Stochastic optimization for large scale optimal transport

Hugo Cisneros

January 7th, 2019

Kantorovitch Formulation of regularized OT :

$$W_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X},\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi||\mu\otimes\nu) \qquad (\mathcal{P}_\varepsilon)$$

$$= \max_{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(u,v) \qquad (\mathcal{D}_\varepsilon)$$

$$= \max_{v\in\mathcal{C}(\mathcal{Y})} H_\varepsilon(v) \triangleq \int_{\mathcal{X}} v^{c,\varepsilon}(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \varepsilon \qquad (\mathcal{S}_\varepsilon)$$

Sinkhorn updates : $u^{\ell+1} = \frac{\mu}{Kv^{\ell}}; v^{\ell+1} = \frac{\nu}{K^T v^{\ell+1}} \rightarrow O(n^2)$

No general solver in the semi-discrete case.

What about large scale ?

Stochastic optimization

Discrete OT :

$$W_\varepsilon(\mu,\nu) = \max_{\boldsymbol{v}\in\mathbb{R}^J} \sum_{i=1}^{I} \overline{h}_\varepsilon(x_i, \boldsymbol{v})\boldsymbol{\mu}_i$$

Gradient aggregation algorithms (SAG and SAGA)

$$v_{k+1} = v_k + \texttt{step} * \Big(\nabla f_i - \nabla f_{[i]} + \frac{1}{I}\sum_i \nabla f_{[i]}\Big)$$

Semi-discrete OT :

$$W_\varepsilon(\mu,\nu) = \max_{v}\ \mathbb{E}_X[h_\varepsilon(X, v)]$$

Stochastic gradient ascent

$$v_{k+1} = v_k + \frac{\texttt{step}}{\sqrt{k}} * \nabla f_i$$

Theoretical analysis

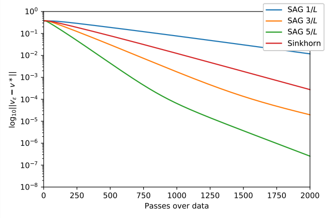
Convergence rates

Stochastic gradient $O(1/\sqrt{k})$ for non strongly convex, $O(1/k)$ for strongly convex

SAG and SAGA $O(1/k)$ for non strongly convex, linear for strongly convex (at the expense of storing gradients)

Numerical findings - Discrete OT

Synthetic data



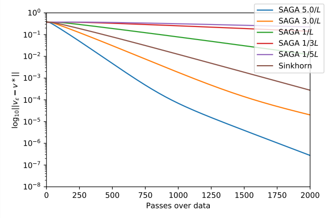
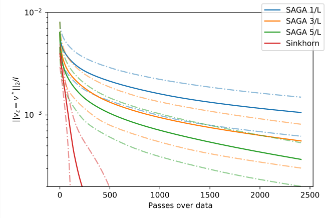
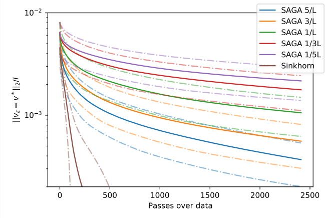
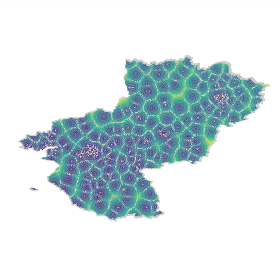


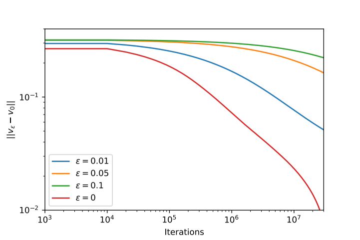
Image retrieval task

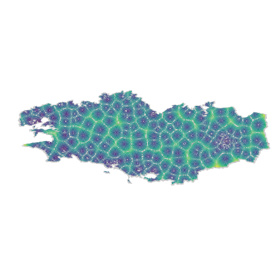


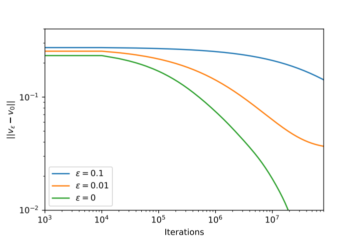


Numerical findings - Semi-discrete OT









Critics

- These methods should be tested in a much larger scale setting to show their real benefit
- Benefits over Sinkhorn of SAG and SAGA wasn't consistently observed and seems to depend on the parameters and structure of the problem

Perspective

- Applications of OT to problems with scales of the order of 10^6 and above
- Applications of semi-discrete OT to high dimensional problems with “exotic” cost functions