An Inexact Penalty Method for Fast Unbalanced Optimal Transport Optimization

Xun Su*1 Hiroyuki Kasai*2

*1 WASEDA University, Department of Communications and Computer Engineering, Graduate School of Fundamental Science and Engineering
*2WASEDA University, Department of Communications and Computer Engineering, School of Fundamental Science and Engineering

With the increasing application of Optimal Transport (OT) in machine learning, the unbalanced optimal transport (UOT) problem, as a variant of optimal transport, has gained attention for its improved generality. There is an urgent need for fast algorithms that can efficiently handle large penalty parameters. In this paper, we propose to use the Inexact penalty to make the Majorize-Minimization algorithm converge quickly even in UOT with large penalties. By using a dynamic scheme, we can successfully compute better and sparser solutions for the large penalty parameter and approach the computational speed of the well-known Sinkhorn algorithm, which sacrifices accuracy by adding an entropy item.

1. Introduction

Optimal transport (OT) has gained popularity in the fields of machine learning and statistical learning due to its ability to measure the distance between two probability measures. New methods that combine OT have outperformed traditional methods in areas such as domain adaptation [f] and generative models [a]. The rise in popularity of OT is largely due to the introduction of Sinkhorn's algorithm [g] for the entropy-regularized Kantorovich formulation problem, which has reduced the computational burden associated with large-scale problems.

Despite its success, the standard OT problem has a limitation in that it only handles *balanced* samples. To accommodate a wider range of applications with *unbalanced* samples, the unbalanced optimal transport (UOT) [c, e] has been proposed. UOT replaces the equality constraints with a KL divergence as a penalty function, and it is solvable by adding an entropic regularization term and using Sinkhorn's algorithm. Although it is fast, scalable, and differentiable, it suffers from instability, larger errors in solution compared to other regularizers.

Recently, [d] proposed a Majorization-Maximization Algorithm to solve the UOT problem without adding an entropy part by considering the mutual connection between the UOT problem and the non-negative matrix factorization problem. The algorithm is computable in GPU form, similar to Sinkhorn's algorithm, but it is still slower, especially for large penalization terms.

In this paper, we propose to combine the inexact penalty method, which was first introduced by [h] in the OT community and has been adapted in Augmented Lagrangian methods for many years, with the MM algorithm, to speed up the optimization process. Our method is simple and effective, and can greatly improve the computational speed of the MM algorithm with larger penalization terms.

Contact: Xun Su, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

2. Algorithm

2.1 Preliminaries

We use $\|\cdot\|_2$ to represent the Euclidean norm. $\mathbb{R}^n$ denotes n-dimensional Euclidean space, and $\mathbb{R}^n_+$ denotes the set of vectors in which all elements are non-negative. $\mathbb{R}^{n\times m}_+$ stands for the set of $n \times m$ matrices in which all elements are non-negative. We present vectors as bold lower-case letters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots$ and matrices as bold-face upper-case letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$. The i-th element of $\boldsymbol{a}$ and the element at the (i, j) position of $\mathbf{A}$ are stated respectively as a_i and $A_{i,j}$. In addition, $\mathbb{1}_n \in \mathbb{R}^n$ is the n-dimensional vector in which all elements are one. For two matrices of the same size $\mathbf{A}$ and $\mathbf{B}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$ is the Frobenius dot-product. We use $\|\boldsymbol{a}\|_2$, $\|\boldsymbol{a}\|_1$, and $\|\boldsymbol{a}\|_\infty$ to represent the ℓ_2-norm, ℓ_1-norm, and ℓ_∞ norm of $\boldsymbol{a}$, respectively.

