On a Combination of Alternating Minimization and Nesterov's Momentum

Sergey Guminov [1 2 3] **Pavel Dvurechensky** [4 2 3] **Nazarii Tupitsa** [1 2 3] **Alexander Gasnikov** [1 2 3]

Abstract

Alternating minimization (AM) procedures are practically efficient in many applications for solving convex and non-convex optimization problems. On the other hand, Nesterov's accelerated gradient is theoretically optimal first-order method for convex optimization. In this paper we combine AM and Nesterov's acceleration to propose an accelerated alternating minimization algorithm. We prove $1/k^2$ convergence rate in terms of the objective for convex problems and $1/k$ in terms of the squared gradient norm for non-convex problems, where k is the iteration counter. Our method does not require any knowledge of neither convexity of the problem nor function parameters such as Lipschitz constant of the gradient, i.e. it is adaptive to convexity and smoothness and is uniformly optimal for smooth convex and non-convex problems. Further, we develop its primal-dual modification for strongly convex problems with linear constraints and prove the same $1/k^2$ for the primal objective residual and constraints feasibility.

1. Introduction

Alternating minimization (AM) optimization algorithms have been known for a long time (Ortega & Rheinboldt, 1970; Bertsekas & Tsitsiklis, 1989). These algorithms assume that the decision variable is divided into several blocks and minimization in each block can be done explicitly. 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 (McCullagh & Nelder, 1989) and sparse recovery (Daubechies et al., 2010). The famous Expectation-maximization (EM) algorithm can also be seen as an AM algorithm (McLachlan & Krishnan, 1996; Andresen & Spokoiny, 2016).

The initial motivation for this paper was accelerating algorithms for optimal transport (OT) applications, which are widespread in the machine learning community (Cuturi, 2013; Cuturi & Doucet, 2014; Arjovsky et al., 2017). The ubiquitous Sinkhorn's algorithm can be seen as an alternating minimization algorithm for the dual to the entropy-regularized optimal transport problem. Recent Greenkhorn algorithm (Altschuler et al., 2017), which is a greedy version of Sinkhorn's algorithm, is a greedy modification of an AM algorithm. For the Wasserstein barycenter (Agueh & Carlier, 2011) problem, the extension of the Sinkhorn's algorithm is known as the Iterative Bregman Projections (IBP) algorithm (Benamou et al., 2015), which can be seen as an alternating minimization procedure (Kroshnin et al., 2019). This motivated us to have a wider look on alternating minimization algorithms and try to accelerate general AM algorithm.

Sublinear $1/k$ convergence rate was proved for general AM algorithm for $n = 2$ in (Beck, 2015). Despite the same convergence rate as for the gradient method, AM-algorithms converge faster in practice as they are free of the choice of the step-size and are adaptive to the local smoothness of the problem. At the same time, there are accelerated gradient methods (AGM) which use a momentum term to have a faster convergence rate of $1/k^2$ (Nesterov, 1983) and use gradient steps rather than block minimization. Our goal in this paper is to combine the idea of alternating minimization and momentum acceleration to propose an accelerated alternating minimization method. As applications of our general approach, we develop accelerated alternating least squares algorithm and apply it to a non-convex collaborative filtering problem, and propose accelerated Sinkhorn's algorithm for OT distances and accelerated Iterative Bregman Projections algorithm for Wasserstein barycenters.

Related work. Besides mentioned above works on AM algorithms, we mention (Beck & Tetruashvili, 2013; Saha & Tewari, 2013; Sun & Hong, 2015), where non-asymptotic convergence rates for AM algorithms for convex problems were proposed and their connection with cyclic coordinate

[1]Moscow Institute of Physics and Technology, Dolgoprudny, Russia [2]Institute for Information Transmission Problems RAS, Moscow, Russia [3]HDI Lab @ National Research University Higher School of Economics, Russian Federation [4]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany. Correspondence to: Nazarii Tupitsa <tupitsa@phystech.edu>, Pavel Dvurechensky <pavel.dvurechensky@wias-berlin.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

descent was discussed, but the analyzed algorithms are not accelerated. Accelerated versions are known for random coordinate descent methods (Nesterov, 2012; Lee & Sidford, 2013; Shalev-Shwartz & Zhang, 2014; Lin et al., 2014; Fercoq & Richtárik, 2015; Allen-Zhu et al., 2016; Nesterov & Stich, 2017; Alacaoglu et al., 2017), cyclic block coordinate descent (Beck & Tetruashvili, 2013), greedy coordinate descent (Lu et al., 2018). These ACD methods are designed for convex problems and use momentum term, but they require knowledge of block-wise Lipschitz constants, i.e. are not parameter-free. A hybrid accelerated random block-coordinate method (AAR-BCD) with exact minimization in the last block was proposed in (Diakonikolas & Orecchia, 2018a) for convex problems. Unlike our greedy choice of the updated block they use random choice and the parameters of the algorithm depend on the block Lipschitz constants, meaning that AAR-BCD algorithm is not parameter-free. An extension providing a two-block accelerated alternating minimization algorithm is available in the updated version (Diakonikolas & Orecchia, 2018b) for the convex case. This method is deterministic and it is explained how to make it parameter-free. At the same time neither of two algorithms from (Diakonikolas & Orecchia, 2018b) have an analysis for non-convex problems or problems with linear constraints, yet it seems that such extensions are possible for their methods. We also underline that our definition of the algorithm parameters, in particular, the sequence a_k in Algorithm 1, is different from theirs.

The summary of the related works on alternating minimization and coordinate methods is presented in the Table 1, where P-F stands for parameter-free, Acc. for accelerated, N-C for non-convex, P-D for primal-dual and B-N for number of blocks.

Table 1. Summary of the related works

	P-F	Acc.	N-C	P-D	B-N
AM [1]	$\surd$	$\times$	$\times$	$\times$	2
AM [2]	$\surd$	$\times$	$\times$	$\times$	any
ACD [3]	$\times$	$\surd$	$\times$	$\surd$	any
AAR-BCD[4]	$\times$	$\surd$	$\times$	$\times$	any
AAM[5]	$\surd$	$\surd$	$\times$	$\times$	2
This paper	$\surd$	$\surd$	$\surd$	$\surd$	any

