**Weierstraß-Institut**
**für Angewandte Analysis und Stochastik**

**Leibniz-Institut im Forschungsverbund Berlin e. V.**

Preprint ISSN 2198-5855

# Gradient methods for problems with inexact model of the objective

Fedor Stonyakin[1,2], Darina Dvinskikh[3], Pavel Dvurechensky[3], Alexey Kroshnin[4,5], Olesya Kuznetsova[1], Artem Agafonov[1], Alexander Gasnikov[1,4], Alexander Tyurin[5], Cesar A. Uribe[6], Dmitry Pasechnyuk[7], Sergei Artamonov[5]

submitted: February 10, 2020

[1] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
E-Mail: gasnikov@yandex.ru
kuznetsova.oa@phystech.edu
agafonov.ad@phystech.edu

[2] V. Vernadsky Crimean Federal University
Simferopol, Crimea
E-Mail: fedyor@mail.ru

[3] Weierstrass Institute, Berlin, Germany
E-Mail: darina.dvinskikh@wias-berlin.de
pavel.dvurechensky@wias-berlin.de

[4] Institute for Information Transmission Problems
Moscow, Russia
E-Mail: kroshnin@phystech.edu

[5] Higher School of Economics
Moscow, Russia
E-Mail: alexandertiurin@gmail.com
sartamonov@hse.ru

[6] Massachusetts Institute of Technology
Cambridge, MA, USA
E-Mail: cauribe@mit.edu

[7] Presidential Physics and Mathematics Lyceum No.239
St. Petersburg, Russia
E-Mail: pasechnyuk2004@gmail.com

No. 2688
Berlin 2020


2010 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25, 65K15.

*Key words and phrases.* Gradient method, inexact oracle, strong convexity, relative smoothness, Bregman divergence.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

# Gradient methods for problems with inexact model of the objective

Fedor Stonyakin, Darina Dvinskikh, Pavel Dvurechensky, Alexey Kroshnin, Olesya Kuznetsova, Artem Agafonov, Alexander Gasnikov, Alexander Tyurin, Cesar A. Uribe, Dmitry Pasechnyuk, Sergei Artamonov

**Abstract**

We consider optimization methods for convex minimization problems under inexact information on the objective function. We introduce inexact model of the objective, which as a particular cases includes inexact oracle [19] and relative smoothness condition [43]. We analyze gradient method which uses this inexact model and obtain convergence rates for convex and strongly convex problems. To show potential applications of our general framework we consider three particular problems. The first one is clustering by electorial model introduced in [49]. The second one is approximating optimal transport distance, for which we propose a Proximal Sinkhorn algorithm. The third one is devoted to approximating optimal transport barycenter and we propose a Proximal Iterative Bregman Projections algorithm. We also illustrate the practical performance of our algorithms by numerical experiments.

## 1 Introduction

In this paper we consider optimization methods for convex problems under inexact information on the objective function. This information is given by an object, which we call *inexact model*. Inexact model generalizes the inexact oracle introduced in [19], where inexactness is assumed to be present in the objective value and its gradient. The authors show that, based on these two objects, it is possible to construct a linear function, which is a lower approximation and, up to a quadratic term, an upper approximation of the objective, and these two approximations are enough to obtain convergence rates for gradient method and accelerated gradient method. We go beyond and assume that the approximations of the objective are given through some function, which is not necessarily linear.

This allows us to construct general gradient-type method which is applicable in for different problem classes and allows to obtain convergence rates in these situations as a corollary of our general theorem. Besides convex problems we focus also on strongly convex objectives and illustrate the application of our general theory by two examples. The first example is data clustering by electoral model [49]. The second example relates to Wasserstein distance and barycenter, which are widely used in data analysis [15, 16].

Many optimization methods use some model of the objective function to define a step by minimization of this model. Usually the model is constructed using exact first-order [46, 21, 52], second-order [51], or higher-order information [11, 48] information on the objective. The influence of inexactness on the convergence of gradient-type methods have being studied at least since [55]. Accelerated first-order methods with inexact oracle are studied in [17, 44, 19, 24, 14]. Some recent works study also non-convex problems in this context [10, 22]. Randomized methods with inexact oracle are also studied

DOI 10.20347/WIAS.PREPRINT.2688 Berlin 2020

in the literature, e.g. coordinate descent in [62, 32], random gradient-free methods and random directional derivative methods in [27, 26]. A method with inexact oracle for variational inequalities can be found in [31].

The contributions of this paper can be summarized as follows.

- ☐ We introduce an inexact model of the objective function for convex optimization problems and strongly convex optimization problems.
- ☐ We introduce and theoretically analyze a gradient-type method for convex and strongly convex problems with an inexact model of the objective function. For the latter case we prove linear rate of convergence.
- ☐ We apply our method to, generally speaking, non-convex optimization problem which arises in clustering model introduced in [49]. To do this we construct an inexact model and apply our general algorithms and convergence theorems.
- ☐ We apply our general framework for Wasserstein distance and barycenter problems and show that it allows to construct a proximal á la [12] version of the Sinkhorn's algorithm [58] and Iterative Bregman Projection algorithm [7].

**Notation.** We define $\mathbf{1} = (1, ..., 1)^T \in \mathbb{R}^n$, $KL(z|t)$ to be the Kullback-Leibler divergence: $KL(z|t) = \sum\limits_{k=1}^{n} z_k \ln(z_k/t_k)$, $\forall z, t \in S_n(1)$, where $S_n(1)$ is the standard simplex in $\mathbb{R}^n$. We also denote by $\odot$ the entrywise product of two matrices.

## 2 Gradient Methods with Inexact Model of the Objective

Consider the convex optimization problem

$$f(x) \to \min\limits_{x \in Q}, \tag{1}$$

where function $f$ is convex and $Q \subseteq \mathbb{R}^n$ is a simple convex compact set. Moreover, assume that $\min_{x \in Q} f(x) = f(x_*)$ for some $x_* \in Q$.

To solve this problem, we introduce a norm $\|\cdot\|$ on $\mathbb{R}^n$ and a prox-function $d(x)$ which is continuous and convex. We underline that, unlike most of the literature, we do not require $d$ to be strongly convex. Without loss of generality, we assume that $\min\limits_{x \in \mathbb{R}^n} d(x) = 0$. Further, we define *Bregman divergence* $V[y](x) := d(x) - d(y) - \langle \nabla d(y), x - y \rangle$. Next we define the inexact model of the objective function, which generalizes the inexact oracle of [19] (see also [24, 10, 28, 35, 61, 63]).

**Definition 1.** Let function $\psi_\delta(x, y)$ be convex in $x \in Q$ and satisfy $\psi_\delta(x, x) = 0$ for all $x \in Q$.

i) We say that $\psi_\delta(x, y)$ is a $(\delta, L)$-model of the function $f$ at a given point $y$ with respect to $V[y](x)$ iff, for all $x \in Q$, the inequality

$$0 \leq f(x) - (f(y) + \psi_\delta(x, y)) \leq LV[y](x) + \delta \tag{2}$$

holds for some $L, \delta > 0$.

ii) We say that $\psi_\delta(x, y)$ is a $(\delta, L, \mu)$-model of the function $f$ at a given point $y$ with respect to $V[y](x)$ iff, for all $x \in Q$, the inequality

$$\mu V[y](x) \leq f(x) - (f(y) + \psi_\delta(x, y)) \leq LV[y](x) + \delta \tag{3}$$

Note that we allow $L$ to depend on $\delta$. We refer to the case i) as convex case and to the case ii) as strongly convex case.

*Remark* 1. In the particular case of function $f$ possessing $(\delta, L)$-oracle [19] at a given point $y$, one has

$$0 \leq f(x) - f(y) - \langle g_\delta(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2 + \delta$$

and $\psi_\delta(x, y) = \langle g_\delta(y), x - y\rangle$. In the same way, if function $f$ is equipped with $(\delta, L, \mu)$-oracle [20], i.e.,

$$\frac{\mu}{2}\|x - y\|^2 \leq f(x) - f(y) - \langle g_{\delta,L,\mu}(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2 + \delta \quad \forall x \in Q,$$

we have $\psi_\delta(x, y) = \langle g_{\delta,L,\mu}(y), x - y\rangle$.

The algorithms we develop are based on solving auxiliary simple problems on each iteration. We assume that these problems can be solved inexactly and, following [6] introduce a definition of inexact solution of a problem.

**Definition 2.** Consider a convex minimization problem

$$\phi(x) \to \min_{x \in Q \subseteq \mathbb{R}^n}. \tag{4}$$

If $\phi$ is smooth, we say that we solve it with $\widetilde{\delta}$-'precision' ($\widetilde{\delta} \geq 0$) if we find $\tilde{x}$ s.t. $\max_{x \in Q}\langle \nabla\phi(\tilde{x}), \tilde{x} - x\rangle = \widetilde{\delta}$. If $\phi$ is general convex, we say that we solve this problem with $\widetilde{\delta}$-'precision' if we find $\tilde{x}$ s.t. $\exists h \in \partial\phi(\widetilde{x}),\ \langle h, x_* - \widetilde{x}\rangle \geq -\widetilde{\delta}$. In both cases we denote this $\tilde{x}$ as $\operatorname{argmin}^{\widetilde{\delta}}_{x \in Q} \phi(x)$.

We notice that the case $\widetilde{\delta} = 0$ corresponds to the case when $\tilde{x}$ is an exact solution of convex optimization problem (4) [6, 46]. The connection of Definition 2 with standard definitions of inexact solution, e.g. in terms of the objective residual, can be found in Appendix G of the full version of the paper [60].

## 2.1 Convex Case

In this subsection we describe a gradient-type method for problems with $(\delta, L)$-model of the objective. This algorithm is a natural extension of gradient method, see [35, 61, 63].

**Algorithm 1** Gradient method with $(\delta, L)$-model of the objective.

1: **Input:** $x_0$ is the starting point, $L > 0$ and $\delta, \widetilde{\delta} > 0$.
2: **for** $k \geq 0$ **do**
3:
$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + LV[x_k](x), \quad x_{k+1} := \arg\min^{\widetilde{\delta}}_{x \in Q} \phi_{k+1}(x).$$
4: **end for**
**Ensure:** $\bar{x}_N = \frac{1}{N}\sum_{k=0}^{N-1} x_{k+1}$

**Theorem 3.** *Let $V[x_0](x_*) \leq R^2$, where $x_0$ is the starting point, and $x_*$ is the nearest minimum point to the point $x_0$ in the sense of Bregman divergence $V[y](x)$. Then, for the sequence, generated by Algorithm 1 the following inequality holds:*

$$f(\bar{x}_N) - f(x_*) \leq \frac{LR^2}{N} + \widetilde{\delta} + \delta,$$

In appendix A of the full version of the paper [60] we prove this theorem and provide an adaptive version of Algorithm 3, which does not require knowledge of the constant $L$.

## 2.2 Strongly Convex Case

In this subsection we consider problem (1) with $(\delta, L, \mu)$-model of the objective function satisfying (3). This more strong assumption allows us to obtain linear rate of convergence of the proposed algorithm. Our algorithm is listed as Algorithm 2 and it is a version of Algorithm 1, which is adaptive to possibly unknown constant $L$.

**Algorithm 2** Adaptive gradient method with an oracle using the $(\delta, L, \mu)$-model

1: **Input:** $x_0$ is the starting point, $\mu > 0$ $L_0 \geq 2\mu$ and $\delta$.
2: Set $S_0 := 0$
3: **for** $k \geq 0$ **do**
4: Find the smallest $i_k \geq 0$ such that

$$f(x_{k+1}) \leq f(x_k) + \psi_\delta(x_{k+1}, x_k) + L_{k+1}V[x_k](x_{k+1}) + \delta,$$

where $L_{k+1} = 2^{i_k - 1}L_k$ for $L_k \geq 2\mu$ and $L_{k+1} = 2^{i_k}L_k$ for $L_k < 2\mu$,
$\alpha_{k+1} := \frac{1}{L_{k+1}}$, $S_{k+1} := S_k + \alpha_{k+1}$.

