Supplementary Material for the paper Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm

In this document, we provide some details on the dual problem for the regularized OT problem, which is used in the analysis of the Sinkhorn's algorithm, details on the efficient implementation of our APDAGD algorithm for the case of the Sinkhorn's kernel being easy to apply, e.g. when the measures are supported on regular grids and C is given by squared Euclidean distance. Finally, we provide the missing proofs for the APDAGD-based approach. This is a separate document and, if not explicitly stated, all the references refer to formulas, algorithms, lemmas and theorems in this document.

1 Details for the Sinkhorn's Algorithm Approach

Below we provide the derivation of the dual problem for the regularized OT problem, which is used in Section 2.

$$\begin{aligned}\min_{X\in\mathcal{U}(r,c)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle &= \min_{X\in\mathbb{R}^{n\times n}_+} \max_{y,z\in\mathbb{R}^n} \langle C, X\rangle + \gamma\langle X, \ln X\rangle + \langle y, X\mathbf{1} - r\rangle + \langle z, X^T\mathbf{1} - c\rangle \\ &= \max_{y,z\in\mathbb{R}^n} -\langle y, r\rangle - \langle z, c\rangle + \min_{X^{ij}\geq 0} \sum_{i,j=1}^n X^{ij}(C^{ij} + \gamma \ln X^{ij} + y^i + z^j) \\ &\left[X^{ij} = \exp\left(-\frac{1}{\gamma}(y^i + z^j + C^{ij}) - 1\right)\right] \\ &\max_{y,z\in\mathbb{R}^n} -\langle y, r\rangle - \langle z, c\rangle - \gamma \sum_{i,j=1}^n \exp\left(-\frac{1}{\gamma}(y^i + z^j + C^{ij}) - 1\right).\end{aligned} \tag{1}$$

Changing the variables $u = -y/\gamma - 1/2$, $v = -z/\gamma - 1/2$, disregarding the constant term -1 and dividing the objective by $-\gamma$, we obtain the dual problem considered in Section 2.

2 Details for the Accelerated Gradient Descent Approach

2.1 Efficient Implementation for Entropic Regularization

In this subsection, we show that the steps of APDAGD can be written in terms of the multiplication of the Sinkhorn kernel matrix $K = \exp(-C/\gamma)$ by a vector. Then, similarly to the Sinkhorn's algorithm the step of APDAGD can be performed faster, if this kernel is easy to apply, e.g. the measures are supported on regular grids and C is given by squared Euclidean distance.

In the particular case of solving the entropy-regularized OT problem by APDAGD, we have $f(x) = \langle C, X\rangle + \gamma\langle X, \ln X\rangle$, $Q = \mathbb{R}^{n^2}_+$, $b^T = (r^T, c^T)$ and $A : \mathbb{R}^{n^2} \to \mathbb{R}^{2n}$ defined by the identity $(A\,\mathrm{vec}(X))^T = ((X\mathbf{1})^T, (X^T\mathbf{1})^T)$, $\lambda = (y, z)$, where y, z are the dual variables, defined in (1). Then, $x(\lambda) = \arg\min_{x\in Q}\left(-f(x) - \langle A^T\lambda, x\rangle\right)$ satisfies $x(\lambda) = \mathrm{vec}(X(y,z))$, where, due to (1), $X(y,z) = e^{-1}\cdot \mathrm{diag}\left(e^{y/\gamma}\right) K \mathrm{diag}\left(e^{z/\gamma}\right)$. At the same time,

$$\nabla\varphi(\lambda) = b - Ax(\lambda) = \begin{pmatrix} r - X(y,z)\mathbf{1} \\ c - X(y,z)^T\mathbf{1}\end{pmatrix}$$

Hence, in order to calculate the gradient $\nabla\varphi(\lambda)$ in step 6 of APDAGD algorithm, one needs to apply the kernel K to the vector $e^{y/\gamma}$ and to the vector $e^{z/\gamma}$, which is easy in the considered situation. Step 8 also involves the kernel K for calculating the value $\varphi(\lambda)$. This is also easy since $\varphi(\lambda)$ can be written as

$$\varphi(\lambda) = \gamma\mathbf{1}^T X(y,z)\mathbf{1} + \langle y, r\rangle + \langle z, c\rangle = \frac{\gamma}{e} e^{y/\gamma} K e^{z/\gamma} + \langle y, r\rangle + \langle z, c\rangle.$$

Again K is used only through matrix-vector multiplication. Other steps operate with dual variables λ, η, $\zeta \in \mathbb{R}^{2n}$ and, thus are also easy. Overall, APDAGD uses the same set of operations as the Sinkhorn's algorithm and, thus can also be implemented in parallel framework.

2.2 Adaptive Primal-Dual Accelerated Gradient Descent (APDAGD) Convergence Analysis

We provide the missing convergence rate proofs for the Adaptive Primal-Dual Accelerated Gradient Descent method for constrained convex optimization problem, which was considered in Section 3 of the main part of the paper.

We consider the problem

$$(P_1) \qquad \min_{x\in Q\subseteq E}\{f(x) : Ax = b\},$$

where $f(x)$ is a γ-strongly convex function on Q. The Lagrange dual problem to Problem (P_1) in a form of a minimization problem is

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

where Λ is the space of Lagrange multipliers and, hence is *unbounded*.

