







Strongly Convex Optimization for the Dual Formulation of Optimal Transport

Nazarii Tupitsa^{1,2,4} , Alexander Gasnikov^{1,2,4} , Pavel Dvurechensky^{2,3} ,
and Sergey Guminov^{1,2} 

¹ Moscow Institute of Physics and Technology, Dolgoprudny, Russia
{tupitsa,sergey.guminov}@phystech.edu

² Institute for Information Transmission Problems RAS, Moscow, Russia
gasnikov@yandex.ru

³ Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
pavel.dvurechensky@wias-berlin.de

⁴ National Research University Higher School of Economics,
Moscow, Russia

Abstract. In this paper we experimentally check a hypothesis, that dual problem to discrete entropy regularized optimal transport problem possesses strong convexity on a certain compact set. We present a numerical estimation technique of parameter of strong convexity and show that such an estimate increases the performance of an accelerated alternating minimization algorithm for strongly convex functions applied to the considered problem.

Keywords: Convex optimization · Optimal transport · Sinkhorn's algorithm · Alternating minimization

1 Introduction

Optimal transport problem has different applications since it allows to define a distance between probability measures including the earth mover's distance [51, 62] and Monge-Kantorovich or Wasserstein distance [61]. These distances play an increasing role in different machine learning tasks, such as unsupervised learning [6, 11], semi-supervised learning [56], clustering [31], text classification [35], as well as in image retrieval, clustering and classification [13, 51, 53], statistics [24, 49], and other applications [33]. In many of these applications the original optimal distances are substituted by entropically regularized optimal transport problem [13] which gives rise to a so-called Sinkhorn divergence.

A close problem arises in transportation research and consists in recovering a matrix of traffic demands between city districts from the information on population and workplace capacities of each district. As it is shown in [28], a

This research was funded by Russian Science Foundation (project 18-71-10108).

© Springer Nature Switzerland AG 2020

Y. Kochetov et al. (Eds.): MOTOR 2020, CCIS 1275, pp. 192–204, 2020.

https://doi.org/10.1007/978-3-030-58657-7_17

natural model of the district’s population dynamics leads to an entropy-linear programming optimization problem for the traffic demand matrix estimation. In this case, the objective function is a sum of an entropy function and a linear function. It is important to note also that the entropy function is multiplied by a regularization parameter γ and the model is close to reality when the regularization parameter is small. The same approach is used in IP traffic matrix estimation [63].

Recent approaches to solving discrete optimal transport problem are based on accelerated primal-dual gradient-based algorithms [21, 30] which in some regimes demonstrate better performance than well-known Sinkhorn’s algorithm [13, 55]. Both these algorithms have complexity polynomially depending on the desired accuracy [3, 21, 40]. Despite, formally, the dual for the optimal transport problem is not strongly convex, it is strongly convex on any bounded subset of any subspace orthogonal to a one-dimensional subspace. In this paper we suggest and check empirically a hypothesis which helps to increase the rate of convergence for the dual problem to optimal transport. The hypothesis is that dual function demonstrates strong convexity on the orthogonal subspace and Sinkhorn’s and other algorithms produce points in this orthogonal subspace meaning that actually the dual problem is strongly convex on the trajectory of the method.

Since we focus mainly on alternating minimization, the related work contains such classical works as [10, 48]. AM algorithms have a number of applications in machine learning problems. For example, iteratively reweighted least squares can be seen as an AM algorithm. Other applications include robust regression [41] and sparse recovery [16]. Famous Expectation Maximization (EM) algorithm can also be seen as an AM algorithm [4, 42]. Sublinear $O(1/k)$ convergence rate was proved for AM algorithm in [8]. AM-algorithms converge faster in practice in comparison to gradient methods as they are free of the choice of the step-size and are adaptive to the local smoothness of the problem. Besides mentioned above works on AM algorithms, we mention [9, 52, 58], where non-asymptotic convergence rates for AM algorithms were proposed and their connection with cyclic coordinate descent was discussed, but the analyzed algorithms are not accelerated. Accelerated versions are known for random coordinate descent methods [2, 23, 26, 27, 36, 37, 44, 47, 54]. These methods use momentum term and block-coordinate steps, rather than full minimization in blocks. A hybrid accelerated random block-coordinate method with exact minimization in the last block and an accelerated alternating minimization algorithm were proposed in [17].

