Convergence of first-order methods via the convex conjugate

Javier Peña*

July 27, 2017

Abstract

This paper gives a unified and succinct approach to the $\mathcal{O}(1/\sqrt{k}), \mathcal{O}(1/k)$, and $\mathcal{O}(1/k^2)$ convergence rates of the subgradient, gradient, and accelerated gradient methods for unconstrained convex minimization. In the three cases the proof of convergence follows from a generic bound defined by the convex conjugate of the objective function.

1 Introduction

The subgradient, gradient, and accelerated gradient methods are icons in the class of first-order algorithms for convex optimization. Under a suitable Lipschitz continuity assumption on the objective function and a judicious choice of step-sizes, the subgradient method yields a point whose objective value is within $\mathcal{O}(1/\sqrt{k})$ of the optimal value after k iterations. In a similar vein, under a suitable Lipschitz continuity assumption on the gradient of the objective function and a judicious choice of step-sizes, the gradient and accelerated gradient methods yield points whose objective values are within $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ of the optimal value respectively after k iterations.

Although the proofs of the $\mathcal{O}(1/\sqrt{k}), \mathcal{O}(1/k)$, and $\mathcal{O}(1/k^2)$ convergence rates for these three algorithms share some common ideas, they are traditionally treated separately. In particular, the known proofs of the $\mathcal{O}(1/k^2)$ convergence rate of the accelerated gradient method, first established by Nesterov in a landmark paper [13], are notoriously less intuitive than those of the $\mathcal{O}(1/\sqrt{k})$ and $\mathcal{O}(1/k)$ convergence rates of the subgradient and gradient methods. Nesterov's accelerated gradient method has had a profound influence in optimization and has led to a vast range of developments. See, e.g., [4, 5, 14, 17, 19] and the many references therein.

Several recent articles [1, 7, 9, 12, 15, 18] have proposed novel approaches that add insight and explain how the accelerated gradient method and some variants achieve a faster convergence rate. This paper makes a contribution of similar spirit. It provides a unified and succinct approach for deriving the convergence rates of the subgradient, gradient, and accelerated gradient algorithms. The crux of the approach is a generic upper bound via the

*Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu

convex conjugate of the objective function. (See Lemma 1 in Section 2.) The construction of the upper bound captures key common features and differences among the three algorithms.

The paper is self-contained and relies only on the basic convex analysis background recalled next. (For further details see [6, 11, 16].) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Endow $\mathbb{R}^n$ with an inner product $\langle \cdot, \cdot \rangle$ and let $\| \cdot \|$ denote the corresponding Euclidean norm. Given a constant $G > 0$, the function f is G-Lipschitz if for all $x, y \in \mathrm{dom}(f) := \{x \in \mathbb{R}^n : f(x) < \infty\}$

$$f(x) - f(y) \leq G\|x - y\|.$$

Observe that if f is convex and G-Lipschitz then for all $x \in \mathrm{int}(\mathrm{dom}(f))$ and $g \in \partial f(x)$

$$g \in \partial f(x) \Rightarrow \|g\| \leq G. \tag{1}$$

Suppose f is differentiable on $\mathrm{dom}(f)$. Given a constant $L > 0$, the gradient ∇f is L-Lipschitz if for all $x, y \in \mathrm{dom}(f)$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Observe that if f is differentiable and ∇f is L-Lipschitz then for all $x, y \in \mathrm{dom}(f)$

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

In particular, if $x \in \mathrm{dom}(f)$ is such that $x - \frac{1}{L}\nabla f(x) \in \mathrm{dom}(f)$ then

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2. \tag{2}$$

Let $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ denote the convex conjugate of f, that is,

$$f^*(z) = \sup_{x \in \mathbb{R}^n} \{\langle z, x \rangle - f(x)\}.$$

The construction of the conjugate readily yields the following property known as *Fenchel's inequality.* For all $z, x \in \mathbb{R}^n$

$$f^*(z) + f(x) \geq \langle z, x \rangle$$

and equality holds if $z \in \partial f(x)$.