Our Adaptive Primal-Dual Accelerated Gradient Descent method can be considered as an Adaptive Accelerated Gradient Descent applied to the dual problem and supplied with a procedure to reconstruct the primal iterate. Since it can be of independent interest, we first, in subsection 2.3 consider Adaptive Accelerated Gradient Descent method for a general convex optimization problem and prove in Theorem 1 its convergence rate in a primal-dual friendly fashion. Then, in subsection 2.4 we use this result to analyze our Adaptive Primal-Dual Accelerated Gradient Descent method.

2.3 Adaptive Accelerated Gradient Descent for Convex Optimization

In this section, we consider a general optimization problem

$$\min_{\lambda \in \Lambda} \varphi(\lambda), \tag{2}$$

where $\Lambda \in H^*$ is a closed convex, generally speaking, *unbounded*, set, $\varphi(\lambda)$ is a convex function with L-Lipschitz-continuous gradient, i.e.

$$\varphi(\eta) \leq \varphi(\lambda) + \langle \nabla \varphi(\lambda), \eta - \lambda \rangle + \frac{L}{2}\|\eta - \lambda\|_{H,*}^2, \quad \forall \eta, \lambda \in H^*. \tag{3}$$

2.3.1 Proximal Setup

In this subsection, we introduce *proximal setup*, which is usually used in proximal gradient methods, see e.g. Ben-Tal and Nemirovski [2015]. We choose some norm $\|\cdot\|$ on the space of vectors λ and a *prox-function* $d(\lambda)$ which is continuous, convex on Λ and

1. admits a continuous in $\lambda \in \Lambda^0$ selection of subgradients $\nabla d(\lambda)$, where $\lambda \in \Lambda^0 \subseteq \Lambda$ is the set of all λ, where $\nabla d(\lambda)$ exists;

2. is 1-strongly convex on Λ with respect to $\|\cdot\|$, i.e., for any $\lambda \in \Lambda^0, \eta \in \Lambda$, $d(\eta) - d(\lambda) - \langle \nabla d(\lambda), \eta - \lambda \rangle \geq \frac{1}{2}\|\eta - \lambda\|^2$.

We define also the corresponding *Bregman divergence* $V[\zeta](\lambda) := d(\lambda) - d(\zeta) - \langle \nabla d(\zeta), \lambda - \zeta \rangle$, $\lambda \in \Lambda, \zeta \in \Lambda^0$. It is easy to see that

$$V[\zeta](\lambda) \geq \frac{1}{2}\|\lambda - \zeta\|^2, \quad \lambda \in \Lambda, \zeta \in \Lambda^0. \tag{4}$$

Standard proximal setups, i.e. Euclidean, entropy, ℓ_1/ℓ_2, simplex, nuclear norm, spectahedron can be found in Ben-Tal and Nemirovski [2015].

2.3.2 Algorithm and Complexity Analysis

In this subsection, we present Adaptive Accelerated Gradient Descent (see Algorithm 1 below) and prove its convergence rate theorem. Our algorithm in its form is very close to [Tseng, 2008, Alg.1] and [Lan et al., 2011, "Variant of Nesterov's algorithm"]. Nevertheless, the algorithms in those two papers assume the Lipschitz constant L to be known and explicitly use it in the algorithm. Our algorithm is free of this drawback. Another distinction of our algorithm is that we prove convergence rate in a primal-dual-friendly manner. As we show in subsection 2.4, this allows us to apply our AAGD to the Lagrange dual problem for (P_1), and reconstruct also primal iterates. In his paper, Tseng obtains primal-dual rates, but only for the case of bounded set Λ. In our case this analysis is inapplicable since the feasible set of the Lagrange dual problem is unbounded. Lan, Lu and Monteiro, consider a special problem of minimizing a linear function and do not prove primal-dual rates for their variant of Nesterov's algorithm.

We denote by $\eta_k, \zeta_k, \lambda_k$ three sequences of iterates of the algorithm and by α_k, β_k two sequences of numbers. The convergence rate is proved for the points η_k.

Lemma 1. *Algorithm 1 is defined correctly in the sense that the inner cycle of checking the inequality* (9) *is finite.*

Proof. Since, before each check of the inequality (9) on the step k, we multiply M_k by 2, after finite number of these multiplications, we will have $M_k \geq L$. Since φ has L-Lipschitz-continuous gradient, due to (3), we obtain that (9) holds after finite number of these repetitions. □

Lemma 2. *Let the sequences* $\{\lambda_k, \eta_k, \zeta_k, \alpha_k, \beta_k\}$, $k \geq 0$ *be generated by Algorithm 1. Then, for all* $\lambda \in \Lambda$, *it holds that*

$$\alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \lambda\rangle \leq \beta_{k+1}(\varphi(\lambda_{k+1}) - \varphi(\eta_{k+1})) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda). \tag{10}$$

