A Summary of Optimization for Optimal Transport and Barycenter Problem

Boyang Liu

November 11, 2019

1 Regularized Discrete Optimal Transport

This section covers original optimal tranpsort (OT), entropy regularization, and sinkhorn algorithms.

1.1 Original Discrete Optimal Transport

Discrete OT problem defines the transport from source p to target q as follows:

$$\begin{aligned} &\arg\min_{\mathbf{T}} < \mathbf{C}, \mathbf{T} > \\ &s.t.\ \mathbf{T}\mathbf{1} = p \\ &s.t.\ \mathbf{T}^T\mathbf{1} = q, \end{aligned} \tag{1}$$

where $\mathbf{T} \in \mathbb{R}^{n_s \times n_t}$ is the transportation plan, $\mathbf{C} \in \mathbb{R}^{n_s \times n_t}$ is pre-known the cost matrix. $p \in \mathbb{R}^{n_s \times 1}$ is the distribution of source while $q \in \mathbb{R}^{n_t \times 1}$ is the distribution of target.

Definition 1.1 (Wasserstein distance)**.** Wasserstein distance between two distribution p and q is $W(p, q)$ and it is defined as $W(p, q) =< C, T^* >$, where T^* is the solution of Eq. 1.

The wasserstein distance (WD) is a measurement of the similarity of two distribution. From above, we can see solving WD is a linear programming problem (network flow). The solver for this LP programming needs $\mathbf{O}(n^3 \log n)$ complexity to solve. Noting this original OT problem cannot guarantee a stable optimal result, since the objective is not strong convex.

2 Entropy Regularized Optimal Transport

The entropy regularized optimal transport is defined as follows:

$$\begin{aligned}
&\arg\min_{\mathbf{T}} < \mathbf{C}, \mathbf{T} > +\gamma \sum_{i,j} T_{i,j} \log T_{i,j} \\
&s.t.\ \mathbf{T}\mathbf{1} = p \\
&s.t.\ \mathbf{T}^T\mathbf{1} = q,
\end{aligned} \tag{2}$$

The regularization has the entropy form, which is called negative entropy regularization. This regularization will make the objective function with strong convexity.

Definition 2.1 (Entropy relaxed Wasserstein distance)**.** Entropy relaxed wasserstein distance between two distribution p and q is $W_\gamma(p,q)$ and it is defined as $W_\gamma(p,q) =< C, T^* > +\gamma \sum_{i,j} T^*_{i,j} \log T^*_{i,j}$, where T^* is the solution of Eq. 2.

The relaxed WD can be solved efficiently by using the sinkhorn algorithms.

3 Sinkhorn Algorithms for Entropy Relaxed OT

Introducing two lagrange multiplier of $\alpha \in \mathbb{R}^{n_i \times 1}$and $\beta \in \mathbb{R}^{n_j \times 1}$, we have the dual objective:

$$\begin{aligned}
&\max_{\alpha,\beta} \min_{T} < C, T > \mathcal{L}(T, \alpha, \beta) \\
&= \max_{\alpha,\beta} \min_{T} < C, T > +\gamma < T, logT > +\alpha^T(p - T\mathbf{1}) + \beta^T(q - T^T\mathbf{1})
\end{aligned} \tag{3}$$

Take the derivative of Eq. 3 respect to T and set it to zero, we have:

$$\begin{aligned}
&\nabla_{T_{i,j}} \mathcal{L}(\alpha, \beta, T) = C_{i,j} + \gamma + \gamma \log T_{i,j} - \alpha_i - \beta_j = 0 \\
&\log T_{i,j} = \frac{\alpha_i + \beta_j - C_{i,j} - \gamma}{\gamma} \\
&T_{i,j} = e^{(\alpha_i - \gamma)/\gamma} e^{\beta_j/\gamma} e^{-C_{i,j}/\gamma}
\end{aligned} \tag{4}$$

Reform Eq. 4, let $u_i = e^{(\alpha_i-\gamma)/\gamma}$, $K_{i,j} = e^{-C_{i,j}/\gamma}$, $v_j = e^{\beta_j/\gamma}$, we have:

$$T = diag(u)Kdiag(v) \tag{5}$$

Recalling we have constrain $T\mathbf{1} = p$ and $T^T\mathbf{1} = q$, we have:

$$diag(u)Kdiag(v)\mathbf{1} = p \tag{6}$$

$$(diag(u)Kdiag(v))^T\mathbf{1} = q \tag{7}$$

$$\tag{8}$$

solve above equations, we have

$$u \otimes Kv = p \tag{9}$$
$$v \otimes K^T u = q \tag{10}$$

Now, we can have the update iterations as follows:

$$u^{k+1} \leftarrow p/Kv^k \tag{11}$$
$$v^{k+1} \leftarrow q/K^T u^{k+1} \tag{12}$$

After u and v converging, the T can be computed using Eq. 5 Noting this iterations only involves matrix-vector multiplication and pointwise divide, which is extremely efficient compared with the original OT.

