Advances in Low-Memory Subgradient Optimization

Pavel E. Dvurechensky, Alexander V. Gasnikov, Evgeni A. Nurminski and Fedor S. Stonyakin

Abstract

This chapter is devoted to the black-box subgradient algorithms with the minimal requirements for the storage of auxiliary results, which are necessary to execute these algorithms. It starts with the original result of N.Z. Shor which open this field with the application to the classical transportation problem. To discuss the fundamentals of non-smooth optimization the theoretical complexity bounds for smooth and non-smooth convex and quasi-convex optimization problems are briefly exposed with the special attention given to adaptive step-size policy. Than this chapter contains descriptions of different modern techniques that allow to solve non-smooth convex optimization problems faster then lower complexity bounds: Netserov smoothing technique, Netserov Universal approach, Legendre (saddle point) representation approach. The new results on Universal Mirror Prox algorithms represent the original parts of the survey. To demonstrate application of non-smooth convex optimization algorithms for solution of huge-scale extremal problems we consider convex optimization problems with non-smooth functional constraints and propose two adaptive Mirror Descent methods. The first method is of primal-dual

Pavel E. Dvurechensky
Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, Berlin, 10117, Germany and Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, build.1, Moscow, 127051, Russia e-mail: pavel.dvurechensky@wias-berlin.de

Alexander V. Gasnikov
Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia e-mail: gasnikov@yandex.ru

Evgeni A. Nurminski
Far Eastern Federal University, Russky ostrov, Vladivostok, 690000, Russia e-mail: nurminskiy.ea@dvfu.ru

Fedor S. Stonyakin
V.I. Vernadsky Crimean Federal University, 4 V. Vernadsky Ave, Simferopol, 295007 and Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701 e-mail: fedyor@mail.ru

variety and proved to be optimal in terms of lower oracle bounds for the class of Lipschitz-continuous convex objective and constraints. The advantages of application of this method to sparse Truss Topology Design problem are discussed in some details. The second method can be applied for solution of convex and quasi-convex optimization problems and is optimal in a sense of complexity bounds. The conclusion part of the survey contains the important references that characterize recent developments of non-smooth convex optimization.

Contents

1 Introduction

We consider a finite-dimensional nondifferentiable convex optimization problem (COP)

$$\min_{x \in E} f(x) = f_\star = f(\mathbf{x}^\star), \mathbf{x}^\star \in X_\star\,, \tag{1}$$

where E denotes a finite-dimensional space of primal variables and $f : E \to \mathbb{R}$ is a finite convex function, not necessarily differentiable. For a given point $\mathbf{x}$ the subgradient oracul returns value of objective function at that point $f(\mathbf{x})$ and subgradient $g \in \partial f(\mathbf{x})$. We do not make any assumption about the choice of g from $\partial f(\mathbf{x})$. As we are interested in computational issues related to solving (1) mainly we assume that this problem is solvable and has nonempty and bounded set of solutions $X_\star$.

This problem enjoys a considerable popularity due to its important theoretical properties and numerous applications in large-scale structured optimization, discrete optimization, exact penalization in constrained optimization, and others. Nonsmooth optimization theory made it possible to solve in an efficient way classical descrete min-max problems [23], l_1-approximation and others, at the same time opening new approaches in bi-level, monotropic programming, two-stage stochastic optimization, to name a few.

As a major steps in :the development of different algorithmic ideas we can start with the subgradient algorithm due to Shor (see [71] for the overview and references to earliest works).

2 Example Application: Transportation Problem

From utilitarian point of view the development of non-smooth (convex) optimization started with the classical transportation problem

$$\begin{array}{c} \min \ \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \\ \sum_{i=1}^m x_{ij} = a_j, \ j = 1,2,\ldots,n; \\ \sum_{j=1}^n x_{ij} = b_i, \ i = 1,2,\ldots,m \\ x_{ij} \geq 0, i = 1,2,\ldots,m; j = 1,2,\ldots,n \end{array} \tag{2}$$

which is widely used in many applications.

By dualizing this problem with respect to balancing constrains we can convert (2) into dual problem of the kind

$$\max \ \Phi(\mathbf{u},\mathbf{v}) \tag{3}$$

where $\mathbf{u} = (u_i, i = 1,2,\ldots,m); \mathbf{v} = (v_j, j = 1,2,\ldots,n)$ are dual variables associated with the balancing constraints in (2) and $\Phi(\mathbf{u},\mathbf{v})$ is the dual function defined as

$$\Phi(\mathbf{u},\mathbf{v}) = \inf_{\mathbf{x} \geq 0} L(\mathbf{x},\mathbf{u},\mathbf{v}) \tag{4}$$

and $L(\mathbf{x},\mathbf{u},\mathbf{v})$ is the Lagrange function of the problem:

$$L(\mathbf{x},\mathbf{u},\mathbf{v}) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{j=1}^{n} u_j(\sum_{i=1}^{m} x_{ij} - a_j) + \sum_{i=1}^{m} v_i(\sum_{j=1}^{n} x_{ij} - b_i).$$

By rearranging terms in this expression we can obtain the following expression for the dual function

$$\begin{gathered}\Phi(\mathbf{u},\mathbf{v}) = -m\textstyle\sum_{j=1}^{n} u_j a_j - n\sum_{i=1}^{m} v_i b_i + \sum_{i=1}^{m}\sum_{j=1}^{n} \inf_{\mathbf{x}\geq 0} x_{ij}\{c_{ij} + u_j + v_i\} = \\ -m\textstyle\sum_{j=1}^{n} u_j a_j - n\sum_{i=1}^{m} v_i b_i - \mathrm{Ind}_D(\mathbf{u},\mathbf{v}),\end{gathered} \tag{5}$$

where

$$\mathrm{Ind}_D(\mathbf{u},\mathbf{v}) = \begin{cases} 0 & \text{when } c_{ij} + u_i + v_j \geq 0; \\ \infty & \text{otherwise.} \end{cases} \tag{6}$$

is the indicator function of the set $D = \{\mathbf{u},\mathbf{v} : c_{ij} + u_j + v_i \geq 0, i = 1,2,\dots,m; j = 1,2,\dots,n\}$ which is the feasible set of the dual problem.

Of course, by explicitely writing feasibility constraints for (3) we obtain the linear dual transportation problem with a fewer variables but with much higher number of constraints. This is bad news for textbook simplex method so many specialized algorithms were developed, one of them was simple-minded method of generalized gradient which started the development of non-smooth optimization.

This method relies on subgradient of concave function $\Phi(\mathbf{u},\mathbf{v})$ which we can transform into convex just by changing signs and replacing inf with sup

$$\begin{gathered}\Phi(\mathbf{u},\mathbf{v}) = m\textstyle\sum_{j=1}^{n} u_j a_j + n\sum_{i=1}^{m} v_i b_i + \\ \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} \sup_{\mathbf{x}\geq 0} x_{ij}\{c_{ij} + u_j + v_i\} = \\ = m\textstyle\sum_{j=1}^{n} u_j a_j + n\sum_{i=1}^{m} v_i b_i + \mathrm{Ind}_D(\mathbf{u},\mathbf{v}),\end{gathered}$$

and ask for its *minimization*.

According to convex analysis [65] the subdifferential $\partial_c\Phi(\mathbf{u},\mathbf{v})$ exists for any $\mathbf{v},\mathbf{u} \in \operatorname{int}\operatorname{dom}(\mathrm{Ind}_D)$, and in this case just equals to the (constant) vector $g_L = (\mathbf{g_u},\mathbf{g_v}) = (\mathbf{a},\mathbf{b})$ of a linear objective in the interior of D. The situation becomes more complicated when $\mathbf{u}, bv$ happens to be at the boundary of D, the subdifferential set ceases to be a singleton and becomes even unbounded, roughly speaking certain linear manifolds are added to g_L but we will not go into details here. The difficulty is that if we mimic gradient method of the kind

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \lambda g_L^u = \mathbf{u}^k - \lambda\mathbf{a}; \mathbf{v}^{k+1} = \mathbf{v}^k - \lambda g_L^v = \mathbf{v}^k - \lambda\mathbf{b}; k = 0,1,\dots \tag{7}$$

with a certain step-size $\lambda > 0$, we inevitably violate the dual feasibility constraints as $\mathbf{a},\mathbf{b} > 0$. Than the dual function (7) becomes undefined and correspondently the subdifferential set becomes undefined as well.

There are at least two simple ways to overcome this difficulty. One is to incorporate in the gradient method certain operations which restore feasibility and the appropriate candidate for it is the orthogonal projection operation where one can make use of the special structure of constraints and sparsity. However it will still require computing projection operator of the kind $B^T(BBT)^{-1}B$ for basis matrices B with rather uncertain number of iteration and of matrices of the size around $(n+m)\times(n+m)$. Neither computers speed nor memory sizes at that time where not up to demands to solve problems of $n+m \approx 10^4$ which was required by GOSPLAN !

The second ingenious way was to take into account that if $\sum_{j=1}^n a_j = \sum_{i=1}^m b_i = V$, which is required anyway for solvability of transportation problem in a closed form. The flow variables may be uniformally bounded by V and the dual function can be redefined as

$$\begin{gathered}\Phi_V(\mathbf{u},\mathbf{v}) = m\sum_{j=1}^n u_j a_j + n\sum_{i=1}^m v_i b_i - \\ \sum_{i=1}^m \sum_{j=1}^n \max_{0\le \mathbf{x}\le V} x_{ij}\{c_{ij}+u_j+v_i\} = \\ = m\sum_{j=1}^n u_j a_j + n\sum_{i=1}^m v_i b_i + P_V(\mathbf{u},\mathbf{v})\end{gathered}$$

where the penalty function $P_V(\mathbf{u},\mathbf{v})$ is easily computed by component-wise maximization:

$$\begin{gathered}P_V(\mathbf{u},\mathbf{v}) = \sum_{i=1}^m \sum_{j=1}^n \max_{x_{ij}\in[0,V]} x_{ij}\{c_{ij}+u_j+v_i\} = \\ \sum_{i=1}^m \sum_{j=1}^n V\{c_{ij}+u_j+v_i\}_+\end{gathered}$$

where $\{\cdot\}_+ = \max\{0,\cdot\}$. Than the dual objective function becomes finite, the optimization problem — unconstrained and we can use simple subgradient method with very low requirements for memory and computations.

Actually even tighter bounds $\mathbf{x}_{ij} \le \min(a_i, b_j)$ can be imposed on the flow variables which may be advantageous for computational reasons.

In both cases there is a fundamental problem of recovering optimal primal $n\times m$ primal solution from $n+m$ dual. This problem was studied by many authors and recent advances in this area can be studied from the excellent paper by A. Nedic and A. Ozdoglar [47]. Theoretically speaking, nonzero positive values of $c_{ij}+u^\star_j+v^\star_i$, where $\mathbf{u}^\star, \mathbf{v}^\star$ are the *exact* optimal solutions of the dual problem (3) signal that the corresponding optimal primal flow $x^\star_{ij}$ is equal to zero. Hopefully after excluding these variables we obtain nondegenerate basis and can compute the remaining variables by simple and efficient linear algebra, especially taking into account the uni-modularity of basis.

However the theoretical gap between zeros and non-zeros is exponentially small even for modest length integer data therefore we need an accuracy unattainable by modern 64-128 bits hardware. Also the real life computations are always accompanied by numerical noise and we face the hard choice in fact guessing which dual constraints are active and which are not.

To connect the transportation problem with non-smooth optimization notice that the penalty function $P_V(\mathbf{u},\mathbf{v})$ is finite with the subdifferential $\partial_c P_V(\mathbf{u},\mathbf{v})$ which can be represented as a set of $n\times m$ matrices

$$\mathbf{g}_{ij} = \begin{cases} V & \text{if } c_{ij} + u_j + v_i > 0 \\ 0 & \text{if } c_{ij} + u_j + v_i < 0 \\ \mathrm{cone}(0, V) & \text{if } c_{ij} + u_j + v_i = 0 \end{cases}$$

so the subdifferential set is a convex hull of up to $2^{(n+m)}$ extreme points — enormous number even for a modest size transportation problem. Nevertheless it is easy to get at least single member of subdifferential and consider the simplest version of subgradient method:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \lambda \bar{\mathbf{g}}^k, k = 0, 1, \dots$$

where $\mathbf{x}^0$ is a given starting point, $\lambda > 0$ — fixed step-size and $\bar{\mathbf{g}}^k = \mathbf{g}^k / \|\mathbf{g}^k\|$ is a normalized subgradient $\mathbf{g}^k \in \partial f(\mathbf{x}^k)$. Of course we assume that $\mathbf{g}^k \neq 0$ otherwise $\mathbf{x}^k$ is already a solution.

Of course, there is no hope of classical convergence result such that $\mathbf{x}^k \to \mathbf{x}^\star \in X_\star$, but the remarkable theorem of Shor [68] establishes that this simplest algorithm determines at least the approximate solution.

3 The First Subgradient Algorithm

As a major step in the development of different algorithmic ideas we can start with the subgradient algorithm due to Shor (see [71] for the overview and references to earliest works). Of course, there is no hope of classical convergence result such that $\mathbf{x}^k \to \mathbf{x}^\star \in X_\star$, but the remarkable theorem of Shor [68] establishes that this very simple algorithm provides an approximate solution of (1) at least theoretically.

