Vol.2021-AVM-112 No.10
2021/2/26

緩和最適輸送問題のためのFrank-Wolfeアルゴリズム 高速化手法と色転写問題への応用

福永 拓海[1,a)]　笠井 裕之[2,3,b)]

Abstract: 確率分布同士の距離を表現可能な最適輸送問題は幅広い分野に応用されている．最適輸送問題では，厳密な質量保存を表す制約条件を有する線形計画問題を解く必要があるが，線形計画問題を高速に解くことが困難であることが一般に知られている．当該問題を解決するため，制約条件の質量保存を緩めた緩和最適輸送問題が提案されており，解法の高速化の実現だけでなく，緩和した制約がより有効に働くいくつかの問題例（色転写問題等）が報告されている．本稿では，そのような緩和問題の中でも凸緩和最適輸送問題に注目し，新たな高速解法を提案し，理論的解析を行う．具体的には，Frank-Wolfe アルゴリズムに基づいた高速最適化手法を提案し，提案した手法の最悪収束反復数の上限値を示す．最後に，色転写問題における数値実験から提案手法の有効性について議論する．

1. Introduction

The Optimal Transport (OT) problem seeks an optimal *transport plan* or *transport matrix* by solving the total minimum transport cost from sources to destinations. This calculation requires *source mass conversation* from one source to targets, and versa, which are represented in transport polytope in formulation. The OT problem can express the distance between two probability distributions, which is known as Wasserstein distance [1], Thus, this has been applied to to a wide variety of machine learning problems such as adversarial risk [2], inference with aggregate data [3], graph optimal transport [4] and domain adaptation [5]. Among the OT problem formulations, the Monge-Kantrovich formulation [1] is represented as a convex linear programming, thus, many dedicated solvers such as an interior-point method and a network-flow method can obtain the solutions.It is, however, still challenging to solve large-scale problems efficiently because its computational cost increases cubically in terms of the data size.

To alleviate this issue, the Sinkhorn algorithm [6], an *entropy-regularized* approach, works effectively on the OT problem, which is faster and enables a parallel implementation. This computation includes a differentiable and unconstrained convex optimization, thus, it is relatively easier to solve. In addition, the resultant OT distances can be applied in many machine learning problems thanks to its differentiability. Besides, addressing its numerical unsuitability and non-robustness against for small values of the regularizer, a stabler variant has been also developed, but it suffers from slow convergence [7]. In order to reduce the runtime, a greedy algorithm of the Sinkhorn algorithm, the Greenkhorn algorithm [8], and its accelerated variant [9] have been proposed. It should noted that these approach produce a *dense* transport matrix because the entropy term is always positive. In another line of directions, a *smooth-regularized* approach exploits strong convexity and Lipschitz continuity [10], where adding smooth terms onto the objective function enables to harness gradient-based approaches and dual formulations. One distinguished feature is that the regularization with the squared Euclidean norm obtain sparser solutions than the entropy-regularized approaches.

Most of the aforementioned work attempt to add regularizers onto the objective function, some works address the fact that the tight mass-conversation constraint in the OT problem does not work well in some applications where weights and mass need not be necessarily preserved. For this particular problem, a *constraint-relaxed* approach has been recently proposed by loosing such strict constraints. This approach has gained a great success on applications such as color transfer [11] and multi-label learning [12]. However, it still exhibits a slow convergence property.