2.2 Backgrounds

The balanced OT problem is defined as

$$\mathrm{OT}(\boldsymbol{a}, \boldsymbol{b}) \quad := \quad \min_{\mathbf{T}\in\mathbb{R}^{n\times m}_+} \langle \mathbf{C}, \mathbf{T} \rangle \qquad (1)$$

$$\text{subject to} \qquad \mathbf{T}\mathbb{1}_n = \boldsymbol{a}, \mathbf{T}^T\mathbb{1}_m = \boldsymbol{b},$$

By adding the KL divergence to penalize the difference, The UOT problem is defined as:

$$\mathrm{UOT}(\boldsymbol{a}, \boldsymbol{b}) := \min_{\mathbf{T}\in\mathbb{R}^{n\times m}_+} \langle \mathbf{C}, \mathbf{T} \rangle \qquad (2)$$

$$+\tau KL(\mathbf{T}\mathbb{1}_n, \boldsymbol{a}) + \tau KL(\mathbf{T}^T\mathbb{1}_n, \boldsymbol{b}). \qquad (3)$$

[d] considered the UOT probelm as a composite optimization problem. They propose the MM algorithm to solve the UOT problem:

$$\mathbf{T}^{(k+1)} \quad = \mathrm{diag}\left(\tfrac{\boldsymbol{a}}{\mathbf{T}^{(k)}\mathbb{1}_m}\right)^{\frac{1}{2}} \left(\mathbf{T}^{(k)} \odot \exp\left(-\tfrac{\mathbf{C}}{2\lambda}\right)\right) \qquad (4)$$

$$\mathrm{diag}\left(\tfrac{\boldsymbol{b}}{\mathbf{T}^{(k)\top}\mathbb{1}_n}\right)^{\frac{1}{2}}, \qquad (5)$$

Algorithm 1 Inexact Penalty Method UOT

```
Input: $\mathbf{T}^0, \mathbf{C}, \tilde{\tau}, \tau, q$
Output: $\mathbf{T}^K$
    $\mathbf{G} = \exp(-\frac{C}{2\tilde{\tau}})$
    for $k = 1$ to $K$ do
        $u = (\frac{a}{\mathbf{T}\mathbb{1}_n})^{\frac{1}{2}}, v = (\frac{b}{\mathbf{T}^T\mathbb{1}_m})^{\frac{1}{2}}$
        $\mathbf{T}^k = \mathbf{T}^k \odot (u^T \mathbf{G} v)$
        $err = \|\mathbf{T}^{k-1} - \mathbf{T}^k\|_2$
        if $err \leq \frac{q}{\tilde{\tau}}$ and $\tilde{\tau} \leq \tau$ then
            $\tilde{\tau} = \min(\tau, 2 * \tilde{\tau})$
        end if
    end for
```

It is noteworthy that the updating formula 4 bears striking similarities with the widely popular Sinkhorn Algorithm, relying solely on matrix multiplication. This allows a GPU acceleration for faster computation.

2.3 Proposed Method

However, we observed that the MM algorithm suffers from performance degradation for larger τ values, as illustrated in **Fig. 2**. Similarly, the Augmented Lagrangian Method (ALM) also uses an additional quadratic penalization term to expedite convergence but is plagued by an ill-conditioned Hessian matrix [b]. To overcome these limitations, we introduce a novel Inexact Penalized MM algorithm (MM-IP) that incorporates a dynamic penalization term, gradually increasing its influence throughout the optimization process. This approach has been previously employed in ALM research, and we demonstrate its effectiveness in our proposed algorithm. In **Algorithm. 1**, we set a small constant q and gradually increase the value of $\tilde{\tau}$ as the optimization error reduces. This inexact optimization process enables our algorithm to obtain a sparser initialization, allowing it to avoid the ill-conditioned Hessian matrix issue encountered in the early stages of optimization.

Since $\tilde{\tau}$ only doubles during the MM-IP algorithm for $O(log(\tau))$ times, the computation burden for recomputing matrix $\mathbf{K}$ is ignorable compared with the MM algorithm.