2 Dual Optimal Transport Problem

In this paper we consider the following discrete-discrete entropically regularized optimal transport problem

$$f(X) = \langle C, X \rangle + \gamma \langle X, \ln X \rangle \rightarrow \min_{X \in \mathcal{U}(r, c)}, \quad (1)$$

$$\mathcal{U}(r, c) = \{X \in \mathbb{R}_+^{N \times N} : X\mathbf{1} = r, X^T\mathbf{1} = c\},$$

where X is the transportation plan, $\ln X$ is taken elementwise, $C \in \mathbb{R}_+^{N \times N}$ is a given cost matrix, $\mathbf{1} \in \mathbb{R}^N$ is the vector of all ones, $r, c \in S_N(\mathbf{1}) := \{s \in \mathbb{R}_+^N : \langle s, \mathbf{1} \rangle = 1\}$ are given discrete measures, and $\langle A, B \rangle$ denotes the Frobenius product of matrices defined as $\langle A, B \rangle = \sum_{i,j=1}^N A_{ij}B_{ij}$.

Next, we consider the dual problem for the above optimal transport problem. First, we note that $\mathcal{U}(r, c) \subset Q := \{X \in \mathbb{R}_+^{N \times N} : \mathbf{1}^T X \mathbf{1} = 1\}$ and the entropy $\langle X, \ln X \rangle$ is strongly convex on Q w.r.t 1-norm, meaning that the dual problem has the objective with Lipschitz-continuous gradient [43]. To be more precise, function f is μ -strongly convex on a set Q with respect to norm $\|\cdot\|$ iff

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2 \quad \forall x, y \in Q.$$

Further, function f is said to have L -Lipschitz-continuous gradient iff, for all $x, y \in Q$, $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$. Here $\|\cdot\|_*$ is the standard conjugate norm for $\|\cdot\|$. The proof that Entropy is 1-strongly convex on the standard simplex w.r.t. to $\|\cdot\|_1$ -norm can be found in [43]. The dual problem is constructed as follows

$$\begin{aligned} & \min_{X \in Q \cap \mathcal{U}(r,c)} \langle C, X \rangle + \gamma \langle X, \ln X \rangle & (2) \\ & = \min_{X \in Q} \max_{y, z \in \mathbb{R}^N} \left\{ \langle C, X \rangle + \gamma \langle X, \ln X \rangle + \langle y, X \mathbf{1} - r \rangle + \langle z, X^T \mathbf{1} - c \rangle \right\} \\ & = \max_{y, z \in \mathbb{R}^N} \left\{ -\langle y, r \rangle - \langle z, c \rangle + \min_{X \in Q} \sum_{i,j=1}^N X^{ij} (C^{ij} + \gamma \ln X^{ij} + y^i + z^j) \right\}. \end{aligned}$$

Note that for all i, j and some small ε

$$X^{ij} (C^{ij} + \gamma \ln X^{ij} + y^i + z^j) < 0$$

for $X^{ij} \in (0, \varepsilon)$ and this quantity approaches 0 as X^{ij} approaches 0. Hence, $X^{ij} > 0$ without loss of generality. Using Lagrange multipliers for the constraint $\mathbf{1}^T X \mathbf{1} = 1$, we obtain the problem

$$\min_{X^{ij} > 0} \max_{\nu} \left\{ \sum_{i,j=1}^N [X^{ij} (C^{ij} + \gamma \ln X^{ij} + y^i + z^j)] - \nu \left[\sum_{i,j=1}^N X^{ij} - 1 \right] \right\}.$$

The solution to this problem is

$$X^{ij} = \frac{\exp\left(-\frac{1}{\gamma} (y^i + z^j + C^{ij}) - 1\right)}{\sum_{i,j=1}^n \exp\left(-\frac{1}{\gamma} (y^i + z^j + C^{ij}) - 1\right)}.$$

With a change of variables $u = -y/\gamma - \frac{1}{2}\mathbf{1}, v = -z/\gamma - \frac{1}{2}\mathbf{1}$ we arrive at the following expression for the dual (minimization) problem

$$\varphi(u, v) = \gamma(\ln(\mathbf{1}^T B(u, v)\mathbf{1}) - \langle u, r \rangle - \langle v, c \rangle) \rightarrow \min_{u, v \in \mathbb{R}^N}, \quad (3)$$

where $[B(u, v)]^{ij} = \exp\left(u^i + v^j - \frac{C^{ij}}{\gamma}\right)$. Let us also define

$$\varphi(y, z) = \varphi\left(-y/\gamma - \frac{1}{2}\mathbf{1}, -z/\gamma - \frac{1}{2}\mathbf{1}\right), \quad (4)$$

i.e. $\varphi(y, z)$ is the dual objective before change of variables. Note that the gradient of this function has the form of two blocks

$$\nabla\varphi(y, z) = \begin{pmatrix} r - \frac{B(-y/\gamma - \mathbf{1}/2, -z/\gamma - \mathbf{1}/2)\mathbf{1}}{\mathbf{1}^T B(-y/\gamma - \mathbf{1}/2, -z/\gamma - \mathbf{1}/2)\mathbf{1}} \\ c - \frac{B(-y/\gamma - \mathbf{1}/2, -z/\gamma - \mathbf{1}/2)^T \mathbf{1}}{\mathbf{1}^T B(-y/\gamma - \mathbf{1}/2, -z/\gamma - \mathbf{1}/2)\mathbf{1}} \end{pmatrix}. \quad (5)$$

Notably, this dual problem is a smooth minimization problem with the objective having Lipschitz continuous gradient with constant $2/\gamma$ [30]. Unfortunately, generally speaking it is not strongly convex since given a point (u_0, v_0) the value of the objective is the same on the whole line $(u_0 + t\mathbf{1}, v_0 - t\mathbf{1})$ parameterized by t . Yet, this function is strongly convex in the subspace orthogonal to these lines [15]. The goal of this paper is to use this strong convexity to accelerate the accelerated alternating minimization method based on Nesterov extrapolation and alternating minimization.