Envisioning to develop a faster solver producing sparser solutions in the OT problem, and particularly addressing its convex *semi-relaxed* formulation, this paper proposes a novel, and to the best of our knowledge, the first block-coordinate Frank-Wolfe (BCFW) algorithm with theoretical analysis. The FW algorithm (a.k.a the conditional gradient method) is a class of linear convex programming methods by calling a linear optimization oracle [13]. The key advantage of this algorithm is that its *projection-free* property is more efficient than the projected gradient method when the dimension of the data is relatively large. The output so-

[1] Department of Computer Science and Engineering, School of Fundamental Science and Engineering, Waseda University, Japan
[2] Department of Communications and Computer Engineering, School of Fundamental Science and Engineering, Waseda University, Japan
[3] Department of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University, Japan
[a)] f_takumi1997@suou.waseda.jp
[b)] hiroyuki.kasai@waseda.p

lutions of the FW algorithm have the desirable sparsity property. However, because this algorithm needs to call linear oracle for all columns of the transport matrix every iteration, it is prohibited when the matrix size is extremely large. Therefore, we further combine the coordinate descent method, which randomly selects one column every iteration, resulting in much smaller computation cost, and also in achieving a faster convergence [14]. Although this architecture has already been discussed in the literature in various problems [10], [15], its concrete convergence for the relaxed OT problem is still not clear. Hence, this paper has theoretical important contributions:

- Our convergence analysis gives an upper-bound of the curvature constantan *without* relying on the oracle as in [15]. Then, we exploit directly a variable block on the semi-relaxed domain and gives the iteration complexities for ϵ-optimality with the FW and BCFW algorithms for the semi-relaxed OT problem. Moreover, their worst iteration only depends on the dimension n, parameter λ and the constant ϵ.
- Our analysis of the duality gap reveals that the *linearization duality gap*, a special case of the Fenchel dual gap, is equivalent to the Lagrangian duality gap. We derive the Lagrangian dual for the semi-relaxed OT problem, and proved this equivalence. This linearization duality gap gives a certificate of the quantity of the current approximation for monitoring the convergence. This can be exploited for the stopping criterion in our proposed algorithms.

2. Preliminary and related work

$\mathbb{R}^n$ is denoted as n-dimensional Euclidean space and $\mathbb{R}^n_+$ is denoted as the set of vectors in which all elements are non-negative. $\mathbb{R}^{m\times n}$ is denoted as the set of $m \times n$ matrices and $\mathbb{R}^{m\times n}_+$ is denoted as the set of $m \times n$ matrices in which all elements are non-negative. We denote vectors as bold lower-case letters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots$ and matrices as bold-face 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 represented as a_i and $A_{i,j}$ respectively. When a matrix $\mathbf{A}$ is denoted as $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$, $\boldsymbol{a}_i$ represents the i-th column vector of $\mathbf{A}$. $\boldsymbol{e}_i$ is the canonical standard unit vector, of which the i-th element is 1, and others are zero. The probability simplex is denoted as $\Delta_m = \{\boldsymbol{a} \in \mathbb{R}^m : \sum_i a_i = 1\}$. $\boldsymbol{\delta_a}$ is the delta function at the vector $\boldsymbol{a}$. $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle_F$ represent the inner product and the Frobenius norm. Given two matrices $\mathbf{A}, \mathbf{B}$, the Frobenius norm is denoted as $\langle \mathbf{A}, \mathbf{B}\rangle_F := \sum_{i=1}^n \langle \boldsymbol{a}_i, b_i\rangle = \sum_{i=1}^m \sum_{j=1}^n A_{i,j}B_{i,j}$.

2.1 Optimal transport (OT)

The OT problem comes from Monge problem, which seeks the optimal mapping between two empirical probability distributions $\boldsymbol{\nu} = \sum_{i=1}^m a_i \boldsymbol{\delta}_{x_i}$, $\boldsymbol{\mu} = \sum_{i=1}^n b_i \boldsymbol{\delta}_{y_i}$ given by

$$\min_T \sum_{i=1}^m d(\boldsymbol{x}_i, T(\boldsymbol{x}_i)), \quad \text{s.t. } b_j = \sum_{i:T(\boldsymbol{x}_i)=\boldsymbol{y}_j} a_i, \forall j \in [m],$$

where $d(\cdot,\cdot)$ is the cost function between two points. Because both mapping and the constraints are discrete, Monge problem is difficult to solve directly. Therefore, Kantorovich proposed the formulation where the constraints are continuous, which is known as the OT problem. Given the cost matrix $\mathbf{C}$, the problem between distributions is defined as:

$$\min_{\mathbf{T}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{T}, \mathbf{C}\rangle_F, \tag{1}$$

where the domain $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$ is defined as

$$\mathcal{U}(\boldsymbol{a}, \boldsymbol{b}) = \{\mathbf{T} \in \mathbb{R}^{m\times n}_+ : \mathbf{T}\mathbf{1}_n = \boldsymbol{a}, \mathbf{T}^T\mathbf{1}_m = \boldsymbol{b}\}. \tag{2}$$