2 First-order methods for unconstrained convex optimization

Throughout the sequel assume $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and consider the problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{3}$$

Let $\bar{f}$ and $\bar{X}$ respectively denote the optimal value and set of optimal solutions to (3).

Algorithm 1 Subgradient/gradient method

input: $x_0 \in \mathbb{R}^n$ and a convex function $f : \mathbb{R}^n \to \mathbb{R}$
for $k = 0, 1, 2, \ldots$ **do**
  pick $g_k \in \partial f(x_k)$ and $t_k > 0$
  $x_{k+1} := x_k - t_k g_k$
end for

Algorithm 2 Accelerated gradient method

input: $x_0 \in \mathbb{R}^n$ and a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$
$y_0 := x_0$, $\theta_0 := 1$
for $k = 0, 1, 2, \ldots$ **do**
  pick $t_k > 0$
  $x_{k+1} := y_k - t_k \nabla f(y_k)$
  let $\theta_{k+1} \in (0, 1)$ be such that $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$
  $y_{k+1} := x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k)$
end for

Algorithm 1 and Algorithm 2 describe respectively the subgradient method and accelerated gradient method for (3). The subgradient method becomes the gradient method when f is differentiable. Algorithm 2 is a variant of Nesterov's original accelerated gradient method [13]. This version has been discussed in [4, 14, 19].

Theorem 1, Theorem 2, and Theorem 3 state well-known convergence properties of Algorithm 1 and Algorithm 2.

Theorem 1. *Suppose f is G-Lipschitz. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 1 satisfies*

$$\frac{\sum_{i=0}^k t_i f(x_i) - \frac{G^2}{2}\sum_{i=0}^k t_i^2}{\sum_{i=0}^k t_i} \leq f(x) + \frac{\|x_0 - x\|^2}{2\sum_{i=0}^k t_i} \tag{4}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $\min_{i=0,1,\ldots,k} f(x_i) - \bar{f} \leq \frac{\operatorname{dist}(x_0,\bar{X})^2 + G^2\sum_{i=0}^k t_i^2}{2\sum_{i=0}^k t_i}$, and $\min_{i=0,1,\ldots,k} f(x_i) - \bar{f} \leq \frac{\operatorname{dist}(x_0,\bar{X})^2 + G^2}{2\sqrt{k+1}}$ for $t_i = \frac{1}{\sqrt{k+1}}$, $i = 0, 1, \ldots, k$.

Theorem 2. *Suppose ∇f is L-Lipschitz and $t_k = \frac{1}{L}$, $k = 0, 1, \ldots$. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 1 satisfies*

$$\frac{f(x_1) + \cdots + f(x_k)}{k} \leq f(x) + \frac{L\|x_0 - x\|^2}{2k} \tag{5}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $f(x_k) - \bar{f} \leq \frac{L\operatorname{dist}(x_0,\bar{X})^2}{2k}$.

Theorem 3. *Suppose f is differentiable, ∇f is L-Lipschitz, and $t_k = \frac{1}{L}$ for $k = 0, 1, \ldots$. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 2 satisfies*

$$f(x_k) \leq f(x) + \frac{L\theta_{k-1}^2\|x_0 - x\|^2}{2} \tag{6}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $f(x_k) - \bar{f} \leq \frac{2L\operatorname{dist}(x_0,\bar{X})^2}{(k+1)^2}$.

The central contribution of this paper is a unified approach to the proofs of Theorem 1, Theorem 2, and Theorem 3. The crux of the approach is the following lemma.

Lemma 1. *There exists a sequence $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ such that for $k = 1, \ldots$ and $\mu_k = \frac{1}{\sum_{i=0}^{k} t_i}$ the left-hand side LHS_k of (4) in Theorem 1 satisfies*

$$\mathsf{LHS}_k \leq -f^*(z_k) + \langle z_k, x_0 \rangle - \frac{\|z_k\|^2}{2\mu_k} = -f^*(z_k) + \min_{u \in \mathbb{R}^n} \left\{ \langle z_k, u \rangle + \frac{\mu_k}{2} \|u - x_0\|^2 \right\}. \quad (7)$$

There also exist sequences $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ such that (7) holds for $\mu_k = \frac{L}{k}$ and the left-hand side LHS_k of (5) in Theorem 2, as well as for $\mu_k = L\theta_{k-1}^2$ and the left-hand side LHS_k of (6) in Theorem 3.

