1 Introduction

This document is intent on finding an $\varepsilon > 0$ solution $\hat{x}$ to the following problem

$$f_* := \min_{x \in X} f(x) \tag{1}$$

where $f : X \to \mathbb{R}$ is a convex, continuous, non-differentiable function with $f_* > -\infty$, and X is a a convex, compact set with diameter $D_x := \max_{x,y \in X} ||x - y|| < \infty$ in the sense that $f(\hat{x}) - f_* < \varepsilon$. Here, $||\cdot||$ denotes the Euclidean norm. To motivate the analysis, we provide a brief overview of previous literature.

Traditional methods of convex minimization primarily rely on the differentiability of f, so we will begin there. Consider a slightly modified version of (1) :

$$h_* := \min_{x \in X} h(x). \tag{2}$$

where h is L-smooth, i.e. h is differentiable and has a Lipschitz continuous gradient with Lipschitz constant $L > 0$. Under these settings, Nesterov presented an algorithm called Accelerated Gradient Descent (AGD) that can find an ε solution to (2) in $\mathcal{O}(\sqrt{LD_x^2/\varepsilon})$ iterations. More specifically, for x_k generated by AGD applied to (2), we have

$$h_* - h(x_k) \leq \mathcal{O}\left(\frac{LD_X^2}{k^2}\right) \tag{3}$$

(see [1]). It was later proved that such a complexity class is optimal for solving problems of this form using first order deterministic methods. The algorithm is described as follows:

Algorithm 1 Nesterov's accelerated gradient descent (AGD)

Start: Choose $x_0 \in X$. Set $\bar{x}_0 := x_0$
for $k = 1, \ldots, N$ **do**

$$\begin{aligned}
\underline{x}_k &= (1 - \gamma_k)\bar{x}_{k-1} + \gamma_k x_{k-1}, \\
x_k &= \operatorname*{argmin}_{u \in X} \langle \nabla h(\underline{x}_k), u \rangle + \frac{\eta_k}{2}\|u - x_{k-1}\|^2, \\
\bar{x}_k &= (1 - \gamma_k)\bar{x}_{k-1} + \gamma_k x_k.
\end{aligned}$$

end for
Output $\bar{x}_N$.

However, the proof of AGD greatly hinges on the fact that h is L-smooth. To solve (1), we must find a way around this issue. One such algorithm called the subgradient method exists for solving non-differentiable problems. Unfortunately, it requires $\mathcal{O}(1/\varepsilon^2)$ iterations to achieve an ε-solution and is quite far from the $\mathcal{O}(1/\sqrt{\varepsilon})$ guaranteed by AGD.

A more natural approach to solving (1) is to find a smooth approximation f_μ to f that is L-smooth. If the approximation is good enough, then we can simply apply AGD to f_μ to achieve desirable performance. The objective is now simple: design a good approximation f_μ of f so that minimizing f_μ is as close to solving (1) as possible.

2 Nesterov's Smoothing

There are many ways to find a smooth approximation of f_μ, but we will focus on Nesterov's technique. Assume that f from (1) can be represented by

$$f(x) = \max_{y \in Y} \langle Ax + b, y \rangle - \phi(y) \tag{4}$$

where $\phi : Y \to \mathbb{R}$ and Y is a convex, compact set. Nesterov proposes the following smooth approximation of f

$$f_\mu(x) = \max_{y \in Y} \langle Ax + b, y \rangle - \phi(y) - \mu d(y) \tag{5}$$

where $d(y)$ is called a prox function that satisfies

- $d(y)$ is continuous and 1-strongly convex on Y;
- $\min_{y\in Y} d(y) = 0$

and $\mu \geq 0$. Before we discuss the properties of f_μ in (5), let us first examine the assumption in (4). Recall that if f is convex, lower semi-continuous, and proper, i.e. $f_* > -\infty$, then $(f^*)^* = f$ where f^* denotes the convex conjugate (or Fenchel transform)

$$f^*(y) = \sup_{x\in X} \langle x, y\rangle - f(x). \tag{6}$$

Thus, for f with properties listed in (1), $f(x)$ admits its Fenchel representation

$$f(x) = \max_{y\in \operatorname{dom} f^*} \langle y, x\rangle - f^*(y). \tag{7}$$

We know previously that f^* defined in (6) is convex. Thus, if we could show that dom f^* is convex, compact, then f always has at least one representation of the form in (4). There are numerous assumptions that ensure this condition. For example, if f itself is Lipschitz continuous, then dom f^* is compact (see [2]). That is, the assumptions made previously are not particularly stringent. We proceed to proving properties of the smooth approximation (5).

Lemma 1. For f_μ defined in (5), we have the following

- $f_\mu(x)$ is continuously differentiable;
- $\nabla f_\mu(x) = A^T y(x)$ where $y(x) = \operatorname{argmax}_{y\in Y} \langle Ax + b, y\rangle - \phi(y) - \mu d(y)$;
- $f_\mu(x)$ is L-smooth with constant $L_\mu := \frac{||A||^2}{\mu}$ where $||A|| := \max_x\{||Ax|| : ||x|| \leq 1\}$.