Theorem 1. *Let f is a finite convex function with a subdifferential ∂f and the sequence $\{\mathbf{x}^k\}$ is obtained by the recursive rule*

$$x^{k+1} = \mathbf{x}^k - \lambda g_\nu^k, k = 0, 1, \dots \tag{8}$$

with $\lambda > 0$ and $g_\nu^k = g^k / \|g^k\|, g^k \in \partial f(\mathbf{x}^k),\ g^k \neq 0$ is a normalized subgradient at the point x^k. Then for any $\varepsilon > 0$ there is an infinite set $Z_\varepsilon \subset Z$ such that for any $k \in Z_\varepsilon$

$$f(\tilde{\mathbf{x}^k}) = f(\mathbf{x}^k) \text{ and } \mathrm{dist}(\tilde{\mathbf{x}^k}, X_\star) \leq \lambda(1 + \varepsilon)/2.$$

The statement of the theorem is illustrated on Fig. 1 together with the idea of the proof. The detailed proof of the theorem goes like following: Let $\mathbf{x}^\star \in X_\star$ and estimate

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 = \|\mathbf{x}^k - \mathbf{x}^\star - \lambda \mathbf{g}_\nu^k\|^2 = \|\mathbf{x}^k - \mathbf{x}^\star\|^2 + \lambda^2 - 2\lambda \bar{\mathbf{g}}^k(\mathbf{x}^k - \mathbf{x}^\star).$$

The last term in fact equals

$$\min_{\mathbf{z} \in H_k} \|\mathbf{x}^\star - \mathbf{z}\|^2 = \|\mathbf{x}^\star - \mathbf{z}^k\|^2 = \delta_k,$$

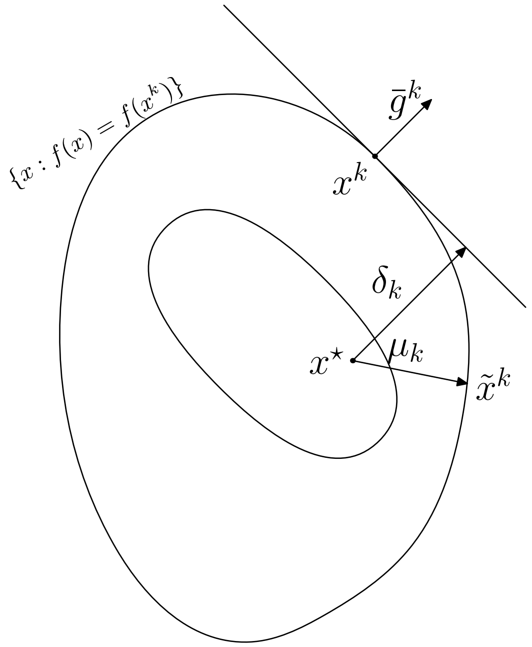

Fig. 1 The statement and the idea of the proof of Shor's theorem

where $H_k = \{\mathbf{z} : \mathbf{z}g_\nu^k = \mathbf{x}^k g_\nu^k$ is a hyperplane, orthogonal to g_ν^k and passing through the point $\mathbf{x}^k$, so

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 = \|\mathbf{x}^k - \mathbf{x}^\star\|^2 + \lambda^2 - 2\lambda\delta_k,\ k = 0,1,2,\dots \tag{9}$$

If $\lambda^2 - 2\lambda\delta_k \leq -\lambda^2\varepsilon$ for any $k \in Z$ then

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \lambda^2\varepsilon,\ k = 0,1,2,\dots \tag{10}$$

therefore

$$0 \leq \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}^0 - \mathbf{x}^\star\|^2 \leq -k\lambda^2\varepsilon \to -\infty \tag{11}$$

when $k \to \infty$. This contradiction proves that there is k_0 such that $\lambda^2 - 2\lambda\delta_{k_0} > -\lambda^2\varepsilon$ or $\delta_{k_0} < \lambda(1+\varepsilon)/2$.

To complete the proof notice that by convexity $f(z^{k_0}) \geq f(\mathbf{x}_{k_0})$ and therefore

$$\min_{z:f(z)=f(\mathbf{x}^{k_0})} \|\mathbf{x}^\star - z\|^2 = \|\mathbf{x}^\star - \bar{z}^{k_0}\|^2 = \min_{z:f(z)\geq f(x^{k_0})} \|\mathbf{x}^\star - z\|^2 \leq \|\mathbf{x}^\star - z^{k_0}\|^2 = \delta_{k_0}. \tag{12}$$

By setting $\tilde{\mathbf{x}}^0 = z^{k_0}$ we obtain $\|\mathbf{x}^\star - \tilde{\mathbf{x}}^0\|^2 < \lambda(1+\varepsilon)/2$.

By replacing x^0 in (11) by $\tilde{\mathbf{x}}^0$ and repeating the reasoning above we obtain $\tilde{x}^1$ such that $\|\mathbf{x}^\star - \tilde{\mathbf{x}}^1\|^2 < \lambda(1+\varepsilon)/2$, then in the same manner $\tilde{\mathbf{x}}^2, \tilde{\mathbf{x}}^3$ and so on with $\|\mathbf{x}^\star - \tilde{\mathbf{x}}^k\|^2 < \lambda(1+\varepsilon)/2, k = 2,3,\dots$ which complete the proof. ■

4 Complexity Results for Convex Optimization

At this section we describe the complexity results for nonsmooth convex optimization problems. Most of the results mentioned below can be found in books [51, 64, 61, 15, 9]. We start with the case when $N \geq n = \dim x$, where N is a number of oracle calls (number of subgradient calculations or/and calculations of separation hyperplane to some simple set).

Let's consider convex optimization problem

$$f(x) \to \min_{x \in Q}, \tag{13}$$

where Q – is a compact and simple set (it's significant here!). We'd like to find such a point x^N that

$$f\left(x^N\right) - f_* \leq \varepsilon,$$

where $f_* = f(x_*)$ is an optimal value of function in (13), x_* – the solution of (13). The lower and the upper bounds for the oracle complexity is (up to a multiplier, depends on Q under logarithm)

$$N \sim n \ln\left(\Delta f / \varepsilon\right),$$

where $\Delta f = \sup\limits_{x,y \in Q} \{f(y) - f(x)\}$. The center of gravity method [46, 25] converges according to this estimate. The center of gravity method in $n = 1$ is a simple binary search method [12]. But in $n > 1$ this method is hard to implement. The complexity of iteration is too high, because we required center of gravity oracle [15]. Wellknown ellipsoid method [69, 51] requires $N = \tilde{\mathrm{O}}\left(n^2 \ln\left(\Delta f / \varepsilon\right)\right)$ oracle calls and $\mathrm{O}\left(n^2\right)$ iteration complexity. Here and below $\tilde{\mathrm{O}}\,()$ means $\mathrm{O}\,()$ up to $\mathrm{O}\left(\ln^{\mathrm{O}(1)} n\right)$-factor (typically this factor is just $\mathrm{O}(\ln n)$). In [76, 15] a special version of cutting plane method was proposed. This method (Vayda's method) requires $N = \tilde{\mathrm{O}}\left(n \ln\left(\Delta f / \varepsilon\right)\right)$ oracle calls and has iteration complexity $\tilde{\mathrm{O}}\left(n^{2.37}\right)$. In the work [45] there proposed a method with $N = \tilde{\mathrm{O}}\left(n \ln\left(\Delta f / \varepsilon\right)\right)$ oracle calls and iteration complexity $\tilde{\mathrm{O}}\left(n^2\right)$. Unfortunately, for the moment it's not obvious that this method is very practical one due to the large log-factors in $\tilde{\mathrm{O}}\,()$.

Based on ellipsoid method in the late 70-th Leonid Khachyan showed [41] that LP is in P in byte complexity. Assume we have to answer is $Ax \leq b$ solvable ($n = \dim x$, $m = \dim b$)? We assume that all elements of A and b are integers. And we'd like to find one of the exact solutions x_*. This problem up to a logarithmic factor in complexity is equivalent to the problem to find the exact solution of LP problem

$\langle c,x\rangle \to \min\limits_{Ax\leq b}$ with integer A, b and c. To find the exact solution of $Ax=b$ one can use polynomial Gauss algorithm $\mathrm{O}\left(n^3\right)$.

What is about $Ax\leq b$? Let's introduce

$$\Lambda = \sum_{i,j=1,1}^{m,n} \log_2 \left|a_{ij}\right| + \sum_{i=1}^{m} \log_2 \left|b_i\right| + \log_2 (mn) + 1.$$

If $Ax\leq b$ is compatible, then there exists such x_* that $\|x_*\|_\infty \leq 2^\Lambda$, $Ax_*\leq b$ otherwise

$$\min_x \|(Ax-b)_+\|_\infty \geq 2^{-(\Lambda-1)}.$$

So one should reformulate $Ax\leq b$ as nonsmooth convex optimization problem

$$\|(Ax-b)_+\|_\infty \to \min_{\|x_*\|_\infty\leq 2^\Lambda}.$$

The approach is to apply ellipsoid method for this problem with $\varepsilon = 2^{-\Lambda}$.

Works in $\mathrm{O}(n\Lambda)$-bit arithmetic with $\tilde{\mathrm{O}}\left(mn+n^2\right)$ cost of PC memory one can find x_* (if it's exist) for $\tilde{\mathrm{O}}\left(n^3\left(n^2+m\right)\Lambda\right)$ a.o. Note, that in the ideal arithmetic with real numbers it is still an open question [10]: is it possible to find the exact solution of LP problem (with real numbers) in polynomial time in ideal arithmetic ($\pi\cdot e$ – costs $\mathrm{O}(1)$).

Table 1 describes (for more details see [9, 15, 61]) optimal estimates for the number of oracle calls for convex optimization problem (13) in the case when $N\leq n$. Now Q is not necessarily compact set.

Table 1 Optimal estimates for the number of oracle calls

$N\leq n$	$\lvert f(y)-f(x)\rvert \leq M\|y-x\|$	$\|\nabla f(y)-\nabla f(x)\|_* \leq L\|y-x\|$
$f(x)$ convex	$\mathrm{O}\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$\mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
$f(x)$ $\mu-$strongly convex in $\|\|$-norm	$\tilde{\mathrm{O}}\left(\frac{M^2}{\mu\varepsilon}\right)$	$\tilde{\mathrm{O}}\left(\sqrt{\frac{L}{\mu}}\left\lceil\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil\right)$ $(\forall N)$

Here R is a "distance" (up to a $\ln n$-factor) between starting point and the nearest solution

$$R = \tilde{\mathrm{O}}\left(\left\|x^0-x_*\right\|\right).$$

Let's describe optimal method in the most simple case: $Q=\mathrm{R}^n$, $\|\| = \|\|_2$ [64, 55]. Define

$$B_2^n(x_*,R) = \{x\in\mathrm{R}^n: \ \|x-x_*\|_2\leq R\}.$$

The main iterative process is (for simplicity we'll denote arbitrary element of $\partial f(x)$ as $\nabla f(x)$)

$$x^{k+1} = x^k - h\nabla f\left(x^k\right). \tag{14}$$

Assume that under $x \in B_2^n\left(x_*, \sqrt{2}R\right)$

$$\|\nabla f(x)\|_2 \leq M, \tag{15}$$

where $R = \left\|x^0 - x_*\right\|_2$.

Hence, from (14), (15) we have

$$\left\|x - x^{k+1}\right\|_2^2 = \left\|x - x^k + h\nabla f\left(x^k\right)\right\|_2^2 =$$

$$= \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2\left\|\nabla f\left(x^k\right)\right\|_2^2 \leq$$

$$\leq \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2M^2.$$

Here we choose $x = x_*$ (if x_* isn't unique, we choose the nearest x_* to x^0)

$$f\left(\frac{1}{N}\sum_{k=0}^{N-1} x^k\right) - f_* \leq \frac{1}{N}\sum_{k=0}^{N-1} f\left(x^k\right) - f(x_*) \leq \frac{1}{N}\sum_{k=0}^{N-1}\left\langle \nabla f\left(x^k\right), x^k - x_*\right\rangle \leq$$

$$\leq \frac{1}{2hN}\sum_{k=0}^{N-1}\left\{\left\|x_* - x^k\right\|_2^2 - \left\|x_* - x^{k+1}\right\|_2^2\right\} + \frac{hM^2}{2} =$$

$$= \frac{1}{2hN}\left(\|x_* - x^0\|_2^2 - \|x_* - x^N\|_2^2\right) + \frac{hM^2}{2}.$$

If

$$h = \frac{R}{M\sqrt{N}}, \quad \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k, \tag{16}$$

then

$$f\left(\bar{x}^N\right) - f_* \leq \frac{MR}{\sqrt{N}}. \tag{17}$$

Note that the precise lower bound for fixed steps first-order methods for the class of convex optimization problems with (15) [26]

$$f\left(x^N\right) - f_* \geq \frac{MR}{\sqrt{N+1}}.$$

Inequality (17) means that (see also Table 1)

$$N = \frac{M^2R^2}{\varepsilon^2}, \quad h = \frac{\varepsilon}{M^2}.$$

So, one can mentioned that if we will use in (14)

$$x^{k+1} = x^k - h_k\nabla f\left(x^k\right), \quad h_k = \frac{\varepsilon}{\|\nabla f(x^k)\|_2^2} \tag{18}$$

the result (17) holds true with [55]

$$\bar{x}^N = \frac{1}{\sum\limits_{k=0}^{N-1} h_k} \sum_{k=0}^{N-1} h_k x^k.$$

If we put in (18),

$$h_k = \frac{R}{\|\nabla f(x^k)\|_2 \sqrt{N}},$$

like in (16), the result analogues to (17) also holds true

$$\min_{k=0,\dots,N-1} f\left(x^k\right) - f_* \leq \frac{MR}{\sqrt{N}}$$

not only for the convex functions, but also for quasi-convex functions [13, 53]:

$$f(\alpha x + (1-\alpha)y) \leq \max\{f(x), f(y)\} \text{ for all } x, y \in Q, \alpha \in [0,1].$$

Note that

$$0 \leq \frac{1}{2hk}\left(\left\|x_* - x^0\right\|_2^2 - \left\|x_* - x^k\right\|_2^2\right) + \frac{hM^2}{2},$$

Hence for all $k = 0, \dots, N$

$$\left\|x_* - x^k\right\|_2^2 \leq \left\|x_* - x^0\right\|_2^2 + h^2M^2k \leq 2\left\|x_* - x^0\right\|_2^2,$$

therefore

$$\left\|x^k - x_*\right\|_2 \leq \sqrt{2}\left\|x^0 - x_*\right\|_2, \quad k = 0, \dots, N. \tag{19}$$

Inequality (19) justifies that we need assumption (15) holds true only with $x \in B_2^n\left(x_*, \sqrt{2}R\right)$.