Concerning the OT problem, the most used algorithm is Sinkhorn's algorithm (Sinkhorn, 1974; Cuturi, 2013). Its complexity for the OT problem was first analyzed in (Altschuler et al., 2017) and improved in (Dvurechensky et al., 2018b). An accelerated gradient descent method in application to OT problem was also analyzed in (Dvurechensky et al., 2018b) with a better dependence on k in the rate, but worse dependence on the dimension of the problem, see also (Lin et al., 2019a). (Altschuler et al., 2017) propose a greedy variant called Greenkhorn together with complexity analysis, which was improved in (Lin et al., 2019a). In an unpublished preprint (Lin et al., 2019b) the authors propose a randomized accelerated version of Sinkhorn's algorithm. We summarize the complexity of existing methods for OT in the Table 2. N is the number of points in the histogram, C is the transportation cost matrix, ε desired accuracy. The complexity of approximating Wasserstein barycenter was analyzed in (Kroshnin et al., 2019), where the complexity by Iterative Bregman Projections algorithm and a variant of accelerated gradient method was obtained. Previous works (Cuturi & Doucet, 2014; Benamou et al., 2015; Staib et al., 2017; Claici et al., 2018) did not give an explicit complexity bounds for approximating barycenter. But there are plenty of algorithms for approximating WB including accelerated gradient method plus Sinkhorn's algorithm (Cuturi & Doucet, 2014), gradient-type methods (Cuturi & Peyré, 2016), accelerated primal-dual gradient descent (Dvurechensky et al., 2018a; Krawtschenko et al., 2020), stochastic gradient descent (Claici et al., 2018; Tiapkin et al., 2020), distributed and parallel gradient descent (Staib et al., 2017; Uribe et al., 2018; Rogozin et al., 2021), alternating direction method of multipliers (ADMM)(Ye et al., 2017; Yang et al., 2018) and interior-point algorithm (Ge et al., 2019). Only recently the question of complexity got some answers. Namely, two approaches for approximating Wasserstein barycenter based on entropic regularization (Cuturi, 2013) were analyzed. The first approach is based on Iterative Bregman Projection (IBP) algorithm (Benamou et al., 2015), which can be considered as a general alternating projections algorithm and also as a generalization of the Sinkhorn's algorithm (Sinkhorn, 1974). The second approach Primal-Dual Accelerateg Gradient Descent (PDAGD) is based on constructing a dual problem and solving it by primal-dual accelerated gradient descent. For both approaches, it was shown, how the regularization parameter should be chosen in order to approximate the original, non-regularized barycenter. In (Lin et al., 2020) the authors proposed a variant of the Iterative Bregman Projection (IBP) algorithm, which they called FastIBP. Very recently (Dvinskikh & Tiapkin, 2021) provided two algorithms to compute Wasserstein barycenter,

Table 2. Summary of OT algorithms

Algorithm	Complexity
Sinkhorn (Cuturi, 2013; Dvurechensky et al., 2018b)	$N^2\|C\|_\infty^2/\varepsilon^2$
Greenkhorn (Altschuler et al., 2017; Lin et al., 2019a)	$N^2\|C\|_\infty^2/\varepsilon^2$
Randkhorn (Lin et al., 2019b)	$N^{7/3}\|C\|_\infty^{4/3}/\varepsilon$
APDA(G/M)D (Dvurechensky et al., 2018b; Lin et al., 2019a)	$N^{5/2}\|C\|_\infty/\varepsilon$
Mirror-Prox (Jambulapati et al., 2019)	$N^2\|C\|_\infty/\varepsilon$
This paper	$N^{5/2}\|C\|_\infty/\varepsilon$

[1](Beck & Tetruashvili, 2013; Beck, 2015)

[2](Saha & Tewari, 2013; Sun & Hong, 2015)

[3](Nesterov, 2012; Lee & Sidford, 2013; Fercoq & Richtárik, 2015; Shalev-Shwartz & Zhang, 2014; Allen-Zhu et al., 2016; Nesterov & Stich, 2017; Beck & Tetruashvili, 2013; Lu et al., 2018; Lin et al., 2014; Alacaoglu et al., 2017)

[4](Diakonikolas & Orecchia, 2018a)

[5](Diakonikolas & Orecchia, 2018b)

one of them has the best theoretical convergence guarantees.

We summarize the known complexity bounds from the literature in Table 3. We underline that despite many advantages of the entropic regularization, in some situations other reguarizations provide more robust results (Blondel et al., 2018). Our proposed method is flexible enough to allow efficient computations with regularizers other than entopic both for OT and WB problems.

Table 3. Summary of algorithms for Wasserstein barycenters

Algorithm	Complexity
IBP (Benamou et al., 2015; Kroshnin et al., 2019)	$mN^2\|C\|_\infty^2/\varepsilon^2$
PDAGD (Kroshnin et al., 2019)	$mN^{5/2}\|C\|_\infty/\varepsilon$
FastIBP (Lin et al., 2020)	$mN^{7/3}\|C\|_\infty^{4/3}/\varepsilon^{4/3}$
Area Convexity (Dvinskikh & Tiapkin, 2021)	$mN^2\|C\|_\infty/\varepsilon$
Mirror-Prox (Dvinskikh & Tiapkin, 2021)	$mN^{5/2}\|C\|_\infty/\varepsilon$
This paper	$mN^{5/2}\|C\|_\infty/\varepsilon$

Our contributions. For objectives with n blocks of variables we introduce an accelerated alternating minimization method with $O(n/k^2)$ convergence rate for the objective values in smooth unconstrained convex problems and $O(n/k)$ convergence rate in terms of the squared norm of the gradient both for convex and non-convex smooth unconstrained problems. Thus, in terms of the dependence on the iteration counter k our algorithm achieves uniformly the best possible rates in convex case (same as for AGM) and in non-convex case (same as for gradient descent (GD)). Moreover, the algorithm automatically adapts to convexity and smoothness: it is completely the same for convex and non-convex settings and does not need to know in advance whether the problem is convex or not, i.e. is uniform for smooth convex and non-convex problems; it does not need to know the Lipschitz constant of the gradient, i.e. is parameter-free. Parameter-free versions exist also for AGM and GD (see, e.g. (Nesterov, 2013)), but they are based on a different idea of backtracking line-search and do not explore the block structure of the problem and block minimization for acceleration in practice.

The main idea of our algorithm is to combine block-wise minimization and the extrapolation (also known as momentum) step which is usually used in accelerated gradient methods. We also show that in the convex setting the proposed method is primal-dual, meaning that if we apply it to a dual problem for a linearly constrained strongly convex problem, we can reconstruct the solution of the primal problem with the same convergence rate. In the follow-up work (Tupitsa et al., 2021) a modification of AAM is proposed and analyzed for strongly convex problems.

To highlight the new properties of our method, the proven convergence rate for non-convex problems and the primal-dual analysis, we consider two particular applications. First, we consider a non-convex collaborative filtering problem and show empirically that our algorithm outperforms the standard alternating least squares algorithm. Second, we apply it to the dual entropy-regularized OT problem to obtain the Accelerated Sinkhorn's algorithm. The Primal-dual analysis is crucial here since the goal is to find the transportation plan, i.e. the primal variable, by solving the dual problem. Our method has complexity comparable to the existing methods and in the experiments, we show that our general method outperforms specific baselines for this problem, including Sinkhorn's algorithm. Importantly, we use a non-standard formulation of the dual entropy-regularized OT problem in the form of minimization of a softmax function. Moreover, *our algorithm is more flexible since it can solve OT problems with other types of regularization, e.g. by squared Euclidean norm.* Finally, in the supplementary, we apply our accelerated primal-dual AM algorithm to the Wasserstein Barycenter (WB) problem and propose an accelerated Iterative Bregman Projection algorithm with the complexity $\widetilde{O}\left(mN^{2.5}/\varepsilon\right)$ to find a barycenter of m histograms of dimension N. This bound is better than the complexity bound for the standard Iterative Bregman Projection algorithm (Kroshnin et al., 2019) $\widetilde{O}\left(mN^2/\varepsilon^2\right)$ in terms of ε. In the follow-up paper (Tupitsa et al., 2020) the AAM method is applied to a more general multimarginal optimal transport problem and complexity estimates are obtained that are better in some regimes than the ones in the literature.