Proof. Note that, from the optimality condition in (7), for any $\lambda \in \Lambda$, we have

$$\langle \nabla V[\zeta_k](\zeta_{k+1}) + \alpha_{k+1}\nabla\varphi(\lambda_{k+1}), \lambda - \zeta_{k+1}\rangle \geq 0. \tag{11}$$

By the definition of $V[\zeta](\lambda)$, we obtain, for any $\lambda \in \Lambda$,

$$\begin{aligned} V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) - V[\zeta_k](\zeta_{k+1}) =& d(\lambda) - d(\zeta_k) - \langle \nabla d(\zeta_k), \lambda - \zeta_k\rangle \\ & - (d(\lambda) - d(\zeta_{k+1}) - \langle \nabla d(\zeta_{k+1}), \lambda - \zeta_{k+1}\rangle) \\ & - (d(\zeta_{k+1}) - d(\zeta_k) - \langle \nabla d(\zeta_k), \zeta_{k+1} - \zeta_k\rangle) \\ =& \langle \nabla d(\zeta_k) - \nabla d(\zeta_{k+1}), \zeta_{k+1} - \lambda\rangle \\ =& \langle -\nabla V[\zeta_k](\zeta_{k+1}), \zeta_{k+1} - \lambda\rangle. \end{aligned} \tag{12}$$

Algorithm 1 Adaptive Accelerated Gradient Descent (AAGD)

Input: starting point $\lambda_0 \in \Lambda^0$, initial guess $0 < L_0 < 2L$, prox-setup: $d(\lambda)$ – 1-strongly convex w.r.t. $\|\cdot\|$, $V[\zeta](\lambda) := d(\lambda) - d(\zeta) - \langle \nabla d(\zeta), \lambda - \zeta \rangle$, $\lambda \in \Lambda, \zeta \in \Lambda^0$.

1: Set $k = 0$, $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0$.
2: **repeat**
3: Set $M_k = L_k/2$.
4: **repeat**
5: Set $M_k = 2M_k$, find α_{k+1} as the largest root of the equation

$$\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2. \tag{5}$$

6:

$$\lambda_{k+1} = \frac{\alpha_{k+1}\zeta_k + \beta_k \eta_k}{\beta_{k+1}}. \tag{6}$$

7:

$$\zeta_{k+1} = \arg\min_{\lambda \in \Lambda} \{V[\zeta_k](\lambda) + \alpha_{k+1}(\varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \lambda - \lambda_{k+1}\rangle)\}. \tag{7}$$

8:

$$\eta_{k+1} = \frac{\alpha_{k+1}\zeta_{k+1} + \beta_k \eta_k}{\beta_{k+1}}. \tag{8}$$

9: **until**

$$\varphi(\eta_{k+1}) \le \varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1}\rangle + \frac{M_k}{2}\|\eta_{k+1} - \lambda_{k+1}\|^2. \tag{9}$$

10: Set $L_{k+1} = M_k/2$, $k = k + 1$.
11: **until** Option 1: $k = k_{max}$.
Option 2: $R^2/\beta_k \le \varepsilon$.
Option 3:

$$\varphi(\eta_k) - \min_{\lambda \in \Lambda : V[\zeta_0](\lambda) \le R^2} \left\{ \sum_{i=0}^{k} \frac{\alpha_i}{\beta_k} \left(\varphi(\lambda_i) + \langle \nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) \right\} \le \varepsilon.$$

Here R is such that $V[\zeta_0](\lambda_*) \le R^2$ and ε is the desired accuracy.

Output: The point η_{k+1}.

Further, for any $\lambda \in \Lambda$,

$$\begin{aligned}
\alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \lambda\rangle &= \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \zeta_{k+1}\rangle + \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_{k+1} - \lambda\rangle \\
&\overset{(11)}{\le} \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \zeta_{k+1}\rangle + \langle -\nabla V[\zeta_k](\zeta_{k+1}), \zeta_{k+1} - \lambda\rangle \\
&\overset{(12)}{=} \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \zeta_{k+1}\rangle + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) - V[\zeta_k](\zeta_{k+1}) \\
&\overset{(4)}{\le} \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \zeta_{k+1}\rangle + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) - \frac{1}{2}\|\zeta_k - \zeta_{k+1}\|^2 \\
&\overset{(6),(8)}{=} \beta_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \lambda_{k+1} - \eta_{k+1}\rangle + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) - \frac{\beta_{k+1}^2}{2\alpha_{k+1}^2}\|\lambda_{k+1} - \eta_{k+1}\|^2 \\
&\overset{(5)}{=} \beta_{k+1}\left(\langle \nabla\varphi(\lambda_{k+1}), \lambda_{k+1} - \eta_{k+1}\rangle - \frac{M_k}{2}\|\lambda_{k+1} - \eta_{k+1}\|^2\right) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) \\
&\overset{(9)}{\le} \beta_{k+1}\left(\varphi(\lambda_{k+1}) - \varphi(\eta_{k+1})\right) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda).
\end{aligned}$$

□

Lemma 3. *Let the sequences $\{\lambda_k, \eta_k, \zeta_k, \alpha_k, \beta_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $\lambda \in \Lambda$, it holds that*

$$\beta_{k+1}\varphi(\eta_{k+1}) - \beta_k\varphi(\eta_k) \leq \alpha_{k+1}\left(\varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \lambda - \lambda_{k+1}\rangle\right) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda). \tag{13}$$