For the general (constrained) case (13) we introduce norm $\|\|$, prox-function $d(x) \geq 0$ ($d\left(x^0\right) = 0$) which is 1-strongly convex due to $\|\|$ and Bregman's divergence [9]

$$V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle.$$

We put $R^2 = V\left(x_*, x^0\right)$, where x_* – is solution of (13) (if x_* isn't unique then we assume that x_* is minimized $V\left(x_*, x^0\right)$). So instead of (15) we'll have ($\|\nabla f(x)\|_* \leq M$) for all $x : V(x, x_*) \leq 2V(x^0, x_*)$.

$$2V\left(x, x^{k+1}\right) \leq 2V\left(x, x^k\right) + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2M^2.$$

Mirror Descent [49, 9] for $k = 0, \dots, N-1$

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h\partial f\left(x^k\right)\right), \quad \mathrm{Mirr}_{x^k}(\mathrm{v}) = \arg\min_{x \in Q}\left\{\left\langle \mathrm{v}, x - x^k\right\rangle + V\left(x, x^k\right)\right\}.$$

And analogues of formulas (16), (17) are also valid

$$f\left(\bar{x}^N\right) - f_* \le \frac{\sqrt{2}MR}{\sqrt{N}}, \quad \left\|x^k - x_*\right\| \le 2\sqrt{V\left(x_*, x^0\right)}, \quad h = \frac{\varepsilon}{M^2}.$$

Typically,

$$\frac{1}{2}\left\|x_* - x^0\right\|^2 \le R^2 \le C\ln n \cdot \left\|x_* - x^0\right\|^2.$$

Example (unit simplex). We have

$$Q = S_n\left(1\right) = \left\{x \in \mathrm{R}_+^n : \sum_{i=1}^n x_i = 1\right\}, \quad \left\|\nabla f\left(x\right)\right\|_\infty \le M_\infty, \quad x \in Q,$$

$$\|\| = \|\|_1, \quad d\left(x\right) = \ln n + \sum_{i=1}^n x_i \ln x_i, \quad h = M_\infty^{-1}\sqrt{2\ln n/N}, \quad x_i^0 = 1/n, \quad i = 1,...,n.$$

For $k = 0,...,N-1$, $i = 1,...,n$

$$x_i^{k+1} = \frac{\exp\left(-h\sum\limits_{r=1}^k \nabla_i f\left(x^r\right)\right)}{\sum\limits_{l=1}^n \exp\left(-h\sum\limits_{r=1}^k \nabla_l f\left(x^r\right)\right)} = \frac{x_i^k \exp\left(-h\nabla_i f\left(x^k\right)\right)}{\sum\limits_{l=1}^n x_l^k \exp\left(-h\nabla_l f\left(x^k\right)\right)}.$$

The main result here is

$$f\left(\bar{x}^N\right) - f_* \le M_\infty\sqrt{\frac{2\ln n}{N}}, \quad \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k.$$

Note, that if we use$\|\|_2$-norm and $d\left(x\right) = \frac{1}{2}\left\|x - x^0\right\|_2^2$ here, we will have more complicated iterations (2-norm projections on unit simplex) and

$$f\left(\bar{x}^N\right) - f_* \le \frac{M_2}{\sqrt{N}}, \quad \left\|\nabla f\left(x\right)\right\|_2 \le M_2, \quad x \in Q.$$

Since typically$M_2 = \mathrm{O}\left(\sqrt{n}M_\infty\right)$, it is worth to use $\|\|_1$-norm.

Assume now that $f(x)$ in (13) is additionally μ-strongly convex in $\|\|_2$ norm:

$$\frac{\mu}{2}\left\|x - y\right\|_2^2 \le f(x) \text{ for all } x, y \in Q.$$

Let

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_k\nabla f\left(x^k\right)\right) = \arg\min_{x\in Q}\left\{h_k\left\langle\nabla f\left(x^k\right), x - x^k\right\rangle + \frac{1}{2}\left\|x - x^k\right\|_2^2\right\},$$

where

$$h_k = \frac{2}{\mu \cdot (k+1)}, \quad d(x) = \frac{1}{2}\left\|x - x^0\right\|_2^2, \quad \|\nabla f(x)\|_2 \le M, \quad x \in Q.$$

Then [67]

$$f\left(\sum_{k=1}^{N} \frac{2k}{k(k+1)} x^k\right) - f_* \le \frac{2M^2}{\mu \cdot (k+1)}.$$

Hence (see also Table 1),

$$N \simeq \frac{2M^2}{\mu\varepsilon}.$$

This bound is also unimprovable up to a constant factor [51, 61].

5 Looking into the Black-Box

In this section we consider how special structure of some non-smooth problems can be used to solve non-smooth optimization problems with the convergence rate $O\left(\frac{1}{k}\right)$, which is faster than the lover bound $O\left(\frac{1}{\sqrt{k}}\right)$ for general class of non-smooth convex problems [51]. Nevertheless, there is no contradiction as additional structure is used and we are looking inside the black-box.

5.1 Nesterov's smoothing

In this subsection, following [54], we consider the problem

$$\min_{\mathbf{x} \in Q_1 \subset E_1} \{f(\mathbf{x}) = h(\mathbf{x}) + \max_{\mathbf{u} \in Q_2 \subset E_2} \{\langle A\mathbf{x}, \mathbf{u}\rangle - \phi(\mathbf{u})\}\}, \tag{20}$$

where $A : E_1 \to E_2^*$ is a linear operator, $\phi(\mathbf{u})$ is a continuous convex function on Q_2, Q_1, Q_2 are convex compacts, h is convex function with L_h-Lipschitz-continuous gradient.

Let us consider an example of $f(x) = \|A\mathbf{x} - \mathbf{b}\|_\infty$ with $A \in \mathbb{R}^{m\times n}$. Then,

$$f(x) = \max_{\mathbf{u} \in \mathbb{R}^m} \{\langle \mathbf{u}, A\mathbf{x} - \mathbf{b}\rangle : \|\mathbf{u}\|_1 \le 1\},$$

$h = 0$, $E_2 = \mathbb{R}^m$, $\phi(\mathbf{u}) = \langle \mathbf{u}, \mathbf{b}\rangle$ and Q_2 is the ball in 1-norm.

The main idea of Nesterov is to add regularization inside the definition of f in (20). More precisely, a prox-function $d_2(\mathbf{u})$ (see definition in Section 4) is introduced for the set Q_2 and a smoothed counterpart $f_\mu(\mathbf{x})$ for f is defined as

$$f_\mu(\mathbf{x}) = h(\mathbf{x}) + \max_{\mathbf{u} \in Q_2} \{\langle A\mathbf{x}, \mathbf{u}\rangle - \phi(\mathbf{u}) - \mu d_2(\mathbf{u})\}$$

and $\mathbf{u}_\mu(\mathbf{x})$ is the optimal solution of this maximization problem.

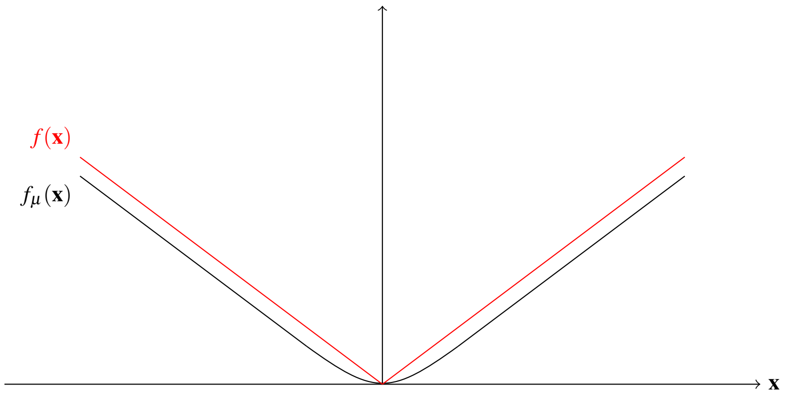

Fig. 2 Function $f_\mu(\mathbf{x})$ is a smooth approximation to non-smooth function $f(\mathbf{x})$.

Theorem 2 ([54]). *The function $f_\mu(\mathbf{x})$ is well defined, convex and continuously differentiable at any $\mathbf{x} \in E_1$ with $\nabla f_\mu(\mathbf{x}) = \nabla h(\mathbf{x}) + A^* \mathbf{u}_\mu(\mathbf{x})$. Moreover, $\nabla f_\mu(\mathbf{x})$ is Lipschitz continuous with constant $L_\mu = L_h + \frac{\|A\|_{1,2}^2}{\mu}$.*

Here the adjoint operator A^* is defined by equality $\langle A\mathbf{x}, \mathbf{u}\rangle = \langle A^*\mathbf{u}, \mathbf{x}\rangle$, $\mathbf{x} \in E_1, \mathbf{u} \in E_2$ and the norm of the operator $\|A\|_{1,2}$ is defined by $\|A\|_{1,2} = \max_{\mathbf{x},\mathbf{u}}\{\langle A\mathbf{x}, \mathbf{u}\rangle : \|\mathbf{x}\|_{E_1} = 1, \|\mathbf{u}\|_{E_2} = 1\}$.

Since Q_2 is bounded, $f_\mu(\mathbf{x})$ is a uniform approximation for the function f, namely, for all $\mathbf{x} \in Q_1$,

$$f_\mu(\mathbf{x}) \leq f(\mathbf{x}) \leq f_\mu(\mathbf{x}) + \mu D_2, \tag{21}$$

where $D_2 = \max\{d_2(\mathbf{u}) : \mathbf{u} \in Q_2\}$.

Then, the idea is to choose μ sufficiently small and apply accelerated gradient method to minimize $f_\mu(\mathbf{x})$ on Q_1. We use accelerated gradient method from [34, 33] which is different from the original method of [54].

Theorem 3 ([34, 33]). *Let the sequences $\{\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k, \alpha_k, C_k\}$, $k \geq 0$ be generated by Algorithm 1. Then, for all $k \geq 0$, it holds that*

$$f(\mathbf{y}^k) - f^* \leq \frac{4LV[\mathbf{z}_0](\mathbf{x}^\star)}{(k+1)^2}. \tag{26}$$

Following the same steps as in the proof of Theorem 3 in [54], we obtain

Theorem 4. *Let Algorithm 1 be applied to minimize $f_\mu(\mathbf{x})$ on Q_1 with $\mu = \frac{2\|A\|_{1,2}}{N+1}\sqrt{\frac{D_1}{D_2}}$, where $D_1 = \max\{d_1(\mathbf{x}) : \mathbf{x} \in Q_1\}$. Then, after N iterations, we have*

Algorithm 1 Accelerated Gradient Method

```
Input: Objective $f(\mathbf{x})$, feasible set $Q$, Lipschitz constant $L$ of the $\nabla f(\mathbf{x})$, starting point $\mathbf{x}^0 \in Q$, ,
    prox-setup: $d(\mathbf{x})$ – 1-strongly convex w.r.t. $\|\cdot\|_{E_1}$, $V[\mathbf{z}](\mathbf{x}) := d(\mathbf{x}) - d(\mathbf{z}) - \langle \nabla d(\mathbf{z}), \mathbf{x}-\mathbf{z} \rangle$.
Set $k = 0$, $C_0 = \alpha_0 = 0$, $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{x}^0$.
for $k = 0, 1, \dots$ do
    Find $\alpha^{k+1}$ as the largest root of the equation
        $C_{k+1} := C_k + \alpha_{k+1} = L\alpha_{k+1}^2.$    (22)
    $\mathbf{x}^{k+1} = \frac{\alpha_{k+1}\mathbf{z}^k + C_k \mathbf{y}^k}{C_{k+1}}.$    (23)
    $\mathbf{z}^{k+1} = \arg\min_{\mathbf{x} \in Q} \{ V[\mathbf{z}^k](\mathbf{x}) + \alpha_{k+1}(f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x} - \mathbf{x}^{k+1} \rangle) \}.$    (24)
    $\mathbf{y}^{k+1} = \frac{\alpha_{k+1}\mathbf{z}^{k+1} + C_k \mathbf{y}^k}{C_{k+1}}.$    (25)
    Set $k = k + 1$.
end for
Output: The point $\mathbf{y}^{k+1}$.
```

$$0 \le f(\mathbf{y}^N) - f_\star \le \frac{4\|A\|_{1,2}\sqrt{D_1 D_2}}{N+1} + \frac{4L_h D_1}{(N+1)^2}. \tag{27}$$

Proof. Applying Theorem 3 to f_μ, and using (21), we obtain

$$\begin{aligned} 0 \le f(\mathbf{y}^N) - f_\star \le f_\mu(\mathbf{y}^N) + \mu D_2 - f_\mu(\mathbf{x}_\mu^\star) &\le \mu D_2 + \frac{4L_\mu D_1}{(N+1)^2} + \frac{4L_h D_1}{(N+1)^2} \\ &= \mu D_2 + \frac{4\|A\|_{1,2}^2 D_1}{\mu(N+1)^2} + \frac{4L_h D_1}{(N+1)^2}. \end{aligned}$$

Substituting the value of μ from the theorem statement, we finish the proof. ■

A generalization of the smoothing technique for the case of non-compact sets Q_1, Q_2, which is especially interesting when dealing with problems dual to problems with linear constraints, can be found in [72]. Ubiquitous entropic regularization of optimal transport [20] can be seen as a particular case of the application of smoothing technique, especially in the context of Wasserstein barycenters [21, 74, 30].