Paper organization. In Sect. 2 we consider the general setting of minimizing a smooth objective function using block minimization. We introduce our uniform accelerated alternating minimization (AAM) method for convex and non-convex problems together with its primal-dual modification for convex linearly constrained problems. In Sect. 3 we study the primal-dual properties of the method. In Sect.4 we discuss the application of our method to the collaborative filtering problem and provide experiments on the Last.fm dataset 360K for the collaborative filtering problem. In Sect. 5 we describe the OT and the WB problems and their entropy-regularized versions, together with the dual for the latters, that are non-standard. Then, we propose the Accelerated versions of Sinkhorn's algorithm and IBP algorithm and obtain their theoretical complexity, and provide the results of numerical experiments on MNIST dataset for both problems and additionally provide experiments for WB problem with Gaussian measures. The proofs of all stated results, the explicit form of algorithms and the application of the proposed methods to the regularized Wasserstein Barycenter problem may be found in the supplement. In Section 6 we provide numerical experiment for least squares problem for linear regression.[6]

[6]Code for all presented algorithms is available at `https://github.com/nazya/AAM`

2. Accelerated Alternating Minimization

In this section we consider the minimization problem $\min_{x\in\mathbb{R}^N} f(x)$, where $f(x)$ is continuously differentiable and, in general *non-convex*, L-smooth function, the latter meaning that its gradient is L-Lipschitz, i.e. $\forall\ x,\ y \in \mathbb{R}^N \quad \|\nabla f(x) - \nabla f(y)\|_2 \leqslant L\|x-y\|_2$. We assume that the space is equipped with the Euclidean norm $\|\cdot\|_2$ and that the problem has at least one solution, denoted by x^*. The set $\{1,\dots,N\}$ of indices of the basis vectors $\{e_i\}_{i=1}^N$ is divided into n disjoint subsets (blocks) I_p, $p\in\{1,\dots,n\}$. Let $S_p(x) = x + \mathrm{span}\{e_i :\ i\in I_p\}$, i.e. the affine subspace containing x and all the points differing from x only over the block p. We use x_i to denote the components of x corresponding to the block i and $\nabla_i f(x)$ to denote the gradient corresponding to the block i. We will further require that for any $p\in\{1,\dots,n\}$ and any $z\in\mathbb{R}^N$ the problem $\min_{x\in S_p(z)} f(x)$ has a solution, and this solution is easily computable.

Algorithm 1 Accelerated Alternating Minimization (AAM)

Input: Starting point x_0.
Output: x^k

1: Set $A_0 = 0$, $x^0 = v^0$.
2: **for** $k \geqslant 0$ **do**
3: Set $\beta_k = \operatorname*{argmin}_{\beta\in[0,1]} f\left(x^k + \beta(v^k - x^k)\right)$
4: Set $y^k = x^k + \beta_k(v^k - x^k)$
5: Choose $i_k = \operatorname*{argmax}_{i\in\{1,\dots,n\}} \|\nabla_i f(y^k)\|_2^2$
6: Set $x^{k+1} = \operatorname*{argmin}_{x\in S_{i_k}(y^k)} f(x)$
7: Find a_{k+1}, $A_{k+1} = A_k + a_{k+1}$ from

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_2^2 = f(x^{k+1})$$

8: Set $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$
9: **end for**

Our accelerated alternating minimization method is listed as Algorithm 1. This algorithm combines AM and Nesterov's momentum and, thus, a full-gradient step 8 is inherited and AM updates are used for faster empirical convergence than AGD. In some sense this is similar to AM compared to gradient descent: theoretical rates are the same, but AM has practical benefits. At the same time, full gradient step 8 is not more expensive than other steps. For example, in the OT applications, full gradient costs nearly the same as block minimization. We underline that Algorithm 1 does not require knowledge of whether the function is convex or non-convex and does not require knowledge of any parameters of the function. The latter is in contrast to standard accelerated gradient descent (Nesterov, 2004), accelerated random coordinate descent (Nesterov, 2012; Lee & Sidford, 2013; Shalev-Shwartz & Zhang, 2014; Lin et al., 2014; Fercoq & Richtárik, 2015; Allen-Zhu et al., 2016; Nesterov & Stich, 2017), accelerated cyclic block coordinate descent (Beck & Tetruashvili, 2013), accelerated greedy coordinate descent (Lu et al., 2018), all of which require the knowledge of either the constant L or block-wise Lipschitz constants. Our method is also different from parameter-free versions of AGM that use a backtracking line-search as, e.g., in (Nesterov, 2013). Parameter-free nature of our method is achieved by applying steps 3 and 7. In standard methods a_k is defined by an equation containing L and β_k is defined based on a_k. We prove that in the case when f is convex and L-smooth, our method has the accelerated $O(n/k^2)$ rate for the objective residual and, for a general setting of possibly non-convex L-smooth functions it guarantees that the squared norm of the gradient decreases as $O(n/k)$. Importantly, the obtained convergence rate in the convex case is n times better than the rate for accelerated random coordinate descent (Nesterov, 2012), which is $O(n^2/k^2)$. The main convergence rate theorem for Algorithm 1 is as follows.

Theorem 1. *a) Assume that f is (possibly non-convex) L-smooth function w.r.t. $\|\cdot\|_2$. Then, after k steps of Algorithm 1,*

$$\min_{i=0,\dots,k}\|\nabla f(y^i)\|_2^2 \leqslant \frac{2nL(f(x^0)-f(x^*))}{k}.$$

b) Assume additionally that f is convex. Then, after k steps of Algorithm 1,

$$f(x^k) - f(x^*) \leqslant \frac{2nL\|x^0-x^*\|_2^2}{k^2}.$$

Proof of Theorem 1, a). L-smoothness of f together with the fact that $x^{k+1} = \operatorname{argmin}_{x\in S_{i_k}(y_k)} f(x)$ where $i_k = \operatorname{argmax}_i \|\nabla_i f(y^k)\|_2^2$ implies

$$f(y^k) - \frac{1}{2L}\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant f(x^{k+1}).$$

Since $i_k = \operatorname{argmax}_i \|\nabla_i f(y^k)\|_2^2$ we have that

$$\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant \frac{1}{n}\|\nabla f(y^k)\|_2^2$$

and

$$f(x^{k+1}) \leqslant f(y^k) - \frac{1}{2nL}\|\nabla f(y^k)\|_2^2 \leqslant f(x^k) - \frac{1}{2nL}\|\nabla f(y^k)\|_2^2.$$

Summing this up for $i = 0,\dots,k$, we obtain

$$f(x^0) - f(x^*) \geqslant f(x^0) - f(x^{N+1}) \geqslant \frac{k}{2nL}\min_{i=0,\dots,k}\|\nabla f(y^i)\|_2^2.$$

Consequently, we may guarantee $\min\limits_{i=0,\dots,k} \|\nabla f(y^i)\|_2^2 \leqslant \frac{2nL(f(x^0)-f(x^*))}{k}$. □

To prove the part b) of Theorem 1 we firstly state an auxiliary lemma. Let us introduce an auxiliary sequence of functions defined as $\psi_0(x) = \frac{1}{2}\|x - x^0\|^2$, $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k\rangle\}$. It is easy to see that $v^k = \operatorname*{argmin}_{x\in\mathbb{R}^N} \psi_k(x)$.