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + L_{k+1}V[x_k](x), \quad x_{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \phi_{k+1}(x).$$

5: **end for**
**Ensure:** $\bar{x}_N = \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$

Let's introduce average parameter $\hat{L}$:

$$1 - \frac{\mu}{\hat{L}} = \sqrt[k+1]{\left(1 - \frac{\mu}{L^{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right)\ldots\left(1 - \frac{\mu}{L_1}\right)}.$$

Note that by $L_i \geq \mu$ $(i = 1, 2, \ldots)$

$$\min_{1 \leq i \leq k+1} L_i \leq \hat{L} \leq \max_{1 \leq i \leq k+1} L_i \leq 2L.$$

The following result holds.

**Theorem 4.** *Let $\psi_\delta(x, y)$ is a $(\delta, L, \mu)$-model for $f$ w.r.t. $V[y](x)$. Then, after $k$ iterations of Algorithm 2, we have*

$$V[x^{k+1}](x_*) \leq \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*),$$

$$f(x^{k+1}) - f(x_*) \leq \frac{4L^2(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + 2L\left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*).$$

The details of proof can be found in Appendix B of the full version of the paper [60]. Note that Algorithm 1 also has linear convergence rate for the strongly convex case. The details can be found in Appendix C of the full version of the paper [60]. The benefit of Algorithm 1 is that there is no need to know the strong convexity parameter $\mu$ for the algorithm to work. On the other hand, this parameter is needed for assessing the quality of the solution returned by the algorithm. The benefit of the adaptive version is that it does not require to know the value of the parameter $L$ and adapts to it. Moreover, the parameter $L$ can be different for the model at different points and the algorithm adapts also for the local value of this parameter.

# 3 Clustering by Electorial Model

In this section we consider clustering model introduced in [49]. In this model voters (data points) choose a party (cluster) in an iterative manner by alternative minimization of the following function.

$$f_{\mu_1,\mu_2}(x=(z,p)) = g(x) + \mu_1 \sum_{k=1}^{n} z_k \ln z_k + \frac{\mu_2}{2}\|p\|_2^2 \to \min_{z\in S_n(1), p\in\mathbb{R}^m_+}, \tag{5}$$

where $\mathbb{R}^m_+$ is a non-negative orthant and $S_n(1)$ is the standard $n$-dimensional simplex in $\mathbb{R}^n$.

The vector $z$ contains probabilities with which voters choose the considered party, and vector $p$ describes the position of the party in the space of voter opinions. The minimized potential is the result of combining two optimization problems into one: voters choose the party whose position is closest to their personal opinion and the party adjusts its position minimizing dispersion and trying not to go too far from its initial position. Yu. Nesterov in [49] used sequential elections process to show that under some natural assumptions the process convergence and gives the clustering of the data-points. This was done for a particular choice of the function $g$ which has limited interpretability. We show, how our framework of inexact model of the objective allows to construct a gradient-type method for the case of general function $g$, which is not necessarily convex.

Assume that $g(x)$ (generally, non-convex) is a function with $L_g$-Lipschitz continuous gradient:

$$\|\nabla g(x) - \nabla g(y)\|_* \le L_g\|x-y\| \quad \forall x, y \in S_n(1)\times\mathbb{R}^m_+, \tag{6}$$

and, following [49], the numbers $\mu_1, \mu_2$ are chosen such that $L_g \le \mu_1$ and $L_g \le \mu_2$.

The norm $\|\cdot\|$ in $S_n(1)\times\mathbb{R}^m_+$ is defined as $\|(z,p)\|^2 = \|z\|_1^2 + \|p\|_2^2$, where $\|z\|_1 = \sum\limits_{k=1}^{n} z_k$ and $\|p\|_2 = \sqrt{\sum\limits_{k=1}^{m} p_k^2}$. The correctness of this definition is proved in Appendix I.

It can be shown that

$$\begin{aligned}\psi_\delta(x,y) = \langle\nabla g(y), x-y\rangle - L_g\cdot KL(z_x|z_y) - \frac{L_g}{2}\|p_x-p_y\|_2^2\\ +\mu_1(KL(z_x|\mathbf{1}) - KL(z_y|\mathbf{1})) + \frac{\mu_2}{2}\left(\|p_x\|_2^2 - \|p_y\|_2^2\right)\end{aligned} \tag{7}$$

is a $(0, 2L_g)$-model of $f_{\mu_1,\mu_2}(x)$ in $x$ with respect to the following Bregman divergence

$$V[y](x) = KL(z_x|z_y) + \frac{1}{2}\|p_x-p_y\|_2^2.$$

The proof is detailed in Appendix I of the full version of the paper [60].

Further, for the case $\min\{\mu_1, \mu_2\} > L_g$ $\psi_\delta(x, y)$ is a strongly convex w.r.t. $V[y](x)$:

$$\psi_\delta(x,y) = \psi_\delta^{lin}(x,y) + (\mu_1 - L_g) \cdot KL(z_x|z_y) + \frac{\mu_2 - L_g}{2}\|p_x - p_y\|_2^2 \geq \tag{8}$$

$$\geq (\min\{\mu_1, \mu_2\} - L_g) \cdot V[y](x),$$

where

$$\psi_\delta^{lin}(x,y) = \langle \nabla g(y), x - y\rangle + \mu_1\langle \nabla KL(z_y|1), z_x - z_y\rangle + \mu_2\langle p_y, p_x - p_y\rangle$$

is linear in $y$. The proof of (8) is given in Appendix I of the full version of the paper [60].

Thus, $\psi_\delta^{lin}(x, y)$ is a $(0, \max\{\mu_1, \mu_2\} + L_g, \min\{\mu_1, \mu_2\} - L_g)$-model of the function $f_{\mu_1,\mu_2}$:

$$f_{\mu_1,\mu_2}(y) + \psi_\delta^{lin}(x,y) + (\min\{\mu_1, \mu_2\} - L_g)V[y](x) \leq f_{\mu_1,\mu_2}(x)$$

and

$$f_{\mu_1,\mu_2}(x) \leq f_{\mu_1,\mu_2}(y) + \psi_\delta^{lin}(x,y) + (\max\{\mu_1, \mu_2\} + L_g)V[y](x).$$

So, we can apply our Algorithms 1 and 2 to the problem (5).

# 4 Proximal Sinkhorn Algorithm for Optimal Transport

In this section we consider the problem of approximating an optimal transport (OT) distance. Recently optimal transport distances has gained a lot of interest in machine learning and statistical applications [4, 8, 18, 33, 40, 54, 59]. To state the OT problem, assume that we are given two discrete probability measures $p, q \in S_n(1)$ and ground cost matrix $C \in \mathbb{R}_+^{n\times n}$, then the optimal transport problem is

$$\langle C, \pi\rangle \to \min_{\pi\in\mathcal{U}(p,q)}, \;\; \mathcal{U}(p,q) = \{\pi \in \mathbb{R}_+^{n\times n} : \pi\mathbf{1} = p, \pi^T\mathbf{1} = q\} \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes Frobenius dot product of matrices, $\pi$ is a transportation plan. The above optimal transport problem is the Kantorovich [37] linear program (LP) formulation of the problem, which goes back to the Monge's problem [45]. The best known theoretical complexity for this linear program is [1] $\widetilde{O}(n^{2.5})$, see [42]. However, there is no known practical implementation of this algorithm. In practice, the simplex method gives complexity $O(n^3 \ln n)$ [53]. We follow the alternative approach based on entropic regularization of the OT problem [15]. We show how our general framework of inexact model of the objective allows to construct Proximal Sinkhorn algorithm with better computational stability in comparison with the standard Sinkhorn algorithm.

For any optimization problem (1), $\psi_\delta(x, y) = f(x) - f(y)$ satisfies Definition 1 with any $L \geq 0$. In this case, our Algorithm 1 becomes inexact *Bregman proximal gradient method*

$$x^{k+1} = \arg\min_{x\in Q}^{\tilde\delta}\left\{f(x) + LV[x^k](x)\right\}.$$

Our idea is to apply this proximal method for the OT problem and approximately find the next iterate $x^{k+1}$ by Sinkhorn's algorithm [58, 15, 2, 29]. The latter is made possible by the choice of $V$ as

[1] Here and below for all (large) $n$: $\widetilde{O}(g(n)) \leq \tilde{C} \cdot (\ln n)^r g(n)$ with some constants $\tilde{C} > 0$ and $r \geq 0$. Typically, $r = 1$, but not in this particular case. If $r = 0$, then $\widetilde{O}(\cdot) = O(\cdot)$.

**Algorithm 3** Sinkhorn's Algorithm

**Require:** Accuracy $\tilde{\varepsilon}$, matrix $K = e^{-C/\gamma}$, marginals $p, q \in S_n(1)$.

Set $t = 0$, $u^0 = \ln p$, $v^0 = \ln q$, $\varepsilon' = \frac{\tilde{\varepsilon}}{4}\left(\max_{i,j} C_{ij} - \min_{i,j} C_{ij} + 2\gamma \ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right)\right)^{-1}$.
**repeat**
  **if** $t \bmod 2 = 0$ **then**
    $u^{t+1} = u^t + \ln p - \ln(B(u^t, v^t)\mathbb{1})$, where $B(u, v) := \operatorname{diag}(e^u) K \operatorname{diag}(e^v)$
    $v^{t+1} = v^t$
  **else**
    $v^{t+1} = v^t + \ln q - \ln(B(u^t, v^t)^T \mathbb{1})$
    $u^{t+1} = u^t$
  **end if**
  $t = t + 1$
**until** $\|B(u^t, v^t)\mathbb{1} - p\|_1 + \left\|B(u^t, v^t)^T \mathbb{1} - q\right\|_1 \leq \varepsilon'$
Find $\hat{\pi}$ as the projection of $B(u^t, v^t)$ on $\mathcal{U}(p, q)$ by Algorithm 2 in [2].

**Ensure:** $\hat{\pi}$.

KL divergence, which makes the problem of finding the point $x^{k+1}$ to be an entropy-regularized OT problem, which, in turn, is efficiently solvable by the Sinkhorn algorithm.

Consider the iterates

$$\pi^0 = pq^T \in \mathcal{U}(p, q), \quad \pi^{k+1} = \arg\min_{\pi \in \mathcal{U}(p,q)}{}^{\varepsilon/2} \left\{\langle C, \pi\rangle + L \cdot KL(\pi|\pi^k)\right\}$$
$$= \arg\min_{\pi \in \mathcal{U}(p,q)}{}^{\varepsilon/2} KL\left(\pi \,\middle|\, \pi^k \odot \exp\left(-\frac{C}{L}\right)\right), \tag{10}$$

which we call outer iterations. On each outer iteration we use Sinkhorn's algorithm 3, which solves the minimization problem in (10) with accuracy $\tilde{\varepsilon}$ in terms of its objective residual. Notice that here $\varepsilon'$ differs from the one from [2, 29] as we need approximated solution to the regularized problem. Moreover, unlike [29] we use a slightly refined theoretical bounds for the Sinkhorn's algorithm not depending on vectors $p$, $q$[2].