Proof. For any $\lambda \in \Lambda$,

$$\begin{aligned}
\alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \lambda_{k+1} - \lambda\rangle &= \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \lambda_{k+1} - \zeta_k\rangle + \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \lambda\rangle \\
&\overset{(5),(6)}{=} \beta_k\langle \nabla\varphi(\lambda_{k+1}), \eta_k - \lambda_{k+1}\rangle + \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \lambda\rangle \\
&\overset{\text{conv-ty}}{\leq} \beta_k\left(\varphi(\eta_k) - \varphi(\lambda_{k+1})\right) + \alpha_{k+1}\langle \nabla\varphi(\lambda_{k+1}), \zeta_k - \lambda\rangle \\
&\overset{(10)}{\leq} \beta_k\left(\varphi(\eta_k) - \varphi(\lambda_{k+1})\right) + \beta_{k+1}\left(\varphi(\lambda_{k+1}) - \varphi(\eta_{k+1})\right) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda) \\
&= \alpha_{k+1}\varphi(\lambda_{k+1}) + \beta_k\varphi(\eta_k) - \beta_{k+1}\varphi(\eta_{k+1}) + V[\zeta_k](\lambda) - V[\zeta_{k+1}](\lambda).
\end{aligned} \tag{14}$$

Rearranging terms, we obtain the statement of the Lemma. □

Theorem 1. *Let the sequences $\{\lambda_k, \eta_k, \zeta_k, \alpha_k, \beta_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $k \geq 0$, it holds that*

$$\beta_k\varphi(\eta_k) \leq \min_{\lambda\in\Lambda}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle \nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) + V[\zeta_0](\lambda)\right\}. \tag{15}$$

The number of inner cycle iterations after an iteration $k \geq 0$ does not exceed

$$4k + 4 + 2\log_2\left(\frac{L}{L_0}\right), \tag{16}$$

where L is the Lipschitz constant for the gradient of φ.

Proof. Let us change the counter in Lemma 2 from k to i and sum all the inequalities for $i = 0, ..., k-1$. Then, for any $\lambda \in \Lambda$,

$$\beta_k\varphi(\eta_k) - \beta_0\varphi(\eta_0) \leq \sum_{i=0}^{k-1}\alpha_{i+1}\left(\varphi(\lambda_{i+1}) + \langle \nabla\varphi(\lambda_{i+1}), \lambda - \lambda_{i+1}\rangle\right) + V[\zeta_0](\lambda) - V[\zeta_k](\lambda). \tag{17}$$

Whence, since $\beta_0 = \alpha_0 = 0$ and $V[\zeta_k](\lambda) \geq 0$,

$$\beta_k\varphi(\eta_k) \leq \sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle \nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) + V[\zeta_0](\lambda), \quad \lambda \in \Lambda. \tag{18}$$

Taking in the right hand side the minimum in $\lambda \in \Lambda$, we obtain the first statement of the Theorem.

The second statement of the Theorem is proved in the same way as in Nesterov and Polyak [2006], but we provide the proof for the reader's convenience. Let us again change the iteration counter in Algorithm 1 from k to i. Let $j_i \geq 1$ be the total number of checks of the inequality (9) on the step $i \geq 0$. Then, $j_0 = 1 + \log_2 \frac{M_0}{L_0}$ and, for $i \geq 1$, $M_i = 2^{j_i - 1} L_i = 2^{j_i - 1} \frac{M_{i-1}}{2}$. Thus, $j_i = 2 + \log_2 \frac{M_i}{M_{i-1}}$, $i \geq 1$. Further, by the same reasoning as in Lemma 2, we obtain that $M_i \leq 2L$, $i \geq 0$. Then, the total number of checks of the inequality (9) is

$$\sum_{i=0}^{k} j_i = 1 + \log_2 \frac{M_0}{L_0} + \sum_{i=1}^{k} \left(2 + \log_2 \frac{M_i}{M_{i-1}} \right) = 2k + 1 + \log_2 \frac{M_k}{L_0} \leq 2k + 2 + \log_2 \frac{L}{L_0}.$$

At the same time, each check of the inequality (9) requires two oracle calls. This proves the second statement of the Theorem. □

Corollary 1. *Let the sequences $\{\lambda_k, \eta_k, \zeta_k, \alpha_k, \beta_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $k \geq 0$, it holds that*

$$\varphi(\eta_k) - \min_{\lambda \in \Lambda} \varphi(\lambda) \leq \frac{V[\zeta_0](\lambda_*)}{\beta_k}, \tag{19}$$

where λ_* is the solution of $\min_{\lambda \in \Lambda} \varphi(\lambda)$ s.t. $V[\zeta_0](\lambda_*)$ is minimal among all the solutions.

Proof. Let λ_* be the solution of $\min_{\lambda \in \Lambda} \varphi(\lambda)$ s.t. $V[\zeta_0](\lambda_*)$ is minimal among all the solutions. Using convexity of φ, from Theorem 1, we obtain

$$\beta_k \varphi(\eta_k) \leq \sum_{i=0}^{k} \alpha_i \varphi(\lambda_*) + V[\zeta_0](\lambda_*).$$

Since $\beta_k = \sum_{i=0}^{k} \alpha_i$, we obtain the statement of the Corollary. □

The following Corollary justifies the stopping criteria in Algorithm 1.