This domain $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$ requires the *mass conversation constraints* between two probabilities $\boldsymbol{a}$ and $\boldsymbol{b}$. The resultant transport matrix $\mathbf{T}^*$ brings a powerful distances between distributions defined as

$$\mathcal{W}_p(\boldsymbol{\nu}, \boldsymbol{\mu}) = \langle \mathbf{T}^*, \mathbf{C}\rangle^{\frac{1}{p}},$$

which is call the p-order *Wasserstein distance* [16]. Especially, when p is equal to 1, the distance is equivalent to the earth mover distance, what is called, geodesic [17]. Many problems appearing in machine learning and statistical learning can be defined in the OT problem. We refer the interested readers to [1] for more comprehensive survey .

2.2 Relaxed optimal transport

As discussed in Section 1, solving large-scale linear programming problems is challenging in terms of the computational costs to obtain solutions[18]. Furthermore, the strict mass conversation constraints may cause dreadful degradation of performances in some application. For example, Ferradans et al. reported that the the tight mass conversation do not reflect the color difference between images in color transfer problem [19]. This subsection introduces two categories of relaxed formulations of the OT problems.

Domain constraint relaxation. One approach is to relax the constraint domain [19]. Ferradans et al. propose to allow each point of $\mathbf{X}$ to be transported to multiple points of $\mathbf{Y}$ and versa. The formula is defined as

$$\min_{\mathbf{T}\in\mathcal{S}_\kappa} \langle \mathbf{T}, \mathbf{C}\rangle,$$

where a relaxed domain $\mathcal{S}_\kappa$ is defined as

$$\mathcal{S}_\kappa = \{\mathbf{T} \in \mathbb{R}^{n\times n}_+ : k_X\mathbf{1}_n \le \mathbf{T}\mathbf{1}_n \le K_X\mathbf{1}_n, k_y\mathbf{1}_n \le \mathbf{T}^T\mathbf{1}_n \le K_Y\mathbf{1}_n, \mathbf{1}_n^T\mathbf{T}\mathbf{1}_n = M\},$$

where constants (k_X, K_X, k_Y, K_Y, M) are hyper-parameters. This method enables the transport matrix to increase or decrease the mass between two points which are low distances. The noteworthy point is that the relaxed domain keeps the linear constraints as the original, thus, existing solvers of linear programming can be used. Rabin et al. extend it to propose the Relaxed Weighted OT, which looses the column constraints [11]. We also have other relaxed formulations considering only $\mathbf{T}\mathbf{1}_n = \boldsymbol{a}$ or $\mathbf{T}^T\mathbf{1}_m = \boldsymbol{b}$ as

$$\min_{\mathbf{T}\mathbf{1}_n = \boldsymbol{a}} \langle \mathbf{T}, \mathbf{C}\rangle \quad \text{or} \quad \min_{\mathbf{T}^T\mathbf{1}_m=\boldsymbol{b}} \langle \mathbf{T}, \mathbf{C}\rangle.$$

Because these optimal solutions are summation of minimum costs of each row or column vector, they can be solved faster than linear programming. In practice, this method is useful for document classification [20], and its extended formulation have recently been developed in context of style transfer [21], [22].

They attempt to define the relaxed earth mover distance (REMD) as the maximum of above formulations, and combine it with neural networks.