**Theorem 5.** *Let $\bar{\pi}^N = \frac{1}{N}\sum_{k=1}^N \pi^k$, where $\pi^k$ are the iterates of* (10)*. Then, after $N = \frac{4L \ln n}{\varepsilon}$ iterations, it holds that $\langle C, \bar{\pi}^N\rangle \leq \min_{\pi \in \mathcal{U}(p,q)} \langle C, \pi\rangle + \varepsilon$. Moreover, the accuracy $\tilde{\varepsilon}$ for the solution of* (10) *is sufficient to be set as $\widetilde{O}(\varepsilon^4/(Ln^4))$ and the complexity of Sinkhorn's Algorithm on $k$-th iteration is bounded as*

$$n^2 \widetilde{O}\left(\min\left\{\exp\left(\frac{\bar{c}_k}{L}\right)\left(\frac{\bar{c}_k}{L} + \ln\frac{\bar{c}_k}{\tilde{\varepsilon}}\right), \frac{\bar{c}_k^2}{L\tilde{\varepsilon}}\right\}\right), \tag{11}$$

*where*[3]

$$\bar{c}_k = \|C\|_\infty + L \ln\left(\frac{\max_{i,j} \pi_{ij}^k}{\min_{i,j} \pi_{ij}^k}\right). \tag{12}$$

[2]One can find the proof in Appendix E of the full version of the paper [60]

[3]This bound is rough and typically $\bar{c}_k$ is smaller in practice. By proper rounding of $\pi^k$ one can guarantee (without loss of generality) that $\pi_{ij}^k \geq \varepsilon/(2n^2\|C\|_\infty)$, which gives

$$\frac{\bar{c}_k}{L} = \frac{\|C\|_\infty}{L} + \ln\left(\frac{2n^2\|C\|_\infty}{\varepsilon}\right).$$

But, in practice there often is no need to make 'rounding' after each outer iteration.

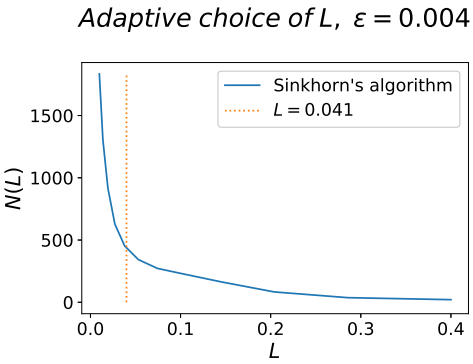


Figure 1: Adaptive choice of L

*Remark* 2. The standard Sinkhorn's method can be seen as a particular case of our algorithm (10) with only one step. To obtain an $\varepsilon$-approximate solution of (9), the regularization parameter $L$ needs to be chosen $O\left(\varepsilon/\ln n\right)$ [2, 29, 36]. This can lead to instability of the Sinkhorn's algorithm [57]. On the opposite, our Proximal Sinkhorn algorithm allows to run Sinkhorn's algorithm with larger regularization parameter. This parameter can be chosen by minimization of the theoretical bound (11), which gives $L = \widetilde{O}(\|C\|_\infty)$. In practice one can choose this constant adaptively since we have a $(\delta, L)$-model for any $L$ and can vary $L$ from iteration to iteration. First, the inner problem (10) is solved with overestimated $L$. Then, we set $L := L/2$ and the problem is solved with the updated value of the parameter and so on until a significant increase (e.g. 10 times) in the complexity of the auxiliary entropy-linear programming problem in comparison with the initial complexity is detected, see Figure 1, where $N(L)$ is a number of required iterations of Sinkhorn algorithm to solve the inner problem with accuracy $\varepsilon$.

From the Theorem 5 and Remark 2 one can roughly estimate the total complexity of Proximal Sinkhorn algorithm as[4] $\widetilde{O}(n^4/\varepsilon^2)$. We also mention several recent complexity bounds[5] for the OT problem $\tilde{O}(n^2/\varepsilon^3)$ [2], $\tilde{O}(n^2/\varepsilon^2)$ and $\tilde{O}(n^{2.5}/\varepsilon)$ [29], $\tilde{O}(n^2/\varepsilon)$ [9, 56], $\tilde{O}(n/\varepsilon^{3+d})$, $d \geq 1$ [1].

## 4.1 Numerical Illustration

In this subsection we provide numerical illustration of the Proximal Sinkhorn algorithm.[6] In the experiments we use a standard MNIST dataset with images scaled to a size $10 \times 10$. The vectors $p$ and $q$ contain the pixel intensities of the first and second images respectively. The value of $c_{ij}$ is equal to the Euclidean distance between the $i$-th pixel from the vector $p$ and the $j$-th pixel from the vector $q$ on the image pixel grid. For experiments with varying number of pixels $n$ the images are resized to be images of $10 \cdot m \times 10 \cdot m$ pixels, where $m \in \mathbb{N}$. We replace all the zero elements in $p$ and $q$ with $10^{-3}$ and, then, normalize these vectors.

[4] Our experiments on MNIST data set show (see Figures 2, 3) that in practice the bound is better.

[5] Strictly speaking for the moment we can not verify all the details of the proof of estimate $\tilde{O}(n^2/\varepsilon)$. Also the proposed in [9, 56] methods are mainly theoretical, like Lee–Sidford's method for OT problem with the complexity $\tilde{O}(n^{2.5})$ [42]. For the moment it is hardly possible to implement these methods such that theirs practical efficiencies correspond to the theoretical ones.

[6] The code is available at `https://github.com/dmivilensky/Proximal-Sinkhorn-algorithm`

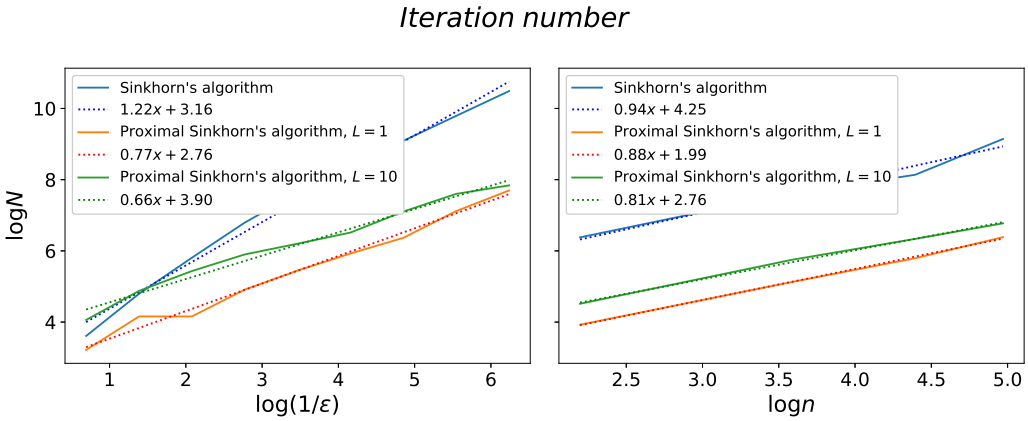


Figure 2: Comparison of iteration number of Sinkhorn's algorithm and total number of Sinkhorn steps in Proximal Sinkhorn's algorithm for different $L$.

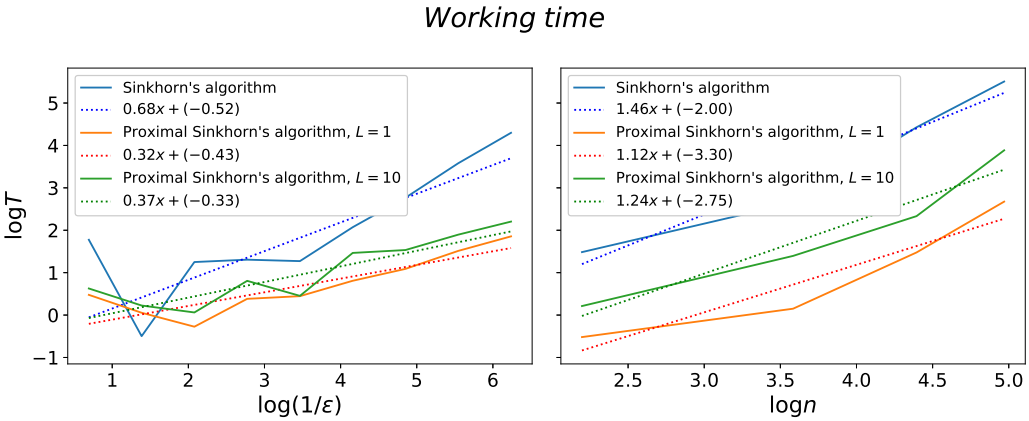


Figure 3: Comparison of working time of Sinkhorn's algorithm and Proximal Sinkhorn's algorithm with different $L$.

Fig. 2 shows that the growth rate of the iteration number with increasing accuracy or size of the problem for the Sinkhorn's algorithm is greater than for the Proximal Sinkhorn's method. At the same time, with a higher value of $L$ in proximal method, the iteration number is greater, and the growth rates with some precision are equal. The same type of dependence on the accuracy and the size of the problem can be seen for the working time (fig. 3).

More experiments can be found in the full version of this paper [60], in particular, on the mean number of inner iterations (Fig. 7 and 8).

# 5 Proximal IBP Algorithm for Wasserstein Barycenter

In this section we consider a more complicated problem of approximating an OT barycenter. OT barycenter is a natural definition of a mean in a space endowed with an OT distance. Such barycenters are used in the analysis of data with geometric structure, e.g. images, and other machine learning applications [16, 7, 38, 54, 39]. For a set of probability measures $\{p_1, \ldots, p_m\}$, cost matrices $C_1, \ldots, C_m \in \mathbb{R}_+^{n \times n}$, and $w \in S_n(1)$, the weighted barycenter of these measures is defined as a

solution of the following convex optimization problem

$$\sum_{l=1}^{m} w_l \min_{\pi_l \in \mathcal{U}(p_l,q)} \langle C_l, \pi_l \rangle \to \min_{q \in S_n(1)} \Longleftrightarrow \sum_{l=1}^{m} w_l \langle C_l, \pi_l \rangle \to \min_{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2},$$

$$\mathcal{C}_1 = \left\{\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m] : \forall l \ \pi_l \mathbb{1} = p_l\right\}, \quad \mathcal{C}_2 = \left\{\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m] : \pi_1^T \mathbb{1} = \cdots = \pi_m^T \mathbb{1}\right\}.$$

The idea is similar to the one in Sec. 4, namely, we use our framework to define a Proximal Iterative Bregman Projections algorithm. The algorithm starts from the point $\boldsymbol{\pi}$ s.t. $\pi_l^0 = \frac{1}{n} p_l \mathbb{1}^T \in \mathcal{U}(p_l, \mathbb{1}/n)$, $l = 1, ..., m$ and iterates

$$\begin{aligned} \boldsymbol{\pi}^{k+1} &= \arg \min_{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2}{}^{\varepsilon/2} \sum_{l=1}^{m} w_l \left\{ \langle C_l, \pi_l \rangle + L \cdot KL(\pi_l | \pi_l^k) \right\} \\ &= \arg \min_{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2}{}^{\varepsilon/2} \sum_{l=1}^{m} w_l KL\left(\pi_l \left| \pi_l^k \odot \exp\left(-\frac{C_l}{L}\right)\right.\right). \end{aligned} \tag{13}$$

These iterations are called outer iterations and on each such iteration, the Iterative Bregman Projections algorithm [7] listed as Algorithm 4 below is used to solve the auxiliary minimization problem.

**Algorithm 4** Iterative Bregman Projection

**Require:** $C_1, \ldots, C_m, p_1, \ldots, p_m, L > 0, \tilde{\varepsilon} > 0$

$u_l^0 := 0, v_l^0 := 0, K_l := \exp\left(-\frac{C_l}{L}\right), l = 1, \ldots, m$

**repeat**

  $v_l^{t+1} := \sum_{k=1}^m w_k \ln K_k^T e^{u_k^t} - \ln K_l^T e^{u_l^t}, \quad \mathtt{u}^{t+1} := \mathtt{u}^t$

  $t := t + 1$

  $u_l^{t+1} := \ln p_l - \ln K_l e^{v_l^t}, \quad \mathtt{v}^{t+1} := \mathtt{v}^t$

  $t := t + 1$

**until** $\sum_{l=1}^m w_l \left\| B_l^T(u_l^t, v_l^t)\mathbb{1} - \bar{q}^t \right\|_1 \leq \frac{\tilde{\varepsilon}}{4 \max_l \|C_l\|_\infty}$, where $B_l(u_l, v_l) = \operatorname{diag}(e^{u_l}) K_l \operatorname{diag}(e^{v_l})$, $\bar{q}^t := \sum_{l=1}^m w_l B_l^T(u_l^t, v_l^t)\mathbb{1}$

$q := \frac{1}{\sum_{l=1}^m w_l \langle \mathbb{1}, B_l \mathbb{1} \rangle} \sum_{l=1}^m w_l B_l^T \mathbb{1}$

Calculate $\hat{\pi}_1, \ldots, \hat{\pi}_m$ by Algorithm 2 from [2] s.t. $\hat{\pi}_l \in \mathcal{U}(p_l, q), \|\hat{\pi}_l - B_l\|_1 \leq \|B_l \mathbb{1} - p_l\|_1 + \left\|B_l^T \mathbb{1} - q\right\|_1$.

