on the produced bias

$$\left|\mathscr{L}(\lambda) - \widehat{\mathscr{L}}^N_{OQ}(\lambda)\right| \le C\, \left\| X^\lambda - X^{\Gamma_N,\lambda} \right\|_2.$$

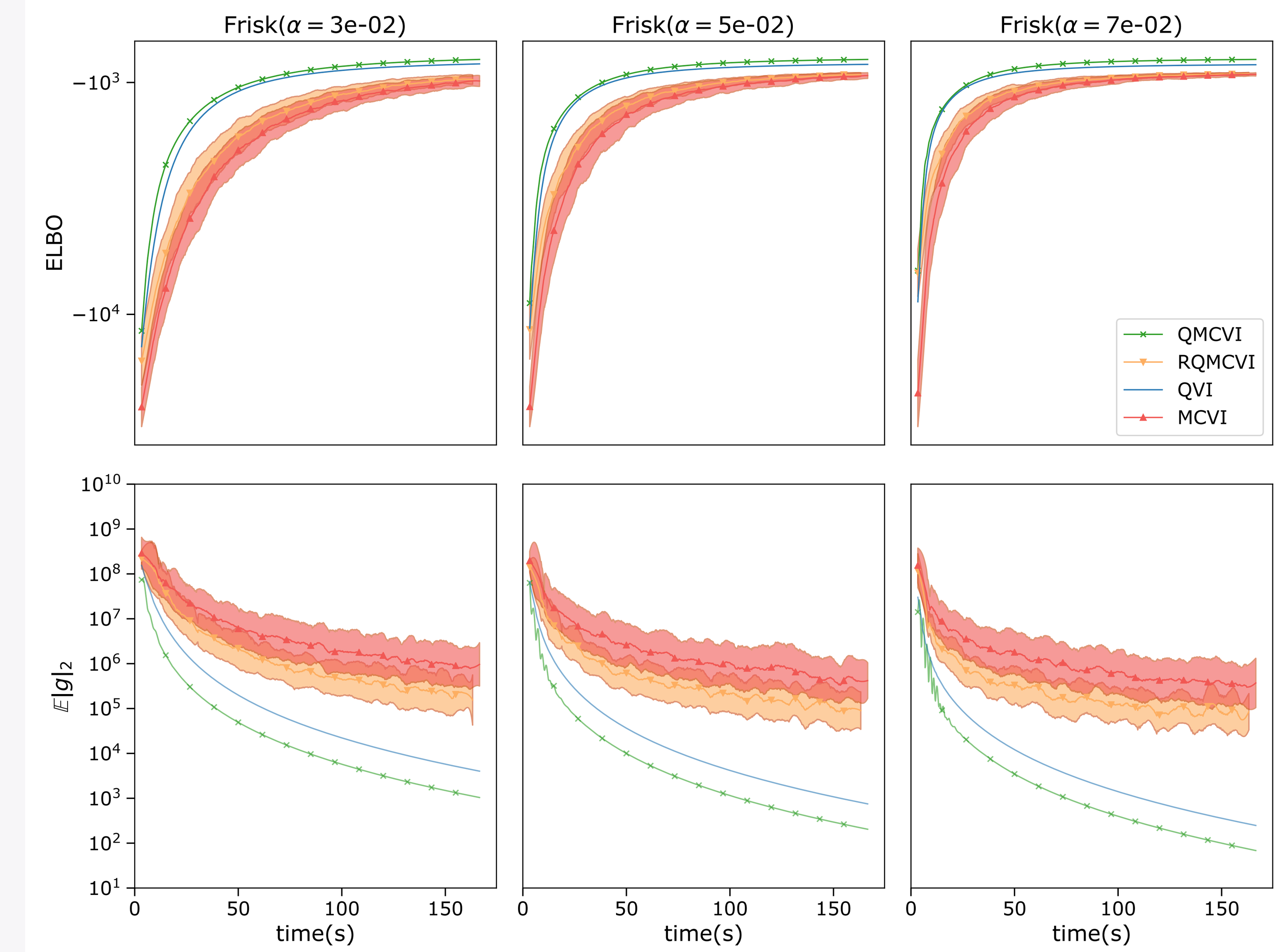
Quantized Variational Inference

Input: y, $p(x,z)$, q_{λ_0}.
Result: Optimal Quantized VI parameters λ_q^*.
while *not converged* **do**
  Get $(X_1^{\Gamma_N,\lambda_k},\dots,X_N^{\Gamma_N,\lambda_k}) \sim q_{\lambda_k}, (w_1^k,\dots,w_N^k)$;
  Compute $\hat{g}^N_{OQ}(\lambda_k) = \nabla_\lambda \sum_{i=1}^N w_i^k H(X_i^{\Gamma_N,\lambda_k})$;
  $\lambda_{k+1} = \lambda_k - \alpha_k \hat{g}^N_{OQ}(\lambda_k)$;
end

EXPERIMENTS



Bayesian Linear Regression.



Hierarchical Poisson Model.

References

Gilles Pagès (2018). "Numerical Probability: An Introduction with Application to Finance". Springer International Publishing

Léon Bottou (2018), Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: SIAM Review 60.2,pp. 223–311.