5.2 *Nemirovski's Mirror Prox*

In his paper [48], Nemirovski considers problem (20) in the following form

$$\min_{\mathbf{x}\in Q_1\subset E_1}\{f(\mathbf{x})=h(\mathbf{x})+\max_{\mathbf{u}\in Q_2\subset E_2}\{\langle A\mathbf{x},\mathbf{u}\rangle+\langle\mathbf{b},\mathbf{u}\rangle\}\}, \tag{28}$$

pointing to the fact that this problem is as general as (20). Indeed, the change of variables $\mathbf{u}\leftarrow(\mathbf{u},t)$ and the feasible set $Q_2\leftarrow\{(\mathbf{u},t):\min_{\mathbf{u}'\in Q_2}\phi(\mathbf{u}')\leq t\leq\phi(\mathbf{u})\}$ allows to make ϕ linear. His idea is to consider problem (28) directly as a convex-concave saddle point problem and associated weak variational inequality (VI).

$$\text{Find}\ \ \mathbf{z}^\star=(\mathbf{x}^\star,\mathbf{u}^\star)\in Q_1\times Q_2\ \ \text{s.t.}\ \ \langle\Phi(\mathbf{z}),\mathbf{z}^\star-\mathbf{z}\rangle\leq 0\ \ \forall\mathbf{z}\in Q_1\times Q_2, \tag{29}$$

where the operator

$$\Phi(\mathbf{z})=\begin{pmatrix}\nabla h(\mathbf{x})+A^*\mathbf{u}\\ -A\mathbf{x}-\mathbf{b}\end{pmatrix} \tag{30}$$

is monotone, i.e. $\langle\Phi(\mathbf{z}_1)-\Phi(\mathbf{z}_2),\mathbf{z}_1-\mathbf{z}_2\rangle\geq 0$, and Lipschitz-continuous, i.e. $\|\Phi(\mathbf{z}_1)-\Phi(\mathbf{z}_2)\|_*\leq L\|\mathbf{z}_1-\mathbf{z}_2\|$. With the appropriate choice of norm on $E_1\times E_2$ and prox-function for $Q_1\times Q_2$, see Section 5 in [48], the Lipschitz constant for Φ can be estimated as $L=2\|A\|_{1,2}\sqrt{D_1D_2}+L_hD_1$.

Algorithm 2 Mirror Prox

```
Input: General VI on a set $Q\subset E$ with operator $\Phi(\mathbf{z})$, Lipschitz constant $L$ of $\Phi(\mathbf{z})$, prox-setup:
    $d(\mathbf{z})$, $V[\mathbf{z}](\mathbf{w})$.
Set $k=0$, $\mathbf{z}^0=\arg\min_{\mathbf{z}\in Q}d(\mathbf{z})$.
for $k=0,1,\ldots$ do
    Calculate
        $\mathbf{w}^k=\arg\min_{\mathbf{z}\in Q}\left\{\langle\Phi(\mathbf{z}^k),\mathbf{z}\rangle+LV[\mathbf{z}^k](\mathbf{z})\right\}$.   (31)
    Calculate
        $\mathbf{z}^{k+1}=\arg\min_{\mathbf{z}\in Q}\left\{\langle\Phi(\mathbf{w}^k),\mathbf{z}\rangle+LV[\mathbf{z}^k](\mathbf{z})\right\}$.   (32)
    Set $k=k+1$.
end for
Output: $\widehat{\mathbf{w}}^k=\frac{1}{k}\sum_{i=0}^{k-1}\mathbf{w}^i$.
```

Theorem 5 ([48]). *Assume that $\Phi(\mathbf{z})$ is monotone and L-Lipschitz-continuous. Then, for any $k\geq 1$ and any $\mathbf{u}\in Q$,*

$$\max_{\mathbf{z}\in Q}\langle\Phi(\mathbf{z}),\widehat{\mathbf{w}}^k-\mathbf{z}\rangle\leq\frac{L}{k}\max_{\mathbf{z}\in Q}V[\mathbf{z}^0](\mathbf{z}). \tag{33}$$

Moreover, if the VI is associated with a convex-concave saddle point problem, i.e.

- $E=E_1\times E_2$,
- $Q=Q_1\times Q_2$ *with convex compact sets* $Q_1\subset E_1$, $Q_2\subset E_2$
- $\Phi(\mathbf{z})=\Phi(\mathbf{x},\mathbf{u})=\begin{pmatrix}\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{u})\\ -\nabla_{\mathbf{u}}f(\mathbf{x},\mathbf{u})\end{pmatrix}$ *for a continuously differentiable function* $f(\mathbf{x},\mathbf{u})$ *which is convex in* $\mathbf{x}\in Q_1$ *and concave in* $\mathbf{u}\in Q_2$,

then

$$[\max_{\mathbf{u}\in Q_2} f(\widehat{\mathbf{x}}^k,\mathbf{u}) - \min_{\mathbf{x}\in Q_1}\max_{\mathbf{u}\in Q_2} f(\mathbf{x},\mathbf{u})] + [\min_{\mathbf{x}\in Q_1}\max_{\mathbf{u}\in Q_2} f(\mathbf{x},\mathbf{u}) - \min_{\mathbf{x}\in Q_1} f(\mathbf{x},\widehat{\mathbf{u}}^k)] \le \frac{L}{k}\max_{\mathbf{z}\in Q} V[\mathbf{z}^0](\mathbf{z}). \tag{34}$$

Proof. Let us fix some iteration $k \ge 0$. By the first-order optimality conditions in (31) and (32), we have, for any $\mathbf{u} \in Q$,

$$\langle \Phi(\mathbf{z}^k) + L\nabla d(\mathbf{w}^k) - L\nabla d(\mathbf{z}^k), \mathbf{u} - \mathbf{w}^k\rangle \ge 0,$$

$$\langle \Phi(\mathbf{w}^k) + L\nabla d(\mathbf{z}^{k+1}) - L\nabla d(\mathbf{z}^k), \mathbf{u} - \mathbf{z}^{k+1}\rangle \ge 0.$$

Whence, for all $\mathbf{u} \in Q$,

$$\begin{gathered}
\langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{u}\rangle = \langle \Phi(\mathbf{w}^k), \mathbf{z}^{k+1} - \mathbf{u}\rangle + \langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle \le \\
\le L\langle \nabla d(\mathbf{z}^k) - \nabla d(\mathbf{z}^{k+1}), \mathbf{z}^{k+1} - \mathbf{u}\rangle + \langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle = \\
= L(d(\mathbf{u}) - d(\mathbf{z}^k) - \langle \nabla d(\mathbf{z}^k), \mathbf{u} - \mathbf{z}^k\rangle) - L(d(\mathbf{u}) - d(\mathbf{z}^{k+1}) - \langle \nabla d(\mathbf{z}^{k+1}), \mathbf{u} - \mathbf{z}^{k+1}\rangle) - \\
- L(d(\mathbf{z}^k) - d(\mathbf{z}^{k+1}) - \langle \nabla d(\mathbf{z}^{k+1}), \mathbf{z}^k - \mathbf{z}^{k+1}\rangle) + \langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle = \\
= LV[\mathbf{z}^k](\mathbf{u}) - LV[\mathbf{z}^{k+1}](\mathbf{u}) - LV[\mathbf{z}^k](\mathbf{z}^{k+1}) + \langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle.
\end{gathered}$$

Further, for all $\mathbf{u} \in Q$,

$$\begin{gathered}
\langle \Phi(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - LV[\mathbf{z}^k](\mathbf{z}^{k+1}) \\
= \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - LV[\mathbf{z}^k](\mathbf{z}^{k+1}) + \langle \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle \le \\
\le \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle + L\langle \nabla d(\mathbf{z}^k) - \nabla d(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - LV[\mathbf{z}^k](\mathbf{z}^{k+1}) = \\
= \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle + L\langle \nabla d(\mathbf{z}^k) - \nabla d(\mathbf{w}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - \\
- L(d(\mathbf{z}^{k+1}) - d(\mathbf{z}^k) - \langle \nabla d(\mathbf{z}^k), \mathbf{z}^{k+1} - \mathbf{z}^k\rangle) = \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - \\
- L(d(\mathbf{w}^k) - d(\mathbf{z}^k) - \langle \nabla d(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^k\rangle) - L(d(\mathbf{z}^{k+1}) - d(\mathbf{w}^k) - \langle \nabla d(\mathbf{w}^k), \mathbf{z}^{k+1} - \mathbf{w}^k\rangle) = \\
= \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - LV[\mathbf{z}^k](\mathbf{w}^k) - LV[\mathbf{w}^k](\mathbf{z}^{k+1}) \le \\
\le \langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1}\rangle - \frac{L}{2}(\|\mathbf{z}^k - \mathbf{w}^k\|^2 + \|\mathbf{z}^{k+1} - \mathbf{w}^k\|^2) \le 0,
\end{gathered}$$

where in the last inequality we used the Lipschitz continuity of Φ.

Further, we obtain, for all $\mathbf{u} \in Q$ and $i \ge 0$,

$$\langle \Phi(\mathbf{w}^i), \mathbf{w}^i - \mathbf{u}\rangle \le LV[\mathbf{z}^i](\mathbf{u}) - LV[\mathbf{z}^{i+1}](\mathbf{u}).$$

Summing up these inequalities for i from 0 to $k-1$, and using the monotonicity of Φ we have:

$$\langle \Phi(\mathbf{u}), \widehat{\mathbf{w}}^k - \mathbf{u}\rangle \le \frac{1}{k}\sum_{i=0}^{k-1}\langle \Phi(\mathbf{w}^i), \mathbf{w}^i - \mathbf{u}\rangle \le \frac{L}{k}(V[\mathbf{z}^0](\mathbf{u}) - V[\mathbf{z}^k](\mathbf{u})).$$

Taking maximum in $\mathbf{u}$, we obtain the first statement of the Theorem. The second statement is straighforward by the definition of Φ for saddle-point problems. ■

Choosing appropriately the norm in the space $E_1 \times E_2$ and applying Mirror Prox algorithm to solve problem (28) as a saddle point problem, we obtain that the saddle point error in the l.h.s. of (34) decays as $\frac{2\|A\|_{1,2}\sqrt{D_1 D_2}+L_h D_1}{k}$. This is slightly worse than the rate in (26) since the accelerated gradient method allows the faster decay for the smooth part $h(\mathbf{x})$. An accelerated Mirror Prox method with the same rate as in (26) can be found in [18].

5.3 Universal Mirror Prox

Now we consider universal analogue of A.S. Nemirovsky's proximal mirror method for variational inequalities with a Holder-continuous operator. Main idea of the this method is the adaptive choice of constants and level of smoothness in minimized prox-mappings at each iteration. These constants are related to the Hölder constant of the operator and this method allows to find a suitable constant at each iteration.

