Stochastic optimization for large scale optimal transport

Project report

Hugo Cisneros

December 3, 2018

This report will study the matter of applying stochastic optimization techniques to solve optimal transport problems in the discrete and semi-discrete settings. In the discrete setting, the standard solver of the regularized OT problem is the Sinkhorn-Knopp algorithm which has a general computational complexity of $O(n^2)$. The problem is notoriously hard to solve, and the complexity of the Sinkhorn-Knopp algorithm is too high for a very large scale setting. Stochastic algorithms can be used to cope with that problem and compute solutions of the regularized OT problem in linear time.

Contents

1 Introduction

Optimal Transport (OT) is well known for its many applications in various domains. It has recently had major successes in Computer Vision or Natural Language Processing

1.1 Optimal transport : problem formulations

[2]

1.1.1 Entropic regularization of OT

We consider two measures $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ defined on metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The cost of moving a unit of mass from $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is defined by the continuous function $c \in \mathcal{C}(\mathcal{X}, \mathcal{Y})$, and written $c(x, y)$. We also define the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$

$$\Pi(\mu, \nu) \triangleq \{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}); \forall (A, B) \subset \mathcal{X} \times \mathcal{Y}, \pi(A, \mathcal{Y}) = \mu(A), \pi(\mathcal{X}, B) = \nu(B)\}$$

The entropic regularized version of the OT problem [1] can be written as a single convex optimization problem in the following form: $\forall(\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$,

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

With $\mathrm{KL}(\pi||\mu \otimes \nu)$ corresponding to the Kullback-Leibler divergence between joint probabilities π and $\mu \otimes \nu$, defined by $\mathrm{KL}(\pi||\xi) \triangleq \int_{\mathcal{X},\mathcal{Y}} \left(\log(\frac{d\pi}{d\xi}(x, y) - 1\right) d\xi(x, y)$.

For $\varepsilon > 0$, the above problem is strongly convex. $(\mathcal{P}_\varepsilon)$ is usually called the primal form of the regularized OT problem, by opposition to the dual and semi-dual form that will be studied further.

Sinkhorn for discrete OT In the discrete setting $\mu = \sum_i^n \delta_{x_i} \boldsymbol{\mu}_i$ and $\nu = \sum_j^m \delta_{x_j} \boldsymbol{\nu}_j$, the sums are finite and the cost is $\mathbf{C} \in \mathbb{R}^{n \times m}$. The structure of the KL divergence gives the optimal solution $\mathbf{P}_\varepsilon \in \Pi(\mu, \nu)$ a convenient structure that makes it possible solving the problem using Sinkhorn's algorithm [1]. There indeed exist two scaling variables $\mathbf{u}_\varepsilon \in \mathbb{R}^n$ and $\mathbf{v}_\varepsilon \in \mathbb{R}^m$ such that

$$\mathbf{P}_\varepsilon = \mathrm{diag}(\mathbf{u}_\varepsilon)\mathbf{K}_\varepsilon \mathrm{diag}(\mathbf{v}_\varepsilon)$$

Where $(\mathbf{K}_\varepsilon)_{i,j} = \exp(-\mathbf{C}_{i,j}/\varepsilon)$ [3]. Those scaling variables can be computed iteratively with the following update at step ℓ,

$$\mathbf{u}_\varepsilon^{\ell+1} \triangleq \frac{\mu}{\mathbf{K}_\varepsilon \mathbf{v}_\varepsilon^\ell} \quad \text{and} \quad \mathbf{v}_\varepsilon^{\ell+1} \triangleq \frac{\nu}{\mathbf{K}_\varepsilon^T \mathbf{u}_\varepsilon^{\ell+1}} \tag{1}$$

Because each step of the algorithm relies on a vector-matrix computation, the overall complexity of the algorithm is $O(nm)$ in the most general configuration.

The algorithm can be used in a large scale setting by using hardware (multiple Wasserstein distances can be computed in parallel on a GPU [**slomp2011gpu**]) and in some other specific cases where the kernel $\mathbf{K}$ is separable or can be expressed as a convolution [3]. In the general case however, the complexity of Sinkhorn's algorithm can be prohibitively large for a large scale problem.

1.1.2 Dual formulation of OT

1.1.3 Semi-dual formulation of OT

2 Stochastic optimization for large scale optimal transport

2.1 Discrete Optimal Transport

2.2 Semi-discrete Optimal Transport

3 Conclusion and Perspectives

References

[1] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, p. 9. URL: `http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf`.

[2] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale, 1781.

[3] Gabriel Peyré and Marco Cuturi. “Computational Optimal Transport”. In: *arXiv:1803.00567 [stat]* (Mar. 1, 2018). arXiv: `1803.00567`. URL: `http://arxiv.org/abs/1803.00567` (visited on 11/22/2018).