Lemma 1 captures some key common features and differences among the subgradient, gradient, and accelerated gradient algorithms. The right-hand side in (7) has the same form in all cases and has the same kind of dependence on the initial point x_0. Furthermore, as Section 3 below details, the construction of the sequences $z_k, \mu_k,\ k = 1, 2 \ldots$ follows the same template for the three algorithms. However, some details of the construction for these sequences need to be carefully tailored to each of the three algorithms.

Proof of Theorem 1, Theorem 2, and Theorem 3. Lemma 1 and Fenchel's inequality imply that for some $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ and all $x \in \mathbb{R}^n$ the left-hand-sides LHS_k of (4), (5), and (6) satisfy

$$\begin{aligned}
\mathsf{LHS}_k &\leq -f^*(z_k) + \min_{u \in \mathbb{R}^n} \left\{ \langle z_k, u \rangle + \frac{\mu_k}{2} \|u - x_0\|^2 \right\} \\
&\leq -f^*(z_k) + \langle z_k, x \rangle + \frac{\mu_k \cdot \|x - x_0\|^2}{2} \\
&\leq f(x) + \frac{\mu_k \cdot \|x - x_0\|^2}{2}.
\end{aligned}$$

To finish, recall that $\mu_k = \frac{1}{\sum_{i=0}^{k} t_i}$ for (4), $\mu_k = \frac{L}{k}$ for (5), and $\mu_k = L\theta_{k-1}^2$ for (6). For the second part of Theorem 2 observe that $f(x_k) \leq \frac{f(x_1) + \cdots + f(x_k)}{k}$ because (2) implies that $f(x_{i+1}) \leq f(x_i) - \frac{1}{2L}\|\nabla f(x_i)\|^2 \leq f(x_i),\ i = 0, 1, \ldots$. For the second part of Theorem 3 observe that a straightforward induction shows that the conditions $\theta_{k+1} \in (0, 1), \theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$, and $\theta_0 = 1$ imply $\theta_{k-1} \leq \frac{2}{k+1}$. □

3 Proof of Lemma 1

Construct the sequences $\mu_k \in \mathbb{R},\ z_k \in \mathbb{R}^n,\ k = 1, 2 \ldots$ as follows. First, choose sequences $\theta_k \in (0, 1), y_k \in \mathbb{R}^n, g_k \in \partial f(y_k),\ k = 1, 2, \ldots$, and two initial values $\mu_0 \in \mathbb{R}_+, z_0 \in \mathbb{R}^n$ or $\mu_1 \in \mathbb{R}_+, z_1 \in \mathbb{R}^n$. Second, let $\mu_k \in \mathbb{R},\ z_k \in \mathbb{R}^n,\ k = 1, 2 \ldots$ be defined by the rules

$$\begin{aligned}
z_{k+1} &= (1 - \theta_k) z_k + \theta_k g_k \\
\mu_{k+1} &= (1 - \theta_k) \mu_k.
\end{aligned}$$

This construction readily implies

$$\begin{aligned}\langle z_{k+1}, x_0\rangle - \tfrac{\|z_{k+1}\|^2}{2\mu_{k+1}} &= (1-\theta_k)\left(\langle z_k, x_0\rangle - \tfrac{\|z_k\|^2}{2\mu_k}\right)\\ &\quad +\theta_k\left(\left\langle g_k, x_0 - \tfrac{z_k}{\mu_k}\right\rangle - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right),\end{aligned}$$

and, by the convexity of f^* and $g_k \in \partial f(y_k)$,

$$\begin{aligned}-f^*(z_{k+1}) &\geq -(1-\theta_k)f^*(z_k) - \theta_k f^*(g_k)\\ &= -(1-\theta_k)f^*(z_k) - \theta_k\left(\langle g_k, y_k\rangle + f(y_k)\right).\end{aligned}$$

Thus

$$\begin{aligned}-f^*(z_{k+1}) + \langle z_{k+1}, x_0\rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} &\geq (1-\theta_k)\left(-f^*(z_k) + \langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right)\\ &\quad +\theta_k\left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + f(y_k) - \frac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right). \quad (8)\end{aligned}$$

To prove (7), proceed by induction. By (8) to show the inductive step k to $k+1$ it suffices to show

$$\mathsf{LHS}_{k+1} - (1-\theta_k)\mathsf{LHS}_k \leq \theta_k\left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + f(y_k) - \frac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right). \quad (9)$$

Next show (9) in each of the three cases.