Algorithm 3 Universal Mirror Prox

```
Input: General VI on a set $Q \subset E$ with operator $\Phi(\mathbf{z})$, accuracy $\varepsilon > 0$, initial guess $M_{-1} > 0$,
    prox-setup: $d(\mathbf{z})$, $V[\mathbf{z}](\mathbf{w})$.
Set $k = 0$, $\mathbf{z}^0 = \arg\min_{\mathbf{z}\in Q} d(\mathbf{z})$.
for $k = 0, 1, \dots$ do
    Set $i_k = 0$
    repeat
        Set $M_k = 2^{i_k - 1} M_{k-1}$.
        Calculate
            $\mathbf{w}^k = \arg\min_{\mathbf{z}\in Q} \left\{ \langle \Phi(\mathbf{z}^k), \mathbf{z} \rangle + M_k V[\mathbf{z}^k](\mathbf{z}) \right\}.$   (35)
        Calculate
            $\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}\in Q} \left\{ \langle \Phi(\mathbf{w}^k), \mathbf{z} \rangle + M_k V[\mathbf{z}^k](\mathbf{z}) \right\}.$   (36)
        $i_k = i_k + 1$.
    until
        $\langle \Phi(\mathbf{w}^k) - \Phi(\mathbf{z}^k), \mathbf{w}^k - \mathbf{z}^{k+1} \rangle \leq \frac{M_k}{2} \left( \|\mathbf{w}^k - \mathbf{z}^k\|^2 + \|\mathbf{w}^k - \mathbf{z}^{k+1}\|^2 \right) + \frac{\varepsilon}{2}.$   (37)
    Set $k = k + 1$.
end for
Output: $\widehat{\mathbf{w}}^k = \frac{1}{k} \sum_{i=0}^{k-1} \mathbf{w}^i$.
```

Theorem 6 ([35]). *For any $k \geq 1$ and any $\mathbf{z} \in Q$,*

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \sum_{i=0}^{k-1} M_i^{-1} \langle \Phi(\mathbf{w}^i), \mathbf{w}^i - \mathbf{z} \rangle \leq \frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} (V[\mathbf{z}^0](\mathbf{z}) - V[\mathbf{z}^k](\mathbf{z})) + \frac{\varepsilon}{2}. \quad (38)$$

Note that if $\max_{\mathbf{z} \in Q} V[\mathbf{z}^0](\mathbf{z}) \leq D$, we can construct the following adaptive stopping criterion for our algorithm

$$\frac{D}{\sum_{i=0}^{k-1} M_i^{-1}} \leq \frac{\varepsilon}{2}.$$

Next, we consider the case of Hölder-continuous operator Φ and show that Algorithm 3 is universal. Assume for some $\nu \in [0,1]$ and $L_\nu \geq 0$

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_* \leq L_\nu \|\mathbf{x} - \mathbf{y}\|^\nu, \quad \mathbf{x}, \mathbf{y} \in Q.$$

holds. The following inequality is a generalization of (72) for VI. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in Q$ and $\delta > 0$,

$$\langle \Phi(\mathbf{y}) - \Phi(\mathbf{x}), \mathbf{y} - \mathbf{z} \rangle \leq \|\Phi(\mathbf{y}) - \Phi(\mathbf{x})\|_* \|\mathbf{y} - \mathbf{z}\| \leq L_\nu \|\mathbf{x} - \mathbf{y}\|^\nu \|\mathbf{y} - \mathbf{z}\| \leq$$

$$\leq \frac{1}{2} \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \left(\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{y} - \mathbf{z}\|^2\right) + \frac{\delta}{2},$$

where

$$L(\delta) = \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}. \quad (39)$$

So, we have

$$\langle \Phi(\mathbf{y}) - \Phi(\mathbf{x}), \mathbf{y} - \mathbf{z} \rangle \leq \frac{L(\delta)}{2} \left(\|\mathbf{y} - \mathbf{x}\|^2 + \|\mathbf{y} - \mathbf{z}\|^2\right) + \delta. \quad (40)$$

Let us consider estimates of the necessary number of iterations are obtained to achieve a given quality of the variational inequality solution.

Corollary 1 (Universal Method for VI). *Assume that the operator Φ is Hölder continuous with constant L_ν for some $\nu \in [0,1]$ and $M_{-1} \leq \left(\frac{2}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$. Also assume that the set Q is bounded. Then, for all $k \geq 0$, we have*

$$\max_{\mathbf{z} \in Q} \langle \Phi(\mathbf{z}), \widehat{w}_k - \mathbf{z} \rangle \leq \frac{(2L_\nu)^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{\mathbf{z} \in Q} V[\mathbf{z}^0](\mathbf{z}) + \frac{\varepsilon}{2} \quad (41)$$

As it follows from (40), if $M_k \geq L(\frac{\varepsilon}{2})$, (37) holds. Thus, for all $i = 0, ..., k-1$, we have $M_i \leq 2 \cdot L(\frac{\varepsilon}{2})$ and

$$\frac{1}{\sum_{i=0}^{k-1} M_i^{-1}} \leq \frac{2L(\frac{\varepsilon}{2})}{k} \leq \frac{(2L_\nu)^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}},$$

(41) holds. Here $L(\cdot)$ is defined in (39). □

Let us add some remarks.

Remark 1. Since the algorithm does not use the values of parameters ν and L_ν, we obtain the following iteration complexity bound

$$2 \inf_{\nu\in[0,1]} \left(\frac{2L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{\mathbf{z}\in Q} V[\mathbf{z}_0](\mathbf{z})$$

to achieve

$$\max_{\mathbf{z}\in Q} \langle \Phi(\mathbf{z}), \widehat{\mathbf{w}}_{\mathbf{k}} - \mathbf{z}\rangle \leq \varepsilon.$$

Remark 2. We can apply this method to convex-concave saddle problems of the form

$$f(x,y) \to \min_{x\in Q_1} \max_{y\in Q_2}, \tag{42}$$

where $Q_{1,2}$ are convex compacts in $\mathbb{R}^n$, f is convex in x and concave in y, there is $\nu \in [0,1]$ and constants $L_{11,\nu}, L_{12,\nu}, L_{21,\nu}, L_{22,\nu} < +\infty$:

$$\|\nabla_x f(x+\Delta x, y+\Delta y) - \nabla_x f(x,y)\|_{1,*} \leq L_{11,\nu}\|\Delta x\|_1^\nu + L_{12,\nu}\|\Delta y\|_2^\nu,$$

$$\|\nabla_y f(x+\Delta x, y+\Delta y) - \nabla_y f(x,y)\|_{2,*} \leq L_{21,\nu}\|\Delta x\|_1^\nu + L_{22,\nu}\|\Delta y\|_2^\nu$$

for all $x, x+\Delta x \in Q_1, y, y+\Delta y \in Q_2$.

It is possible to achieve an acceptable approximation $(\widehat{x}, \widehat{y}) \in Q_1 \times Q_2$:

$$\max_{y\in Q_2} f(\widehat{x}, y) - \min_{x\in Q_1} f(x, \widehat{y}) \leq \varepsilon \tag{43}$$

for the saddle point $(x_*, y_*) \in Q_1 \times Q_2$ of the (42) problem in no more than

$$O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2}{1+\nu}}\right)$$

iterations, which indicates the optimality of the proposed method, at least for $\nu = 0$ and $\nu = 1$. However, in practice experiments show that (43) can be achieved much faster due to the adaptability of the method.

6 Non-Smooth Optimization in Large Dimensions

The optimization of non-smooth functionals with constraints attracts widespread interest in large-scale optimization and its applications [8, 62]. Subgradient methods for nonsmooth optimization have a long history starting with the method for deterministic unconstrained problems and Euclidean setting in [70] and the generalization for constrained problems in [63], where the idea of steps switching between the direction of subgradient of the objective and the direction of subgradient of the constraint was suggested. Non-Euclidean extension, usually referred to as Mirror Descent, originated in [49, 51] and was later analyzed in [6]. An extension for constrained problems was proposed in [51], see also recent version in [5]. To prove faster convergence rate of Mirror Descent for strongly convex objective in an unconstrained case, the restart technique [50, 51, 52] was used in [38]. Usually, the stepsize and stopping rule for Mirror Descent requires to know the Lipschitz constant of the objective function and constraint, if any. Adaptive stepsizes, which do not require this information, are considered in [49] for problems without inequality constraints, and in [5] for constrained problems.

Formally speaking, we consider the following convex constrained minimization problem

$$\min\{f(\mathbf{x}): \quad \mathbf{x} \in X \subset E, \quad g(\mathbf{x}) \leq 0\}, \tag{44}$$

where X is a convex closed subset of a finite-dimensional real vector space E, $f : X \to \mathbb{R}$, $g : E \to \mathbb{R}$ are convex functions.

We assume g to be a non-smooth Lipschitz-continuous function and the problem (4) to be regular. The last means that there exists a point $\bar{\mathbf{x}}$ in relative interior of the set X, such that $g(\bar{\mathbf{x}}) < 0$.

Note that, despite problem (44) contains only one inequality constraint, considered algorithms allow to solve more general problems with a number of constraints given as $\{g_i(\mathbf{x}) \leq 0, i = 1, ..., m\}$. The reason is that these constraints can be aggregated and represented as an equivalent constraint given by $\{g(\mathbf{x}) \leq 0\}$, where $g(\mathbf{x}) = \max_{i=1,...,m} g_i(\mathbf{x})$.

We consider some adaptive Mirror Descent methods [4] for the problem (44). Both considered methods have complexity $O\left(\frac{1}{\varepsilon^2}\right)$ and optimal.

We consider algorithms, which are based on Mirror Descent method. Thus, we start with the description of proximal setup and basic properties of Mirror Descent step. Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of a linear function $g \in E^*$ at $\mathbf{x} \in E$ by $\langle g, \mathbf{x} \rangle$. Let $\|\cdot\|_E$ be some norm on E, $\|\cdot\|_{E,*}$ be its dual, defined by $\|g\|_{E,*} = \max\limits_{\mathbf{x}} \big\{ \langle g, \mathbf{x} \rangle, \|\mathbf{x}\|_E \leq 1 \big\}$. We use $\nabla f(\mathbf{x})$ to denote any subgradient of a function f at a point $\mathbf{x} \in \mathrm{dom} f$.

Given a vector $\mathbf{x} \in X^0$, and a vector $p \in E^*$, the Mirror Descent step is defined as

$$\mathbf{x}^+ = \mathrm{Mirr}[\mathbf{x}](p) := \arg\min_{\mathbf{z} \in X} \big\{ \langle p, \mathbf{z} \rangle + V[\mathbf{x}](\mathbf{z}) \big\} = \arg\min_{\mathbf{z} \in X} \big\{ \langle p, \mathbf{z} \rangle + d(\mathbf{z}) - \langle \nabla d(\mathbf{x}), \mathbf{z} \rangle \big\}. \tag{45}$$

We make the simplicity assumption, which means that $\mathrm{Mirr}[\mathbf{x}](p)$ is easily computable.

The following lemma [9] describes the main property of the Mirror Descent step.

Lemma 1. *Let f be some convex function over a set X, $h > 0$ be a stepsize, $\mathbf{x} \in X^0$. Let the point $\mathbf{x}^+$ be defined by $\mathbf{x}^+ = \mathrm{Mirr}[\mathbf{x}](h \cdot (\nabla f(\mathbf{x})))$. Then, for any $\mathbf{z} \in X$,*

$$h \cdot \big(f(\mathbf{x}) - f(\mathbf{z})\big) \le h \cdot \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle \\ \le \frac{h^2}{2} \|\nabla f(\mathbf{x}) + V[\mathbf{x}](\mathbf{z}) - V[\mathbf{x}^+](\mathbf{z}). \quad (46)$$

The following analogue of Lemma (1) for δ-subgradients $\nabla_\delta f$ holds.

Lemma 2. *Let f be some convex function over a set X, $h > 0$ be a stepsize, $\mathbf{x} \in X^0$. Let the point $\mathbf{x}^+$ be defined by $\mathbf{x}^+ = \mathrm{Mirr}[\mathbf{x}](h \cdot (\nabla_\delta f(\mathbf{x})))$. Then, for any $\mathbf{z} \in X$,*

$$h \cdot \big(f(\mathbf{x}) - f(\mathbf{z})\big) \le h \cdot \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle + h \cdot \delta \\ \le \frac{h^2}{2} \|\nabla_\delta f(\mathbf{x})\| + h \cdot \delta + V[\mathbf{x}](\mathbf{z}) - V[\mathbf{x}^+](\mathbf{z}). \quad (47)$$

We consider problem (44) in two different settings, namely, non-smooth Lipschitz-continuous objective function f and general objective function f, which is not necessarily Lipschitz-continuous, e.g. a quadratic function. In both cases, we assume that g is non-smooth and is Lipschitz-continuous

$$|g(\mathbf{x}) - g(\mathbf{y})| \le M_g \|\mathbf{x} - \mathbf{y}\|_E, \quad \mathbf{x}, \mathbf{y} \in X. \quad (48)$$

Let $\mathbf{x}_*$ be a solution to (44). We say that a point $\tilde{\mathbf{x}} \in X$ is an ε-*solution* to (44) if

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}_*) \le \varepsilon, \quad g(\tilde{\mathbf{x}}) \le \varepsilon. \quad (49)$$

All considered in this item methods are applicable in the case of using δ-subgradients instead of usual subgradients. For this case we can get an ε-solution $\tilde{\mathbf{x}} \in X$:

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}_*) \le \varepsilon + O(\delta), \quad g(\tilde{\mathbf{x}}) \le \varepsilon + O(\delta). \quad (50)$$

The methods we describe are based on the of Polyak's switching subgradient method [63] for constrained convex problems, also analyzed in [56], and Mirror Descent method originated in [51]; see also [49].

6.1 Convex Non-Smooth Objective Function

In this subsection, we assume that f is a non-smooth Lipschitz-continuous function

$$|f(\mathbf{x}) - f(\mathbf{y})| \le M_f \|\mathbf{x} - \mathbf{y}\|_E, \quad \mathbf{x}, \mathbf{y} \in X. \quad (51)$$

Let $\mathbf{x}_*$ be a solution to (44) and assume that we know a constant $\Theta_0 > 0$ such that

$$d(\mathbf{x}_*) \le \Theta_0^2. \quad (52)$$

For example, if X is a compact set, one can choose $\Theta_0^2 = \max_{\mathbf{x}\in X} d(\mathbf{x})$.