3. Experiments

We use Random Generated Gaussian distribution to test our method. In **Fig. 1**, we present the results of our experiment, where we generate five pairs of 100-dimensional random Gaussian distributions. We apply linear programming to obtain the analytical optimal solution $\mathbf{T}^*$ for the mass-equal pairs. For our experiments, we set $\tau = 1000$, and we use Sinkhorn's algorithm (UOT) with a regularizer value of $\epsilon = 1e-3$ to compute the solution. We also set the initial value of $\tilde{\tau} = 0.1$ and $q = 10^{-4}$ for our Inexact Penalized MM algorithm (MM-IP). We further apply Nesterov acceleration to our algorithm to obtain AMM-IP.

2 experiments demonstrate that the MM algorithm struggles to optimize the transport cost for the large penalization parameter, resulting in a significantly higher error than other methods. This can be observed in **Fig. 2**, where the large value of τ forces the optimization algorithm to converge to a dense solution, which is worse than both Sinkhorn algorithm and our proposed methods.

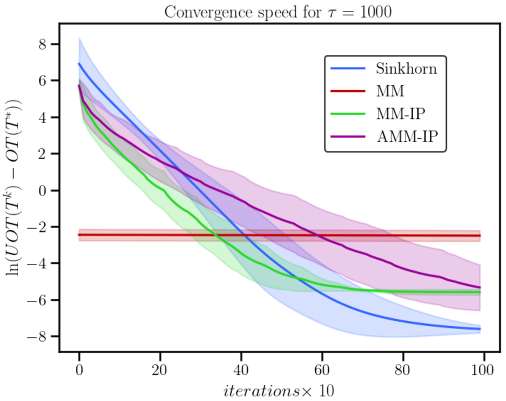

Figure 1: Comparing of the convergence speed for different projection methods.

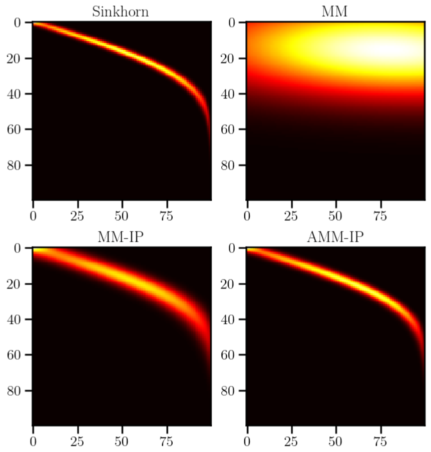

Figure 2: Comparing of the solutions obtained using different optimization methods, MM algorithm fails to accurately capture the sparse tendency of the solution in the UOT problems. MM-IP and AMM-IP methods perform significantly better, producing solutions similar to Sinkhorn's.

4. Conclusion

Overall, our experimental results illustrate the effectiveness of our proposed Inexact Penalized MM algorithm (MM-IP). Compared to the MM algorithm, our proposed method can effectively handle the challenge of larger τ values by utilizing an inexact penalization process that avoids poor initialization. This results in a superior solution quality that is competitive with the widely-known Sinkhorn algorithm. In the future, we plan to incorporate our expertise in the field of ALM to further accelerate the MM algorithm.

References

[a] Wasserstein Generative Adversarial Networks, pp. 214–223, PMLR (2017)

[b] Practical Augmented Lagrangian Methods for Constrained Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA (2014)

[c] Free boundaries in optimal transport and Monge-Ampère obstacle problems, year, Annals of Mathematics, pp. 673–730 (2010)

[d] Unbalanced Optimal Transport through Non-negative Penalized Linear Regression, NIPS (2021)

[e] Scaling Algorithms For Unbalanced Transport Problems, arXiv preprint: arXiv:1607.05816 (2017)

[f] Optimal Transport for Domain Adaptation, year, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1853–1865 (2017)

[g] year, Sinkhorn Distances: Lightspeed Computation of Optimal Transport, NIPS (2013)

[h] year, year, year, A Fast Proximal Point Method for Computing Exact Wasserstein Distance, pp. 433–453, PMLR (2020)