Quantized Variational Inference

Amir Dib

Centre Borelli
Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, SNCF, ITNOVEM.

Conference on Neural Information Processing Systems, December 2020.

Variationnal Inference

Given data y, a model $p(y,z)$ with latent variable z, we want to approximate the distribution $p(z|y)$. Given a variational distribution q_λ, the following decomposition can be obtained [3]

$$\log p(y) = \underbrace{\mathbb{E}_{z \sim q_\lambda}\left[\log \frac{p(z,y)}{q_\lambda(z)}\right]}_{\text{ELBO } \mathscr{L}(\lambda)} + \underbrace{\mathrm{KL}\left(q_\lambda(z) \| p(z|y)\right)}_{\text{KL-divergence}}. \tag{1}$$

Given data y, a model $p(y,z)$ with latent variable z, we want to approximate the distribution $p(z|y)$. Given a variational distribution q_λ, the following decomposition can be obtained [3]

$$\log p(y) = \underbrace{\mathbb{E}_{z\sim q_\lambda}\left[\log \frac{p(z,y)}{q_\lambda(z)}\right]}_{\text{ELBO } \mathscr{L}(\lambda)} + \underbrace{\mathrm{KL}\left(q_\lambda(z) \| p(z|y)\right)}_{\text{KL-divergence}}. \tag{1}$$

Using the reparametrization trick [1] with noise parameter $X \sim q$ and denoting $X^\lambda = h_\lambda(X)$, the inference problem can be rewritten as finding λ^* such as

$$\lambda^* \in \operatorname{argmax} \mathbb{E}_q\left[f(X^\lambda)\right]. \tag{2}$$

Optimization Procedure

Given a sample $(X_1, \ldots, X_N)$ of size N, typical Monte Carlo Variational Inference (MCVI) consists of a Gradient descent at each step k

$$\lambda_{k+1} = \lambda_k - \alpha_k \underbrace{\frac{1}{N} \sum_{i=1}^{N} \nabla_\lambda f\left(X_i^{\lambda_k}\right)}_{\widehat{g}_{\mathrm{MC}}^N}. \tag{3}$$

Gradient descent descent speed crucially depends on the following quantity

$$\mathbb{E}|g|_{\ell_2}^2 = \operatorname{tr} \mathbb{V} g + |\mathbb{E} g|_{\ell_2}^2. \tag{4}$$

Optimization Procedure

Our approach consists of considering alternative sampling instead of the traditional MC. Precisely, we consider the optimal quantizer [2] at level N, $X^{\Gamma_N,\lambda}$, resulting in the following gradient descent scheme

$$\lambda_{k+1} = \lambda_k - \alpha_k \nabla_\lambda \sum_{i=1}^{N} \omega_i^k f\left(X_i^{\Gamma_N,\lambda_k}\right) \tag{5}$$

with $\omega_i^k = \mathbb{P}\left(X^{\Gamma_N^k,\lambda_k} = x_i^k\right)$.

Experiments

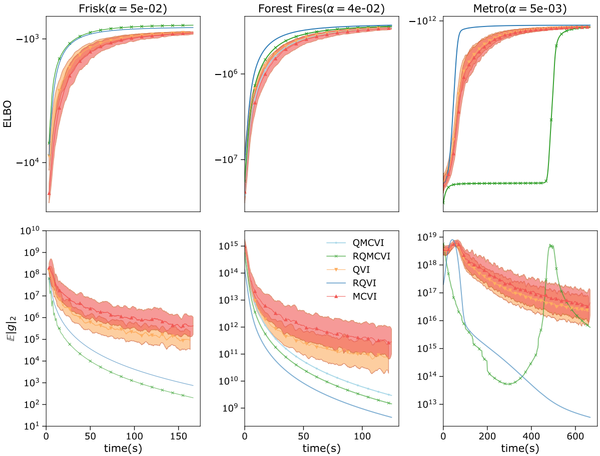

Figure: ELBO (first row, log scale) and expect gradient norm (second row, log scale) during the optimization procedure for various models: Poisson Generalized Linear Model (left), Bayesian Linear Regression (center) and Bayesian Neural Network (right) as function of time.

Bibliography

[1] Durk P Kingma, Tim Salimans, and Max Welling. “Variational Dropout and the Local Reparameterization Trick”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 2575–2583.

[2] Gilles Pagès. *Numerical Probability: An Introduction with Applications to Finance*. en. Universitext. Springer International Publishing, 2018.

[3] L. K. Saul, T. Jaakkola, and M. I. Jordan. “Mean Field Theory for Sigmoid Belief Networks”. en. In: *Journal of Artificial Intelligence Research* 4 (Mar. 1996), pp. 61–76.