Regularized constraint relaxation. In another line of attempts, the penalty of the domains defined in (2) is added into the objective function [10]. Relaxing both marginal constraints in (2) yields the following relaxed formulation:

$$\min_{\mathbf{T}\geq\mathbf{0}} \langle \mathbf{T}, \mathbf{C}\rangle + \frac{1}{2}\Phi(\mathbf{T}\mathbf{1}_n, \boldsymbol{a}) + \frac{1}{2}\Phi(\mathbf{T}^T\mathbf{1}_m, \boldsymbol{b}),$$

where $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is a smooth divergence measure.

We also have an alternative formulation, which relaxes one of the two constraints in (2). This is called a *semi-relaxed* problem and is defined as the following:

$$\min_{\mathbf{T}\geq\mathbf{0}, \mathbf{T}^T\mathbf{1}_m=\boldsymbol{b}} \langle \mathbf{T}, \mathbf{C}\rangle + \Phi(\mathbf{T}\mathbf{1}_n, \boldsymbol{a}). \tag{3}$$

Benamou proposes a similar formulation, and is solved by use of augmented Lagrangian[23]. Ferradans et al. propose a regularized and relaxed problem specifically focusing on both color transfer and barycenter [19]. They use the proximal splitting method and the coordinate descent method. Rabin et al. also propose the weighted regularization term $\| \kappa - \mathbf{1}_n \|_1$ as well as Relaxed Weighted OT so that the ratio of the source image becomes close to that of the target image [11]. Moreover, using the Kullback-Leibler (KL) divergence as $\Phi(\boldsymbol{x}, \boldsymbol{y})$, a multi-label prediction problem is solved by use of Sinkhorn-like algorithm because the entropy regularization is added to the objective function [12]. However, the KL divergence and is not unstable because of diverging at zero [7]. Furthermore, there are some relaxed methods which address cardinality penalized problems. Because they are in NP-hard and difficult to solve, most methods loose the regularization term. In instead of cardinality of solutions, Carli et al. approximate them by exploiting the rank regularization, sum-of-norm relaxation and maximum norm relaxation for an effective clustering [24].

2.3 Frank-Wolfe and block-coordinate algorithms

The Frank-Wolfe (FW) algorithm is one of the constraint convex optimization methods, and is known to be a linear approximation algorithm that uses conditional gradient [25]. Although FW is known to converge to optimal solutions in a *sublinear* rate, its *projection-free* property is preferable in the case where the convex constraint is simple and the feasible point can be found easily. More specifically, at every iteration, the feasible point $\boldsymbol{s}$ is first found by minimizing the *linearization* of f over the convex feasible set $\mathcal{M}$. To find the feasible point $\boldsymbol{s}$, we have to solve the following subproblem :

$$\boldsymbol{s} = \arg\min_{\boldsymbol{s}'\in\mathcal{M}} \langle \boldsymbol{s}', \nabla f(\boldsymbol{x}^{(k)})\rangle \tag{4}$$

where $\boldsymbol{x}^{(k)}$ represents the k-th current point. Since the domain $\mathcal{M}$ is the convex set and the objective is linear for $\boldsymbol{s}$, it is possible to solve (4) by linear programming. Finally, the next iterate $\boldsymbol{x}^{(k+1)}$ can be obtained by a convex combination as $\boldsymbol{x}^{(k+1)} = (1-\gamma)\boldsymbol{x}^{(k)} + \gamma\boldsymbol{s}$ where γ is a stepsize. Thus, the generated iterates can be maintained inside the feasible set $\mathcal{M}$ if the initial point $\boldsymbol{x}^{(0)}$ is in $\mathcal{M}$.