4 The Relationship of Regularized OT and Bregman Projection

Considering the entropy relaxed OT for Eq. 2, we have:

$$\begin{aligned}
&< C, T > +\gamma < T, \log T > \\
&= \sum_{i,j} T_{i,j} C_{i,j} + \gamma T_{i,j} \log T_{i,j} \\
&= \gamma \sum_{i,j} T_{i,j} \log(T_{i,j} e^{\frac{C_{i,j}}{\gamma}}) \\
&= \gamma \sum_{i,j} T_{i,j} \log \frac{T_{i,j}}{e^{\frac{-C_{i,j}}{\gamma}}} \\
&= \gamma \mathbf{KL}(T || e^{\frac{C}{\gamma}}) && (13) \\
&s.t.\ T\mathbf{1} = p, T^T\mathbf{1} = q && (14)
\end{aligned}$$

From Eq. 14, we know the optimal value is $e^{\frac{C}{\gamma}}$. In order to get the optimal solution in feasible set, we iteratively project this optimal solution to the polytope of constrains, which can be proved by non-expansive properties.

4.1 The projection opretors

The polytope of constrain is defined as:

$$\begin{aligned}
&\mathcal{C}_1 : T \in \mathbb{R}_+^{n_i \times n_j}; T\mathbf{1} = p \\
&\mathcal{C}_2 : T \in \mathbb{R}_+^{n_i \times n_j}; T^T\mathbf{1} = q && (15)
\end{aligned}$$

The projection two these two polytopes are as follows:

$$P^{KL}_{C_1}(T) = diag(p/T\mathbf{1})T \tag{16}$$
$$P^{KL}_{C_2}(T) = Tdiag(q/T^T\mathbf{1}) \tag{17}$$

This projection is equivalent to the sinkhorn iterations. This part gives a bridge between bregman projection and sinkhorn algorithms.

5 Wasserstein Barycenter Problem

Wasserstein barycenter problem (WBP) is a interpolation method between serveral probability distribution. Given a normalized weights $\lambda \in \mathbb{R}^{K\times 1}$, the WBP is defined as the optimization problem as follows:

$$\min_{q\in\mathbb{R}^{n_j\times 1}} \left(\sum_{k}^{K} \lambda_k W_\gamma(p_k, q)\right) \tag{18}$$
$$s.t. T_k\mathbf{1} = p, T_k^T\mathbf{1} = q, \forall k = 1, ..., K \tag{19}$$

The feasible set of two polytopes can be defined as follows:

$$\mathcal{C}_1 : T_k \in \mathbb{R}_+^{n_i\times n_j}; \forall k,\ \ T_k\mathbf{1} = p_k; \tag{20}$$
$$\mathcal{C}_2 : T_k \in \mathbb{R}_+^{n_i\times n_j}; \forall k, \exists q,\ \ T_k^T\mathbf{1} = q \tag{21}$$

The WBP problem can be changed as follows:

$$\min_{T_k,q} \left(\sum_{k=1}^{K} \lambda_k \mathbf{KL}(T_k||\xi_k)\right), T_k \in C_1 \bigcap C_2 \tag{22}$$

where $\forall k$, $\xi_k \coloneqq \xi \coloneqq e^{-\dfrac{C}{\gamma}}$. Since the constrains are all convex polytope, we can also use bregman projection to solve this problem.

The projector to polytope $\mathcal{C}_1$ is the same as the relaxed OT projector to marginal distribution p.

The projector to polytope $\mathcal{C}_2$ hard to solve since we do not know what is the distribution q. If we know the distribution q, the projector will be same as the relaxed OT projector.

The projector to the polytope $\mathcal{C}_2$ is defined as:

$$P(T_k) = diag(\frac{q}{T_k\mathbf{1}})T_k, \tag{23}$$

where

$$q = \prod_{k=1}^{K} (T_k\mathbf{1})^{\lambda^k} \tag{24}$$

The Eq. 17 is just normalized to the marginal distribution q. The Eq. refgeometricmean is the geometric mean of the projection result to multiple marginal distribution p_k. The proof is as follows:

Introducing the lagrange multiplier for $\mathcal{C}_2$, for T_k, we have

$$\mathcal{L}(\alpha_k^T, T_k, q) = \mathbf{KL}(T_k^{s+1} || T_k^s) + \alpha_k^T (q - (T_k^s)^T \mathbf{1}) \tag{25}$$

Take the derivative of q and T_k, we have:

$$\nabla_{T_k} \mathcal{L}(\alpha_k^T, T_k, q) = \lambda_k \log(\frac{T_K^{S+1}}{T_K}) + \alpha_k \mathbf{1}^T = 0, \ \forall k \tag{26}$$

$$\nabla_q \mathcal{L}(\alpha_k^T, T_k, q) = \sum_k \alpha_k = 0 \tag{27}$$

Recalling we have $T_{k+1}^T \mathbf{1} = q$, we can solve the equations and get the result in Eq. 17, 24.