Algorithm 4 Adaptive Mirror Descent (Non-Smooth Objective)

```
Input: accuracy $\varepsilon > 0$; $\Theta_0$ s.t. $d(\mathbf{x}_*) \le \Theta_0^2$.
$\mathbf{x}^0 = \arg\min_{\mathbf{x}\in X} d(\mathbf{x})$.
Initialize the set $I$ as empty set.
Set $k = 0$.
repeat
    if $g(\mathbf{x}^k) \le \varepsilon$ then
        $M_k = \|\nabla f(\mathbf{x}^k)\|_{E,*}$,
        $h_k = \frac{\varepsilon}{M_k^2}$
        $\mathbf{x}^{k+1} = \mathrm{Mirr}[\mathbf{x}^k](h_k \nabla f(\mathbf{x}^k))$ (”productive step”)
        Add $k$ to $I$.
    else
        $M_k = \|\nabla g(\mathbf{x}^k)\|_{E,*}$
        $h_k = \frac{\varepsilon}{M_k^2}$
        $\mathbf{x}^{k+1} = \mathrm{Mirr}[\mathbf{x}^k](h_k \nabla g(\mathbf{x}^k))$ (”non-productive step”)
    end if
    Set $k = k + 1$.
until $\sum\limits_{j=0}^{k-1} \frac{1}{M_j^2} \ge \frac{2\Theta_0^2}{\varepsilon^2}$
Output: $\bar{\mathbf{x}}^k := \frac{\sum_{i\in I} h_i \mathbf{x}^i}{\sum_{i\in I} h_i}$
```

Theorem 7. *Assume that inequalities* (48) *and* (51) *hold and a known constant* $\Theta_0 > 0$ *is such that* $d(\mathbf{x}_*) \le \Theta_0^2$*. Then, Algorithm 4 stops after not more than*

$$k = \left\lceil \frac{2\max\{M_f^2, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{53}$$

iterations and $\bar{\mathbf{x}}^k$ *is an* ε*-solution to* (44) *in the sense of* (49).

Let us now show that Algorithm 4 allows to reconstruct an approximate solution to the problem, which is dual to (44). We consider a special type of problem (44) with g given by

$$g(\mathbf{x}) = \max_{i\in\{1,\ldots,m\}} \big\{ g_i(\mathbf{x}) \big\}. \tag{54}$$

Then, the dual problem to (44) is

$$\varphi(\lambda) = \min_{\mathbf{x}\in X} \left\{ f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \right\} \to \max_{\lambda_i \ge 0, i=1,\ldots,m} \varphi(\lambda), \tag{55}$$

where $\lambda_i \ge 0, i = 1, \ldots, m$ are Lagrange multipliers.

We slightly modify the assumption (52) and assume that the set X is bounded and that we know a constant $\Theta_0 > 0$ such that

$$\max_{\mathbf{x}\in X} d(\mathbf{x}) \le \Theta_0^2.$$

As before, denote $[k] = \{j \in \{0,...,k-1\}\}$, $J = [k]\setminus I$. Let $j \in J$. Then a subgradient of $g(\mathbf{x})$ is used to make the j-th step of Algorithm 4. To find this subgradient, it is natural to find an active constraint $i \in 1,...,m$ such that $g(\mathbf{x}^j) = g_i(\mathbf{x}^j)$ and use $\nabla g(\mathbf{x}^j) = \nabla g_i(\mathbf{x}^j)$ to make a step. Denote $i(j) \in 1,...,m$ the number of active constraint, whose subgradient is used to make a non-productive step at iteration $j \in J$. In other words, $g(\mathbf{x}^j) = g_{i(j)}(\mathbf{x}^j)$ and $\nabla g(\mathbf{x}^j) = \nabla g_{i(j)}(\mathbf{x}^j)$. We define an approximate dual solution on a step $k \ge 0$ as

$$\bar{\lambda}_i^k = \frac{1}{\sum\limits_{j\in I} h_j} \sum_{j\in J, i(j)=i} h_j, \quad i \in \{1,...,m\}. \tag{56}$$

and modify Algorithm 4 to return a pair $(\bar{\mathbf{x}}^k, \bar{\lambda}^k)$.

Theorem 8. *Assume that the set X is bounded, the inequalities* (48) *and* (51) *hold and a known constant $\Theta_0 > 0$ is such that $d(\mathbf{x}_*) \le \Theta_0^2$. Then, modified Algorithm 4 stops after not more than*

$$k = \left\lceil \frac{2\max\{M_f^2, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil$$

iterations and the pair $(\bar{\mathbf{x}}^k, \bar{\lambda}^k)$ returned by this algorithm satisfies

$$f(\bar{\mathbf{x}}^k) - \varphi(\bar{\lambda}^k) \le \varepsilon, \quad g(\bar{\mathbf{x}}^k) \le \varepsilon. \tag{57}$$

6.2 Truss topology design problem: primal-duality and sparsity

Now we consider interesting example of huge-scale problem [58, 62] with a sparse structure. We would like to illustrate two important ideas. Firstly, consideration of the dual problem can simplify the solution, if it is possible to reconstruct the solution of the primal problem by solving the dual problem. Secondly, for a special sparse non-smooth piece-wise linear functions we suggest a very efficient implementation of one subgradient iteration [58]. In such cases simple subgradient methods (for example, Algorithm 4) can be useful due to the relatively inexpensive cost of iterations.

Recall (see e.g. [62]) that Truss Topology Design problem consists in finding the best mechanical structure resisting to an external force with an upper bound for the total weight of construction. Its mathematical formulation looks as follows:

$$\min_{w\in R_+^m} \{\langle \overline{f}, \mathbf{z}\rangle : A(w)\mathbf{z} = \overline{f},\, \langle e, w\rangle = T\}, \tag{58}$$

where $\overline{f}$ is a vector of external forces, $\mathbf{z} \in R^{2n}$ is a vector of virtual displacements of n nodes in R^2, w is a vector of m bars, and T is the total weight of construction. The compliance matrix $A(w)$ has the following form:

$$A(w) = \sum_{i=1}^{m} w_i a_i a_i^T,$$

where $a_i \in R^{2n}$ are the vectors describing the interactions of two nodes connected by an arc. These vectors are very sparse: for 2D-model they have at most 4 nonzero elements.

Let us rewrite the problem (58) as a Linear Programming problem.

$$\begin{aligned}
&\min_{\mathbf{z},w}\{\langle \overline{f},\mathbf{z}\rangle : A(w)\mathbf{z} = \overline{f},\, w \geq 0,\, \langle e,w\rangle = T\} = \\
&= \min_{w}\{\langle \overline{f}, A^{-1}(w)\overline{f}\rangle : w \in \triangle(T) = \{w \geq 0,\, \langle e,w\rangle = T\}\} = \\
&= \min_{w\in\triangle(T)} \max_{\mathbf{z}}\{2\langle \overline{f},\mathbf{z}\rangle - \langle A(w)\mathbf{z},\mathbf{z}\rangle\} \geq \max_{\mathbf{z}} \min_{w\in\triangle(T)}\{2\langle \overline{f},\mathbf{z}\rangle - \langle A(w)\mathbf{z},\mathbf{z}\rangle\} = \\
&= \max_{\mathbf{z}}\{2\langle \overline{f},\mathbf{z}\rangle - T\max_{1\leq i\leq m}\langle a_i,\mathbf{z}\rangle^2\} = \max_{\lambda,\mathbf{y}}\{2\lambda\langle \overline{f},\mathbf{y}\rangle - \lambda^2 T\max_{1\leq i\leq m}\langle a_i,\mathbf{y}\rangle^2\} = \\
&= \max_{\mathbf{y}} \frac{\langle \overline{f},\mathbf{y}\rangle^2}{T\max\limits_{1\leq i\leq m}\langle a_i,\mathbf{y}\rangle^2} = \frac{1}{T}\left(\max_{\mathbf{y}}\{\langle \overline{f},\mathbf{y}\rangle : \max_{1\leq i\leq m}|\langle a_i,\mathbf{y}\rangle| \leq 1\}\right)^2.
\end{aligned} \tag{59}$$

Note that for the inequality in the third line we do not need any assumption.

Denote by $\mathbf{y}^*$ the optimal solution of the optimization problem in the brackets. Then there exist multipliers $\mathbf{x}^* \in R^m_+$ such that

$$\overline{f} = \sum_{i\in J_+} a_i\mathbf{x}_i^* - \sum_{i\in J_-} a_i\mathbf{x}_i^*, \qquad \mathbf{x}_i^* = 0,\, i \notin J_+ \bigcap J_-, \tag{60}$$

where $J_+ = \{i : \langle a_i,\mathbf{y}^*\rangle = 1\}$, and $J_- = \{i : \langle a_i,\mathbf{y}^*\rangle = -1\}$. Multiplying the first equation in (60) by $\mathbf{y}^*$, we get

$$\langle \overline{f},\mathbf{y}^*\rangle = \langle e,\mathbf{x}^*\rangle. \tag{61}$$

Note that the first equation in (60) can be written as

$$\overline{f} = A(\mathbf{x}^*)\mathbf{y}^*. \tag{62}$$

Let us reconstruct now the solution of the primal problem. Denote

$$w^* = \frac{T}{\langle e,\mathbf{x}^*\rangle}\cdot\mathbf{x}^*, \qquad \mathbf{z}^* = \frac{\langle e,\mathbf{x}^*\rangle}{T}\cdot\mathbf{y}^*. \tag{63}$$

Then, in view of (62) we have $\overline{f} = A(w^*)\mathbf{z}^*$, and $w^* \in \triangle(T)$. Thus, the pair (63) is feasible for the primal problem. On the other hand,

$$\langle \bar{f}, \mathbf{z}^* \rangle = \langle \bar{f}, \frac{\langle e, \mathbf{x}^* \rangle}{T} \cdot \mathbf{y}^* \rangle = \frac{1}{T} \cdot \langle e, \mathbf{x}^* \rangle \cdot \langle \bar{f}, \mathbf{y}^* \rangle = \frac{1}{T} \cdot \langle \bar{f}, \mathbf{y}^* \rangle^2.$$

Thus, the duality gap in the chain (59) is zero, and the pair $(w^*, \mathbf{z}^*)$, defined by (63) is the optimal solution of the primal problem.

The above discussion allows us to concentrate on the following (dual) Linear Programming problem:

$$\max_{\mathbf{y}} \{ \langle \bar{f}, \mathbf{y} \rangle : \max_{1 \le i \le m} \langle \mathscr{P}m\mathbf{a}^i, \mathbf{y} \rangle \le 1 \}, \tag{64}$$

which we can solve by the primal-dual Algorithm 4.

Assume that we have *local* truss: each node is connected only with few neighbors. It allows to apply the property of *sparsity* for vectors $\mathbf{a}^i$ $(1 \le i \le m)$. In this case the computational cost of each iteration grows as $O(\log_2 m)$ [58, 62].

In [58] a special class of huge-scale problems with sparse subgradients was considered. According to [58] for smooth functions this is a very rare feature. For example, for quadratic function $f(\mathbf{y}) = \frac{1}{2}\langle A\mathbf{y}, \mathbf{y} \rangle$ the gradient $\nabla f(\mathbf{y}) = A\mathbf{y}$ usually is dense even for a sparse matrix A.

However, the subgradients of non-smooth function $f(\mathbf{y}) = \max_{1 \le i \le m} \langle \mathbf{a}^i, \mathbf{y} \rangle$ (see (64) above) are sparse provided that all vectors $\mathbf{a}_i$ share this property. This fact is based on the following observation. For the function $f(\mathbf{y}) = \max_{1 \le i \le m} \langle \mathbf{a}_i, \mathbf{y} \rangle$ with sparse matrix $A = (\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^m)$ the vector $\nabla f(\mathbf{y}) = \mathbf{a}^{i(\mathbf{y})}$ is a subgradient at point $\mathbf{y}$. Then the standard subgradient step

$$\mathbf{y}_+ = \mathbf{y} - h \cdot \nabla f(\mathbf{y})$$

changes only a few entries of vector $\mathbf{y}$ and the vector $\mathbf{z}_+ = A^T \mathbf{y}_+$ differs from $\mathbf{z} = A^T \mathbf{y}$ also in a few positions only. Thus, the function value $f(\mathbf{y}_+)$ can be easily updated provided that we have an efficient procedure for recomputing the maximum of m values.

Note the objective functional in (64) is linear and the costs of iteration of Algorithm 4 and considered in [58] switching simple subgradient scheme is comparable. At the same time, the step productivity condition is simpler for Algorithm 4 as considered in [58] switching subgradient scheme. Therefore main observations for [58] are correct for Algorithm 4.

6.3 General Convex and Quasi-Convex Objective Functions

In this subsection, we assume that the objective function f in (44) might not satisfy (51) and, hence, its subgradients could be unbounded. One of the examples is a quadratic function. We also assume that inequality (52) holds.

We further consider ideas in [56, 60] and adapt them for problem (44), in a way that our algorithm allows to use non-Euclidean proximal setup, as does Mirror Descent, and does not require to know the constant M_g. Following [56], given a function f for each subgradient $\nabla f(\mathbf{x})$ at a point $\mathbf{y} \in X$, we define

$$v_f[\mathbf{y}](\mathbf{x}) = \begin{cases} \left\langle \dfrac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_{E,*}}, \mathbf{x} - \mathbf{y} \right\rangle, & \nabla f(\mathbf{x}) \neq 0 \\ 0 & \nabla f(\mathbf{x}) = 0 \end{cases}, \quad \mathbf{x} \in X. \tag{65}$$

Algorithm 5 Adaptive Mirror Descent (General Convex Objective)

```
Input: accuracy $\varepsilon > 0$; $\Theta_0$ s.t. $d(\mathbf{x}_*) \le \Theta_0^2$.
$\mathbf{x}^0 = \arg\min_{\mathbf{x} \in X} d(\mathbf{x})$.
Initialize the set $I$ as empty set.
Set $k = 0$.
repeat
    if $g(\mathbf{x}^k) \le \varepsilon$ then
        $h_k = \frac{\varepsilon}{\|\nabla f(\mathbf{x}^k)\|_{E,*}}$
        $\mathbf{x}^{k+1} = \mathrm{Mirr}[\mathbf{x}^k](h_k \nabla f(\mathbf{x}^k))$ (”productive step”)
        Add $k$ to $I$.
    else
        $h_k = \frac{\varepsilon}{\|\nabla g(\mathbf{x}^k)\|_{E,*}^2}$
        $\mathbf{x}^{k+1} = \mathrm{Mirr}[\mathbf{x}^k](h_k \nabla g(\mathbf{x}^k))$ (”non-productive step”)
    end if
    Set $k = k + 1$.
until $|I| + \sum_{j \in J} \frac{1}{\|\nabla g(\mathbf{x}^j)\|_{E,*}^2} \ge \frac{2\Theta_0^2}{\varepsilon^2}$
Output: $\bar{\mathbf{x}}^k := \arg\min_{\mathbf{x}^j, j \in I} f(\mathbf{x}^j)$
```

The following result gives complexity estimate for Algorithm 5 in terms of $v_f[\mathbf{x}_*](\mathbf{x})$. Below we use this theorem to establish complexity result for smooth objective f.