The variables in the dual problem (3) naturally decompose into two blocks u and v . Moreover, minimization over any one block may be performed analytically.

Lemma 1. *The iterations*

$$u^{k+1} \in \operatorname{argmin}_{u \in \mathbb{R}^N} \varphi(u, v^k), \quad v^{k+1} \in \operatorname{argmin}_{v \in \mathbb{R}^N} \varphi(u^{k+1}, v),$$

can be written explicitly as

$$\begin{aligned} u^{k+1} &= u^k + \ln r - \ln(B(u^k, v^k)\mathbf{1}), \\ v^{k+1} &= v^k + \ln c - \ln(B(u^{k+1}, v^k)^T \mathbf{1}). \end{aligned}$$

This lemma implies that an alternating minimization method applied to the dual formulation is a natural algorithm. In fact, this is the celebrated Sinkhorn's algorithm [13, 55] in one of its forms [3] listed as Algorithm 1. This algorithm may also be implemented more efficiently as a matrix-scaling algorithm, see [13]. For the reader's convenience, we prove this lemma here.

Algorithm 1. Sinkhorn's Algorithm

Output: x^k

for $k \geq 1$ **do**

$$u^{k+1} = u^k + \ln r - \ln(B(u^k, v^k)\mathbf{1})$$

$$v^{k+1} = v^k$$

$$u^{k+2} = u^{k+1}$$

$$v^{k+2} = v^{k+1} + \ln c - \ln(B(u^{k+1}, v^{k+1})^T \mathbf{1})$$

end for

Proof. From optimality conditions, for u to be optimal, it is sufficient to have $\nabla_u \varphi(u, v) = 0$, or

$$r - (\mathbf{1}^T B(u, v^k) \mathbf{1})^{-1} B(u, v^k) \mathbf{1} = 0. \quad (6)$$

Now we check that it is, indeed, the case for $u = u^{k+1}$ from the statement of this lemma. We check that

$$\begin{aligned} B(u^{k+1}, v^k) \mathbf{1} &= \text{diag}(e^{(u^{k+1} - u^k)}) B(u^k, v^k) \mathbf{1} \\ &= \text{diag}(e^{\ln r - \ln(B(u^k, v^k) \mathbf{1})}) B(u^k, v^k) \mathbf{1} \\ &= \text{diag}(r) \text{diag}(B(u^k, v^k) \mathbf{1})^{-1} B(u^k, v^k) \mathbf{1} = \text{diag}(r) \mathbf{1} = r \end{aligned}$$

and the conclusion then follows from the fact that

$$\mathbf{1}^T B(u^{k+1}, v^k) \mathbf{1} = \mathbf{1}^T r = 1.$$

The optimality of v^{k+1} can be proved in the same way.

3 Accelerated Sinkhorn's Algorithm

In this section, we describe accelerated alternating minimization method from [59], which originates from [29, 30, 46], where the latter preprint [30] describes accelerated alternating minimization for non-strongly functions. Our goal is to use the algorithm which has a possibility to use strong convexity. Formally, the dual OT problem (3) is not strongly convex on the whole space. It is strongly convex on any bounded subset of the subspace orthogonal to lines $(u_0 + t\mathbf{1}, v_0 - t\mathbf{1})$. For non-strongly convex problems algorithm (2) has the following sublinear convergence rate $f(x^k) - f(x_*) \leq \frac{4nLR^2}{k^2}$. The proof can be found in [30]. The following Algorithm 2 requires the knowledge of the parameter μ of strong convexity. Notice, that this algorithm run with $\mu = 0$ coincides with its modification for non-strongly functions from [30]. But actually, we were able to outperform the algorithm from [30] by estimating a parameter of strong convexity, but only in iterations.

Algorithm 2. Accelerated Alternating Minimization 2

Input: Starting point x_0 .

