First-Order Optimization Methods

Weston Jackson

April 2017

Abstract

First-order methods are central to many algorithms in convex optimization. For any differentiable function, first-order methods can be used to iteratively approach critical points. This paper defines and describes the properties of a variety of first-order methods, primarily focusing on gradient descent, mirror descent, and stochastic gradient descent. The discussion includes descriptions of the classical algorithms as well as some recent breakthroughs used to accelerate these methods. Central to these breakthroughs is the use of momentum, linear coupling, and variance reduction. This survey gives concise descriptions of the main ideas behind these recent developments, explaining proofs of convergence from a conceptual standpoint.

Contents

1 Introduction

1.1 Overview

Convex optimization, a field devoted to minimizing a convex function over a convex set, has applications in a variety of fields including operations research, machine learning, and economics. Convex optimization problems have a specific form for a convex set K and a convex function f

$$\min_{x \in K} f(x)$$

There are a variety of ways to solve convex optimization problems. Often, we choose to frame optimization problems as *linear programs* or *integer programs*, which can be solved with methods such as the simplex method and ellipsoid method. Other times, it is often more useful to employ iterative first and second-order techniques in order to achieve a fast approximation. For this reason, innovations in powerful first-order methods such as gradient descent, mirror descent, and stochastic gradient descent can play a crucial role in quickly approximating a variety of problems.

1.2 Basic Convexity

The subject matter in this survey relies on the basic notions of *convex sets* and *convex functions*. For our purpose, we will use the definitions from Nisheeth K. Vishnoi's textbook on optimization methods [1].

Definition 1 (Convex Set [1])**.** *A set $K \subset \mathbb{R}^n$ is convex if, for every two points in K, the line segment connecting them is contained in K. Equivalently,*

$$\lambda x + (1 - \lambda) y \in K$$

for $x, y \in K, \lambda \in [0, 1]$.

Definition 2 (Convex Function [1])**.** *A function $f : K \to \mathbb{R}$ is convex if*

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

for $x, y \in K, \lambda \in [0, 1]$.

Intuitively, a convex set has the property that any point that lies between two points $x, y \in K$, is also in K. Similarly, a convex function f has the property that any line segment connecting two points on f is always above f. In this way, the space above a convex function f defines a convex set. Convex functions are important for their shared first-order and second-order properties. Because this paper focuses on first-order methods, we will predominantly need the following first-order property of convex functions:

Claim 3 (First-order convexity [1])**.** *If f is differentiable, then it is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

for $x, y \in K$.

Proof. We first show convexity implies the first-order condition. By definition, we have that:

$$f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda)f(x), \forall \lambda \in [0,1], x, y \in K \tag{1}$$

Rewriting, we have

$$f(y) \geq f(x) + \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \tag{2}$$

As $\lambda \to 0$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

Similarly, the first-order condition implies f is convex. Let $z = \lambda x + (1-\lambda)y$. Then we get the following equations:

$$\begin{aligned} f(x) &\geq f(z) + \langle \nabla f(z), x - z \rangle & (3) \\ f(y) &\geq f(z) + \langle \nabla f(z), y - z \rangle & (4) \end{aligned}$$

Multiply 3 by λ and 4 by $1-\lambda$, we get

$$\begin{aligned} \lambda f(x) + (1-\lambda) f(y) &\geq f(z) + \langle \nabla f(z), y - z \rangle & (5) \\ &\geq f(z) + \langle \nabla f(z), \lambda x + (1-\lambda)y - z \rangle & (6) \end{aligned}$$

Because $\langle \nabla f(z), \lambda x + (1-\lambda)y - z \rangle = \langle \nabla f(z), z - z \rangle = 0$, the result is the definition of convexity:

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y) \tag{7}$$

□

We note that this claim implies that any first-order approximation of f is an underestimate for any other point y in the convex set K. Additionally, we also use the following important claim from Vishnoi on critical points for convex function.

Claim 4 (Convex optimum [1])**.** *For a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a point x^*, the following are equivalent:*

a. x^* is a global minimum of f

b. x^* is a local minimum of f

c. $\nabla f(x^*) = 0$

Proof. We know that $a \to b$ is true for any global minimum. We also know that $b \to c$ is true for any local minimum. To show $c \to a$ is true, we assume that $\nabla f(x^*) = 0$. Then we have that $f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x \rangle = f(x^*)$. Thus, if x^* has $\nabla f(x^*) = 0$, we have that all other points y have $f(y) \geq f(x^*)$. Thus, x^* is a global minimum [1].

□

Using these properties of convexity, we have the tools that we need to describe first-order methods for minimizing convex functions.

2 Gradient Descent

2.1 Overview

A natural solution to finding the minimum of a convex function is to decrease the objective function f until we arrive at a minimum.