Corollary 2. *Let λ_* be a solution of $\min_{\lambda \in \Lambda} \varphi(\lambda)$ such that $V[\zeta_0](\lambda_*)$ is minimal among all the solutions. Let R be such that $V[\zeta_0](\lambda_*) \leq R^2$ and ε be the desired accuracy. Let the sequences $\{\lambda_k, \eta_k, \zeta_k, \alpha_k, \beta_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, if one of the following inequalities holds*

$$R^2 / \beta_k \leq \varepsilon, \tag{20}$$

$$\varphi(\eta_k) - \min_{\lambda \in \Lambda : V[\zeta_0](\lambda) \leq R^2} \left\{ \sum_{i=0}^{k} \frac{\alpha_i}{\beta_k} \left(\varphi(\lambda_i) + \langle \nabla \varphi(\lambda_i), \lambda - \lambda_i \rangle \right) \right\} \leq \varepsilon, \tag{21}$$

then

$$\varphi(\eta_k) - \min_{\lambda \in \Lambda} \varphi(\lambda) \leq \varepsilon. \tag{22}$$

Proof. If the inequality (20) holds, the statement of the Corollary follows from inequality $V[\zeta_0](\lambda_*) \leq R^2$ Corollary 1.

Since $V[\zeta_0](\lambda_*) \leq R^2$, the point λ_* is a feasible point in the problem

$$\min_{\lambda \in \Lambda : V[\zeta_0](\lambda) \leq R^2} \left\{ \sum_{i=0}^{k} \frac{\alpha_i}{\beta_k} \left(\varphi(\lambda_i) + \langle \nabla \varphi(\lambda_i), \lambda - \lambda_i \rangle \right) \right\}.$$

Then, by convexity of φ, we obtain

$$\begin{aligned} \min_{\lambda \in \Lambda : V[\zeta_0](\lambda) \leq R^2} \left\{ \sum_{i=0}^{k} \frac{\alpha_i}{\beta_k} \left(\varphi(\lambda_i) + \langle \nabla \varphi(\lambda_i), \lambda - \lambda_i \rangle \right) \right\} &\leq \sum_{i=0}^{k} \frac{\alpha_i}{\beta_k} \left(\varphi(\lambda_i) + \langle \nabla \varphi(\lambda_i), \lambda_* - \lambda_i \rangle \right) \\ &\leq \varphi(\lambda_*). \end{aligned}$$

This and (21) finishes the proof. □

Let us now obtain the lower bound for the sequence β_k, $k \geq 0$, which will give the rate of convergence for Algorithm 1.

Lemma 4. *Let the sequence $\{\beta_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $k \geq 1$ it holds that*

$$\beta_k \geq \frac{(k+1)^2}{8L}, \tag{23}$$

where L is the Lipschitz constant for the gradient of φ.

Proof. As we mentioned in the proof of Theorem 1, $M_k \leq 2L$, $k \geq 0$. For $k = 1$, since $\alpha_0 = 0$ and $A_1 = \alpha_0 + \alpha_1 = \alpha_1$, we have from (5)

$$\beta_1 = \alpha_1 = \frac{1}{M_1} \geq \frac{1}{2L}.$$

Hence, (23) holds for $k = 1$.

Let us now assume that (23) holds for some $k \geq 1$ and prove that it holds for $k+1$. From (5) we have a quadratic equation for α_{k+1}

$$M_k \alpha_{k+1}^2 - \alpha_{k+1} - \beta_k = 0.$$

Since we need to take the largest root, we obtain,

$$\begin{aligned} \alpha_{k+1} = \frac{1 + \sqrt{1 + 4M_k\beta_k}}{2M_k} &= \frac{1}{2M_k} + \sqrt{\frac{1}{4M_k^2} + \frac{\beta_k}{M_k}} \geq \frac{1}{2M_k} + \sqrt{\frac{\beta_k}{M_k}} \\ &\geq \frac{1}{4L} + \frac{1}{\sqrt{2L}} \frac{k+1}{2\sqrt{2L}} = \frac{k+2}{4L}, \end{aligned}$$

where we used the induction assumption that (23) holds for k. Using the obtained inequality, from (5) and (23) for k, we get

$$\beta_{k+1} = \beta_k + \alpha_{k+1} \geq \frac{(k+1)^2}{8L} + \frac{k+2}{4L} \geq \frac{(k+2)^2}{8L}.$$

□

Corollary 3. *Let the sequences* $\{\lambda_k, \eta_k, \zeta_k\}$, $k \geq 0$ *be generated by Algorithm 1. Then, for all* $k \geq 1$, *it holds that*

$$\varphi(\eta_k) - \min_{\lambda \in \Lambda} \varphi(\lambda) \leq \frac{8LV[\zeta_0](\lambda_*)}{(k+1)^2}, \tag{24}$$

where λ_* *is the solution of* $\min_{\lambda \in \Lambda} \varphi(\lambda)$ *s.t.* $V[\zeta_0](\lambda_*)$ *is minimal among all the solutions.*

2.4 Adaptive Primal-Dual Accelerated Gradient Descent for Constrained Convex Optimization

In this section, we return to the constrained convex optimization problem, which was considered in Section 3 of the main part of the paper. For the reader's convenience, we repeat the problem statement and some details.