Lemma 2. *After k steps of Algorithm 1 it holds that*

$$A_k f(x^k) \leqslant \min_{x\in\mathbb{R}^N} \psi_k(x) = \psi_k(v^k). \tag{1}$$

Moreover, $A_k \geqslant \frac{k^2}{4Ln}$, where n is the number of blocks.

Proof of Theorem 1 b). From the convexity of $f(x)$ we have $\frac{1}{A_k}\sum_{i=0}^{k-1} a_{k+1}(f(y^k)+\langle\nabla f(y^k), x-y^k\rangle) \leqslant f(x^*)$. From Lemma 2, using the standard argument (Nesterov, 2005), we have

$$\begin{aligned} &A_k f(x^k) \leqslant \psi_k(v^k) \leqslant \psi_k(x^*) \\ &= \frac{1}{2}\|x^* - x^0\|_2^2 + \sum_{i=0}^{k-1} a_{i+1}(f(y^i) + \langle \nabla f(y^i), x^* - y^i\rangle) \\ &\qquad\qquad \leqslant A_k f(x^*) + \frac{1}{2}\|x^* - x^0\|_2^2. \end{aligned}$$

Since $A_k \geqslant \frac{k^2}{4nL}$, we finally obtain the statement of the theorem $f(x^k) - f(x^*) \leqslant \frac{2nL\|x^*-x^0\|_2^2}{k^2}$. □

The obtained rate leads to complexity $O(\sqrt{n/\varepsilon})$ to achieve accuracy ε in terms of the objective. As we show below, for the collaborative filtering problem and optimal transport problem $n = 2$ and our accelerated method provides acceleration from complexity $O(1/\varepsilon)$ of existing AM methods to the better complexity $O(1/\sqrt{\varepsilon})$.

3. Primal-Dual Extension

In this section we consider the primal-dual (up to a sign) pair of minimization problems

$$(P_1) \min_{x\in Q\subseteq E} \left\{ f(x) : \mathbf{A}x = b \right\},$$

$$(P_2) \min_{\lambda\in\Lambda} \left\{ \phi(\lambda) = \langle \lambda, b\rangle + \max_{x\in Q}\left(-f(x) - \langle \mathbf{A}^T\lambda, x\rangle\right) \right\},$$

where E is a finite-dimensional real vector space, Q is a simple closed convex set, f is a γ-strongly convex function, $\mathbf{A}$ is a given linear operator from E to some finite-dimensional real vector space H, $b \in H$ is given, $\Lambda = H^*$ is the conjugate space.

Algorithm 2 Primal-Dual AAM

1: $A_0 = a_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **for** $k \geqslant 0$ **do**
3: Set $\beta_k = \operatorname*{argmin}_{\beta\in[0,1]} \phi\left(\eta^k + \beta(\zeta^k - \eta^k)\right)$
4: Set $\lambda^k = \beta_k\zeta^k + (1-\beta_k)\eta^k$
5: Choose $i_k = \operatorname*{argmax}_{i\in\{1,\dots,n\}} \|\nabla_i\phi(\lambda^k)\|_2^2$
6: Set $\eta^{k+1} = \operatorname*{argmin}_{\eta\in S_{i_k}(\lambda^k)} \phi(\eta)$
7: Find a_{k+1}, $A_{k+1} = A_k + a_{k+1}$ from

$$\phi(\lambda^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})}\|\nabla\phi(\lambda^k)\|_2^2 = \phi(\eta^{k+1})$$

8: Set $\zeta^{k+1} = \zeta^k - a_{k+1}\nabla\phi(\lambda^k)$
9: Set $\hat{x}^{k+1} = \frac{a_{k+1}x(\lambda^k) + A_k\hat{x}^k}{A_{k+1}}$.
10: **end for**
Output: The points $\hat{x}^{k+1}$, η^{k+1}.

Since f is convex, $\phi(\lambda)$ is a convex function and, by Danskin's theorem, its subgradient is equal to

$$\nabla\phi(\lambda) = b - \mathbf{A}x(\lambda), \tag{2}$$

where $x(\lambda)$ is some solution of the convex problem

$$\max_{x\in Q}\left(-f(x) - \langle\mathbf{A}^T\lambda, x\rangle\right). \tag{3}$$

In what follows, we assume that H is equipped with the Euclidean norm, $\phi(\lambda)$ is L-smooth and that the problem (P_2) has a solution λ^* and there exist some $R > 0$ such that $\|\lambda^*\|_2 \leqslant R$. We underline that the quantity R will be used only in the convergence analysis, but not in the algorithm itself. Our primal-dual algorithm based on Algorithm 1 for the pair (P_1)-(P_2) is listed as Algorithm 2.

The key result for this method is that it guarantees convergence in terms of the constraints and the duality gap for the primal problem, provided that the primal objective is strongly convex. The rate of convergence and complexity remain the same as for Algorithm 1.

Theorem 3. *Let the objective $f(x)$ in the problem (P_1) be γ-strongly convex w.r.t. $\|\cdot\|_E$, and let $\|\lambda^*\| \leqslant R$. Then, for the sequences $\hat{x}^k, \eta^k$, $k \geqslant 0$, generated by Algorithm 2,*

$$|\phi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{8n\|\mathbf{A}\|_{E\to H}^2 R^2}{\gamma k^2}, \tag{4}$$

$$\|\mathbf{A}\hat{x}^k - b\|_2 \leqslant \frac{8n\|\mathbf{A}\|_{E\to H}^2 R}{\gamma k^2}, \tag{5}$$

$$\|\hat{x}^k - x^*\|_E \leqslant \frac{4n\|\mathbf{A}\|_{E\to H} R}{\gamma k} \tag{6}$$

where $\|\mathbf{A}\|_{E\to H}$ is the norm of $\mathbf{A}$ as a linear operator from E to H, i.e. $\|\mathbf{A}\|_{E\to H} =$

$\max_{u,v} \{\langle Au, v\rangle : \|u\|_E = 1, \|v\|_H = 1\}$, *and* $\|\cdot\|_H = \|\cdot\|_2$.