**Ensure:** $q, \hat{\boldsymbol{\pi}} = [\hat{\pi}_1, \ldots, \hat{\pi}_m]$.

**Theorem 6.** *Let $\bar{\boldsymbol{\pi}}^N = \frac{1}{N} \sum_{k=1}^N \boldsymbol{\pi}^k$, where $\boldsymbol{\pi}^k$ are the iterates of* (13)*. Then, after $N = \frac{4Lm \ln n}{\varepsilon}$ iterations, it holds that*

$$\sum_{l=1}^{m} w_l \langle C_l, \bar{\pi}_l^N \rangle \leq \min_{\boldsymbol{\pi} \in \mathcal{C}_1 \cap \mathcal{C}_2} \sum_{l=1}^{m} w_l \langle C_l, \pi_l \rangle + \varepsilon.$$

*Moreover, the accuracy $\tilde{\varepsilon}$ for the solution of* (13) *is sufficient to be set as $\tilde{\varepsilon} = \widetilde{O}(\varepsilon^2/(mn^3))$ and the complexity of IBP on $k$-th iteration is bounded as*

$$mn^2 \widetilde{O}\left(\min\left\{\exp\left(\frac{\bar{c}_k}{L}\right) \ln \frac{\bar{c}_k}{\tilde{\varepsilon}}, \frac{\bar{c}_k^2}{L\tilde{\varepsilon}}\right\}\right),$$

$$\bar{c}_k = O\left(\max_{l=1,\ldots,m}\left[\|C_l\|_\infty + L \ln\left(\frac{\max_{i,j}[\pi_l^k]_{ij}}{\min_{i,j}[\pi_l^k]_{ij}}\right)\right]\right).$$

The proof of Theorem 6 is based on Theorem 3 and [38]. All the remarks from Section 4 for Proximal Sinkhorn algorithm also hold for Proximal IBP. In [38] it was shown that complexity of IBP is $\widetilde{O}\left(n^2/\varepsilon^2\right)$. Despite the theoretical complexity of Proximal IBP is worse than this bound, we show in the next section that in practice Proximal IBP beats the standard IBP algorithm. As an alternative to the IBP algorithm we mention primal-dual accelerated gradient descent [23, 64].

## 5.1 Numerical Illustration

In this section, we present preliminary computational results for the numerical performance analysis of the Proximal Iterative Bregman Projection (ProxIBP) method discussed above asthe iterates (13).

Initially, we show the results for the computation of a non-regularized Wasserstein barycenter of a set of $10$ truncated Gaussian distributions with finite support. For the finite support

$$x = [-5, -4.9, -4.8, \ldots, -0.1, 0, 0.1, \ldots, 4.8, 4.9, 5]$$

, we set the finite distribution $p_l$ such that $p_l(i) = \mathcal{N}(x_i; \mu_i, \sigma_i)$, that is, the value at coordinate $i$ of the distribution $p_l$, for $1 \leq l \leq m$, is the value of the Normal distribution with mean $\mu_i$ and standard deviation $\sigma_i$. The values $\{\mu_i\} \sim$ Uniform$[-5, 5]$, are uniformly chosen in the line segment $[-5, 5]$, and the values are selected as $\{\mu_i\} \sim$ Uniform$[0.25, 1.25]$. For simplicity of exposition, we select uniform weighting for all distributions, i.e., $w_l = 1/m$.

Figure 4 shows the numerical results for a number of comparative scenarios between the Iterative Bregman Projection (IBP) algorithm proposed in [7] and its Proximal variant in (13). For both algorithms, we show the function values achieved by the generated iterates, and the final approximated barycenter. The results for the IBP algorithm are shown in Figure 4(a) and Figure 4(b). Figure 4(a) shows the weighted distance between the generated barycenter and the original distributions for three different desired accuracy values. It is clear that a bigger $\varepsilon$ generates a faster convergence, but the final cost is slightly higher than in other cases. Figure 4(b) shows the resulting barycenter for the three values of the accuracy parameter. For higher accuracy, the effects of the regularization constant are smaller and thus we obtain a "spikier" barycenter. Figure 4(c) and Figure 4(d) shows a similar analysis for the proposed Proximal IBP in (13), in Figure 4(c) we observe the function value of the generated barycenter, for a fixed number of inner loop iterations, and changing values of $L$, note that here $L$ is not a regularization parameter but the weight on the Bregman function. For larger values of $L$, the inner loop problem is easier to solve, requires less iterations to achieve certain accuracy, with the price in a larger number of iterations in the outer loop. For the particular problem studied, $200$ iterations in the outer loop are sufficient to achieve good performance even with relavively smaller values of $L$. Figure 4(c) shows the generated barycenters for the Proximal IBP algorithm. Finally, Figure 4(e) and Figure 4(f) show the results, for the analogous adaptive stopping condition described in Line $11$ of Algorithm 3 with $\varepsilon = 1 \cdot 10^{-10}$. We test two different values of the parameter $L$, namely $1$ and $0.1$. Additionally, we explore the suggested adaptive search procedure, where one decreases the value of the parameter $L$ at each iteration, until the inner problem has become particularly hard to solve. This last approach is shown a fast convergence as it reaches a comparable value in around $10$ iterations. Figure 4(f) shows the resulting barycenters.

Again, we refer to the full version [60] for additional experiments e.g. on computing Wasserstein barycenters of images from MNIST dataset (Fig. 5 and 6).

## Conclusions

In this paper we consider gradient methods with inexact information of the objective given by inexact model of this objective. We analyze a gradient-type method for this type of problems and provide its convergence rate. To illustrate the applications, we consider optimization problems in optimal transport and a clustering model. Notably, our framework allows to solve non-convex problems which have a convex inexact model, which is illustrated in the section devoted to clustering model.

## References

[1] J. Altschuler, F. Bach, A. Rudi, and J. Weed. Approximating the quadratic transportation metric in near-linear time. *arXiv:1810.10046*, 2018.

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

[3] A. S. Anikin, A. V. Gasnikov, P. E. Dvurechensky, A. I. Tyurin, and A. V. Chernov. 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.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

[5] A. Beck. 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.

[6] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski, 2015.

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

[8] J. Bigot, T. Klein, et al. Consistent estimation of a population barycenter in the wasserstein space. *arXiv:1212.2562*, 2012.

[9] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *arXiv:1810.07717*, 2018.

[10] L. Bogolubsky, P. Dvurechensky, A. Gasnikov, G. Gusev, Y. Nesterov, A. Raigorodskii, A. Tikhonov, and M. Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. *NIPS 2016*, 2016. http://papers.nips.cc/paper/6565-learning-supervised-pagerank-with-gradient-based-and-gradient-free-optimization-methods.pdf.

[11] C. Cartis, N. I. M. Gould, and P. L. Toint. Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *arXiv:1708.04044*, 2018.

[12] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

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

[14] M. B. Cohen, J. Diakonikolas, and L. Orecchia. On acceleration with noise-corrupted gradients. *arXiv:1805.12591*, 2018.

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

[16] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[17] A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. on Optimization*, 19(3):1171–1183, Oct. 2008.

[18] E. Del Barrio, H. Lescornel, and J.-M. Loubes. A statistical analysis of a deformation model with wasserstein barycenters : estimation procedure and goodness of fit test. *arXiv:1508.06465*, 2015.

[19] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

[20] O. Devolder, F. Glineur, Y. Nesterov, et al. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016, 2013.

[21] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *arXiv:1610.03446*, 2016.

[22] P. Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization. *arXiv:1703.09180*, 2017.

[23] P. Dvurechensky, D. Dvinskikh, A. Gasnikov, C. A. Uribe, and A. Nedić. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pages 10783–10793. Curran Associates, Inc., 2018. arXiv:1802.04367.

[24] P. Dvurechensky and A. Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

[25] P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik. 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*, pages 584–595, 2016.

[26] P. Dvurechensky, A. Gasnikov, and E. Gorbunov. An accelerated directional derivative method for smooth stochastic convex optimization. *arXiv:1804.02394*, 2018.

[27] P. Dvurechensky, A. Gasnikov, and E. Gorbunov. An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv:1802.09022*, 2018.

[28] P. Dvurechensky, A. Gasnikov, and D. Kamzolov. Universal intermediate gradient method for convex problems with inexact oracle. *arXiv:1712.06036*, 2017.

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

[30] P. Dvurechensky, A. Gasnikov, S. Omelchenko, and A. Tiurin. Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport. 2017.

[31] P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Generalized Mirror Prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters. 2018.

[32] P. Dvurechensky, A. Gasnikov, and A. Tiurin. Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method). 2017.

[33] J. Ebert, V. Spokoiny, and A. Suvorikova. Construction of non-asymptotic confidence sets in 2-Wasserstein space. 2017.

[34] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717 – 735, 1989.

[35] A. Gasnikov. Universal gradient descent. 2017.

[36] A. Gasnikov, P. Dvurechensky, D. Kamzolov, Y. Nesterov, V. Spokoiny, P. Stetsyuk, A. Suvorikova, and A. Chernov. Universal method with inexact oracle and its applications for searching equillibriums in multistage transport problems. 2015.

[37] L. Kantorovich. On the translocation of masses. *Doklady Acad. Sci. USSR (N.S.)*, 37(7-8):227–229, 1942.

[38] A. Kroshnin, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, N. Tupitsa, and C. Uribe. On the complexity of approximating Wasserstein barycenter. 2019.

[39] A. Kroshnin, V. Spokoiny, and A. Suvorikova. Statistical inference for bures-Wasserstein barycenters. 2019.

[40] T. Le Gouic and J.-M. Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[41] Y. LeCun. The MNIST database of handwritten digits. 1998.

[42] Y. T. Lee and A. Sidford. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433, 2014.

[43] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[44] J. Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

[45] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

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

[47] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[48] Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. CORE Discussion Papers 2018005, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Mar. 2018.

[49] Y. Nesterov. Soft clustering by convex electoral model. CORE Discussion Papers 2018001, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Jan. 2018.

[50] Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. 2018.

[51] Y. Nesterov and B. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[52] P. Ochs, J. Fadili, and T. Brox. Non-smooth Non-convex Bregman Minimization: Unification and new Algorithms. *Journal of Optimization Theory and Applications*, 181(1):244–278, 2019.

[53] O. Pele and M. Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009.

[54] G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[55] B. Polyak. *Introduction to Optimization*. New York, Optimization Software, 1987.

[56] K. Quanrud. Approximating Optimal Transport With Linear Programs. In *2nd Symposium on Simplicity in Algorithms (SOSA 2019)*, volume 69, pages 6:1–6:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[57] B. Schmitzer. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. 2016.

[58] R. Sinkhorn. Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. II. *Proc. Amer. Math. Soc.*, 45(2):195–198, 1974.

[59] J. Solomon, R. M. Rustamov, L. Guibas, and A. Butscher. Wasserstein Propagation for Semi-supervised Learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32, pages 306–314. PMLR, 2014.