Proof. The proofs of the first two bullets follow directly from results in convex optimization. In particular, the subgradient set $\partial f_\mu(x) = A^T \operatorname{argmax}_{y\in Y} \langle Ax + b, y\rangle - \phi(y) - \mu d(y)$. Since $\mu > 0$ and $d(y)$ is strongly convex, the maximization problem has unique solution and thus the subgradient set is a single set, i.e. $\partial f_\mu(x) = \nabla f_\mu(x)$. These results are not our primary focus in this document, so we refer the reader to [3] for additional details regarding subgradients and differentiability of convex functions. To prove the last claim, let $x_1, x_2 \in X$ and $y_1 = \operatorname{argmax}_{y\in Y} \langle x_1, y\rangle - f^*(y)$ and $y_2 = \operatorname{argmax}_{y\in Y} \langle x_2, y\rangle - f^*(y)$. Assume for brevity and without loss of generality that ϕ and d are differentiable. Then by the optimality conditions of these two optimization problems, we have

$$\begin{aligned}\langle Ax_1 + b - \nabla\phi(y_1) - \mu\nabla d(y_1), y_1 - y_2\rangle \geq 0\\ \langle Ax_2 + b - \nabla\phi(y_2) - \mu\nabla d(y_2), y_2 - y_1\rangle \geq 0.\end{aligned}$$

Adding them together and applying the convexity, strong convexity of ϕ and d respectively, we obtain

$$\begin{aligned}\langle A(x_1 - x_2), y_1 - y_2\rangle &\geq \langle \nabla\phi(y_1) - \nabla\phi(y_2) + \mu(\nabla d(y_1) - \nabla d(y_2)), y_1 - y_2\rangle\\ &\geq \mu\langle \nabla d(y_1) - \nabla d(y_2), y_1 - y_2\rangle\\ &\geq \mu\,||y_1 - y_2||^2.\end{aligned}$$

Applying Cauchy-Schwarz, we continue with

$$||y_1 - y_2|| \leq \frac{||A||}{\mu}\,||x_1 - x_2||.$$

After noting the above and the definition of $\nabla f_\mu(x)$, we conclude the proof. □

Lemma 2. For any $\mu \geq 0$, let $D_Y^2 = \max_{y\in Y} d(y)$. We have

$$f(x) - \mu D_Y^2 \leq f_\mu(x) \leq f(x).$$

Proof. This is an immediate consequence of $0 \leq d(y) \leq D_Y^2$. □

With this approximation $f_\mu(x)$ in hand, we can apply AGD to $f_\mu(x)$ to solve for an $\varepsilon-$solution.

Theorem 1. Apply AGD to minimize $f_\mu(x)$ to obtain an approximate solution x_k. Then the error must satisfy

$$f(x_k) - f_* \leq \mathcal{O}\left(\frac{||A||^2 D_X^2}{\mu k^2} + \mu D_Y^2\right) \tag{8}$$

Proof. Note that the error can be decomposed into an approximation and optimization error as follows

$$f(x_k) - f_* = f(x_k) - f_\mu(x_k) + f_\mu(x_k) - f_* \leq (f(x_k) - f_\mu(x_k)) + (f_\mu(x_k) - f_{\mu,*})$$

where $f_{\mu,*} := \min_{x \in X} f_\mu(x)$. From Lemma 2, the approximation error $f(x_k) - f_\mu(x_k)$ is bounded above by μD_Y^2. Applying AGD to f_μ, equation (3) implies that the optimization error is $\mathcal{O}\left(\frac{||A||^2 D_X^2}{\mu k^2}\right)$ since $L_\mu = \frac{||A||^2}{\mu}$. We sum these two errors to conclude (8). □

Corollary 1. Let $\mu = \mathcal{O}(\varepsilon / D_Y^2)$ in Theorem 1. Then

$$f(x_k) - f_* \leq \mathcal{O}\left(\frac{||A||^2 D_X^2 D_Y^2}{\varepsilon t^2}\right) + \varepsilon$$

and the total number of iterations needed to compute an ε-solution is at most $N_\varepsilon = \mathcal{O}\left(\frac{||A|| D_X D_Y}{\varepsilon}\right)$.

3 Concluding Remarks and Numerical Experiments

We have just seen how to construct an L-smooth approximation to a (lower semi-) continuous convex function in theory. However, a few remarks are still in order before the practicality of the algorithm can be determined. Recall that we assumed f can be represented of the form in (4). While it is argued that such a representation very likely exists (such as the Fenchel representation), finding $f^*(y)$ is not always easy. In fact, there are many cases when the representation in (4) is not only not unique, but much easier to compute if the Fenchel representation is not the one used. For example[1], let $f(x) = \max_{1 \leq i \leq m} \left|a_i^T x - b_i\right|$ with $a_i, b_i \in \mathbb{R}^n$ for all $1 \leq i \leq m$. Computing f^* is by no means a trivial task, but

$$f(x) = \max_{y \in \mathbb{R}^m} \{(a_i^T x - b_i) y_i : \sum_i |y_i| \leq 1\}$$

is of the form in (4).