Theorem 9. *Assume that inequality* (48) *holds and a known constant* $\Theta_0 > 0$ *is such that* $d(\mathbf{x}_*) \le \Theta_0^2$. *Then, Algorithm 5 stops after not more than*

$$k = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{66}$$

iterations and it holds that $\min_{i \in I} v_f[\mathbf{x}_*](\mathbf{x}^i) \le \varepsilon$ *and* $g(\bar{\mathbf{x}}^k) \le \varepsilon$.

To obtain the complexity of our algorithm in terms of the values of the objective function f, we define non-decreasing function

$$\omega(\tau) = \begin{cases} \max\limits_{\mathbf{x} \in X} \{f(\mathbf{x}) - f(\mathbf{x}_*) : \|\mathbf{x} - \mathbf{x}_*\|_E \le \tau\} & \tau \ge 0, \\ 0 & \tau < 0. \end{cases} \tag{67}$$

and use the following lemma from [56].

Lemma 3. *Assume that f is a convex function. Then, for any $\mathbf{x} \in X$,*

$$f(\mathbf{x}) - f(\mathbf{x}_*) \leqslant \omega(\nu_f[\mathbf{x}_*](\mathbf{x})). \tag{68}$$

Corollary 2. *Assume that the objective function f in* (44) *is given as $f(\mathbf{x}) = \max_{i \in \{1,\ldots,m\}} f_i(\mathbf{x})$, where $f_i(\mathbf{x})$, $i = 1,\ldots,m$ are differentiable with Lipschitz-continuous gradient*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_{E,*} \leq L_i \|\mathbf{x} - \mathbf{y}\|_E \quad \forall \mathbf{x}, \mathbf{y} \in X, \quad i \in \{1,\ldots,m\}. \tag{69}$$

Then $\bar{\mathbf{x}}^k$ is $\widetilde{\varepsilon}$-solution to (44) *in the sense of* (49)*, where*

$$\widetilde{\varepsilon} = \max\{\varepsilon, \varepsilon \max_{i=1,\ldots,m} \|\nabla f_i(\mathbf{x}_*)\|_{E,*} + \varepsilon^2 \max_{i=1,\ldots,m} L_i/2\}.$$

Remark 3. According to [53, 61] main lemma 3 holds for quasi-convex objective functions [13] too:

$$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \max\{f(\mathbf{x}), f(\mathbf{y})\} \text{ for all } \mathbf{x}, \mathbf{y}, \alpha \in [0,1].$$

This means that results of this subsection are valid for quasi-convex objectives.

Remark 4. In view of the Lipschitzness and, generally speaking, non-smoothness of functional limitations, the obtained estimate for the number of iterations means that the proposed method is optimal from the point of view of oracle evaluations: $O\left(\frac{1}{\varepsilon^2}\right)$ iterations are sufficient for achieving the required accuracy ε of solving the problem for the class of target functionals considered in this section of the article. Note also that the considered algorithm 4 applies to the considered classes of problems with constraints with convex objective functionals of different smoothness levels. However, the non-fulfillment, generally speaking, of the Lipschitz condition for the objective functional f does not allow one to substantiate the optimality of the algorithms 4 in the general situation (for example, with a Lipschitz-continuous gradient). More precisely, situations are possible when the productive steps of the norm (sub)gradients of the objective functional $\|\nabla f(\mathbf{x}^k)\|_*$ are large enough and this will interfere with the speedy achievement of the stopping criterion of the 4.

7 Universal Methods

In this section we consider problem

$$\min_{\mathbf{x} \in Q \subseteq E} f(\mathbf{x}), \tag{70}$$

where Q is a convex set and f is a convex function with Hölder-continuous subgradient

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_* \leq L_\nu \|\mathbf{x}_1 - \mathbf{x}_2\|^\nu \tag{71}$$

with $\nu \in [0,1]$. The case $\nu = 0$ corresponds to non-smooth optimization and the case $\nu = 1$ corresponds to smooth optimization. The goal of this section is to present the Universal Accelerated Gradient method first proposed by Nesterov [59]. This method is a black-box method which does not require the knowledge of constants ν, L_ν and works in accordance with the lower complexity bound $O\left(\left(\frac{L_\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+3\nu}}\right)$ obtained in [51].

The main idea is based on the observation that a non-smooth convex function can be upper bounded by a quadratic objective function slightly shifted above. More precisely, for any $\mathbf{x}, \mathbf{y} \in Q$,

$$\begin{aligned} f(\mathbf{y}) &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L_\nu}{1+\nu} \|\mathbf{y} - \mathbf{x}\|^{1+\nu} \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L(\delta)}{2} \|\mathbf{y} - \mathbf{x}\|^2 + \delta, \end{aligned} \tag{72}$$

where

$$L(\delta) = \left(\frac{1-\nu}{1+\nu}\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}.$$

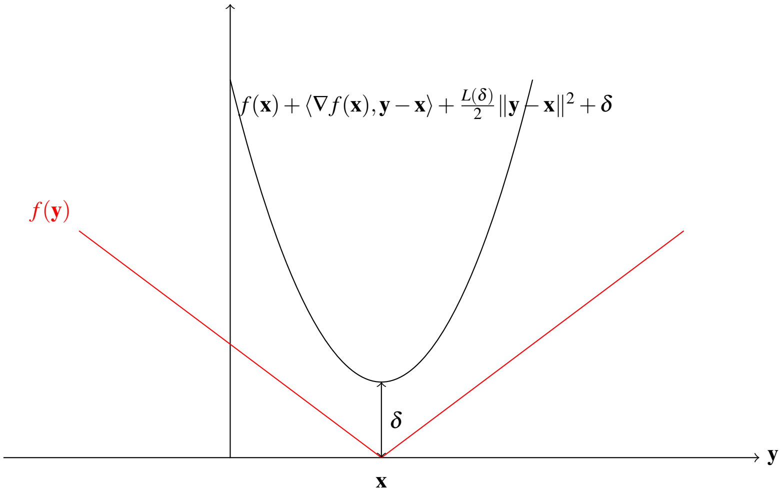
$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L(\delta)}{2} \|\mathbf{y} - \mathbf{x}\|^2 + \delta$
$f(\mathbf{y})$
δ
$\mathbf{y}$
$\mathbf{x}$

Fig. 3 Quadratic majorant of a non-smooth function $f(\mathbf{x})$.

The next idea is to apply an accelerated gradient method with backtracking procedure to adapt for the unknown $L(\delta)$ with appropriately chosen δ. The method we present is based on accelerated gradient method from [34, 33] and, thus is different from the original method of [59].

Inequality (72) guarantees that the backtracking procedure in the inner cycle is finite.

Algorithm 6 Universal Accelerated Gradient Method

```
Input: Accuracy $\varepsilon$, starting point $\mathbf{x}^0 \in Q$, initial guess $L_0 > 0$, prox-setup: $d(\mathbf{x})$ – 1-strongly convex w.r.t. $\|\cdot\|_E$, $V[\mathbf{z}](\mathbf{x}) := d(\mathbf{x}) - d(\mathbf{z}) - \langle \nabla d(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle$.
Set $k = 0$, $C_0 = \alpha_0 = 0$, $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{x}^0$.
for $k = 0, 1, \ldots$ do
    Set $M_k = L_k/2$.
    repeat
        Set $M_k = 2M_k$, find $\alpha_{k+1}$ as the largest root of the equation
            $$C_{k+1} := C_k + \alpha_{k+1} = M_k \alpha_{k+1}^2. \tag{73}$$
        $$\mathbf{x}^{k+1} = \frac{\alpha_{k+1}\mathbf{z}^k + C_k \mathbf{y}^k}{C_{k+1}}. \tag{74}$$
        $$\mathbf{z}^{k+1} = \arg\min_{\mathbf{x} \in Q} \{ V[\mathbf{z}^k](\mathbf{x}) + \alpha_{k+1}(f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x} - \mathbf{x}^{k+1} \rangle) \}. \tag{75}$$
        $$\mathbf{y}^{k+1} = \frac{\alpha_{k+1}\mathbf{z}^{k+1} + C_k \mathbf{y}^k}{C_{k+1}}. \tag{76}$$
    until
        $$f(\mathbf{y}^{k+1}) \le f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{M_k}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|^2 + \frac{\alpha_{k+1}\varepsilon}{2C_{k+1}}. \tag{77}$$
    Set $L_{k+1} = M_k/2$, $k = k+1$.
end for
Output: The point $\mathbf{y}^{k+1}$.
```

Theorem 10 ([59]). *Let f satisfy* (71). *Then,*

$$f(\mathbf{y}^{k+1}) - f_\star \le \left(\frac{2^{2+4\nu} L_\nu^2}{\varepsilon^{1-\nu} k^{1+3\nu}} \right)^{\frac{1}{1+\nu}} V[\mathbf{x}^0](\mathbf{x}^\star) + \frac{\varepsilon}{2}. \tag{78}$$

Moreover, the number of oracle calls is bounded by

$$4(k+1) + 2\log_2 \left((2V[\mathbf{x}^0](\mathbf{x}^\star))^{\frac{1-\nu}{1+3\nu}} \left(\frac{1}{\varepsilon}\right)^{\frac{3(1-\nu)}{1+3\nu}} L_\nu^{\frac{4}{1+3\nu}} \right).$$

Translating this rate of convergence to the language of complexity, we obtain that to obtain a solution with an accuracy ε the number of iterations is no more than

$$O\left(\inf_{\nu \in [0,1]} \left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+3\nu}} \left(V[\mathbf{x}^0](\mathbf{x}^\star) \right)^{\frac{1+\nu}{1+3\nu}} \right),$$

i.e. is optimal.

In his paper, Nesterov considers a more general composite optimization problem

$$\min_{\mathbf{x}\in Q\subseteq E} f(\mathbf{x}) + h(\mathbf{x}), \tag{79}$$

where h is a simple convex function, and obtains the same complexity guarantees. Universal methods were extended for the case of strongly convex problems by a restart technique in [66], for non-convex optimization in [36] and for the case of non-convex optimization with inexact oracle in [29]. As we can see from (72), universal accelerated gradient method is connected to smooth problems with inexact oracle. The study of accelerated gradient methods with inexact oracle was first proposed in [22] and was very well developed in [24, 31, 11, 29] including stochastic optimization problems and strongly convex problems. A universal method with inexact oracle can be found in [32]. Experiments show [59] that universal method accelerates to $O\left(\frac{1}{k}\right)$ rate for non-smooth problems with a special ”smoothing friendly” (see Section 5) structure. This is especially interesting for traffic flow modelling problems, which possess such structure [3].

8 Concluding remarks

Modern numerical methods for non-smooth convex optimization problems are typically based on the structure of the problem. We start with one of the most powerful example of such type. For geometric median search problem there exists efficient method that significantly outperform described above lower complexity bounds [19]. In Machine Learning we typically meet the problems with hidden affine structure and small effective dimension (SVM) that allow us to use different smoothing techniques [1]. Description of one of these techniques (Nesterov’s smoothing technique) one can find in this survey. The other popular technique is based on averaging of the function around the small ball with the center at the point in consideration [28]. A huge amount of data since applications lead to composite optimization problems with non smooth composite (LASSO). For this class of problems accelerated (fast) gradient methods are typically applied [7], [57], [42]. This approach (composite optimization) have been recently expanded for more general class of problems [73]. In different Image Processing applications one can find a lot of non-smooth problems formulations with saddle-point structure. That is the goal function has Legendre representation. In this case one can apply special versions of accelerated (primal-dual) methods [16], [17], [44]. Universal Mirror Prox method described above demonstrates the alternative approach which can be applied in rather general context. Unfortunately, the most of these tricks have proven to be beyond the scope of this survey. But we include in the survey the description of the Universal Accelerated Gradient Descent algorithm [73] which in the general case can also be applied to a wide variety of problems.

Another important direction in Nonsmooth Convex Optimization is huge-scale optimization for sparse problems [58]. The basic idea that reduce huge dimension to nonsmoothness is as follows:

$$\langle \mathbf{a}_k, \mathbf{x} \rangle - b_k \leq 0, \quad k = 1, \dots m, \quad m \gg 1$$

is equivalent to the single nonsmooth constraint:

$$\max_{k=1,\dots m} \{\langle \mathbf{a}_k, \mathbf{x} \rangle - b_k\} \leq 0.$$

We demonstrated this idea above on Truss Topology Design example.

One should note that we concentrate in this survey only on deterministic convex optimization problems, but the most beautiful things in non smooth optimization is that stochasticity [51], [27], [39], [40] and online context [37] in general doesn't change (up to a logarithmic factor in the strongly convex case) anything in complexity estimates. As an example, of stochastic (randomized) approach one can mentioned the work [2] where one can find reformulation of Google problem as non smooth convex optimization problem. Special randomized Mirror Descent algorithm allows to solve this problem almost independently on the number of vertexes.

