CME307&E311 Course Project IV: Computing Wasserstein Barycenter via Linear Programming

Yinyu Ye

February 11, 2022

In this project, we study the computation the Wasserstein barycenter of a set of discrete probability measures. Given support points of probability measures in a metric space and a transportation cost function (e.g. the Euclidean distance), Wasserstein distance defines a distance between two measures as the minimal transportation cost between them. Given a set of measures in the same space, the p-Wasserstein barycenter is defined as the measure minimizing the sum of p-Wasserstein distances to all measures in the set. Note that computing the barycenter of a set of discrete measures can be formulated by linear programming.

In this project, we focus on the case of $p = 2$ and compare the performance of different interior-point methods in solving the barycenter problem. We refer the notations and model setup to [1]. Instead of running experiments on MNIST dataset, we first restrict our attention to the algorithmic side and consider the following way in specifying the distributions $\mathcal{P}^{(t)}$.

- Generate m_t samples from normal distribution $\mathcal{N}(\mu_t, \sigma_t^2)$ and construct $\mathcal{P}^{(t)}$ as the empirical distribution on the m_t samples.
- In the following experiments, you should vary the choice of the number of samples m_t, the number of distributions N, and the parameters (μ_t, σ_t^2).

In this way, the objective becomes finding the Wasserstein barycenter of N normal distributions. Additionally, we first focus on the case of *Pre-specified Support Problem* (See [1]) and choose the support of the barycenter distribution $\mathcal{P}$ be the union of the supports of $\mathcal{P}^{(t)}$'s.

Question 1: Implement central-path method for this problem. Alternatively, you may also implement steepest descent method instead. Clearly state the linear program, include pseudo-code for your implementation, and report the parameters you use in the optimization algorithm such as step size.

Question 2: Implement the predictor-corrector interior point method [2] for this problem with the

single low-rank regularization method (SLRM) and double low-rank regularization method (DLRM) in [1]. These two methods aim to reduce the cost of solving the Newton equations in interior point method. How does this algorithm compare to the vanilla implementation of interior point method in Question 1? Plot the barycenter distribution $\mathcal{P}$ for two to three problem instances (specifications of N, m_t, etc.).

Now we consider the general *Free Support Problem* where the support of distribution $\mathcal{P}$ is also a decision variable. In [3] and [4], the authors proposed an entropy-smoothed version of Wasserstein distance for both regularization and computation purpose. The new distance replaces the summand $\langle D^{(t)}, \Pi^{(t)} \rangle$ in (3) and (4) in [1] with

$$\langle D^{(t)}, \Pi^{(t)} \rangle - \frac{1}{\lambda} h(\Pi^{(t)})$$

where $h(\cdot)$ is the entropy function. Intuitively, the entropy function will encourage the dispersion of the distribution $\Pi^{(t)}$ and avoid concentrations on a few points. Computationally, this new formulation enables a cheap computation of the gradients with respect to both the probability distribution parameters $(a_1, ..., a_m)$ and the support $\boldsymbol{X} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_m)$ (in the language of [1]).

Question 3: Implement Algorithm 3 in [4] with different choice of the regularization parameter γ. How does the resultant barycenter distribution compare with the ones obtained from interior-point methods? Note that Algorithm 3 in [4] considers a free-support setting while the two proceeding questions consider a pre-specified support. For a fair comparison, you may implement the free-support version of the interior-point method in [1]. Specifically, it will alternate between solving a linear program for the distribution parameter $(a_1, ..., a_m)$ and solving a quadratic program for the support $\boldsymbol{X} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_m)$. The quadratic program features for an analytical solution as (7) in [1].

As noted in these papers, the free support problem is then a non-convex problem. Now we are interested in how the gradient-based algorithm (Algorithm 3 in [4]) compares to the interior point method in respect with escaping saddle points and local minima.

Question 4: Implement MAAIPM algorithm in [1] and compare it against Algorithm 3 in [4] in respect with the original objective function value

$$\sum_{t=1}^{N} \langle D^{(t)}, \Pi^{(t)} \rangle.$$

While implementing Algorithm 3 in [4], you may want to periodically increase the regularization parameter γ to mitigate the effect of the additional penalty term. Please report the runtime, the number of iterations, and the objective value under both algorithms.

Question 5: Now you may migrate the experiments to the MNIST dataset[1] and Fashion MNIST dataset[2]. The advantage of using these datasets is that it can provide a more meaningful visualization of

[1] http://yann.lecun.com/exdb/mnist/

[2] https://www.kaggle.com/zalando-research/fashionmnist

the barycenter distribution.

References

[1] Dongdong Ge, Haoyue Wang, Zikai Xiong, Yinyu Ye. Interior-Point Methods Strike Back: Solving the Wasserstein Barycenter Problem. https://arxiv.org/abs/1905.12895.

[2] Mizuno, Shinji, Michael J. Todd, and Yinyu Ye. On adaptive-step primal-dual interior-point algorithms for linear programming. Mathematics of Operations research 18.4 (1993): 964-981.

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 2013.

[4] Marco Cuturi, Arnaud Doucet. Fast computation of Wasserstein barycenters. Proceedings of the International Conference on Machine Learning 2014.