2.4.1 Preliminaries

We consider convex optimization problem of the following form

$$(P_1) \qquad \min_{x \in Q \subseteq E} \{f(x) : Ax = b\},$$

where $f(x)$ is a γ-strongly convex function on Q with respect to some chosen norm $\|\cdot\|_E$ on E and $A : E \to H$ is a linear operator, $b \in H$.

The Lagrange dual problem to Problem (P_1) is

$$(D_1) \qquad \max_{\lambda \in \Lambda} \left\{ -\langle \lambda, b \rangle + \min_{x \in Q} \left(f(x) + \langle A^T \lambda, x \rangle \right) \right\}.$$

Here we denote $\Lambda = H^*$ the space of Lagrange multipliers. It is convenient to rewrite Problem (D_1) in the equivalent form of a minimization problem

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

It is obvious that

$$Opt[D_1] = -Opt[P_2], \tag{25}$$

where $Opt[D_1]$, $Opt[P_2]$ are the optimal function value in Problem (D_1) and Problem (P_2) respectively. The following inequality follows from the weak duality

$$Opt[P_1] \geq Opt[D_1]. \tag{26}$$

We denote

$$\varphi(\lambda) = \langle \lambda, b \rangle + \max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right). \tag{27}$$

Since f is strongly convex, $\varphi(\lambda)$ is a smooth function and its gradient is equal to (see e.g. Nesterov [2005])

$$\nabla \varphi(\lambda) = b - Ax(\lambda), \tag{28}$$

where $x(\lambda)$ is the unique solution of the strongly-convex problem

$$\max_{x \in Q} \left(-f(x) - \langle A^T \lambda, x \rangle \right). \tag{29}$$

Note that $\nabla \varphi(\lambda)$ is Lipschitz-continuous (see e.g. Nesterov [2005]) with constant

$$L \leq \frac{\|A\|^2_{E \to H}}{\gamma}.$$

We also assume that the dual problem (D_1) has a solution λ^* and there exists some $R > 0$ such that

$$\|\lambda^*\|_2 \leq R < +\infty. \tag{30}$$

2.4.2 Adaptive Primal-Dual Accelerated Gradient Descent

Now we are ready to apply Algorithm 1 to the problem (P_2) and incorporate in the algorithm a procedure, which allows to reconstruct also an approximate solution of the problem (P_1). We choose Euclidean proximal setup, which means that we introduce euclidean norm $\|\cdot\|_2$ in the space of vectors λ and choose the prox-function $d(\lambda) = \frac{1}{2}\|\lambda\|_2^2$. Then, we have $V[\zeta](\lambda) = \frac{1}{2}\|\lambda - \zeta\|_2^2$. We state here as Algorithm 2 a more detailed version of Algorithm 3 in the main part of the paper. The first difference is that here we do not introduce an auxiliary sequence $\tau_k = \alpha_{k+1}/\beta_{k+1}$. The second difference is that here we use an equivalent form

$$\zeta_{k+1} = \arg\min_{\lambda \in \Lambda} \left\{ \frac{1}{2}\|\lambda - \zeta_k\|_2^2 + \alpha_{k+1}(\varphi(\lambda_{k+1}) + \langle \nabla \varphi(\lambda_{k+1}), \lambda - \lambda_{k+1} \rangle) \right\}$$

of the step

$$\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_{k+1}).$$

The third difference consists in the observation that, since, by definition, $\beta_k = \sum_{i=0}^k \alpha_k$,

$$\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_{k+1}) + \beta_k \hat{x}_k}{\beta_{k+1}} = \frac{1}{\beta_{k+1}} \sum_{i=0}^{k+1} \alpha_i x(\lambda_i).$$

Finally, here we use a stronger stopping rule

$$|f(\hat{x}_{k+1}) + \varphi(\eta_{k+1})| \leq \varepsilon_f$$

than in the main part of the paper. The reason is that, to obtain complexity result for approximating OT distance, it is enough to satisfy

$$f(\hat{x}_{k+1}) + \varphi(\eta_{k+1}) \leq \varepsilon$$

with a special choice of ε.

Algorithm 2 Adaptive Primal-Dual Accelerated Gradient Descent (APDAGD)

Input: starting point $\lambda_0 = 0$, initial guess $L_0 > 0$, accuracy $\varepsilon_f, \varepsilon_{eq} > 0$.

1: Set $k = 0$, $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

2: **repeat**

3: Set $M_k = L_k/2$.