[60] F. Stonyakin, D. Dvinskikh, P. Dvurechensky, A. Kroshnin, O. Kuznetsova, A. Agafonov, A. Gasnikov, A. Tyurin, C. A. Uribe, D. Pasechnyuk, and S. Artamonov. Gradient methods for problems with inexact model of the objective. *arXiv:1902.09001*, 2019.

[61] F. Stonyakin, A. Gasnikov, A. Tyurin, D. Pasechnyuk, A. Agafonov, P. Dvurechensky, D. Dvinskikh, and V. Piskunova. Inexact Model: A Framework for Optimization and Variational Inequalities. 2019.

[62] R. Tappenden, P. Richtárik, and J. Gondzio. Inexact Coordinate Descent: Complexity and Preconditioning. *Journal of Optimization Theory and Applications*, 170(1):144–176, Jul 2016.

[63] A. Tyurin and A. Gasnikov. Fast gradient descent method for convex optimization problems with an oracle that generates a $(\delta, L)$-model of a function in a requested point. *Computational Mathematics and Mathematical Physics. (Accepted)*, 2019.

[64] C. A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedić. Distributed Computation of Wasserstein Barycenters Over Networks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6544–6549, 2018.

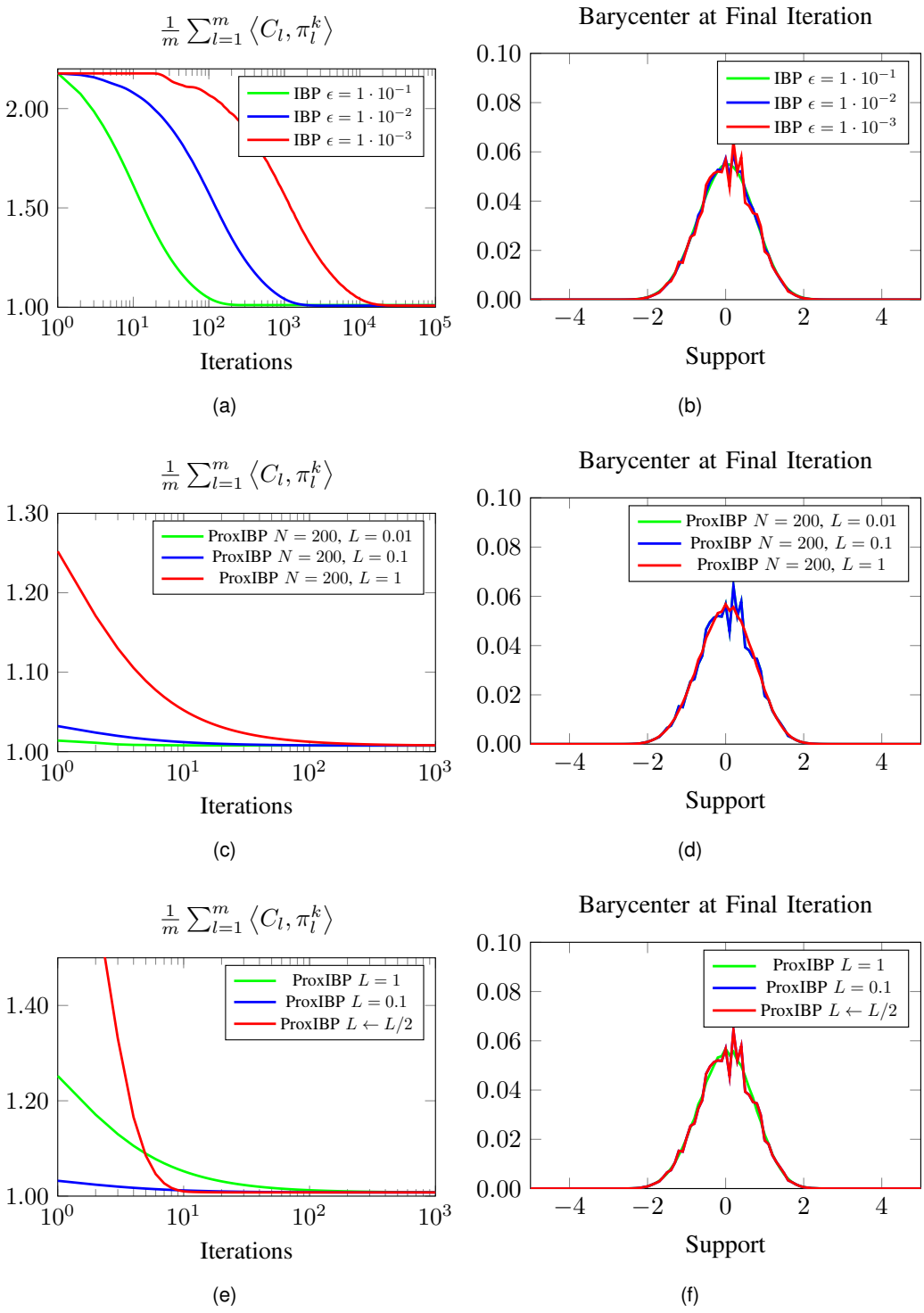


Figure 4: Numerical results for the computation of the barycenter of $10$ truncated Gaussian random variables with finite support for the IBP Algorithm and the Proximal IBP algorithm. Both function value and final resulting barycenter are shown for an number of simulation scenarios.

# A Adaptive gradient method and proof of Theorem 3

**Algorithm 5** Adaptive gradient method with inexact model of the objective

1: **Input:** $x_0$ is the starting point, $L_0 > 0$ and $\delta > 0$.
2: Set $S_0 := 0$
3: **for** $k \geq 0$ **do**
4: Find the smallest $i_k \geq 0$ such that

$$f(x_{k+1}) \leq f(x_k) + \psi_\delta(x_{k+1}, x_k) + L_{k+1}V[x_k](x_{k+1}) + \delta, \tag{14}$$

where $L_{k+1} = 2^{i_k - 1}L_k$, $S_{k+1} := S_k + \frac{1}{L_{k+1}}$.

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + L_{k+1}V[x_k](x), \quad x_{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \phi_{k+1}(x). \tag{15}$$

5: **end for**
**Ensure:** $\bar{x}_N = \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$

Before we prove the theorem we should note that that we can reduce Algorithm 5 to Algorithm 1. Indeed, let us always choose $L_{k+1} = L$ instead of $L_{k+1} = 2^{i_k - 1}L_k$ in Algorithm 5. In this case it is guaranteed that we exit the inner loop in Algorithm 5 after the first step due to $(\delta, L)$-model definition. Moreover, with this choice of $L_{k+1}$ Algorithm 5 generates the same sequences as in Algorithm 1. We prove Thus, Theorem 3 is a corollary of Theorem 7.

**Theorem 7.** *Let $V[x_0](x_*) \leq R^2$, where $x_0$ is the starting point, and $x_*$ is the nearest minimum point to the point $x_0$ in the sense of Bregman divergence $V[y](x)$. Then, for the sequence, generated by Algorithm 5 the following inequality holds:*

$$f(\bar{x}_N) - f(x_*) \leq \frac{R^2}{S_N} + \widetilde{\delta} + \delta \leq \frac{2LR^2}{N} + \widetilde{\delta} + \delta.$$

*Moreover, for Algorithm 5 the total number of attempts to solve* (15) *is bounded by* $2N + \log_2 \frac{L}{L_0}$.

First, we need to prove two lemmas in order to obtain the final result. Let us prove Lemma 1.

**Lemma 1.** *Let $\psi(x)$ be a convex function and*

$$y = \arg\min_{x \in Q}^{\widetilde{\delta}} \{\psi(x) + \beta V[z](x)\},$$

*where $\beta \geq 0$. Then*

$$\psi(x_*) + \beta V[z](x_*) \geq \psi(y) + \beta V[z](y) + \beta V[y](x_*) - \widetilde{\delta}.$$

*Proof.* Using Definition 2, we have:

$$\exists g \in \partial\psi(y), \ \ \langle g + \beta\nabla_y V[z](y), x_* - y\rangle \geq -\widetilde{\delta}.$$

Then inequality

$$\psi(x_*) - \psi(y) \geq \langle g, x_* - y\rangle \geq \langle \beta \nabla_y V[z](y), y - x_*\rangle - \widetilde{\delta}$$

and equality

$$\begin{aligned}\langle \nabla_y V[z](y), y - x_*\rangle = \langle \nabla d(y) - \nabla d(z), y - x_*\rangle = d(y) - d(z) - \langle \nabla d(z), y - z\rangle + \\ + d(x_*) - d(y) - \langle \nabla d(y), x_* - y\rangle - d(x_*) + d(z) + \langle \nabla d(z), x_* - z\rangle = \\ = V[z](y) + V[y](x_*) - V[z](x_*)\end{aligned}$$

complete the proof. □

**Lemma 2.** *We have the following inequality:*

$$\frac{f(x_{k+1})}{L_{k+1}} - \frac{f(x_*)}{L_{k+1}} \leq V[x_k](x_*) - V[x_{k+1}](x_*) + \frac{\widetilde{\delta}}{L_{k+1}} + \frac{\delta}{L_{k+1}}.$$

*Proof.* From the stopping criterion (14), we have:

$$f(x_{k+1}) \leq f(x_k) + \psi_\delta(x_{k+1}, x_k) + L_{k+1} V[x_k](x_{k+1}) + \delta.$$

Using Lemma 1 with $\psi(x) = \psi_\delta(x, x_k)$ and $\beta = L_{k+1}$, we obtain:

$$f(x_{k+1}) \leq f(x_k) + \psi_\delta(x_*, x_k) + L_{k+1} V[x_k](x_*) - L_{k+1} V[x_{k+1}](x_*) + \widetilde{\delta} + \delta.$$

In view of the model definition (2), we have:

$$f(x_{k+1}) \leq f(x_*) + L_{k+1} V[x_k](x_*) - L_{k+1} V[x_{k+1}](x_*) + \widetilde{\delta} + \delta.$$

□

*Remark* 3. Let us show that $L_k \leq 2L \quad \forall k \geq 0$. For $k = 0$ this is true from the fact that $L_0 \leq L$. For $k \geq 1$ this follows from the fact that we leave the inner cycle earlier than $L_k$ will be greater than $2L$. The exit from the cycle is guaranteed by the condition that there is an $(\delta, L)$-model for $f(x)$ at any point $x \in Q$.

Finally, we prove the theorem.

*Proof.* Let us sum up the inequality from Lemma 2 from $0$ to $N - 1$:

$$\sum_{k=0}^{N-1} \frac{f(x_{k+1})}{L_{k+1}} - S_N f(x_*) \leq V[x_0](x_*) - V[x_N](x_*) + S_N \widetilde{\delta} + S_N \delta.$$

Since $V[x_N](x_*) \geq 0$ and $V[x_0](x_*) \leq R^2$, we obtain inequality

$$\sum_{k=0}^{N-1} \frac{f(x_{k+1})}{L_{k+1}} - S_N f(x_*) \leq R^2 + S_N \widetilde{\delta} + S_N \delta.$$

Let us divide both parts by $S_N$.

$$\frac{1}{S_N} \sum_{k=0}^{N-1} \frac{f(x_{k+1})}{L_{k+1}} - f(x_*) \leq \frac{R^2}{S_N} + \widetilde{\delta} + \delta.$$

Using the convexity of $f(x)$ we can show that

$$f(\bar{x}_N) - f(x_*) \leq \frac{R^2}{S_N} + \widetilde{\delta} + \delta.$$

Remains only to prove that

$$\frac{1}{S_N} \leq \frac{2L}{N}.$$

As it follows from Definition 1 and Remark 3 for all $k \geq 0$ $L_k \leq 2L$. Thus, we have that

$$\frac{1}{L_k} \geq \frac{1}{2L}$$

and

$$S_N = \sum_{k=0}^{N} \frac{1}{L_k} \geq \frac{N}{2L}.$$

The total number of attempts to solve (15) is bounded in the same way as in [47]. □

In the same way as it is done in [3, 13, 25, 30, 50], one can show that the proposed method is primal-dual.