Output: x^k

 1: Set $A_0 = 0$, $x^0 = v^0$, $\tau_0 = 1$

 2: **for** $k \geq 0$ **do**

3: Set

$$\beta_k = \operatorname{argmin}_{\beta \in [0,1]} f(x^k + \beta(v^k - x^k)) \quad (7)$$

 4: Set $y^k = x^k + \beta_k(v^k - x^k)$ {Extrapolation step}

 5: Choose $i_k = \operatorname{argmax}_{i \in \{1, \dots, n\}} \|\nabla_i f(y^k)\|_2^2$

 6: Set $x^{k+1} = \operatorname{argmin}_{x \in S_{i_k}(y^k)} f(x)$ {Block minimization}

 7: If L is known choose a_{k+1} s.t. $\frac{a_{k+1}^2}{(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} = \frac{1}{Ln}$
 If L is unknown, find largest a_{k+1} from the equation

$$f(y^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|\nabla f(y^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|v^k - y^k\|_2^2 = f(x^{k+1}) \quad (8)$$

 8: Set $A_{k+1} = A_k + a_{k+1}$, $\tau_{k+1} = \tau_k + \mu a_{k+1}$

 9: Set $v^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^N} \psi_{k+1}(x)$ {Update momentum term}

 10: **end for**

Theoretical justification is given by the following theorem proved in [59].

Theorem 1 [[59] Theorem 1]. *After k steps of Algorithm 2 it holds that*

$$f(x^k) - f(x_*) \leq nLR^2 \min \left\{ \frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu}{nL}}\right)^{k-1} \right\}, \quad (9)$$

where R is an estimate for $\|x_0 - x_*\|$ satisfying $\|x_0 - x_*\| \leq R$.

Applying Algorithm 2 to the dual entropy-regularized optimal transport problem (3) with the objective (4), and using the estimate $L = 2/\gamma$ and $R \leq \sqrt{n/2} \left(\|C\|_\infty - \frac{\gamma}{2} \ln \min_{i,j} \{r_i, c_j\} \right)$ [30], we obtain the following Corollary.

Corollary 1. *Let the histograms r, c be slightly modified, s.t. $\min_{i,j} \{r_i, c_j\} \geq \varepsilon$.*

For example, one can set $(\tilde{r}, \tilde{c}) = (1 - \frac{\varepsilon}{8}) \left((r, c) + \frac{\varepsilon}{n(8-\varepsilon)} (\mathbf{1}, \mathbf{1}) \right)$. Let Algorithm 2 be applied to the dual entropy-regularized optimal transport problem (3) with the objective (4). Let this dual problem have μ -strongly convex objective. Then, after k steps of Algorithm 2 it holds that

$$\varphi(y, z) - \varphi(y_*, z_*) \leq \frac{2n}{\gamma} \left(\|C\|_\infty - \frac{\gamma}{2} \ln \varepsilon \right)^2 \min \left\{ \frac{4}{k^2}, \left(1 - \sqrt{\frac{\mu\gamma}{4}}\right)^{k-1} \right\}. \quad (10)$$

The specification of Algorithm 2 for the dual entropy regularized optimal transport problem (3) with the objective (4) is listed below as Algorithm 3. Each variable has two blocks that naturally correspond to the variables (y, z) in (4).

Algorithm 3. Accelerated Sinkhorn with Strong Convexity

Input: Starting point x_0 .

Output: x^k

1: Set $A_0 = 0$, $x^0 = w^0$, $\tau_0 = 1$

2: **for** $k \geq 0$ **do**

3: Set

$$\beta_k = \operatorname{argmin}_{\beta \in [0,1]} \varphi(x^k + \beta(w^k - x^k)) \quad (11)$$

4: Set $s^k = x^k + \beta_k(w^k - x^k)$ {Extrapolation step}

5: Choose $i_k = \operatorname{argmax}_{i \in \{1,2\}} \|\nabla_i \varphi(s^k)\|_2^2$, where $\nabla \varphi(\cdot)$ is given in (5).

6: **if** $i_k = 1$ **then**

7: $x_1^{k+1} = s_1^k + \ln r - \ln(B(s_1^k, s_2^k) \mathbf{1})$, $x_2^{k+1} = s_2^k$

8: **else**

9: $x_2^{k+1} = s_2^k + \ln c - \ln(B(s_1^k, s_2^k)^T \mathbf{1})$, $x_1^{k+1} = s_1^k$

10: **end if**

11: If L is known choose a_{k+1} s.t. $\frac{a_{k+1}^2}{(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} = \frac{1}{2L}$

If L is unknown, find largest a_{k+1} from the equation

$$\varphi(s^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|\nabla \varphi(s^k)\|_2^2 + \frac{\mu \tau_k a_{k+1}}{2(A_k + a_{k+1})(\tau_k + \mu a_{k+1})} \|w^k - s^k\|_2^2 = \varphi(x^{k+1}) \quad (12)$$

12: Set $A_{k+1} = A_k + a_{k+1}$, $\tau_{k+1} = \tau_k + \mu a_{k+1}$

13: Set $w^{k+1} = w^k - a_{k+1} \nabla \varphi(s^k)$ {Update momentum term}

14: **end for**

We point out that usually, the goal is to solve the primal OT problem. For simplicity, we consider only dual OT problem since the solution of the primal can be reconstructed via standard primal-dual analysis [5, 12, 21, 22] applied to the discussed methods.