4. Application to Non-convex Optimization

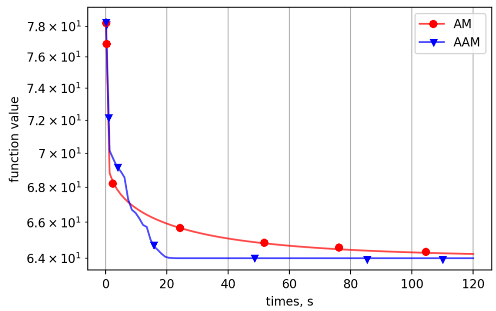

Figure 1. Performance of AM and Algorithm 1 on the problem (7)

In this section we apply our general accelerated AM method to a non-convex collaborative filtering problem. The problem consists of completion of the user-item preferences matrix with estimated values based on a small number of observed ratings made by other users. This is a particular case of the matrix completion problem. The unknown ratings $\hat{r}_{ui}$ associated with the user u and the item i are sought as a product $x_u^\top y_i$, where the vectors x_u and y_i are the optimized variables. We assume that we are given r_{ui} – observed preference rates associated with some users and items. The confidence c_{ui} for an observation r_{ui} is defined as $c_{ui} = 1 + 5r_{ui}$, and the binarized rating p_{ui} is defined as $p_{ui} = 1$ if $r_{ui} > 0$ and $p_{ui} = 0$ if $r_{ui} = 0$. Following the approach in (Hu et al., 2008), we minimize the data fitting term with a regularizer

$$\min_{x,y} \; F(x,y) = \sum_{\text{observed } u,i} c_{ui}\left(r_{ui} - x_u^\top y_i\right)^2 + \lambda\left(\sum_u \|x_u\|_2^2 + \sum_i \|y_i\|_2^2\right). \quad (7)$$

This function can be explicitly minimized over x for fixed y and vice-versa, which motivates the use of alternating minimization procedures.

The considered objective function is not convex, but has Lipchitz continuous gradient (by Theorem 1 from (Khenissi & Nasraoui, 2019)), so the minimization via Algorithm 1 is possible. We use the standard AM algorithm as a baseline. We generate the matrix $\{r_{ui}\}_{u,i}$ from Last.fm dataset 360K with ratings given by listeners to certain artists. There were 70 users and 100 artists observed, and the sparsity coefficient of the matrix was approximately 2%. The regularization coefficient was set to $\lambda = 0.1$ In Figure 1 we compare the performance of AM and Algorithm 1 applied to the problem (7).

5. Application to Optimal Transport and Wasserstein Barycenter

In this section we apply the developed methods to solve the discrete-discrete optimal transportation problem

$$\min_{X\in\mathcal{U}(r,c)} f(X) = \langle C, X\rangle \quad (8)$$
$$\mathcal{U}(r,c) = \{X \in \mathbb{R}_+^{N\times N} : X\mathbf{1} = r, X^T\mathbf{1} = c\},$$

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

Optimal transport distances lead to the concept of Wasserstein barycenter (WB). Given two probability measures p, q and a cost matrix $C \in \mathbb{R}_+^{N\times N}$ we define optimal transportation distance between them as

$$W_C(p,q) = \min_{X\in\mathcal{U}(p,q)} \langle X, C\rangle.$$

For a given set of probability measures p_i and cost matrices C_i we define their weighted barycenter with weights $w \in S_m(1)$ as a solution of the following convex optimization problem:

$$\min_{q\in S_N(1)} \sum_{i=1}^m w_i W_{C_i}(p_i, q).$$

The key aspect to apply our method is the strong convexity of the function to minimize. To ensure this, we introduce a *general* strongly convex regularizer $\mathcal{R}(X)$, e.g. entropy (Cuturi, 2013) or squared Euclidean norm (Essid & Solomon, 2018). Since the $f(X)$ is strongly convex, we are in the situation of Section 3. We underline that our method is able to solve OT problems with *general* regularizers, but, next we focus on a special case of entropic regularization as the most used in practice. In this case $\mathcal{R}(X) = \langle X, \ln X\rangle$ with $\ln X$ taken elementwise. The detailed derivations and proofs for this subsection can be found in the supplementary.

Using the entropic regularization we define the regularized OT-distance for $\gamma > 0$:

$$W_{C,\gamma}(p,q) = \min_{\pi\in\mathcal{U}(p,q)} \langle \pi, C\rangle + \gamma\mathcal{R}(\pi),$$

and the regularized barycenter which is the solution to the following problem:

$$\min_{q\in\in S_N(1)} \sum_{l=1}^m w_l \mathcal{W}_{C_l,\gamma}(p_l, q). \quad (9)$$

Importantly, the entropy $\langle X, \ln X\rangle$ is *not* strongly convex on $\mathbb{R}_+^{N\times N}$. Thus, if we just take $Q = \mathbb{R}_+^{N\times N}$ in Section

3, we will get a standard dual problem (Altschuler et al., 2017)[Sect. 3.3] in the form of minimization of a sum of exponents. This objective *does not* have Lipschitz-continuous gradient as the gradient grows exponentially. Previous works (Dvurechensky et al., 2018b; Lin et al., 2019a;b) do not take this into account and apply accelerated gradient methods to the dual problem, which makes their complexity results not completely correct.

To resolve this problem, 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 this new set Q in 1-norm. Thus, we introduce an additional constraint $\mathbf{1}^T X \mathbf{1} = 1$ into the problem. Since this constraint is a corollary of the constraint $X \in \mathcal{U}(r,c)$, the solution of the problem remains the same. The gain is that the gradient in the dual now becomes Lipschitz continuous and we can apply our primal-dual AAM.

Introducing the dual variables y, z, we derive in the supplementary the dual entropy OT problem

$$\min_{y,z\in\mathbb{R}^N} \gamma \ln\left(\sum_{i,j=1}^N \exp\left(\frac{-(y^i+z^j+C^{ij})}{\gamma}\right)\right) + \langle y, r\rangle + \langle z, c\rangle, \tag{10}$$

and the dual (minimization) problem of (9)

$$\min_{\substack{u,v\\ \sum_{l=1}^m w_l v_l = 0}} \gamma \sum_{l=1}^m w_l \ln \sum_{i,j=1}^N \exp \frac{-(u_l^i + v_l^j + C_l^{ij})}{\gamma} - \gamma \sum_{l=1}^m w_l \langle u_l, p_l\rangle \tag{11}$$

The variables in the dual problem (10), (11) naturally decompose into two blocks. Moreover, minimization over any one block can be made explicitly and the expressions are the same as for the Sinkhorn's algorithm in the form of (Altschuler et al., 2017) and IBP from (Kroshnin et al., 2019). The detailed proof of this fact may be found in the corresponding section of the supplement.

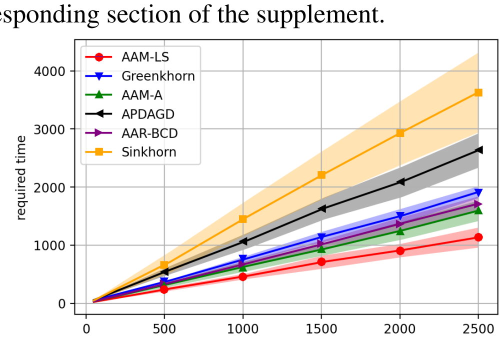

Figure 2. Performance comparison on MNIST dataset. Filled in area corresponds to 1 standard deviation.