# B Analysis of Algorithm 2 in the case of $(\delta, L, \mu)$-model

Now we consider a proof of Theorem 4. To analyze Algorithm 2, assume that it works for $k$ iterations. By Lemma 1, for each $x \in Q$:

$$-\widetilde{\delta} \leq \psi_\delta(x, x^k) - \psi_\delta(x^{k+1}, x^k) + L^{k+1}V[x^k](x) - L^{k+1}V[x^{k+1}](x) - L^{k+1}V[x^k](x^{k+1}).$$

It means that

$$\begin{aligned} L_{k+1}V[x^{k+1}](x) \leq \widetilde{\delta} + \psi_\delta(x, x^k) - \psi_\delta(x^{k+1}, x^k) + \\ + L_{k+1}V[x^k](x) - L_{k+1}V[x^k](x^{k+1}). \end{aligned} \tag{16}$$

Further, $\psi_\delta(x, y)$ is a $(\delta, L)$-model w.r.t. $V[y](x)$ and from

$$f(x^{k+1}) \leq f(x^k) + \psi_\delta(x^{k+1}, x^k) + L_{k+1}V[x^k](x^{k+1}) + \delta,$$

we get

$$-L_{k+1}V[x^k](x^{k+1}) \leq \delta - f(x^{k+1}) + f(x^k) + \psi_\delta(x^{k+1}, x^k).$$

Now (16) means

$$L_{k+1}V[x^{k+1}](x) \leq \widetilde{\delta} + \delta - f(x^{k+1}) + f(x^k) + \psi_\delta(x, x^k) + L_{k+1}V[x^k](x). \tag{17}$$

Since $\psi_\delta(x, y)$ is a $(\delta, L, \mu)$-model for $f$, we have:

$$f(x^k) + \psi_\delta(x, x^k) \leq f(x) - \mu V[x^k](x).$$

Considering (17), we obtain:

$$L_{k+1}V[x^{k+1}](x) \leq \widetilde{\delta} + \delta + f(x) - f(x^{k+1}) + (L_{k+1} - \mu)V[x^k](x).$$

Set $x = x_*$. Since $L_0 \leq 2L$, we have $L_{k+1} \leq 2L$ for each $k \geq 0$. We also assume in Algorithm 2 that $L_{k+1} \geq \mu$. Thus, we have

$$\frac{1}{2L} \leq \frac{1}{L_{k+1}} \leq \frac{1}{\mu} \quad (\forall k = 0, 1, 2 \ldots).$$

Then we have $\forall i \in \mathbb{N} : i < k$

$$\left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) \ldots \left(1 - \frac{\mu}{L_{k-i}}\right) \leq \left(1 - \frac{\mu}{2L}\right)^{i+1}. \tag{18}$$

Therefore, we obtain:

$$V[x^{k+1}](x_*) \leq \frac{f(x_*) - f(x^{k+1}) + \delta + \widetilde{\delta}}{L^{k+1}} + \left(1 - \frac{\mu}{L_{k+1}}\right) V[x^k](x_*),$$

and

$$\begin{aligned}
&\frac{f(x^{k+1}) - f(x_*)}{L_{k+1}} + V[x^{k+1}](x_*) \leq \\
&\leq \frac{\delta + \widetilde{\delta}}{L_{k+1}} + \left(1 - \frac{\mu}{L_{k+1}}\right) V[x^k](x_*) \leq (\delta + \widetilde{\delta}) \left(\frac{1}{L_{k+1}} + \frac{1}{L_k}\left(1 - \frac{\mu}{L_{k+1}}\right)\right) + \\
&+ \left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) V[x^k](x_*) \leq \ldots \leq (\delta + \widetilde{\delta}) \left(\frac{1}{L_{k+1}} + \frac{1}{L_k}\left(1 - \frac{\mu}{L_k}\right) + \right. \\
&+ \frac{1}{L_{k-1}}\left(1 - \frac{\mu}{L_k}\right)\left(1 - \frac{\mu}{L_{k-1}}\right) + \ldots + \left. \frac{1}{L_1}\left(1 - \frac{\mu}{L_k}\right)\left(1 - \frac{\mu}{L_{k-1}}\right) \ldots \left(1 - \frac{\mu}{L_1}\right)\right) + \\
&+ \left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) \ldots \left(1 - \frac{\mu}{L_1}\right) V[x^0](x_*).
\end{aligned}$$

For further reasoning we introduce average parameter $\hat{L}$:

$$1 - \frac{\mu}{\hat{L}} = \sqrt[k+1]{\left(1 - \frac{\mu}{L^{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) \ldots \left(1 - \frac{\mu}{L_1}\right)}.$$

Note that by $L_i \geq \mu$ $(i = 1, 2, \ldots)$

$$\min_{1 \leq i \leq k+1} L_i \leq \hat{L} \leq \max_{1 \leq i \leq k+1} L_i \leq 2L.$$

Now, taking into account (18), we have:

$$\frac{f(x^{k+1}) - f(x_*)}{L_{k+1}} + V[x^{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{\mu} \sum_{i=0}^{k} \left(1 - \frac{\mu}{2L}\right) + \left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*) \leq \tag{19}$$

$$\leq \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*). \tag{20}$$

Finally, we have

$$V[x^{k+1}](x_*) \leq \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*).$$

and by (19) – (20) and $L^{k+1} \leq 2L$ means:

$$f(x^{k+1}) - f(x_*) \leq \frac{4L^2(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + 2L\left(1 - \frac{\mu}{\hat{L}}\right)^{k+1} V[x^0](x_*).$$

# C Analysis of Algorithm 1 in the case of $(\delta, L, \mu)$-model

**Theorem 8.** *Let $\psi_\delta(x, y)$ be a $(\delta, L, \mu)$-model for $f$ w.r.t. $V[y](x)$ and $y_k = \mathrm{argmin}_{i=1,\dots,k}(f(x_i))$. Then, after $k$ iterations of Algorithm 1, we have*

$$V[x^{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{\mu} + \left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*).$$

*and*

$$f(y_{k+1}) - f(x_*) \leq L\left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*) + \delta + \widetilde{\delta}.$$

*Proof.* Clearly, $f(x_*) \leq f(x^{k+1})$ and

$$LV[x^{k+1}](x_*) \leq \widetilde{\delta} + \delta + (L - \mu)V[x^k](x_*),$$

i.e.

$$V[x^{k+1}](x_*) \leq \frac{1}{L}(\delta + \widetilde{\delta}) + \left(1 - \frac{\mu}{L}\right) V[x^k](x_*).$$

Further,

$$V[x^{k+1}](x_*) \leq \frac{1}{L}(\delta + \widetilde{\delta}) + \left(1 - \frac{\mu}{L}\right)\left(\frac{1}{L}(\delta + \widetilde{\delta}) + \left(1 - \frac{\mu}{L}\right) V[x^{k-1}](x_*)\right) \leq \dots \leq$$
$$\leq \frac{1}{L}(\widetilde{\delta} + \delta)\left(1 + \left(1 - \frac{\mu}{L}\right) + \dots + \left(1 - \frac{\mu}{L}\right)^k\right) + \left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*).$$

Therefore, taking into account the following fact

$$\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i < \frac{1}{1 - \left(1 - \frac{\mu}{L}\right)} = \frac{L}{\mu},$$

we obtain

$$V[x^{k+1}](x_*) \leq \frac{\delta + \widetilde{\delta}}{\mu} + \left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*).$$

Now we consider the question on convergence by function:

$$V[x^{k+1}](x_*) \leq \left(f(x_*) - f(x^{k+1}) + \delta + \widetilde{\delta}\right)\frac{1}{L} + \left(1 - \frac{\mu}{L}\right) V[x^k](x_*) \leq$$
$$\leq \left(f(x_*) - f(x^{k+1}) + \delta + \widetilde{\delta}\right)\frac{1}{L} +$$
$$+ \left(1 - \frac{\mu}{L}\right)\left(\left(f(x_*) - f(x^k) + \delta + \widetilde{\delta}\right)\frac{1}{L} + \left(1 - \frac{\mu}{L}\right) V[x^{k-1}](x_*)\right) \leq$$
$$\leq \dots \leq \left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*) + \frac{1}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i \left(f(x_*) - f\left(x^{k+1-i}\right) + \delta + \widetilde{\delta}\right).$$

Therefore, we have

$$\frac{1}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i (f(x^{k+1-i}) - f(x_*)) \leq \left(1 - \frac{\mu}{L}\right)^{k+1} V[x^0](x_*) + \frac{1}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i (\delta + \widetilde{\delta}).$$

Table 1: Results for Example 9.

| | Non-adaptive | | Adaptive | |
|---|---|---|---|---|
| **k** | **Time** | **Estimate** | **Time** | **Estimate** |
| 160 | 0:01:19 | 0.19827 | 0:05:25 | 0.02110 |
| 180 | 0:01:27 | 0.16220 | 0:05:55 | 0.01258 |
| 200 | 0:01:36 | 0.13264 | 0:07:11 | 0.00750 |
| 220 | 0:01:55 | 0.10849 | 0:07:19 | 0.00474 |
| 240 | 0:01:57 | 0.08873 | 0:07:56 | 0.00282 |

Denote by $y_k = \operatorname{argmin}_{i=1,\ldots,k}(f(x_i))$. Then, taking into account

$$\frac{1}{L}\sum_{i=0}^{k}\left(1-\frac{\mu}{L}\right)^i = \frac{1}{\mu}\left(1-\left(1-\frac{\mu}{L}\right)^{k+1}\right) \geq \frac{1}{L},$$

we obtain

$$f(y_{k+1}) - f(x_*) \leq \mu\frac{\left(1-\frac{\mu}{L}\right)^{k+1}}{1-\left(1-\frac{\mu}{L}\right)^{k+1}}V[x^0](x_*) + \delta + \widetilde{\delta} \leq$$
$$\leq L\left(1-\frac{\mu}{L}\right)^{k+1}V[x^0](x_*) + \delta + \widetilde{\delta}.$$

□

# D Some Numerical Tests for Algorithms 1 and 2

We consider two numerical examples for Algorithms 1 and 2 for minimizing $\mu$-strongly convex objective function of $N$ variables on a unit ball $B_1(0)$ with center at zero with respect to the standard Euclidean norm. It is clear that such functions admit $(\delta, L, \mu)$-model of the standard form $\psi_\delta(x, y) = \langle \nabla f(y), x - y\rangle$ for the case of Lipschitz-continuous gradient $\nabla f$. In the first of the considered examples, it is easy to estimate $L$ and $\mu$, and the ratio $\frac{\mu}{L}$ is not very small, which ensures a completely acceptable rate of convergence of the non-adaptive method (see Table 1 below). In the second example, the objective is ill-conditioned meaning that the ratio $\frac{\mu}{L}$ so small that the computer considers the value $1-\frac{\mu}{L}$ to be equal to 1 and Theorem 8 for the non-adaptive algorithm does not allow to estimate the rate of convergence at all. In this case, the use of adaptive Algorithm 2 leads to noticeable results (see the Table 2 below).

**Example 9.** Consider a function

$$f(x) = x_1^2 + 2x_2^2 + 3x_3^2 + \ldots + Nx_N^2,$$

where $N = 100$ and input data

$$x^0 = \frac{(0.2, \ldots, 0.2)}{||(0.2, \ldots, 0.2)||} \text{ is the initial approximation ,}$$

$\mu = 2$, $L^0 = 2\mu$, $L = 2N$.

The results of the comparison of the work of algorithms 1 and 3 are presented in the comparative Table 1, where $k$ is the number of iterations of these algorithms. Time presented in the format of "HOURS:MINUTES:SECONDS".