4 Estimating a Parameter of Strong Convexity

We build an initial estimate of strong convexity parameter μ by searching the value $\hat{\mu}$ from $[0, \hat{L}]$ which gives the minimum objective value after 10 iterations. \hat{L} is an upper bound on the parameter of Lipschitz continuity of the gradient.

Dependence of the objective value after 10 iterations on μ is presented on Fig. 1.

Then we restart the algorithm from the best point with $\mu = [2\hat{\mu}, \hat{\mu}, \hat{\mu}/2]$ every 10 iterations.

The significant implementation detail is connected with the accumulation of the momentum term (vector w) by Algorithm 2. If we restart the algorithm

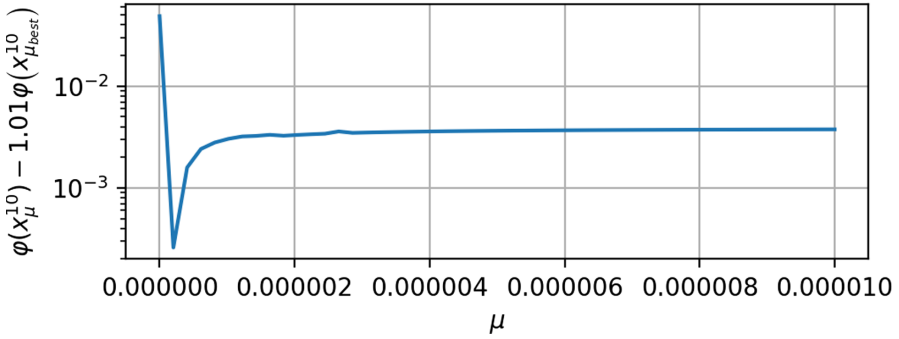


Fig. 1. Empirical dependence of the progress after 10 iterations $h(\mu) = \varphi(x_\mu^{10})$ on the strong convexity parameter μ used in Algorithm 2. The initial value of μ is chosen as a point of minimum of this dependence.

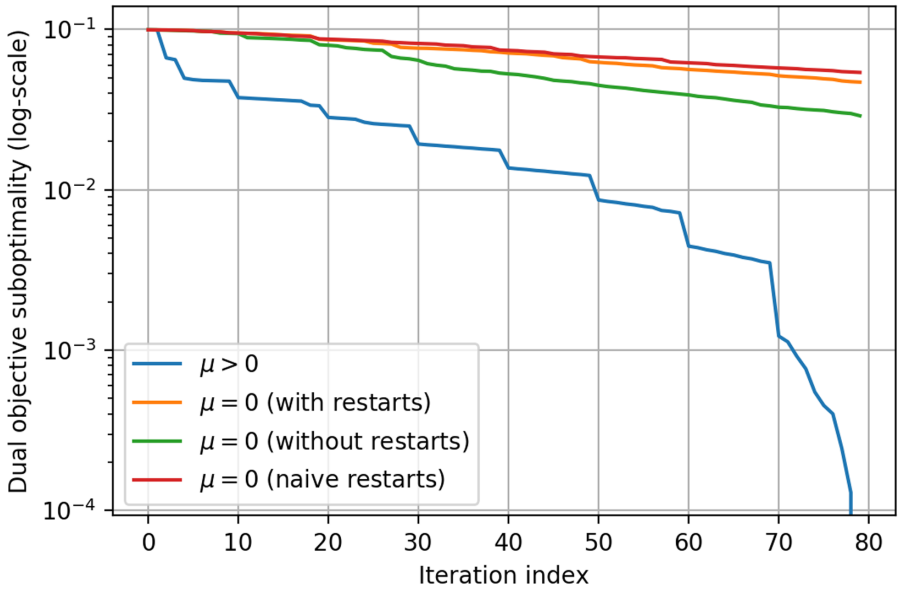


Fig. 2. Performance of Algorithm 2 with the optimal choice of parameter μ on the dual entropy regularized optimal transport problem (3).

naively (with $w^0 = x^0$), we will lose all accumulated information. That is why, we restart the algorithm with w^0 obtained from the last iteration of the previous restart. In order to compare the difference we bring to comparison the case of naive restarts.

As we can see from (Fig. 2), the value of the dual objective decreases faster when one uses the method with positive strong convexity parameter than when one uses the method with $\mu = 0$.