Concerning **OT problem**, the goal is to approximate the non-regularized OT distance, the regularization parameter has to be chosen small, which leads to instabilities for the matrix-scaling Sinkhorn's algorithm of (Cuturi, 2013).

We obtained the final bound of the complexity to find an ε-approximation for the non-regularized OT problem to be $O\left(\frac{N^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right)$. Compared to the same bound for the Sinkhorn's algorithm, which is $O\left(\frac{N^2 \ln N \|C\|_\infty^2}{\varepsilon^2}\right)$, the new result for our accelerated algorithm is better in terms of ε. Detailed derivations can be found in the supplementary.

In Figure 2, we provide a numerical comparison of our methods with Sinkhorn's algorithm, the AAR-BCD method (Diakonikolas & Orecchia, 2018a), the APDA(G/M)D method (Dvurechensky et al., 2018b; Lin et al., 2019a) and with the Greenkhorn algorithm (Altschuler et al., 2017). We do not provide numerical comparison with Area Convexity algorithm from (Jambulapati et al., 2019) because the authors **did not** implement their algorithm. Instead of this the authors "implemented their algorithm as an instance of mirror prox". For this instance "there is not a known proof of convergence with an area-convex regularizer". So it's impossible to know the moment of time when the desired accuracy is reached. The AAM-LS method is the Accelerated Sinkhorn algorithm based on Algorithm 2, while the AAM-A is the Accelerated Sinkhorn algorithm based on the APDAGD method. Pseudocode of both these methods may be found in the supplementary. We performed experiments using randomly chosen images from MNIST dataset. We slightly modified the smaller values in the measures corresponding to the images as in (Dvurechensky et al., 2018b). We choose several values of accuracy $\varepsilon \in [0.0004, 0.002]$, sampled 5 pairs of images and ran the methods until the desired accuracy was reached, which is ensured using computable stopping criteria (Dvurechensky et al., 2018b). Our AAM algorithms outperform the other methods and also have much lower variance in performance compared to the Sinkhorn's algorithm. Probably the large variance in the results for Sinkhorn's algorithm is caused by its instability for small γ, which corresponds to small ε.

For **WB problem**, we add to the comparison recently presented algorithm from (Dvinskikh & Tiapkin, 2021). All presented algorithms have convergence guarantees on the value of non-regularized primal function, e.g. they guarantee that $\sum_{l=1}^m w_l W(p_l, \bar{q}^t) - \sum_{l=1}^m w_l W(p_l, q^*) \leqslant \varepsilon$ after t number of iterations (see Table 3), where $\bar{q}^t = \sum_{l=1}^m w_l q_l^t$ and $q_l^t = (X^t)^T \mathbf{1}$, X^t is an approximation of a tansportation plan at iteration t. But the particular implementation of Area Convexity algorithm from (Dvinskikh & Tiapkin, 2021) is supposed to work faster than theoretical analysis allows, because alternating minimization procedure for calculation of a prox-mapping has different stopping criterion, which is more easy to satisfy. To compare actual convergence, we took on 5 randomly chosen images from

MNIST dataset and plotted in Figure 5 and Figure 6 the rate of decay of primal function from a transportation plan, which is projected on the feasible set with Algorithm 2 from (Altschuler et al., 2017). We divided visualisation into two figures because of the scaling issues: Area-Convexity and Mirror-Prox were much slower than the others. IBP appears twice for a reference. Parameter of entropic regularization $\gamma = 5e-4$.

Figure 3 and Figure 4 illustrate the results obtained after 500s by the proposed algorithms.

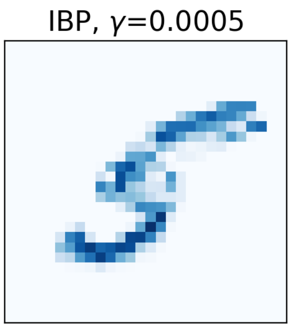

Figure 3. Approximate barycenter

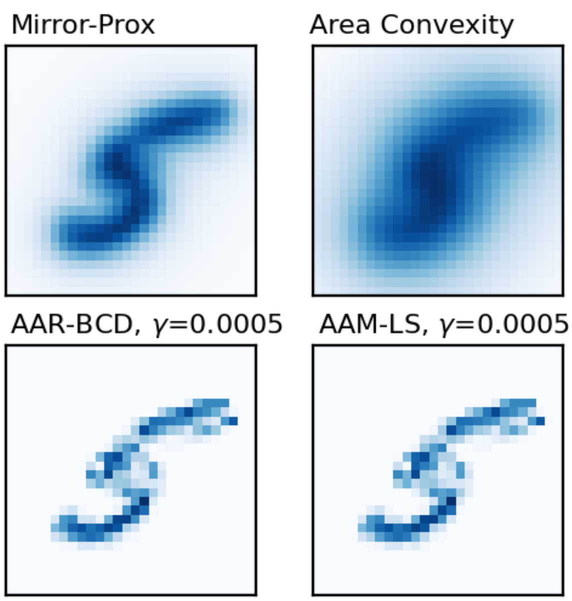

Figure 4. Approximate barycenter

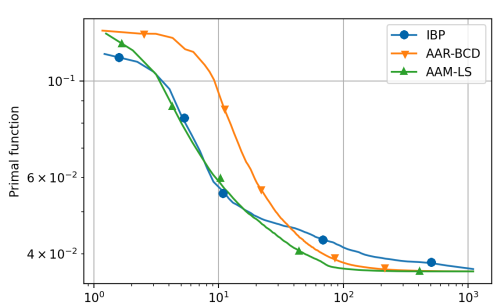

Figure 5. Performance comparison on MNIST dataset.

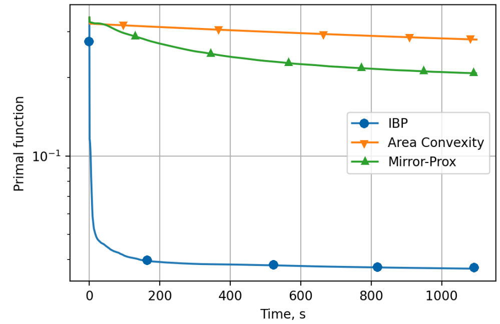

Figure 6. Performance comparison on MNIST dataset.

We also compare the performance of algorithms in terms of $\sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1$ which is used as stopping criterion for IBP algorithm, in Figure 7.

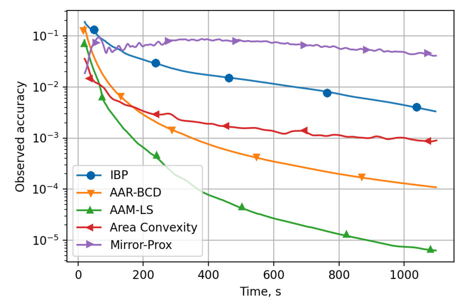

Figure 7. Performance comparison on MNIST dataset.

One may be interested in convergence to a true barycenter. To show the convergence we conducted experiments with random Gaussian measures. For this setup one has analytic expression for a Wasserstein barycenter.

In Figure 8 we compare the performance of algorithms in terms of $\|\bar{q}^t - q^*\|_1$, where q^* is a true barycenter. Parameter of entropic regularization $\gamma = 5e-5$.