Table 2: Results for Example 10.

| | Adaptive | |
|---|---|---|
| **k** | **Time** | **Estimate** |
| 50 | 0:07:37 | 0.71273 |
| 100 | 0:14:27 | 0.51241 |
| 150 | 0:23:00 | 0.372301 |
| 200 | 0:28:07 | 0.27334 |
| 250 | 0:34:32 | 0.19699 |
| 300 | 0:43:10 | 0.14456 |

As we can see from the Table 1, in the previous example the non-adaptive method converges no worse than the adaptive one. However, it is possible that $\frac{\mu}{L}$ is too small, which leads to $1 - \frac{\mu}{L} \approx 1$. In this case, Theorem 8 cannot estimate the rate of convergence of the method. We give another example.

**Example 10.** Consider the target functional

$$f(x_1, \ldots, x_N) = \sum_{k=1}^{N} (kx_k^2 + e^{-kx_k}).$$

It is easy to verify that for such a function one can choose $\mu = 2 + \frac{1}{e}$ and $L = 2N + N^2 e$ and the program calculates the value of $1 - \frac{\mu}{L}$ equal to 1. However, applying Algorithm 2 with adaptive tuning to the constant $L$ and Theorem 4 we obtain meaningful results, which we present in Table 2.

Experiments were performed using CPython 3.7 software on a computer with a 3-core AMD Athlon II X3 450 processor with a clock frequency of 3.2 GHz. The computer's RAM was 8 GB.

# E Complexity Analysis of Sinkhorn's Algorithm

Let us consider regularized optimal transport problem

$$\langle C, \pi \rangle + \gamma \sum_{i,j} \pi_{ij} \ln \pi_{ij} \to \min_{\pi \in \mathcal{U}(p,q)}. \tag{21}$$

Recall that the dual problem to (21) is equivalent to

$$f(u, v) := \langle \mathbb{1} B(u, v) \mathbb{1} \rangle - \langle u, p \rangle - \langle q, v \rangle \to \min_{u,v \in \mathbb{R}^n}, \tag{22}$$

where $B(u, v) := \operatorname{diag}(e^u) e^{-C/\gamma} \operatorname{diag}(e^v)$ [29]. Below we present a slightly refined complexity analysis of Algorithm 3 based on the same approach as in [29]. First, we prove that Sinkhorn's iterations are contractant for $e^{u^t - u^*}$ and $e^{v^t - v^*}$ in Hilbert's projective metric (cf. [34]).

**Lemma 3.** *Let us define*

$$R_t := \begin{cases} \max_j(v_j^t - v_j^*) - \min_j(v_j^t - v_j^*), & t \bmod 2 = 0, \\ \max_i(u_i^t - u_i^*) - \min_i(u_i^t - u_i^*), & t \bmod 2 = 1, \end{cases}$$

*where* $(u^*, v^*)$ *is the solution of problem* (22)*. Then for any* $t \geq 0$ *it holds* $R_{t+1} \leq R_t$*.*

*Proof.* W.l.o.g. consider even $t$. Let us denote $\pi^* = B(u^*, v^*) \in \mathcal{U}(p, q)$. Then for any $i$

$$u_i^{t+1} - u_i^* = u_i^t - u_i^* + \ln p_i - \ln \left( \sum_j e^{u_i^t - u_i^*} \pi_{ij}^* e^{v_j^t - v_j^*} \right) = - \ln \left( \sum_j \frac{\pi_{ij}^*}{p_i} e^{v_j^t - v_j^*} \right),$$

and since $\sum_j \frac{\pi_{ij}^*}{p_i} = 1$ one obtains

$$e^{\min_j (v_j^t - v_j^*)} \leq \sum_j \frac{\pi_{ij}^*}{p_i} e^{v_j^t - v_j^*} \leq e^{\max_j (v_j^t - v_j^*)},$$

therefore,

$$R_{t+1} \leq R_t.$$

□

Now repeating the proof of Theorem 1 from [29] we obtain the following complexity bound.

**Theorem 11.** *The inner cycle of Algorithm 3 stops in number of iterations*

$$N = O \left( \frac{R_0}{\varepsilon'} \right),$$

*and*

$$R_0 \leq \frac{\max_{i,j} C_{ij} - \min_{i,j} C_{ij}}{\gamma}.$$

Notice that now we require an approximated solution of regularized problem (21), thus the choice of $\varepsilon'$ in Algorithm 3 differs from the one from [2, 29].

**Theorem 12.** *Algorithm 3 returns $\hat{\pi} \in \mathcal{U}(p, q)$ s.t.*

$$\langle C, \hat{\pi} \rangle + \gamma \sum_{i,j} \hat{\pi}_{ij} \ln \hat{\pi}_{ij} \leq \langle C, \pi^* \rangle + \gamma \sum_{i,j} \pi_{ij}^* \ln \pi_{ij}^* + \tilde{\varepsilon},$$

*where $\pi^*$ is the solution of problem* (21)*.*

*Proof.* Notice that for any $\pi \in \mathcal{U}(B(u, v)\mathbb{1}, B(u, v)^T \mathbb{1})$ it holds

$$\langle C, B(u, v) \rangle + \gamma \sum_{i,j} B(u, v)_{ij} \ln B(u, v)_{ij} \leq \langle C, \pi \rangle + \gamma \sum_{i,j} \pi_{ij} \ln \pi_{ij}.$$

It is easy to see that for any pair $\pi, \tilde{\pi} \in S_{n \times n}(1)$

$$\left| \sum_{i,j} \pi_{ij} \ln \pi_{ij} - \sum_{i,j} \tilde{\pi}_{ij} \ln \tilde{\pi}_{ij} \right| \leq n^2 h \ln \frac{1}{h} + \|\pi - \tilde{\pi}\|_1 \ln \frac{1}{h} \quad \forall h \in (0, e^{-1}),$$

thus

$$\left| \sum_{i,j} \pi_{ij} \ln \pi_{ij} - \sum_{i,j} \tilde{\pi}_{ij} \ln \tilde{\pi}_{ij} \right| \leq 2 \|\pi - \tilde{\pi}\|_1 \ln \left( \frac{n^2}{\|\pi - \tilde{\pi}\|_1} \right).$$

Now, for any $\pi \in S_{n\times n}(1)$ and $r, c \in S_n(1)$ there exists $\tilde{\pi} \in \mathcal{U}(r,c)$ given by Algorithm 2 from [2] s.t. $\|\pi - \tilde{\pi}\|_1 \leq \|\pi\mathbb{1} - r\|_1 + \left\|\pi^T\mathbb{1} - c\right\|_1$. Combining all these facts together we obtain for $\hat{\pi}$ defined in Algorithm 3 the following estimate

$$\langle C, \hat{\pi}\rangle + \gamma\sum_{i,j}\hat{\pi}_{ij}\ln\hat{\pi}_{ij} \leq \langle C, \pi^*\rangle + \gamma\sum_{i,j}\pi^*_{ij}\ln\pi^*_{ij} + 2\left(\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{n^2}{\varepsilon'}\right)\right)\varepsilon'$$

Substituting

$$\varepsilon' = \frac{\tilde{\varepsilon}}{4\left(\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right)\right)}$$

we obtain

$$\begin{aligned}
&\langle C, \hat{\pi}\rangle + \gamma\sum_{i,j}\hat{\pi}_{ij}\ln\hat{\pi}_{ij} - \left[\langle C, \pi^*\rangle + \gamma\sum_{i,j}\pi^*_{ij}\ln\pi^*_{ij}\right] \\
&\quad\leq 2\left(\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{n^2}{\varepsilon'}\right)\right)\frac{\tilde{\varepsilon}}{4\left(\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right)\right)} \\
&\quad\leq \frac{\tilde{\varepsilon}}{2}\frac{\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right) + 2\gamma\ln\frac{\tilde{\varepsilon}}{4\gamma\varepsilon'}}{\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right)} \\
&\quad\leq \frac{\tilde{\varepsilon}}{2}\left(1 + \frac{\tilde{\varepsilon}/(2e\varepsilon')}{\max_{i,j}C_{ij} - \min_{i,j}C_{ij} + 2\gamma\ln\left(\frac{4\gamma n^2}{\tilde{\varepsilon}}\right)}\right) \leq \tilde{\varepsilon}.
\end{aligned}$$

□

# F Additional Experiments for Prox-Sinkhorn Algorithm

Figure 5 shows the dependence of the mean inner method iteration number upon accuracy and size of the vector $\pi$. With the growth of $L$, there is a decrease in the mean inner method iteration number. However, the type of dependence from the accuracy or size of the problem is the same.

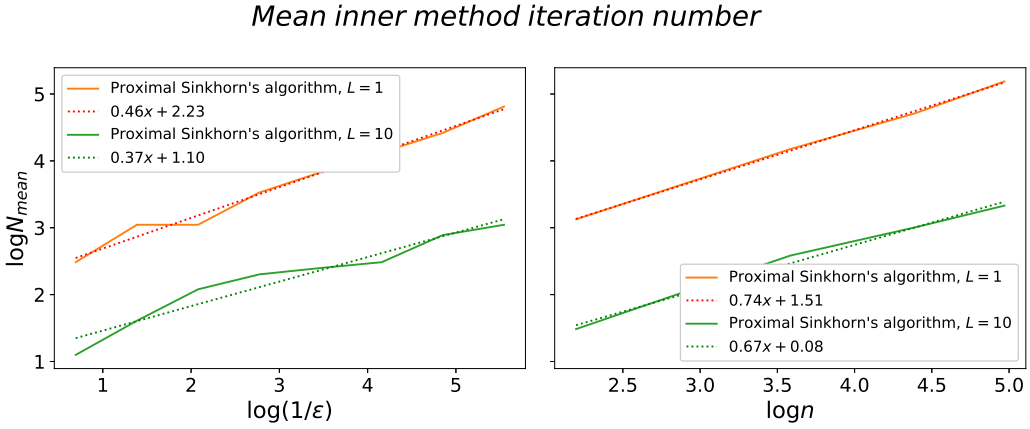


Figure 5: Comparison of inner method iteration number of proximal Sinkhorn's algorithm for different $L$.

Consider a graph of change of the auxiliary problem solution complexity with increasing external method iteration number (fig. 6). Note that at the first interval there is an increase in the inner method iteration number with two peaks at different levels. On subsequent iterations of the external method the complexity of the solution of the auxiliary problem decreases, approaching a constant.

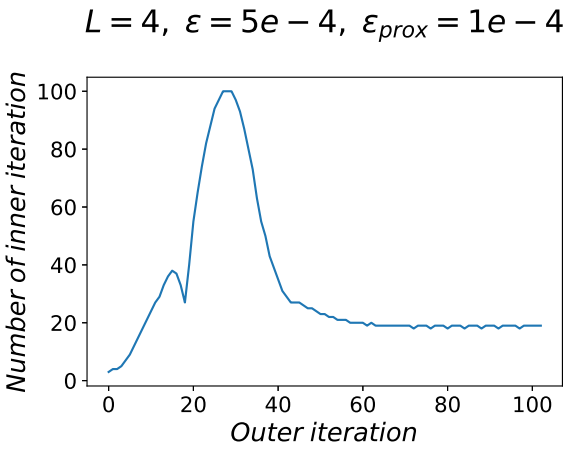


Figure 6: The dependence of the total *Number of inner iteration* from the number of *Outer iteration* of external method

## G On Inexact Solution of Auxiliary Subproblems

Our goal is to provide a relation between the accuracy of the solution of an optimization problem in terms of the objective residual and $\widetilde{\delta}$-'precision' in the sense of Definition 2.