First, for (4) take $\theta_k = \frac{t_{k+1}}{\sum_{i=0}^{k+1} t_i}$, $y_k = x_{k+1}$, and initial values $\mu_0 = \frac{1}{t_0}$, $z_0 = \frac{t_0 g_0}{t_0} = g_0$. Then $\mu_k = \frac{1}{\sum_{i=0}^{k} t_i}$, $\frac{\theta_k}{(1-\theta_k)\mu_k} = t_{k+1}$, and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$. Therefore

$$\begin{aligned}\mathsf{LHS}_{k+1} - (1-\theta_k)\mathsf{LHS}_k &= \tfrac{t_{k+1}f(x_{k+1}) - \frac{G^2}{2}t_{k+1}^2}{\sum_{i=0}^{k+1} t_i}\\ &= \theta_k\left(f(x_{k+1}) - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}G^2\right)\\ &\leq \theta_k\left(\left\langle g_k, x_0 - y_k - \tfrac{z_k}{\mu_k}\right\rangle + f(y_k) - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right).\end{aligned}$$

The inequality in the last step follows from (1).

Second, for (5) take $\theta_k = \frac{1}{k+1}$, $y_k = x_k$, and initial values $\mu_1 = L, z_1 = \nabla f(x_0)$. Then $\mu_k = \frac{L}{k}$, $\frac{\theta_k}{(1-\theta_k)\mu_k} = L$, and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$. Therefore

$$\begin{aligned}\mathsf{LHS}_{k+1} - (1-\theta_k)\mathsf{LHS}_k &= \tfrac{f(x_{k+1})}{k+1}\\ &\leq \theta_k(f(x_k) - \tfrac{1}{2L}\|\nabla f(x_k)\|^2)\\ &= \theta_k\left(f(y_k) - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right)\\ &= \theta_k\left(\left\langle g_k, x_0 - y_k - \tfrac{z_k}{\mu_k}\right\rangle + f(y_k) - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2\right).\end{aligned}$$

The inequality in the second step follows from $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ and (2).

Third, for (6) take θ_k, y_k as in Algorithm 2 and initial values $\mu_1 = L, z_1 = \nabla f(x_0)$. A separate induction argument shows that $\mu_k = L\theta_{k-1}^2$, $\frac{\theta_k^2}{(1-\theta_k)\mu_k} = \frac{1}{L}$, and

$$\begin{aligned} y_k &= (1-\theta_k)x_k + \theta_k(x_0 - \tfrac{z_k}{\mu_k}) \\ x_{k+1} &= (1-\theta_k)x_k + \theta_k(x_0 - \tfrac{z_{k+1}}{\mu_{k+1}}) \end{aligned}$$

for $k = 1, 2, \ldots$. In particular,

$$(1-\theta_k)(y_k - x_k) = \theta_k \left(x_0 - y_k - \frac{z_k}{\mu_k} \right). \tag{10}$$

Therefore

$$\begin{aligned} \mathsf{LHS}_{k+1} - (1-\theta_k)\mathsf{LHS}_k &= f(x_{k+1}) - (1-\theta_k)f(x_k) \\ &\leq f(y_k) - \tfrac{1}{2L}\|\nabla f(y_k)\|^2 - (1-\theta_k)(f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle) \\ &= (1-\theta_k)\langle g_k, y_k - x_k \rangle + \theta_k f(y_k) - \tfrac{1}{2L}\|g_k\|^2 \\ &= \theta_k \left(\left\langle g_k, x_0 - y_k - \tfrac{z_k}{\mu_k} \right\rangle + f(y_k) - \tfrac{\theta_k}{2(1-\theta_k)\mu_k}\|g_k\|^2 \right). \end{aligned}$$

The inequality in the second step follows from $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$ and (2), and from the convexity of f. The fourth step follows from (10).