5 Conclusion

In this work we have investigated, how strong convexity can be used to accelerate the accelerated Sinkhorn’s algorithm for the dual entropy-regularized optimal transport problem. As we see, the accelerated alternating minimization method in its particular version of accelerated Sinkhorn’s algorithm with strong convexity can utilize an estimated value of the strong convexity parameter to converge faster. We underline that it is not clear how one can incorporate this information in the standard Sinkhorn’s algorithm to accelerate it. As future work we would like to note the study of automatic strong convexity adaptation procedures like in [25, 50], which are now adapted for gradient methods and coordinate descent methods, rather than for alternating minimization methods. Among other extensions, it would be interesting to understand whether restricted strong convexity improves convergence rates of the methods for approximating Wasserstein barycenter [1, 14, 19, 34, 38, 60] and related distributed optimization methods [18]. Another direction is an application to similar optimization problems, which arise in transportation research in connection to equilibrium in congestion traffic models and traffic demands matrix estimation [7, 20] and multimarginal optimal transport [39]. Finally, we use regularization for the OT problem to make the dual problem have Lipschitz gradient. It would be interesting to use universal methods [32, 45, 57] for the dual OT problem.

References

1. Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43**(2), 904–924 (2011)
2. Allen-Zhu, Z., Qu, Z., Richtarik, P., Yuan, Y.: Even faster accelerated coordinate descent using non-uniform sampling. In: Balcan, M.F., Weinberger, K.Q. (eds.) *Proceedings of The 33rd International Conference Machine Learning. Proceedings of Machine Learning Research*, vol. 48, pp. 1110–1119. PMLR, New York, New York, USA (2016), <http://proceedings.mlr.press/v48/allen-zhuc16.html>, [arXiv:1512.09103](https://arxiv.org/abs/1512.09103)
3. Altschuler, J., Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30, pp. 1961–1971. Curran Associates, Inc. (2017). [arXiv:1705.09634](https://arxiv.org/abs/1705.09634)
4. Andresen, A., Spokoiny, V.: Convergence of an alternating maximization procedure. *JMLR* **17**(63), 1–53 (2016). <http://jmlr.org/papers/v17/15-392.html>
5. Anikin, A.S., Gasnikov, A.V., Dvurechensky, P.E., Tyurin, A.I., Chernov, A.V.: Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. *Comput. Math. Math. Phys.* **57**(8), 1262–1276 (2017). <https://doi.org/10.1134/S0965542517080048>
6. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017). [arXiv:1701.07875](https://arxiv.org/abs/1701.07875)
7. Baimurzina, D.R., et al.: Universal method of searching for equilibria and stochastic equilibria in transportation networks. *Comput. Math. Math. Phys.* **59**(1), 19–33 (2019). <https://doi.org/10.1134/S0965542519010020>, [arXiv:1701.02473](https://arxiv.org/abs/1701.02473)

8. Beck, A.: On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM J. Optim.* **25**(1), 185–209 (2015)
9. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. *SIAM J. Optim.* **23**(4), 2037–2060 (2013)
10. Bertsekas, D.P., Tsitsiklis, J.N.: *Parallel and Distributed Computation: Numerical Methods*, vol. 23. Prentice Hall Englewood Cliffs, Upper Saddle River (1989)
11. Bigot, J., Gouet, R., Klein, T., López, A.: Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.* **53**(1), 1–26 (2017)
12. Chernov, A., Dvurechensky, P., Gasnikov, A.: Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In: Kochetov, Y., Khachay, M., Beresnev, V., Nurminski, E., Pardalos, P. (eds.) *DOOR 2016*. LNCS, vol. 9869, pp. 391–403. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44914-2_31
13. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 26, pp. 2292–2300. Curran Associates, Inc. (2013)
14. Cuturi, M., Doucet, A.: Fast computation of Wasserstein barycenters. In: Xing, E.P., Jebara, T. (eds.) *Proceedings of the the 31st International Conference Machine Learning*, vol. 32, pp. 685–693. PMLR, Beijing, China (2014)
15. Cuturi, M., Peyré, G.: A smoothed dual approach for variational Wasserstein problems. *SIAM J. Imaging Sci.* **9**(1), 320–343 (2016)
16. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. *Commun. Pure Appl. Math.* **63**(1), 1–38 (2010)
17. Diakonikolas, J., Orecchia, L.: Alternating randomized block coordinate descent. In: Dy, J., Krause, A. (eds.) *Proceedings of the the 35th International Conference Machine Learning*, vol. 80, pp. 1224–1232. PMLR, Stockholm, Sweden (2018). <http://proceedings.mlr.press/v80/diakonikolas18a.html>
18. Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., Uribe, C.A.: On primal and dual approaches for distributed stochastic convex optimization over networks. In: 2019 IEEE 58th Conference on Decision and Control, pp. 7435–7440 (2019). <https://doi.org/10.1109/CDC40024.2019.9029798>, [arXiv:1903.09844](https://arxiv.org/abs/1903.09844)
19. Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C.A., Nedić, A.: Decentralize and randomize: faster algorithm for Wasserstein barycenters. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 31, pp. 10783–10793. *Neural Information Processing Systems 2018*, Curran Associates, Inc. (2018). [arXiv:1806.03915](https://arxiv.org/abs/1806.03915)
20. Dvurechensky, P., Gasnikov, A., Gasnikova, E., Matsievsky, S., Rodomanov, A., Usik, I.: Primal-dual method for searching equilibrium in hierarchical congestion population games. In: *Supplementary Proceedings of the International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016)*, Vladivostok, Russia, 19–23 September 2016, pp. 584–595 (2016). [arXiv:1606.08988](https://arxiv.org/abs/1606.08988)
21. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm. In: Dy, J., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 1367–1376. PMLR (2018). [arXiv:1802.04367](https://arxiv.org/abs/1802.04367)