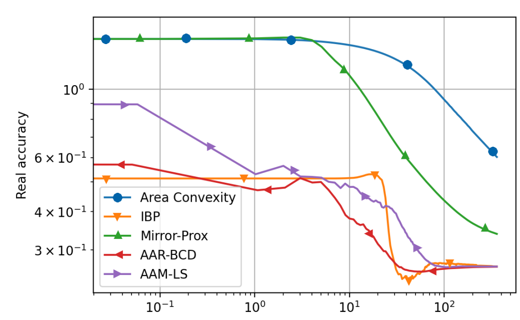

Figure 8. Performance comparison on Gaussian measures.

In Figure 9 we compare the performance of algorithms in terms of $\sum_{l=1}^{m} w_l \|q_l^t - \bar{q}^t\|_1$ in order to show a relation between Real accuracy and Observed accuracy.

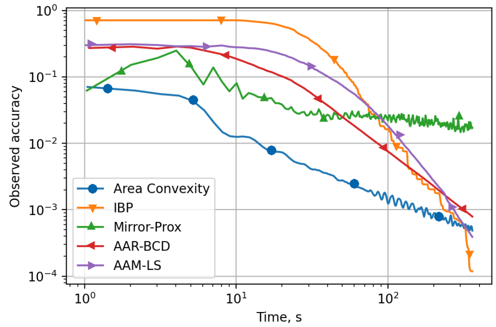

Figure 9. Performance comparison on Gaussian measures.

6. Application to Least Squares

We also illustrate the results by solving the alternating least squares problem on the Blog Feedback Data Set (Buza, 2014) obtained from UCI Machine Learning Repository. The data set contains 280 attributes and 52,396 data points. The attributes correspond to various metrics of crawled blog posts. The data is labeled, and the labels correspond to the number of comments that were posted within 24 hours from a fixed basetime. The goal of a regression method is to predict the number of comments that a blog post receives.

We partition the data into n blocks of the same size sequentially, e.g. we group the first N/n coordinates into the first block, the second N/n coordinates into the second block, and so on. We present comparison with block sizes N/n are 5 and 20, corresponding to $n = 56$ and $n = 14$.

The comparison for the linear regression is presented in Figure 10 and in Figure 11.

7. Conclusions

In this paper we propose an accelerated alternating minimization algorithm that combines greedy block-wise updates with full relaxation and Nesterov's moment. The method automatically adapts to the gradient Lipschitz constant and convexity of the problem. It achieves in the convex case $O(n/k^2)$ convergence rate for the objective and in the non-convex case $O(n/k)$ convergence rate for the squared norm of the gradient. We also propose a primal-dual extension of this algorithm for minimizing strongly convex functions under linear constraints. The practical efficiency of the algorithm is demonstrated by a series of numerical experiments.

Acknowledgements

We are grateful to the anonymous referees for their helpful comments and suggestions. We are also grateful to Jelena Diakonikolas for discussions related to this work. This research was funded by Russian Science Foundation (project 18-71-10108).

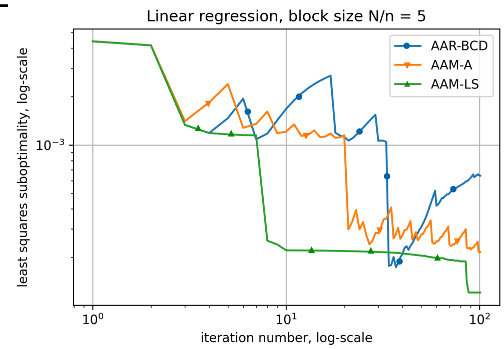

Figure 10. Performance comparison for the linear regression

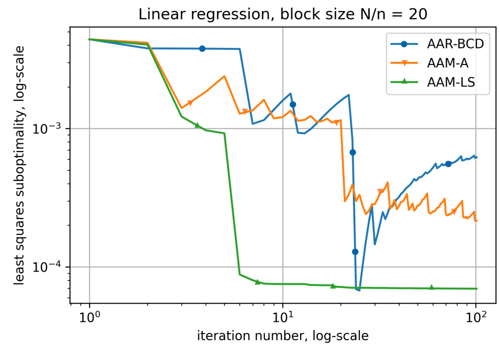

Figure 11. Performance comparison for the linear regression

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Alacaoglu, A., Tran Dinh, Q., Fercoq, O., and Cevher, V. Smooth primal-dual coordinate descent algorithms for nonsmooth convex optimization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5852–5861. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7167-smooth-primal-dual-coordinate-descent-algorithms-for-nonsmooth-convex-optimization.pdf.

Allen-Zhu, Z., Qu, Z., Richtarik, P., and Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1110–1119, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/allen-zhuc16.html. First appeared in arXiv:1512.09103.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approxfimation algorithms for optimal transport via sinkhorn iteration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1961–1971. Curran Associates, Inc., 2017. arXiv:1705.09634.

Andresen, A. and Spokoiny, V. Convergence of an alternating maximization procedure. *Journal of Machine Learning Research*, 17(63):1–53, 2016. URL http://jmlr.org/papers/v17/15-392.html.

Anikin, A. S., Gasnikov, A. V., Dvurechensky, P. E., Tyurin, A. I., and Chernov, A. V. Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. *Computational Mathematics and Mathematical Physics*, 57(8):1262–1276, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Beck, A. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

Beck, A. and Tetruashvili, L. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bertsekas, D. P. and Tsitsiklis, J. N. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 880–889. PMLR, 09–11 Apr 2018. URL http://proceedings.mlr.press/v84/blondel18a.html.

Buza, K. Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pp. 145–152. Springer, 2014.

Chernov, A., Dvurechensky, P., and Gasnikov, A. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Kochetov, Y., Khachay, M., Beresnev, V., Nurminski, E., and Pardalos, P. (eds.), *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, pp. 391–403. Springer International Publishing, 2016.

Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 999–1008. PMLR, 2018. URL http://proceedings.mlr.press/v80/claici18a.html.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2292–2300. Curran Associates, Inc., 2013.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/cuturi14.html.

Cuturi, M. and Peyré, G. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Daubechies, I., DeVore, R., Fornasier, M., and Güntürk, C. S. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010. doi: 10.1002/cpa.20303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20303.

Diakonikolas, J. and Orecchia, L. Alternating randomized block coordinate descent. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1224–1232, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR. URL http://proceedings.mlr.press/v80/diakonikolas18a.html.

Diakonikolas, J. and Orecchia, L. Alternating randomized block coordinate descent. *arXiv:1805.09185*, 2018b.

Dvinskikh, D. and Tiapkin, D. Improved complexity bounds in wasserstein barycenter problem. In Banerjee, A. and

Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1738–1746. PMLR, 2021. URL http://proceedings.mlr.press/v130/dvinskikh21a.html.

Dvinskikh, D., Gorbunov, E., Gasnikov, A., Dvurechensky, P., and Uribe, C. A. On primal and dual approaches for distributed stochastic convex optimization over networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 7435–7440, 2019. doi: 10.1109/CDC40024.2019.9029798. arXiv:1903.09844.