One of disadvantages of the FW algorithm is that solving the minimization problem needs to be performed in each iteration. For this issue, if the variable $\mathcal{M}$ can be *block-separable* as a cartesian product $\mathcal{M} = \mathcal{M}^{(1)} \times \mathcal{M}^{(2)} \times \cdots \times \mathcal{M}^{(n)} \subset \mathbb{R}^m$ over $n \geq 1$, we can perform a *single cheaper* update on only $\mathcal{M}^{(i)}$ instead of on an entire of $\mathcal{M}$. In this line of algorithms, the block-coordinate Frank-Wolfe (BCFW) algorithm has been proposed, for example, in the structural SVM problem in [15] and in the MAP inference [26]. This algorithm can be applied to the constrained convex problem of the form

$$\min_{\boldsymbol{x}\in\mathcal{M}^{(1)}\times\mathcal{M}^{(2)}\times\cdots\times\mathcal{M}^{(n)}} f(\boldsymbol{x}).$$

We assume that each factor $\mathcal{M}^{(i)}$ is convex, with $m = \sum_{i=1}^n m_i$. We solve the subproblem on the factor which is selected randomly. As a result, the BCFW algorithm can be implemented in cheaper iteration. When $n = 1$, this algorithm is reduced to the FW algorithm.

3. Block-coordinate Frank-Wolfe algorithm for semi-relaxed optimal transport

This paper particularly addresses the semi-relaxed problem of (3) with $\Phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2\lambda} \| \boldsymbol{x} - \boldsymbol{y} \|_2^2$ because it is not only smooth but also convex. The problem of interest is formally defined as

$$\min_{\substack{\mathbf{T}\geq\mathbf{0},\\ \mathbf{T}^T\mathbf{1}_m=\boldsymbol{b}}} \left\{ f(\mathbf{T}) := \langle \mathbf{T}, \mathbf{C}\rangle + \frac{1}{2\lambda} \| \mathbf{T}\mathbf{1}_n - \boldsymbol{a} \|_2^2 \right\}, \tag{5}$$

where λ is a relax parameter. The domain is transformed into

$$\mathcal{M} = b_1\Delta_m \times b_2\Delta_m \times \cdots \times b_n\Delta_m,$$

where $b_i\Delta_m$ represents the simplex of the summation b_i.

We first consider the FW algorithm for this problem, and then propose a faster block-coordinate Frank-Wolfe algorithm.

3.1 Frank-Wolfe algorithm (FW) for semi-relaxed OT problem

The gradient $\nabla f(\mathbf{T}) \in \mathbb{R}^{mn}$ is given as

$$\nabla f(\mathbf{T}) = \begin{pmatrix} \boldsymbol{c}_1 \\ \vdots \\ \boldsymbol{c}_i \\ \vdots \\ \boldsymbol{c}_n \end{pmatrix} + \frac{1}{\lambda} \begin{pmatrix} \mathbf{T}\mathbf{1}_n - \boldsymbol{a} \\ \vdots \\ \mathbf{T}\mathbf{1}_n - \boldsymbol{a} \\ \vdots \\ \mathbf{T}\mathbf{1}_n - \boldsymbol{a} \end{pmatrix},$$

where $\nabla f_i(\mathbf{T}) := \boldsymbol{c}_i + 1/\lambda \cdot (\mathbf{T}\mathbf{1}_n - \boldsymbol{a}) \in \mathbb{R}^m$ represents the gradient on the i-th variable block $b_i\Delta_m$. The subproblem (4) is equivalent to

$$b_i\boldsymbol{e}_j = b_i \arg\min_{\boldsymbol{e}_k\in\Delta_m, k\in[m]} \langle \boldsymbol{e}_k, \nabla_i f(\mathbf{T}^k)\rangle. \tag{6}$$

where j is in $[m]$ and $\boldsymbol{e}_j$ is the extreme point on probability simplex [27]. In other words, we should find the index of the minimal elements of the gradient of the variable blocks. The computational cost of the subproblem (6) is greatly improved..