**Theorem 13.** *Assume that we find a point $\tilde{x}$ such that $\phi(\tilde{x}) - \phi(\tilde{x}^*) \leq \widetilde{\varepsilon}$, where $\tilde{x}^*$ is an exact solution of problem* (4)*. Assume also that $\phi$ has $\tilde{L}$-Lipschitz continuous gradient in $Q$.*

*If $\nabla\phi(\widetilde{x}^*) = 0$, then $\tilde{x} = \operatorname{argmin}^{\widetilde{\delta}}_{x\in Q} \phi(x)$ with $\widetilde{\delta} = \widetilde{R}\sqrt{2\tilde{L}\widetilde{\varepsilon}}$, where $\widetilde{R} = \max_{x,y\in Q} \|y - x\|$.*

*If $\phi$ is $\mu$-strongly convex on $Q$, then $\tilde{x} = \operatorname{argmin}^{\widetilde{\delta}}_{x\in Q} \phi(x)$ with*

$$\widetilde{\delta} = (\tilde{L}\widetilde{R} + \|\nabla\phi(\tilde{x}^*)\|_*)\sqrt{2\widetilde{\varepsilon}/\mu}, \tag{23}$$

*Proof.* 1. Assume that $\nabla\phi(\tilde{x}^*) = 0$. Then

$$\frac{1}{2\tilde{L}}\|\nabla\phi(\tilde{x})\|^2_* \leq \phi(\tilde{x}) - \phi(\tilde{x}^*) \leq \widetilde{\varepsilon},$$

$$\widetilde{\delta} = \max_{x\in Q}\langle\nabla\phi(\tilde{x}), \tilde{x} - x\rangle \leq \|\nabla\phi(\tilde{x})\|_* \max_{x\in Q}\|\tilde{x} - x\| \leq \sqrt{2\tilde{L}\widetilde{\varepsilon}}\max_{x\in Q}\|\tilde{x} - x\|.$$

2. Let us now assume that $\nabla\phi(\tilde{x}^*) \neq 0$. For strongly convex function $\phi(x)$ we have

$$\frac{\mu}{2}\|\tilde{x} - \tilde{x}^*\|^2 \leq \phi(\tilde{x}) - \phi(\tilde{x}^*) \leq \widetilde{\varepsilon}.$$

Hence,

$$\|\tilde{x} - \tilde{x}^*\| \leq \sqrt{\frac{2}{\mu}\widetilde{\varepsilon}}. \tag{24}$$

Using this and Lipschitz gradient condition, we obtain

$$\|\nabla\phi(\tilde{x}) - \nabla\phi(\tilde{x}^*)\|_* \leq \tilde{L}\|\tilde{x} - \tilde{x}^*\| \leq \tilde{L}\sqrt{\frac{2}{\mu}\widetilde{\varepsilon}}. \tag{25}$$

Hence,

$$\begin{aligned}
\widetilde{\delta} &= \max_{x\in Q}\langle\nabla\phi(\tilde{x}), \tilde{x} - x\rangle = \max_{x\in Q}\langle\nabla\phi(\tilde{x}) - \nabla\phi(\tilde{x}^*), \tilde{x} - x\rangle + \max_{x\in Q}\langle\nabla\phi(\tilde{x}^*), \tilde{x} - x\rangle \\
&\overset{(25)}{\leq} \tilde{L}\sqrt{\frac{2}{\mu}\widetilde{\varepsilon}}\max_{x\in Q}\|\tilde{x} - x\| + \max_{x\in Q}\langle\nabla\phi(\tilde{x}^*), \tilde{x}^* - x\rangle + \max_{x\in Q}\langle\nabla\phi(\tilde{x}^*), \tilde{x} - \tilde{x}^*\rangle \\
&\overset{(24)}{\leq} \tilde{L}\sqrt{\frac{2}{\mu}\widetilde{\varepsilon}}\max_{x\in Q}\|\tilde{x} - x\| + \|\nabla\phi(\tilde{x}^*)\|_*\sqrt{\frac{2}{\mu}\widetilde{\varepsilon}}.
\end{aligned}$$

□

# H ProxSinkhorn Theorem

Proof of Theorem 5.

*Proof.* The estimate for the number of iterations $N$ follows from Theorem 3 since $V[\pi_0](\pi_*) \leq \ln n^2$ as $\pi \in S_{n^2}(1)$. The first component of (11) is proved in [34], and the second component basically follows from [5, 29]. Proofs of the second component and bound on $\bar{c}_k$ (12) are provided in Appendix E (Theorem 11). Let us show that it is sufficient to solve minimization problem (10) on each iteration with accuracy $\tilde{\varepsilon} = \widetilde{O}(\varepsilon^4/(Ln^4))$ in terms of the objective residual to guarantee $\tilde{\delta} = \varepsilon/2$ accuracy in terms of Definition 2.

To prove this fact, we use relation (23) in Theorem 13 of Appendix G with $\|\cdot\| = \|\cdot\|_1$, $\tilde{R} = 2$, $\mu = L$. To bound $\Delta = \tilde{L}\widetilde{R} + \|\nabla\phi(\tilde{x}^*)\|_*$ (in notations of Theorem 13) we modify $\mathcal{U}(p, q)$ by adding constraints: $\pi_{ij} \geq \varepsilon/(4n^2)$, $i, j = 1, ..., n$. The solution of the changed problem is still an $O(\varepsilon)$-solution of the original problem. For the modified problem $\Delta = 5Ln^2\tilde{R}/\varepsilon$. According to (23) one should solve auxiliary problem with accuracy by function value $\widetilde{\varepsilon}$, which is chosen such that $\varepsilon/2 = \tilde{\delta} = (5Ln^2/\varepsilon)\tilde{R}\sqrt{2\widetilde{\varepsilon}/L}$. The only problem is that now we cannot directly apply Sinkhorn's algorithm. This problem can be solved by trivial affine transformation of $\pi$-space. This transformation reduces modified polyhedral to the standard one and we can use Sinkhorn's Algorithm. Such a transformation doesn't change (in terms of $O(\ )$) the requirements to the accuracy. □

# I On Clustering by Electoral model

Let us show accurately first that the norm $||\cdot||$ introduced as $||(z, p)||^2 = ||z||_1^2 + ||p||_2^2$ is indeed a norm:

1 $||\alpha x|| = \sqrt{||\alpha z||_1^2 + ||\alpha p||_2^2} = \sqrt{\alpha^2(||z||_1^2 + ||p||_2^2)} = |\alpha| \cdot ||x||$

2 $||x|| = 0 \iff ||z||_1^2 + ||p||_2^2 = 0 \iff ||z||_1 = 0$ and $||p||_2 = 0 \iff x = 0$

3 $||x + y|| = \sqrt{||z_x + z_y||_1^2 + ||p_x + p_y||_2^2} \leq \sqrt{(||z_x||_1 + ||z_y||_1)^2 + (||p_x||_2 + ||p_y||_2)^2} \leq$
$\leq \sqrt{||z_x||_1^2 + ||p_x||_2^2} + \sqrt{||z_y||_1^2 + ||p_y||_2^2} = ||x|| + ||y||$

The last inequality holds since $\sqrt{(a+b)^2+(c+d)^2} \leq \sqrt{a^2+c^2}+\sqrt{b^2+d^2}$ for each $a, b, c, d \geq 0$.

Now prove that (7) satisfies the definition of (0, $2L_g$)-model of the function (5).

It is easy to see that $\psi_\delta(x,x) = 0$. Let us show, that inequality (2) holds for $\psi_\delta(x,y)$. For the function $g(x)$ satisfying (6) we have:

$$|g(x) - g(y) - \langle \nabla g(y), x-y\rangle| \leq \frac{L_g}{2}\|x-y\|^2. \tag{26}$$

It means that $f_{\mu_1,\mu_2}(x) - f_{\mu_1,\mu_2}(y) - \psi_\delta(x,y) =$

$$= g(x) - g(y) - \langle \nabla g(y), x-y\rangle + \mu_1 \cdot KL(z_x|\mathbf{1}) - \mu_1 \cdot KL(z_y|\mathbf{1}) +$$
$$+\frac{\mu_2}{2}\|p_x\|_2^2 - \frac{\mu_2}{2}\|p_y\|_2^2 - \mu_1 \cdot KL(z_x|\mathbf{1}) + \mu_1 \cdot KL(z_y|\mathbf{1}) + L_g \cdot KL(z_x|z_y) +$$
$$+\frac{L_g}{2}\|p_x - p_y\|_2^2 - \frac{\mu_2}{2}\|p_x\|_2^2 + \frac{\mu_2}{2}\|p_y\|_2^2 =$$

$$= g(x) - g(y) - \langle \nabla g(y), x-y\rangle + L_g \cdot KL(z_x|z_y) + \frac{L_g}{2}\|p_x - p_y\|_2^2.$$

Along with (26) and $KL(z_x|z_y) \geq \frac{\|z_x - z_y\|_1^2}{2}$ it leads to

$$f_{\mu_1,\mu_2}(x) - f_{\mu_1,\mu_2}(y) - \psi_\delta(x,y) \leq \frac{L_g}{2}\|x-y\|^2 + L_g \cdot KL(z_x|z_y) + \frac{L_g}{2}\|p_x - p_y\|_2^2,$$
$$f_{\mu_1,\mu_2}(x) - f_{\mu_1,\mu_2}(y) - \psi_\delta(x,y) \geq -\frac{L_g}{2}\|x-y\|^2 + L_g \cdot KL(z_x|z_y) + \frac{L_g}{2}\|p_x - p_y\|_2^2.$$

Finally, by definition of the norm $\|\cdot\|$, we have

$$0 \leq f_{\mu_1,\mu_2}(x) - f_{\mu_1,\mu_2}(y) - \psi_\delta(x,y) \leq 2L_g \cdot KL(z_x|z_y) + L_g\|p_x - p_y\|_2^2 = 2L_g V[y](x),$$

i.e. $\psi_\delta(x,y)$ is a $(0, 2L_g)$-model of the function $f_{\mu_1,\mu_2}$.

Finally, let us show that inequality (8) holds, i.e. $\psi_\delta(x,y)$ is strongly convex w.r.t. $V[y](x)$. Indeed,

$$\psi_\delta(x,y) = \langle \nabla g(y), x-y\rangle + \mu_1\langle \nabla KL(z_y|1), z_x - z_y\rangle + \mu_2\langle p_y, p_x - p_y\rangle -$$
$$-L_g \cdot KL(z_x|z_y) - \frac{L_g}{2}\|p_x - p_y\|_2^2 + \mu_1\left(\cdot KL(z_x|\mathbf{1}) - KL(z_y|\mathbf{1}) - \langle \nabla KL(z_y|1), z_x - z_y\rangle\right) +$$
$$+\frac{\mu_2}{2}\left(\|p_x\|_2^2 - \|p_y\|_2^2 - \langle 2\cdot p_y, p_x\rangle - p_y\right) =$$
$$= \psi_\delta^{lin}(x,y) + (\mu_1 - L_g)\cdot KL(z_x|z_y) + \frac{\mu_2 - L_g}{2}\cdot\|p_x - p_y\|_2^2$$

and inequality is proven.

# J Figures 5 and 6


(a) Iterates (b) Euclidean Mean (c) Function Value

Figure 7: Computation of the barycenter of a set of $20$ images of the number $7$ from the MNIST dataset [41].

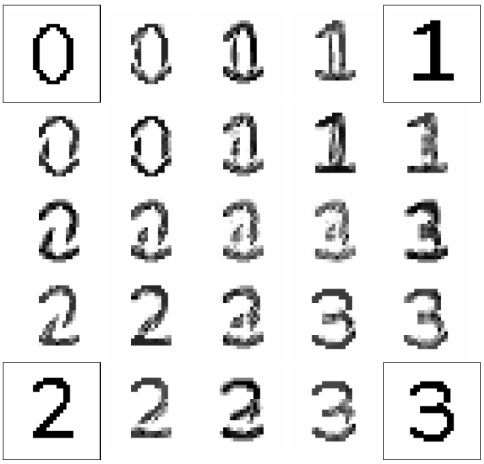

Figure 8: Approximate barycenter computed by the Prox IBP algorithm, for different weighting combinations between four original images (marked by the black boxes).