Dvurechensky, P., Gasnikov, A., Gasnikova, E., Matsievsky, S., Rodomanov, A., and Usik, I. Primal-dual method for searching equilibrium in hierarchical congestion population games. In *Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, September 19 - 23, 2016*, pp. 584–595, 2016. arXiv:1606.08988.

Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedić, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pp. 10783–10793. Curran Associates, Inc., 2018a. URL http://papers.nips.cc/paper/8274-decentralize-and-randomize-faster-algorithm-for-wasserstein-barycenters.pdf. arXiv:1802.04367.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1367–1376, 2018b. arXiv:1802.04367.

Dvurechensky, P., Gasnikov, A., Omelchenko, S., and Tiurin, A. A stable alternative to Sinkhorn's algorithm for regularized optimal transport. In Kononov, A., Khachay, M., Kalyagin, V. A., and Pardalos, P. (eds.), *Mathematical Optimization Theory and Operations Research*, pp. 406–423, Cham, 2020. Springer International Publishing. ISBN 978-3-030-49988-4.

Essid, M. and Solomon, J. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018. arXiv:1704.08200.

Fercoq, O. and Richtárik, P. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015. First appeared in arXiv:1312.5799.

Ge, D., Wang, H., Xiong, Z., and Ye, Y. Interior-point methods strike back: Solving the wasserstein barycenter problem. In *Advances in Neural Information Processing Systems 32*, pp. 6894–6905. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8913-interior-point-methods-strike-back-solving-the-wasserstein-barycenter-problem.pdf.

Guminov, S. V., Nesterov, Y. E., Dvurechensky, P. E., and Gasnikov, A. V. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Doklady Mathematics*, 99(2): 125–128, Mar 2019.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 263–272, USA, 2008. IEEE Computer Society. ISBN 9780769535029. doi: 10.1109/ICDM.2008.22. URL https://doi.org/10.1109/ICDM.2008.22.

Jambulapati, A., Sidford, A., and Tian, K. A direct tilde$o(1/\varepsilon)$ iteration parallel algorithm for optimal transport. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11359–11370. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9313-a-direct-tildeo1epsilon-iteration-parallel-algorithm-for-optimal-transport.pdf.

Khenissi, S. and Nasraoui, O. Modeling and counteracting exposure bias in recommender systems, Dec 2019. URL https://doi.org/10.18297/etd/3182.

Krawtschenko, R., Uribe, C. A., Gasnikov, A., and Dvurechensky, P. Distributed optimization with quantization for computing wasserstein barycenters. *arXiv:2010.14325*, 2020. doi: 10.20347/WIAS.PREPRINT.2782. WIAS preprint 2782.

Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P. E., Gasnikov, A., and Uribe, C. A. On the complexity of approximating wasserstein barycenters. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 3530–3540, 2019. URL http://proceedings.mlr.press/v97/kroshnin19a.html.

Lee, Y. T. and Sidford, A. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, pp. 147–156, Washington, DC, USA, 2013. IEEE Computer Society. ISBN 978-0-7695-5135-7. doi: 10.1109/FOCS.2013.24. URL http://dx.doi.org/10.1109/FOCS.2013.24. First appeared in arXiv:1305.1922.

Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3059–3067. Curran Associates, Inc., 2014. First appeared in arXiv:1407.1296.

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3982–3991, Long Beach, California, USA, 09–15 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v97/lin19a.html.

Lin, T., Ho, N., and Jordan, M. I. On the efficiency of the Sinkhorn and Greenkhorn algorithms and their acceleration for optimal transport. *arXiv:1906.01437*, 2019b.

Lin, T., Ho, N., Chen, X., Cuturi, M., and Jordan, M. I. Computational Hardness and Fast Algorithm for Fixed-Support Wasserstein Barycenter. *arXiv e-prints*, art. arXiv:2002.04783, February 2020.

Lu, H., Freund, R., and Mirrokni, V. Accelerating greedy coordinate descent methods. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3257–3266, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/lu18b.html.

McCullagh, P. and Nelder, J. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989. ISBN 9780412317606.

McLachlan, G. and Krishnan, T. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 1996.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL https://doi.org/10.1137/100802001. First appeared in 2010 as CORE discussion paper 2010/2.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013. First appeared in 2007 as CORE discussion paper 2007/76.

Nesterov, Y. and Stich, S. U. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182. URL https://doi.org/10.1137/16M1060182. First presented in May 2015 http://www.mathnet.ru:8080/PresentFiles/11909/7_nesterov.pdf.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, pp. 1–28, 2020. doi: 10.1080/10556788.2020.1731747. URL https://doi.org/10.1080/10556788.2020.1731747. arXiv:1809.05895.

Ortega, J. and Rheinboldt, W. *Iterative Solution of Nonlinear Equations in Several Variables*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1970. ISBN 9780898714616. URL https://books.google.de/books?id=UEFRfUZBpUEC.

Rogozin, A., Beznosikov, A., Dvinskikh, D., Kovalev, D., Dvurechensky, P., and Gasnikov, A. Decentralized distributed optimization for saddle point problems. *arXiv:2102.07758*, 2021.

Saha, A. and Tewari, A. On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 64–72, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/shalev-shwartz14.html. First appeared in arXiv:1309.2375.

Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. II. *Proc. Amer. Math. Soc.*, 45:195–198, 1974.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming wasserstein barycenters. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2647–2658. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6858-parallel-streaming-wasserstein-barycenters.pdf.

Sun, R. and 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 - Volume 1*, NIPS'15, pp. 1306–1314, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969385.

Tiapkin, D., Gasnikov, A., and Dvurechensky, P. Stochastic saddle-point optimization for wasserstein barycenters. *arXiv:2006.06763*, 2020.

Tupitsa, N., Dvurechensky, P., Gasnikov, A., and Uribe, C. A. Multimarginal optimal transport by accelerated alternating minimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 6132–6137, 2020. doi: 10.1109/CDC42340.2020.9304010. URL https://ieeexplore.ieee.org/document/9304010. arXiv:2004.02294.

Tupitsa, N., Dvurechensky, P., Gasnikov, A., and Guminov, S. Alternating minimization methods for strongly convex optimization. *Journal of Inverse and Ill-posed Problems*, 2021. doi: doi:10.1515/jiip-2020-0074. URL https://doi.org/10.1515/jiip-2020-0074. WIAS Preprint No. 2692, arXiv:1911.08987.

Uribe, C. A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Nedić, A. Distributed computation of Wasserstein barycenters over networks. In *2018 IEEE 57th Annual Conference on Decision and Control (CDC)*, 2018. Accepted, arXiv:1803.02933.

Yang, L., Li, J., Sun, D., and Toh, K.-C. A Fast Globally Linearly Convergent Algorithm for the Computation of Wasserstein Barycenters. *arXiv e-prints*, art. arXiv:1809.04249, September 2018.

Ye, J., Wu, P., Wang, J. Z., and Li, J. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *Trans. Sig. Proc.*, 65(9):2317–2332, May 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2659647. URL https://doi.org/10.1109/TSP.2017.2659647.