After finding the points $\mathbf{S}$, we search the optimal stepsize γ. One general way to calculate the stepsizes in the FW algorithm

is a *diminishing* stepsize, where $\gamma = k/(k+2)$. In another approach, we solve $\arg\min_{\gamma\in[0,1]} f((1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{s})$ directly if the objective is quadratic. Fortunately, the objective of semi-relaxed problem is quadratic, which makes it solvable. Hence, the optimal stepsize γ is calculated as

$$\gamma = \frac{\lambda\langle \mathbf{T}, \mathbf{C}\rangle_F + \langle \mathbf{T}\mathbf{1}_n - \mathbf{S}\mathbf{1}_n, \mathbf{T}\mathbf{1}_n - \boldsymbol{a}\rangle}{\lambda \parallel \mathbf{T}\mathbf{1}_n - \mathbf{S}\mathbf{1}_n \parallel^2}.$$

As for the stopping criterion, we monitor the duality $g(\mathbf{T})$, and stop the algorithm when $g(\mathbf{T}) < \epsilon$, where ϵ is an approximation parameter.

3.2 Block-Coordinate Frank-Wolfe (BCFW) for semi-relaxed OT problem

We now consider the application of the block-coordinate Frank-Wolfe algorithm to semi-relaxed problem because the feasible set $\mathcal{M}$ can be separable as the cartesian product. The procedure of the algorithm is most similar to that of the FW algorithm, but they are a little different. It is necessary to solve the subproblem on the variable block selected randomly at every iteration. The problem is re-formulated as

$$\boldsymbol{s}_i = \underset{\boldsymbol{s}' \in b_i\Delta_m}{\arg\min} \langle \boldsymbol{s}'_i, \boldsymbol{c}_i + \frac{1}{\lambda}(\mathbf{T}^{(k)}\mathbf{1}_n - \boldsymbol{a})\rangle, \tag{7}$$

where the index i is selected randomly. As for the stepsize calculation we can use the formula $\gamma = 2n/(k+2n)$, which is also required for the convergence guarantee. We can also solve $\arg\min_{\gamma\in[0,1]} f((1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{s})$ and calculate γ directly like the FW algorithm since the objective of semi-relaxed problem is quadratic. Nevertheless, the optimal stepsize in the BCFW algorithm is different from that of the FW algorithm because the BCFW algorithm requires the update of the column vector on the variable block which are selected randomly. Concretely, all the elements of $\mathbf{T}$ is equal to those of $\mathbf{S}$ except for the elements on the variable factor selected. The Frobenius product $\langle \mathbf{T}-\mathbf{S}, \mathbf{C}\rangle_F$ is transformed into the inner product $\langle \boldsymbol{t}_i - \boldsymbol{s}_i, \boldsymbol{c}_i\rangle$ and the vector $\mathbf{T}\mathbf{1}_n - \mathbf{S}\mathbf{1}_n$ into $\boldsymbol{t}_i - \boldsymbol{s}_i$. Hence, the optimal stepsize γ is calculated as

$$\gamma = \frac{\lambda\langle \boldsymbol{t}_i^{(k)} - \boldsymbol{s}_i, \boldsymbol{c}_i\rangle + \langle \boldsymbol{t}_i^{(k)} - \boldsymbol{s}_i, \mathbf{T}^{(k)}\mathbf{1}_n - \boldsymbol{a}\rangle}{\lambda \parallel \boldsymbol{t}_i^{(k)} - \boldsymbol{s}_i \parallel_2^2},$$

where $\boldsymbol{s}_i$ is the solution of the subproblem on the selected factor.

4. Theoretical analysis

We prove the convergence analysis of the FW and BCFW algorithms proposed in the previous section. We then discuss the relationship between the linearization duality gap, as a special case of the Fenchel duality gap, and Lagrange duality gap. This provides the equivalence between them in this semi-relaxed OT problem. Finally, we also investigate the computational complexity of them. In the presentation, we will present them.

5. Numerical evaluations in color transfer problem

We have proposed in this paper a faster Frank-Wolfe algorithm and its block-coordinate variant for a convex semi-relaxed optimal transport problem. In the presentation, we will present some numerical evaluation results in color transfer problem.

