Kernelized Wasserstein Natural Gradient

Arbel, M., Gretton, A., Li, W., and Montúfar, G. (2019).

Léon Zheng

M2 MVA "Mathématiques, Vision, Apprentissage"
Computational Optimal Transport 2019

leon.zheng@polytechnique.org

January 7, 2020

Presentation of the problem: efficient estimation of NG

Machine learning problem formulation: $\Theta \in \mathbb{R}^q$, ρ_θ distribution on $\Omega \in \mathbb{R}^d$

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) := \mathcal{F}(\rho_\theta) \tag{1}$$

Conditioning problem in optimization problems

- SGD methods are effective, but suffer from ill-conditioning
- Natural gradient (NG): pull-back of a given metric on the distributions space. Requires to invert the metric tensor.

Goal: estimate the natural gradient with a good trade-off between accuracy and computational cost.

Contributions of the paper

- Estimate the Wasserstein natural gradient using kernel estimators
- Can be used to optimize deep neural network on `Cifar100` data set.

Proposed method: estimate KWNG with kernel estimators

$G_W(\theta)$ represents the pulled-back metric of Wasserstein 2 by $\theta \mapsto \rho_\theta$.

Wasserstein natural gradient (WNG):

$$\nabla^W \mathcal{L}(\theta) := -\arg\min_{u \in \mathbb{R}^q} \mathcal{L}(\theta) + \nabla \mathcal{L}(\theta)^T u + \frac{1}{2} u^T G_W(\theta) u \tag{2}$$

1. Formulate a functional optimization problem for $G_W(\theta)$ using duality
2. Restriction of this problem on a RKHS $\mathcal{H}$ associated to kernel k
3. Add regularization terms and obtain a quadratic saddle point problem
4. Define the Kernelized WNG (KWNG) as the solution of this problem
5. Estimate the KWNG with Nyström methods: low-rank approximation of kernels with N samples and M basis points
6. Obtain a closed form of this estimator of KWNG called $\widehat{\nabla^W \mathcal{L}(\theta)}$

Theoretical analysis

Estimator of KWNG: C, K, T include partial derivatives of the kernel k, evaluated at samples $(X_n)_{1\leq n\leq N}$ and basis points $(Y_m)_{1\leq m\leq M}$.

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(D(\theta)^{-1} - D(\theta)^{-1}T^T\left(TD(\theta)^{-1}T^T + \lambda\epsilon K + \frac{\epsilon}{N}CC^T\right)^{\dagger} TD(\theta)^{-1}\right)\widehat{\nabla \mathcal{L}(\theta)}$$

with $D(\theta)$, λ, ϵ regularization terms.

Theorem (Consistency of the estimator)

For N large enough, $M \sim dN^{-\frac{1}{2b+1}}\log(N)$, with probability $1-\delta$:

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}(N^{-\frac{2b}{2b+1}}) \tag{3}$$

where $b = \min(1, \alpha + \frac{1}{2})$ and $\alpha \geq 0$ defines the smoothness of the problem.

Discussions

- Choice of the kernel: Gaussian? Sigmoid? Laplacian? Polynomial?
- Computational cost: $\mathcal{O}(M^2 dN + qM^2 + M^3)$

Numerical findings: consistency of the estimated KWNG

Experimental context:

- Multivariate normal model with diagonal covariance matrix
- Wasserstein squared distance as loss function with target ρ_θ^*

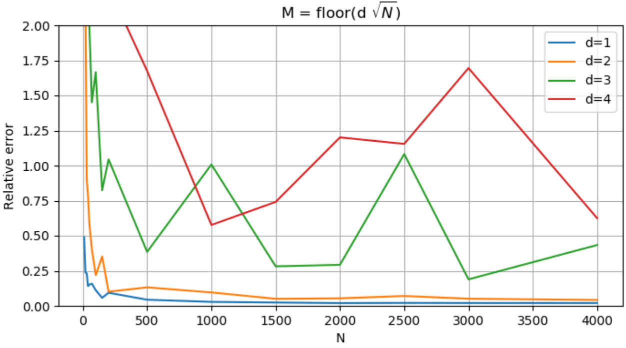

Figure: Consistency of the estimator, with Gaussian kernel. N number of samples, M number of basis points, d the dimension of the problem.

Numerical findings: estimated KWNG in descent method

Steepest descent method:

$$\theta_{t+1} = \theta_t - \alpha_{t+1} \widetilde{\nabla} \mathcal{L}(\theta_t) \tag{4}$$

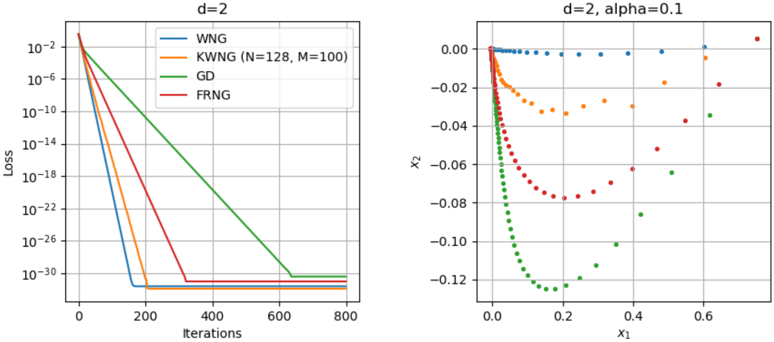

Figure: Comparison of steepest descent methods with different type of gradients: exact Wasserstein natural gradient (WNG), estimator of KWNG (KWNG), Euclidean gradient (GD) and exact Fisher-Rao natural gradient (FRNG).

Critics: choice of the kernel in practice

Importance of the choice of the kernel in practice. The paper lacks a theory about the practicle choice of the kernel.

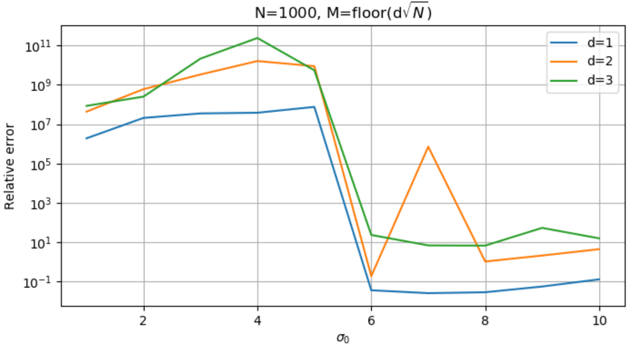

Figure: Fine-tuning of Gaussian kernel of parameter $\sigma = \sigma_0 \sigma_{N,M}$ for computing the estimator of KWNG. N number of samples, M number of basis points, d the dimension of the problem.

Conclusion, perspective

Advantage of the method

- Robustness to ill-conditioning
- Good trade-off between accuracy and computational cost
- General method valid for other metrics

Possible improvements

- Less sensitivity to the choice of kernel
- More robustness in high dimensions
- Other approximation for faster computation

Future works

- Why Wasserstein metric is more powerful than Fisher-Rao metric for natural gradient?
- Derive an estimator of the kernelized Fisher-Rao NG + comparison with KFAC, EFKAC