Furthermore, the computation of x_k in Algorithm 1 requires the computation of the gradient of the smoothed approximation $f_\mu = A^T y(x)$. However, $y(x)$ is not necessarily easily computable either. This smoothing technique proposed by Nesterov (called NEST-S) works best if $y(x)$ can be computed quickly. For many commonly used non-differentiable functions, such as $||\cdot||_p$, this can be done. Additionally, one can appropriately choose the prox function $d(y)$ to allow for easier computation of $y(x)$. The most common setting is $d(y) = \frac{1}{2}||y||^2$, but there exists many other functions satisfying the assumptions to make use of.

Assuming that the above discussion is not an issue, we have seen that NEST-S produces an $\varepsilon-$solution in $\mathcal{O}(1/\varepsilon)$ iterations which is much faster than the $\mathcal{O}(1/\varepsilon^2)$ guaranteed by the subgradient method. Let us take a look at a numerical example. Consider the following optimization problem

$$F_* := \min_{x \in X} f(x) + h(x)$$

where $X = B(0, r)$, $f(x) = \frac{1}{2}\langle Qx, x\rangle - \langle q, x\rangle$, $h(x) = ||Kx - b||$. Here, r, Q, q and K are taken from [5], but are not of specific interest. From this formulation, we see that both f and h are continuous, convex functions, but h is not differentiable. Thus, if we are to apply AGD to solve the optimization problem,

[1] example taken from [4]

we must first smooth h using Nesterov's technique. We will use the Fenchel representation in (7). Note that

$$\begin{aligned} h^*(y) &= \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - ||x-b|| \\ &= \langle b, y\rangle + \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - ||x|| \\ &= \langle b, y\rangle + g^*(y) \end{aligned}$$

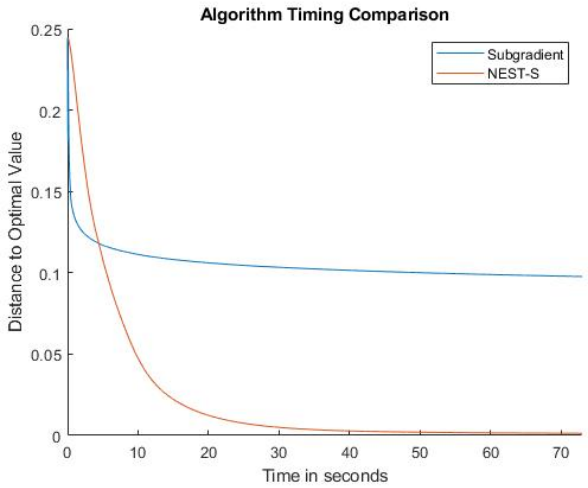
where $g^*(y)$ is the convex conjugate of $g(x) = ||x||$. We know from class that $g^*(y) = \begin{cases} 0 & ||y|| \le 1 \\ \infty & ||y|| > 1 \end{cases}$ so $h^*(y) = \langle b, v\rangle$ with $\operatorname{dom} h^* = Y := \{y : ||y|| \le 1\}$. Since $\operatorname{dom} h^*$ is compact the Fenchel representation

$$h(x) = \max_{y\in Y} \langle y, x\rangle - \langle b, y\rangle$$

satisfies (4). Thus, to continue with AGD, we simply need to compute $y(x)$ for any x and $\mu > 0$. The problem can be reformulated as

$$\begin{aligned} y(x) &= \operatorname*{argmax}_{y\in Y} \langle x-b, y\rangle - \frac{\mu}{2}||y||^2 \\ &= \operatorname*{argmin}_{y\in Y} \langle b-x, y\rangle + \frac{\mu}{2}||y||^2 \\ &= \operatorname*{argmin}_{y\in Y} \left|\left| y - \frac{1}{\mu}(x-b)\right|\right|^2 \end{aligned}$$

by completing the square. This is simply a projection onto Y. Moreover, $y(x) = \frac{(x-b)/\mu}{||(x-b)/\mu||}$. With an analytical solution to $y(x)$, we can apply AGD to solve our problem. Numerical experiments are shown in Figure 3. Notice that although the algorithms are initially comparable, to achieve any reasonable accuracy $\varepsilon > 0$, NEST-S is far more desirable.



Bibliography

[1] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Massachusetts, 2004.

[2] R Tyrrell Rockafellar. *Convex analysis*. Princeton University Press (Princeton, NJ), 1970.

[3] Stephen Boyd Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Niao He. Smoothing techniques.

[5] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pages 1–35, 2019.