22. Dvurechensky, P., Gasnikov, A., Omelchenko, S., Tiurin, A.: A stable alternative to Sinkhorn's algorithm for regularized optimal transport. In: Kononov, A., et al. (eds.) *Mathematical Optimization Theory and Operations Research (MOTOR 2020)*. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49988-4_28, [arXiv:1706.07622](https://arxiv.org/abs/1706.07622)
23. Dvurechensky, P., Gasnikov, A., Tiurin, A.: Randomized similar triangles method: a unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method) (2017). [arXiv:1707.08486](https://arxiv.org/abs/1707.08486)
24. Ebert, J., Spokoiny, V., Suvorikova, A.: Construction of non-asymptotic confidence sets in 2-Wasserstein space (2017). [arXiv:1703.03658](https://arxiv.org/abs/1703.03658)
25. Fercoq, O., Qu, Z.: Restarting the accelerated coordinate descent method with a rough strong convexity estimate. *Comput. Optim. Appl.* **75**(1), 63–91 (2020). <https://doi.org/10.1007/s10589-019-00137-2>
26. Fercoq, O., Richtárik, P.: Accelerated, parallel, and proximal coordinate descent. *SIAM J. Optimiz.* **25**(4), 1997–2023 (2015)
27. Gasnikov, A., Dvurechensky, P., Usmanova, I.: On accelerated randomized methods. *Proc. Moscow Inst. Phys. Technol.* **8**(2), 67–100 (2016), (in Russian). [arXiv:1508.02182](https://arxiv.org/abs/1508.02182)
28. Gasnikov, A., Gasnikova, E., Mendel, M., Chepurchenko, K.: Evolutionary derivations of entropy model for traffic demand matrix calculation. *Matematicheskoe Modelirovanie* **28**(4), 111–124 (2016). (in Russian)
29. Guminov, S.V., Nesterov, Y.E., Dvurechensky, P.E., Gasnikov, A.V.: Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Doklady Math.* **99**(2), 125–128 (2019)
30. Guminov, S., Dvurechensky, P., Tupitsa, N., Gasnikov, A.: Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated iterative bregman projections (2019). [arXiv:1906.03622](https://arxiv.org/abs/1906.03622)
31. Ho, N., Nguyen, X., Yurochkin, M., Bui, H.H., Huynh, V., Phung, D.: Multilevel clustering via Wasserstein means. In: Precup, D., Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1501–1509. PMLR (2017)
32. Kamzolov, D., Dvurechensky, P., Gasnikov, A.V.: Universal intermediate gradient method for convex problems with inexact oracle. *Optim. Methods Softw.* 1–28 (2020). <https://doi.org/10.1080/10556788.2019.1711079>, [arXiv:1712.06036](https://arxiv.org/abs/1712.06036)
33. Kolouri, S., Park, S.R., Thorpe, M., Slepcev, D., Rohde, G.K.: Optimal mass transport: signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **34**(4), 43–59 (2017)
34. Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Uribe, C.: On the complexity of approximating Wasserstein barycenters. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 97, pp. 3530–3540. PMLR, Long Beach, California, USA (2019). [arXiv:1901.08686](https://arxiv.org/abs/1901.08686)
35. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: *Proceedings of the the 32nd International Conference Machine Learning (ICML 2015)*, vol. 37, pp. 957–966. PMLR (2015)
36. Lee, Y.T., Sidford, A.: Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In: *Proceedings of the 2013 IEEE 54th FOCS (FOCS 2013)*, pp. 147–156. IEEE Computer Society, Washington, DC, USA (2013). [arXiv:1305.1922](https://arxiv.org/abs/1305.1922)