4: **repeat**

5: Set $M_k = 2M_k$, find α_{k+1} as the largest root of the equation

$$\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2. \tag{31}$$

6: Calculate

$$\lambda_{k+1} = \frac{\alpha_{k+1}\zeta_k + \beta_k \eta_k}{\beta_{k+1}}. \tag{32}$$

7: Calculate

$$\zeta_{k+1} = \arg\min_{\lambda \in \Lambda} \left\{ \frac{1}{2}\|\lambda - \zeta_k\|_2^2 + \alpha_{k+1}(\varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \lambda - \lambda_{k+1}\rangle) \right\}. \tag{33}$$

8: Calculate

$$\eta_{k+1} = \frac{\alpha_{k+1}\zeta_{k+1} + \beta_k \eta_k}{\beta_{k+1}}. \tag{34}$$

9: **until**

$$\varphi(\eta_{k+1}) \leq \varphi(\lambda_{k+1}) + \langle \nabla\varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1}\rangle + \frac{M_k}{2}\|\eta_{k+1} - \lambda_{k+1}\|_2^2. \tag{35}$$

10: Set

$$\hat{x}_{k+1} = \frac{1}{\beta_{k+1}} \sum_{i=0}^{k+1} \alpha_i x(\lambda_i) = \frac{\alpha_{k+1} x(\lambda_{k+1}) + \beta_k \hat{x}_k}{\beta_{k+1}}.$$

11: Set $L_{k+1} = M_k/2$, $k = k + 1$.

12: **until** $|f(\hat{x}_{k+1}) + \varphi(\eta_{k+1})| \leq \varepsilon_f$, $\|A_1 \hat{x}_{k+1} - b_1\|_2 \leq \varepsilon_{eq}$.

Output: The points $\hat{x}_{k+1}$, η_{k+1}.

Theorem 2. *Assume that the objective in the problem (P_1) is γ-strongly convex and that the dual solution λ^* satisfies $\|\lambda^*\|_2 \leq R$. Then, for $k \geq 1$, the points $\hat{x}_k$, η_k in Algorithm 2 satisfy*

$$-\frac{16LR^2}{k^2} \leq f(\hat{x}_k) - Opt[P_1] \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{16LR^2}{k^2}, \tag{36}$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{16LR}{k^2}, \tag{37}$$

$$\|\hat{x}_k - x^*\|_E \leq \frac{8}{k}\sqrt{\frac{LR^2}{\gamma}}, \tag{38}$$

where x^* and $Opt[P_1]$ are respectively an optimal solution and the optimal value in the problem (P_1), and $L \leq \frac{\|A\|_{E\to H}^2}{\gamma}$. Moreover, the stopping criterion in step 11 is correctly defined and

the number of inner cycle iterations after an iteration $k \geq 0$ does not exceed

$$4k + 4 + 2\log_2\left(\frac{L}{L_0}\right), \tag{39}$$

where $L \leq \frac{\|A\|^2_{E\to H}}{\gamma}$ is the Lipschitz constant for the gradient of φ.

Proof. From Theorem 1 with specific choice of the Bregman divergence, since $\zeta_0 = 0$, we have, for all $k \geq 0$,

$$\beta_k \varphi(\eta_k) \leq \min_{\lambda\in\Lambda}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) + \frac{1}{2}\|\lambda\|_2^2\right\} \tag{40}$$

Let us introduce a set $\Lambda_R = \{\lambda : \|\lambda\|_2 \leq 2R\}$ where R is given in (30). Then, from (40), we obtain

$$\begin{aligned}
\beta_k\varphi(\eta_k) &\leq \min_{\lambda\in\Lambda}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) + \frac{1}{2}\|\lambda\|_2^2\right\} \\
&\leq \min_{\lambda\in\Lambda_R}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) + \frac{1}{2}\|\lambda\|_2^2\right\} \\
&\leq \min_{\lambda\in\Lambda_R}\left\{\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right)\right\} + 2R^2.
\end{aligned} \tag{41}$$

On the other hand, from the definition (27) of $\varphi(\lambda)$, we have

$$\begin{aligned}
\varphi(\lambda_i) &= \langle\lambda_i, b\rangle + \max_{x\in Q}\left(-f(x) - \langle A^T\lambda_i, x\rangle\right) \\
&= \langle\lambda_i, b\rangle - f(x(\lambda_i)) - \langle A^T\lambda_i, x(\lambda_i)\rangle.
\end{aligned}$$

Combining this equality with (28), we obtain

$$\begin{aligned}
\varphi(\lambda_i) - \langle\nabla\varphi(\lambda_i), \lambda_i\rangle &= \varphi(\lambda) - \langle\nabla\varphi(\lambda), \lambda_i\rangle \\
&= \langle\lambda_i, b\rangle - f(x(\lambda_i)) - \langle A^T\lambda_i, x(\lambda_i)\rangle \\
&\quad - \langle b - Ax(\lambda_i), \lambda_i\rangle = -f(x(\lambda_i)).
\end{aligned}$$

Summing these inequalities from $i = 0$ to $i = k$ with the weights $\{\alpha_i\}_{i=1,\dots k}$, we get, using the convexity of f,

$$\begin{aligned}
&\sum_{i=0}^{k}\alpha_i\left(\varphi(\lambda_i) + \langle\nabla\varphi(\lambda_i), \lambda - \lambda_i\rangle\right) \\
&= -\sum_{i=0}^{k}\alpha_i f(x(\lambda_i)) + \sum_{i=0}^{k}\alpha_i\langle b - Ax(\lambda_i), \lambda\rangle \\
&\leq -\beta_k f(\hat{x}_k) + \beta_k\langle b - A\hat{x}_k, \lambda\rangle.
\end{aligned}$$