To complete the proof of (7) by induction it only remains to verify that (7) holds for $k = 0$ or $k = 1$ in each of the three cases. For (4) observe that $f(x_0) = \langle z_0, x_0 \rangle - f^*(z_0)$ because $z_0 = g_0 \in \partial f(x_0)$. From (1) and $\mu_0 = \frac{1}{t_0}$ it follows that

$$\mathsf{LHS}_0 = \frac{t_0 f(x_0) - \frac{G^2}{2}t_0^2}{t_0} = f(x_0) - \frac{t_0}{2}G^2 \leq -f^*(z_0) + \langle z_0, x_0 \rangle - \frac{\|z_0\|^2}{2\mu_0}.$$

For both (5) and (6) observe that $f(x_0) = \langle z_1, x_0 \rangle - f^*(z_1)$ because $z_1 = \nabla f(x_0)$. From (2) and $\mu_1 = L$, it follows that

$$\mathsf{LHS}_1 = f(x_1) = f\left(x_0 - \frac{1}{L}\nabla f(x_0) \right) \leq f(x_0) - \frac{1}{2L}\|\nabla f(x_0)\|^2 = -f^*(z_1) + \langle z_1, x_0 \rangle - \frac{\|z_1\|^2}{2\mu_1}.$$

4 Potential extensions

This section sketches some potential extensions that will a topic for future work.

4.1 Proximal iterations

There are various first-order methods defined via proximal iterations [4,5,8,10,19]. Suppose $f = \phi + \psi$, where $\phi, \psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex functions such that the proximal map

$$\mathsf{Prox}_t(x) := \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ \psi(y) + \frac{1}{2t}\|x - y\|^2 \right\}$$

is computable. If ϕ is differentiable, then Algorithm 2 extends (see, e.g., [4]) by replacing step 5 with

$$5' : \; x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla \phi(y_k)).$$

Algorithm 1 also extends in a similar fashion.

A suitable extended version of Lemma 1 would readily yield a unified proof of the corresponding extended versions of Theorem 1, Theorem 2, and Theorem 3. The author conjectures that this is indeed the case if the right hand side in (7) is replaced with the following expression

$$-\phi^*(z_k) + \min_{u\in\mathbb{R}^n}\left\{\psi(u) + \langle z_k, u\rangle + \frac{\mu_k}{2}\|u - x_0\|^2\right\}.$$

4.2 Stronger convergence results

The convex conjugate approach developed in this paper may also yield alternative proofs of other stronger convergence properties of first-order methods. In particular, the $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence rates of the gradient and the accelerated gradient methods the can be strengthened to $o(1/k)$ and $o(1/k^2)$ respectively as shown in [3,10]. It is also known that the sequence of iterates generated by the gradient and accelerated gradient methods converge weakly to a minimizer as discussed in [2, 8]. The convex conjugate approach introduced in this paper may lead to succinct and unified derivations of these and possibly other results.

Acknowledgements

This research has been funded by NSF grant CMMI-1534850.

References

[1] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[2] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, pages 1–53, 2016.

[3] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[6] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*. Springer, New York, 2000.

[7] S. Bubeck, Y. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[8] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166(3):968–982, 2015.

[9] N. Flam and F. Bach. From averaging to acceleration, there is only a step-size. In *COLT*, pages 658–695, 2015.

[10] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[11] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer–Verlag, Berlin, 1993.

[12] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[13] Y. Nesterov. A method for unconstrained convex minimization problem with rate of convergence $\mathcal{O}(1/k^2)$. Doklady AN SSSR (in Russian). *(English translation. Soviet Math. Dokl.)*, 269:543–547, 1983.

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

[15] B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

[16] T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[17] S. Sra, S. Nowozin, and S. Wright. *Optimization for machine learning*. MIT Press, 2012.

[18] W. Su, S. Boyd, and E. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[19] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, 2008.