References

[1] Peyre, G. and Cuturi, M.: Computational Optimal Transport, *Foundations and Trends in Machine Learning*, Vol. 11, No. 5-6, pp. 355–607 (2019).
[2] Pydi, M. S. and Jog, V.: Adversarial Risk via Optimal Transport and Optimal Couplings, *ICML* (2020).
[3] Singh, R., Haasler, I., Zhang, Q., Karlsson, J. and Chen, Y.: Inference with Aggregate Data: An Optimal Transport Approach (2020).
[4] Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L. and Liu, J.: Graph Optimal Transport for Cross-Domain Alignment, *ICML* (2020).
[5] Redko, I., Courty, N., Flamary, R. and Tuia, D.: Optimal Transport for Multi-source Domain Adaptation under Target Shift, *AISTATS* (2019).
[6] Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances (2013).
[7] Chizat, L., Peyré, G., Schmitzer, B. and Vialard, F.-X.: Scaling Algorithms for Unbalanced Transport Problems (2017).
[8] Altschuler, J., Weed, J. and Rigollet, P.: Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration (2018).
[9] Lin, T., Ho, N. and Jordan, M. I.: On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms (2020).
[10] Blondel, M., Seguy, V. and Rolet, A.: Smooth and Sparse Optimal Transport (2018).
[11] Rabin, J., Ferradans, S. and Papadakis, N.: Adaptive color transfer with relaxed optimal transport, *IEEE International Conference on Image Processing (ICIP)* (2014).
[12] Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M. and Poggio, T.: Learning with a Wasserstein Loss (2015).
[13] Frank, M. and Wolfe, P.: An algorithm for quadratic programming, *Naval Research Logstics Quarterly*, Vol. 3, pp. 95–110 (1956).
[14] Wright, S. J.: Coordinate descent algorithms, *Mathematical Programming*, Vol. 151, pp. 3–34 (2015).
[15] Lacoste-Julien, S., Jaggi, M., Schmidt, M. and Pletscher, P.: Block-Coordinate Frank-Wolfe Optimization for Structural SVMs, *ICML*, Atlanta, United States (2013).
[16] Villani, C.: *Optimal transport: Old and new*, Springer (2008).
[17] Levina, E. and Bickel, P. J.: The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics., *ICCV*, pp. 251–256 (2001).
[18] Bonneel, N., van de Panne, M., Paris, S. and Heidrich, W.: Displacement Interpolation Using Lagrangian Mass Transport, *ACM Trans. Graph.*, Vol. 30, No. 6, pp. 1–12 (2011).
[19] Ferradans, S., Papadakis, N., Peyré, G. and Aujol, J.-F.: Regularized Discrete Optimal Transport (2013).
[20] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K.: From Word Embeddings To Document Distances, *ICML*, Vol. 37, pp. 957–966 (2015).
[21] Kolkin, N., Salavon, J. and Shakhnarovich, G.: Style Transfer by Relaxed Optimal Transport and Self-Similarity (2019).
[22] Qiu, T., Ni, B., Liu, Z. and Chen, X.: Fast Optimal Transport Artistic Style Transfer, *MultiMedia Modeling* (2021).
[23] Benamou, Jean-David: Numerical resolution of an "unbalanced" mass transport problem, *ESAIM: M2AN*, Vol. 37, No. 5, pp. 851–868 (2003).
[24] Carli, F. P., Ning, L. and Georgiou, T. T.: Convex Clustering via Optimal Mass Transport (2013).
[25] Dostl, Z.: *Optimal Quadratic Programming Algorithms: With Applications to Variational Inequalities*, Springer Publishing Company, Incorporated, 1st edition (2009).
[26] Swoboda, P. and Kolmogorov, V.: MAP inference via Block-Coordinate Frank-Wolfe Algorithm, *CVPR* (2019).
[27] Clarkson, K. L.: Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm, *ACM Transactions on Algorithms*, Vol. 6, No. 4 (2010).