Finally, let's note that in the decentralized distributed non smooth (stochastic) convex optimization for the last few years there appear optimal methods [43], [75], [14].

Acknowledgements To all good people and especialy to Russian Foundation for Basic Research grant 18-29-03071.

References

1. Allen-Zhu, Z., Hazan, E.: Optimal black-box reductions between optimization objectives. In: Advances in Neural Information Processing Systems, pp. 1614–1622 (2016)
2. Anikin, A., Gasnikov, A., Gornov, A., Kamzolov, D., Maximov, Y., Nesterov, Y.: Efficient numerical methods to solve sparse linear equations with application to pagerank. arXiv preprint arXiv:1508.07607 (2015)
3. Baimurzina, D., Gasnikov, A., Gasnikova, E., Dvurechensky, P., Ershov, E., Kubentaeva, M., Lagunovskaya, A.: Universal similar triangulars method for searching equilibriums in traffic flow distribution models. arXiv:1701.02473 (2017)
4. Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F., Titov, A.: Mirror descent and convex optimization problems with non-smooth inequality constraints. In: P. Giselsson, A. Rantzer (eds.) Large-Scale and Distributed Optimization, chap. 8, pp. 181–215. Springer International Publishing (2018). DOI 10.1007/978-3-319-97478-1_8. ArXiv:1710.06612
5. Beck, A., Ben-Tal, A., Guttmann-Beck, N., Tetruashvili, L.: The comirror algorithm for solving nonsmooth constrained convex problems. Operations Research Letters **38**(6), 493 – 498 (2010). DOI https://doi.org/10.1016/j.orl.2010.08.005. URL http://www.sciencedirect.com/science/article/pii/S0167637710001094
6. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003). DOI 10.1016/S0167-6377(02)00231-6. URL http://dx.doi.org/10.1016/S0167-6377(02)00231-6
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009). DOI 10.1137/080716542. URL https://doi.org/10.1137/080716542

8. Ben-Tal, A., Nemirovski, A.: Robust truss topology design via semidefnite programming. SIAM J. Optim. **7**(4), 991 – 1016 (1997)
9. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization (Lecture Notes). Personal web-page of A. Nemirovski (2015). URL `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`
10. Blum, L., Cucker, F., Shub, M., Smale, S.: Complexity and real computation. Springer Science & Business Media (2012)
11. Bogolubsky, L., Dvurechensky, P., Gasnikov, A., Gusev, G., Nesterov, Y., Raigorodskii, A.M., Tikhonov, A., Zhukovskii, M.: Learning supervised pagerank with gradient-based and gradient-free optimization methods. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 4914–4922. Curran Associates, Inc. (2016). ArXiv:1603.00717
12. Brent, R.: Algorithms for Minimization Without Derivatives. Dover Books on Mathematics. Dover Publications (1973). URL `https://books.google.de/books?id=6Ay2biHG-GEC`
13. B.T.Polyak: Minimization of nonsmooth functionals. USSR Computational Mathematics and Mathematical Physics **9**(3), 14–29 (1969). URL `https://www.sciencedirect.com/science/article/abs/pii/0041555369900615`
14. Bubeck, K.S.F.B.S., Lee, Y.T., Massoulie, L.: Optimal algorithms for non-smooth distributed optimization in networks
15. Bubeck, S.: Convex optimization: algorithms and complexity. Foundations and Trends in Machine Learning **8**(3–4), 231–357 (2015). URL `https://arxiv.org/pdf/1405.4980.pdf`
16. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision **40**(1), 120–145 (2011)
17. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. SIAM Journal on Optimization **24**(4), 1779–1814 (2014)
18. Chen, Y., Lan, G., Ouyang, Y.: Accelerated schemes for a class of variational inequalities. Mathematical Programming **165**(1), 113–149 (2017)
19. Cohen, M.B., Lee, Y.T., Miller, G., Pachocki, J., Sidford, A.: Geometric median in nearly linear time. In: Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, pp. 9–21. ACM (2016)
20. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 26, pp. 2292–2300. Curran Associates, Inc. (2013)
21. Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. In: E.P. Xing, T. Jebara (eds.) Proceedings of the 31st International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 32, pp. 685–693. PMLR, Bejing, China (2014). URL `http://proceedings.mlr.press/v32/cuturi14.html`
22. d'Aspremont, A.: Smooth optimization with approximate gradient. SIAM J. on Optimization **19**(3), 1171–1183 (2008). DOI 10.1137/060676386. URL `http://dx.doi.org/10.1137/060676386`
23. Demyanov, A., Demyanov, V., Malozemov, V.: Minmaxmin problems revisited. Optimization Methods and Software **17**(5), 783–804 (2002). DOI 10.1080/1055678021000060810. URL `https://doi.org/10.1080/1055678021000060810`
24. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming **146**(1), 37–75 (2014). DOI 10.1007/s10107-013-0677-5. URL `http://dx.doi.org/10.1007/s10107-013-0677-5`
25. D.J., N.: Location of the maximum on unimodal surfaces. Journal of the Association for Computing Machinery **12**, 395–398 (1965)
26. Drori Y., T.M.: An optimal variants of kelley's cutting-plane method. Mathematical Programming **160**(1–2), 321–351 (2016)
27. Duchi, J.: Introductory lectures on stochastic optimization. Park City Mathematics Institute, Graduate Summer School Lectures (2016)

28. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM Journal on Optimization **22**(2), 674–701 (2012)
29. Dvurechensky, P.: Gradient method with inexact oracle for composite non-convex optimization. arXiv:1703.09180 (2017)
30. Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C.A., Nedić, A.: Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In: Proceedings of the 32th Conference on Neural Information Processing Systems, NIPS'18 (2018). (Accepted), arXiv:1802.04367
31. Dvurechensky, P., Gasnikov, A.: Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. Journal of Optimization Theory and Applications **171**(1), 121–145 (2016). DOI 10.1007/s10957-016-0999-6. URL `http://dx.doi.org/10.1007/s10957-016-0999-6`
32. Dvurechensky, P., Gasnikov, A., Kamzolov, D.: Universal intermediate gradient method for convex problems with inexact oracle. arXiv:1712.06036 (2017)
33. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In: J. Dy, A. Krause (eds.) Proceedings of the 35th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 80, pp. 1367–1376 (2018). ArXiv:1802.04367
34. Dvurechensky, P., Gasnikov, A., Omelchenko, S., Tiurin, A.: Adaptive similar triangles method: a stable alternative to sinkhorn's algorithm for regularized optimal transport. arXiv:1706.07622 (2017)
35. Dvurechensky, P., Gasnikov, A., Stonyakin, F., Titov, A.: Generalized Mirror Prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters. arXiv:1806.05140 (2018)
36. Ghadimi, S., Lan, G., Zhang, H.: Generalized uniformly optimal methods for nonlinear programming. arXiv:1508.07384 (2015). URL `https://arxiv.org/abs/1508.07384`
37. Hazan, E., et al.: Introduction to online convex optimization. Foundations and Trends® in Optimization **2**(3-4), 157–325 (2016)
38. Juditsky, A., Nemirovski, A.: First order methods for non-smooth convex large-scale optimization, i: General purpose methods. In: S.W. Suvrit Sra Sebastian Nowozin (ed.) Optimization for Machine Learning, pp. 121–184. Cambridge, MA: MIT Press (2012)
39. Juditsky, A., Nemirovski, A., et al.: First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. Optimization for Machine Learning pp. 121–148 (2011)
40. Juditsky, A., Nemirovski, A., et al.: First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. Optimization for Machine Learning pp. 149–183 (2011)
41. Khachiyan, L.G.: A polynomial algorithm in linear programming. In: Doklady Academii Nauk SSSR, vol. 244, pp. 1093–1096 (1979)
42. Lan, G.: Gradient sliding for composite optimization. Mathematical Programming **159**(1), 201–235 (2016). DOI 10.1007/s10107-015-0955-5. URL `https://doi.org/10.1007/s10107-015-0955-5`
43. Lan, G., Lee, S., Zhou, Y.: Communication-efficient algorithms for decentralized and stochastic optimization. arXiv preprint arXiv:1701.03961 (2017)
44. Lan, G., Ouyang, Y.: Accelerated gradient sliding for structured convex optimization. arXiv preprint arXiv:1609.04905 (2016)
45. Lee, Y.T., Sidford, A., Wong, S.C.w.: A faster cutting plane method and its implications for combinatorial and convex optimization. In: Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on, pp. 1049–1065. IEEE (2015)
46. Levin, A.Y.: On an algorithm for the minimization of convex functions. Soviet Math. Doklady (1965)
47. Nedić, A., Ozdaglar, A.: Approximate primal solutions and rate analysis for dual subgradient methods. SIAM Journal on Optimization **19**(4), 1757–1780 (2009). DOI 10.1137/070708111. URL `https://doi.org/10.1137/070708111`

48. Nemirovski, A.: Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization **15**(1), 229–251 (2004)
49. Nemirovskii, A.: Efficient methods for large-scale convex optimization problems. Ekonomika i Matematicheskie Metody **15** (1979). In Russian
50. Nemirovskii, A., Nesterov, Y.: Optimal methods of smooth convex minimization. USSR Computational Mathematics and Mathematical Physics **25**(2), 21 – 30 (1985). DOI https://doi.org/10.1016/0041-5553(85)90100-4. URL `http://www.sciencedirect.com/science/article/pii/0041555385901004`
51. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. J. Wiley & Sons, New York (1983)
52. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Soviet Mathematics Doklady **27**(2), 372–376 (1983)
53. Nesterov, Y.: Effective methods in nonlinear programming. Moscow (1989)
54. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming **103**(1), 127–152 (2005)
55. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Mathematical Programming **120**(1), 221–259 (2009). DOI 10.1007/s10107-007-0149-x. URL `https://doi.org/10.1007/s10107-007-0149-x`. First appeared in 2005 as CORE discussion paper 2005/67
56. Nesterov, Y.: Introduction to Convex Optimization. Moscow, MCCME (2010)
57. Nesterov, Y.: Gradient methods for minimizing composite functions. Mathematical Programming **140**(1), 125–161 (2013). First appeared in 2007 as CORE discussion paper 2007/76
58. Nesterov, Y.: Subgradient methods for huge-scale optimization problems. Mathematical Programming **146**(1), 275–297 (2014). DOI 10.1007/s10107-013-0686-4. URL `https://doi.org/10.1007/s10107-013-0686-4`. First appeared in 2012.
59. Nesterov, Y.: Universal gradient methods for convex optimization problems. Mathematical Programming **152**(1), 381–404 (2015). DOI 10.1007/s10107-014-0790-0. URL `http://dx.doi.org/10.1007/s10107-014-0790-0`
60. Nesterov, Y.: Subgradient methods for convex functions with nonstandard growth properties (2016). Http://www.mathnet.ru:8080/PresentFiles/16179/growthbm_nesterov.pdf
61. Nesterov, Y.: Lectures on Convex Optimization. Springer International Publishing (2018)
62. Nesterov, Y., Shpirko, S.: Primal-dual subgradient method for huge-scale linear conic problems. SIAM Journal on Optimization **24**(3), 1444–1457 (2014). DOI 10.1137/130929345. URL `https://doi.org/10.1137/130929345`
63. Polyak, B.: A general method of solving extremum problems. Soviet Mathematics Doklady **8**(3), 593–597 (1967)
64. Polyak, B.: Introduction to Optimization. New York, Optimization Software (1987)
65. Rockafellar, R.: Convex Analysis. Priceton University, Princeton (1970)
66. Roulet, V., d'Aspremont, A.: Sharpness, restart and acceleration. arXiv:1702.03828 (2017)
67. S. Lacost-Julien, M.S., Bach, F.: A simpler approach to obtaining $o(1/t)$ convergence rate for the projected stochastic subgradient method. arxiv preprint arxiv:1212.2002 (2012). URL `http://arxiv.org/pdf/1212.2002v2.pdf`
68. Shor, N.: Minimization of Nondifferentiable Functions. Naukova Dumka (1979)
69. Shor, N.: Minimization Methods for Non-Differentiable Functions. Springer-Verlag Berlin Heidelberg (1985)
70. Shor, N.Z.: Generalized gradient descent with application to block programming. Kibernetika **3**(3), 53–55 (1967)
71. Shor, N.Z., Kiwiel, K.C., Ruszczynski, A.: Minimization Methods for Non-Differentiable Functions, *Springer Series in Computational Mathematics*, vol. 3. Springer Berlin Heidelberg (2012)
72. Tran-Dinh, Q., Fercoq, O., Cevher, V.: A smooth primal-dual optimization framework for nonsmooth composite convex minimization. SIAM Journal on Optimization **28**(1),

96–134 (2018). DOI 10.1137/16M1093094. URL https://doi.org/10.1137/16M1093094. ArXiv:1507.06243

73. Tyurin, A., Gasnikov, A.: Fast gradient descent method for convex optimization problems with an oracle that generates a model of a function in a requested point. arXiv preprint arXiv:1711.02747 (2017)
74. Uribe, C.A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Nedić, A.: Distributed computation of Wasserstein barycenters over networks. In: 2018 IEEE 57th Annual Conference on Decision and Control (CDC) (2018). Accepted, arXiv:1803.02933
75. Uribe, C.A., Lee, S., Gasnikov, A., Nedić, A.: Optimal algorithms for distributed optimization. arXiv preprint arXiv:1712.00232 (2017)
76. Vaidya, P.M.: Speeding-up linear programming using fast matrix multiplication. In Foundations of Computer Science, 1989, 30th Annual Symposium on p. 332–337 (1989)