Substituting this inequality to (41), we obtain

$$\beta_k\varphi(\eta_k) \leq -\beta_k f(\hat{x}_k) + \beta_k \min_{\lambda\in\Lambda_R}\{\langle b - A\hat{x}_k, \lambda\rangle\} + 2R^2.$$

Finally, since

$$\max_{\lambda\in\Lambda_R}\{\langle -b + A_1\hat{x}_k, \lambda\rangle\} = 2R\|A\hat{x}_k - b\|_2,$$

we obtain

$$\varphi(\eta_k) + f(\hat{x}_k) + 2R\|A\hat{x}_k - b\|_2 \leq \frac{2R^2}{\beta_k}. \tag{42}$$

Since λ^* is an optimal solution of Problem (D_1), we have, for any $x \in Q$

$$Opt[P_1] \leq f(x) + \langle\lambda^*, Ax - b\rangle.$$

Using the assumption (30), we get

$$f(\hat{x}_k) \geq Opt[P_1] - R\|A\hat{x}_k - b\|_2. \tag{43}$$

Hence,

$$\begin{aligned}\varphi(\eta_k) + f(\hat{x}_k) &= \varphi(\eta_k) - Opt[P_2] + Opt[P_2] + Opt[P_1] - Opt[P_1] + f(\hat{x}_k)\\ &\overset{(25)}{=} \varphi(\eta_k) - Opt[P_2] - Opt[D_1] + Opt[P_1] - Opt[P_1] + f(\hat{x}_k)\\ &\overset{(26)}{\geq} -Opt[P_1] + f(\hat{x}_k) \overset{(43)}{\geq} -R\|A\hat{x}_k - b\|_2.\end{aligned} \tag{44}$$

This and (42) give

$$R\|A\hat{x}_k - b\|_2 \leq \frac{2R^2}{\beta_k}. \tag{45}$$

Hence, we obtain

$$\varphi(\eta_k) + f(\hat{x}_k) \overset{(44),(45)}{\geq} -\frac{2R^2}{\beta_k}. \tag{46}$$

On the other hand, we have

$$\varphi(\eta_k) + f(\hat{x}_k) \overset{(42)}{\leq} \frac{2R^2}{\beta_k}. \tag{47}$$

Combining (45), (46), (47), we conclude

$$\begin{aligned}\|A\hat{x}_k - b\|_2 &\leq \frac{2R}{\beta_k},\\ |\varphi(\eta_k) + f(\hat{x}_k)| &\leq \frac{2R^2}{\beta_k}.\end{aligned} \tag{48}$$

At the same time,

$$\begin{aligned}\varphi(\eta_k) + Opt[P_1] &= \varphi(\eta_k) - Opt[P_2] + Opt[P_2] + Opt[P_1] \\ &\overset{(25)}{=} \varphi(\eta_k) - Opt[P_2] - Opt[D_1] + Opt[P_1] \overset{(26)}{\geq} 0.\end{aligned}$$

Hence,

$$f(\hat{x}_k) - Opt[P_1] \leq f(\hat{x}_k) + \varphi(\eta_k). \tag{49}$$

From (48), (49), by Lemma 4, stating that, for any $k \geq 0$, $\beta_k \geq \frac{(k+1)^2}{8L}$, we obtain inequalities (36) and (37) in the Theorem statements.

It remains to prove inequality (38). By the optimality condition for Problem (P_1), we have

$$\langle \nabla f(x^*) + A^T\lambda^*, \hat{x}_k - x^* \rangle \geq 0, \quad Ax^* = b,$$

where $\nabla f(x^*) \in \partial f(x^*)$. Then

$$\begin{aligned}\langle \nabla f(x^*), \hat{x}_k - x^* \rangle &\geq -\langle A^T\lambda^*, \hat{x}_k - x^* \rangle \\ &\geq -\langle \lambda^{*(1)}, A\hat{x}_k - b \rangle \\ &\geq -R\|A\hat{x}_k - b\|_2 \overset{(45)}{\geq} -\frac{2R^2}{\beta_k},\end{aligned} \tag{50}$$

where we used the same reasoning as while deriving (43). Using this inequality and γ strong convexity of f, we obtain

$$\frac{\gamma}{2}\|\hat{x}_k - x^*\|_E^2 \leq f(\hat{x}_k) - Opt[P_1] - \langle \nabla f(x^*), \hat{x}_k - x^* \rangle \overset{(48),(49)}{\leq} \frac{4R^2}{\beta_k}.$$

Since, by Lemma 4, for any $k \geq 0$, $\beta_k \geq \frac{(k+1)^2}{8L}$, we obtain inequality (38).

□

References

Aaron Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015. URL `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`.

Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126 (1):1–29, 2011.

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

Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. ISSN 1436-4646. doi: 10.1007/s10107-006-0706-8. URL http://dx.doi.org/10.1007/s10107-006-0706-8.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, MIT, 2008. URL http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.