37. Lin, Q., Lu, Z., Xiao, L.: An accelerated proximal coordinate gradient method. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 27, pp. 3059–3067. Curran Associates, Inc. (2014). [arXiv:1407.1296](https://arxiv.org/abs/1407.1296)
38. Lin, T., Ho, N., Chen, X., Cuturi, M., Jordan, M.I.: Computational hardness and fast algorithm for fixed-support Wasserstein barycenter (2020). [arXiv:2002.04783](https://arxiv.org/abs/2002.04783)
39. Lin, T., Ho, N., Cuturi, M., Jordan, M.I.: On the Complexity of Approximating Multimarginal Optimal Transport (2019). [arXiv:1910.00152](https://arxiv.org/abs/1910.00152)
40. Lin, T., Ho, N., Jordan, M.: On efficient optimal transport: an analysis of greedy and accelerated mirror descent algorithms. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 97, pp. 3982–3991. PMLR, Long Beach, California, USA (2019)
41. McCullagh, P., Nelder, J.: *Generalized Linear Models*. In: *Monographs Statistics and Applied Probability Series*, 2nd edn. Chapman & Hall (1989)
42. McLachlan, G., Krishnan, T.: *The EM Algorithm and Extensions*. In: *Wiley Series in Probability and Statistics*. Wiley (1996)
43. Nesterov, Y.: Smooth minimization of non-smooth functions. *Math. Program.* **103**(1), 127–152 (2005)
44. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optimiz.* **22**(2), 341–362 (2012)
45. Nesterov, Y.: Universal gradient methods for convex optimization problems. *Math. Program.* **152**(1), 381–404 (2015). <https://doi.org/10.1007/s10107-014-0790-0>
46. Nesterov, Y., Gasnikov, A., Guminov, S., Dvurechensky, P.: Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optim. Methods Softw.* 1–28 (2020). <https://doi.org/10.1080/10556788.2020.1731747>, [arXiv:1809.05895](https://arxiv.org/abs/1809.05895)
47. Nesterov, Y., Stich, S.U.: Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM J. Optim.* **27**(1), 110–123 (2017)
48. Ortega, J., Rheinboldt, W.: *Iterative Solution of Nonlinear Equations in Several Variables*. *Classics in Applied Mathematics*. SIAM (1970)
49. Panaretos, V.M., Zemel, Y.: Amplitude and phase variation of point processes. *Ann. Statist.* **44**(2), 771–812 (2016)
50. Roulet, V., d’Aspremont, A.: Sharpness, restart and acceleration. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30, pp. 1119–1129. Curran Associates, Inc. (2017). <http://papers.nips.cc/paper/6712-sharpness-restart-and-acceleration.pdf>
51. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover’s distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40**(2), 99–121 (2000). <https://doi.org/10.1023/A:1026543900054>
52. Saha, A., Tewari, A.: On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM J. Optim.* **23**(1), 576–601 (2013)
53. Sandler, R., Lindenbaum, M.: Nonnegative matrix factorization with earth mover’s distance metric for image analysis. *IEEE TPAMI* **33**(8), 1590–1602 (2011)
54. Shalev-Shwartz, S., Zhang, T.: Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In: Xing, E.P., Jebara, T. (eds.) *Proceedings of the 31st International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 32, pp. 64–72. PMLR, Beijing, China (2014). <http://proceedings.mlr.press/v32/shalev-shwartz14.html>, [arXiv:1309.2375](https://arxiv.org/abs/1309.2375)
55. Sinkhorn, R.: Diagonal equivalence to matrices with prescribed row and column sums. II. *Proc. Am. Math. Soc.* **45**, 195–198 (1974)

56. Solomon, J., Rustamov, R.M., Guibas, L., Butscher, A.: Wasserstein propagation for semi-supervised learning. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), vol. 32, pp. I-306-I-314. PMLR (2014)
57. Stonyakin, F.S., et al.: Gradient methods for problems with inexact model of the objective. In: Khachay, M., Kochetov, Y., Pardalos, P. (eds.) Mathematical Optimization Theory and Operations Research (MOTOR 2019). Lecture Notes in Computer Science, vol. 11548, pp. 97–114. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22629-9_8, [arXiv:1902.09001](https://arxiv.org/abs/1902.09001)
58. Sun, R., Hong, M.: Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), vol. 1, pp. 1306–1314. MIT Press, Cambridge, MA, USA (2015)
59. Tupitsa, N., Dvurechensky, P., Gasnikov, A., Guminov, S.: Alternating Minimization Methods for Strongly Convex Optimization (2019). [arXiv:1911.08987](https://arxiv.org/abs/1911.08987)
60. Uribe, C.A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Nedić, A.: Distributed computation of Wasserstein barycenters over networks. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 6544–6549 (2018). [arXiv:1803.02933](https://arxiv.org/abs/1803.02933)
61. Villani, C.: Optimal Transport: Old and New, vol. 338. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-71050-9>
62. Werman, M., Peleg, S., Rosenfeld, A.: A distance metric for multidimensional histograms. *Comput. Vis. Graph. Image Process.* **32**(3), 328–336 (1985)
63. Zhang, Y., Roughan, M., Lund, C., Donoho, D.L.: Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach. *IEEE/ACM Trans. Netw.* **13**(5), 947–960 (2005)