Assuming f is convex, we can use its first derivative to minimize it over a convex set. Consider the problem where f is differentiable and we want to find a sequence of points $x_1 \dots x_T$ where $f(x_1) \geq f(x_2) \dots \geq f(x_T) \geq f(x^*)$. Since $\nabla f(x_1)$ points in the direction where f grows the fastest at x_1, we can use $-\nabla f(x_1)$ to find the direction where f *decreases* the fastest. In particular, subtracting the first derivative as follows moves us in the approximate direction of the minimum:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

This method of using ∇f to approach the optimum of a function f is what is commonly referred to as *gradient descent* [2]. Gradient descent cannot find the exact minimum, but we use it iteratively to find an ϵ-close approximate solution. Additionally, gradient descent is independent of the dimension of the problem, making the convergence rate much more efficient in high dimensional space.

We use η to denote the learning rate, or the rate at which approach the approximate minimum of f. The value of η must be chosen carefully, as a large learning rate may cause cause us to skip over the optimum while a small learning rate can lead to many more iterations before achieving convergence.

We show the full gradient descent method from Nesterov's 2004 lectures in Algorithm 1 [2].

Algorithm 1 GD(x_0, η, T) [2]

1: Let $x_1 = x_0$
2: **for** $t = 1 \dots T$ **do**
3: $\quad x_{t+1} = x_t - \eta \nabla f(x_t)$
4: **end for**
5: **return** x_T

2.2 Convergence

In order to bound the number of steps it takes for gradient descent to converge on f, we must have some knowledge of the function's gradient. Although there are a variety conditions that can be imposed on f in order to show that gradient descent converges, the standard proof relies on $||x_0 - x^*|| \leq D$, for starting point x_1, and the function f being L-smooth.

Definition 5 (L-smooth [1])**.** *Assume that f is differentiable such that*

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

for any $x, y \in K$. Then we say f is L-Smooth.

We give a variation of Vishnoi's proof, which uses the definition of L-smoothness to prove the following theorem for the gradient descent algorithm.

Theorem 6 (Vishnoi [1]). *Gradient descent with learning rate* $0 \leq \eta \leq \frac{1}{L}$ *needs* $T = O(\frac{DL}{\epsilon})$ *iterations to achieve* $f(x_T) - f(x^*) \leq \epsilon$.

Proof Sketch. We start using a simple inequality that can be derived from f being L-Smooth:

$$f(x) - f(y) \leq \langle y - x, \nabla f(y) \rangle + L||x - y||^2 \tag{1}$$

Substituting x_t and x_{t+1} into this inequality, we get the following after each iteration:

$$f(x_{t+1}) - f(x_t) \leq \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + L||x_{t+1} - x_t||^2 \tag{2}$$

From the gradient update rule, we know that $x_{t+1} - x_t = -\eta \nabla f(x_t)$. We then assume $\eta = \frac{1}{2L}$:

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{4L}||\nabla f(x_t)||^2 \tag{3}$$

This inequality shows that the objective decreases by $-\frac{1}{4L}||\nabla f(x_t)||^2$ every step, under the L-smoothness assumption. Intuitively, it suggests that the larger the gradient, the faster we approach the optimum. Establishing this bound is central to proving the theorem, and a variant of this inequality is needed in almost every gradient descent related proof. Then, using the definition of convexity, we can bound the size of the gradient $||\nabla f(x_t)||$ by $\frac{f(x_t) - f(x^*)}{||x_t - x^*||}$, leading to the following:

$$f(x_{t+1}) - f(x_t) \leq -\frac{(f(x_t) - f(x^*))^2}{4L||x_t - x^*||^2} \tag{4}$$

Let $D = ||x_0 - x^*||^2$ and $\Theta = f(x_0) - f(x^*)$. Notice that to halve the distance to x^* from $\frac{\Theta}{2^i}$ to $\frac{\Theta}{2^{i+1}}$, we need to perform at most $O(\frac{LD^2 2^i}{\Theta})$ steps. The total amount of halves we need to achieve an ϵ-approximation is $\log \frac{\Theta}{\epsilon}$. This gives us the following summation:

$$\sum_{i=1}^{\log \frac{\Theta}{\epsilon}} O(\frac{LD^2 2^i}{\Theta}) = O(\frac{LD^2}{\epsilon}) \tag{5}$$

Thus, we need $T = O(\frac{LD^2}{\epsilon})$ iterations to converge [1].

□

Throughout this survey, many convergence proofs will follow this same format. We first use a bound on the gradient to ensure progress at each step, then telescope over all T steps in order to bound the error $f(x_T) - f(x^*)$ in terms of T. While the previous proof uses the L-smooth bound to show convergence, it is known that gradient descent converges even faster with stronger bounds. In particular, using the notion of σ-strong convexity, we can achieve an even faster convergence rate.

Definition 7 (σ-strong convexity [1]). *Assume that* f *is twice differentiable such that:*

$$\nabla^2 f(x) \succeq \sigma$$

$\forall x \in K$. *Then we say* f *is* σ*-strongly convex.*

Theorem 8 (Vishnoi [1]). *If f is σ-strong convex and L-smooth, needs $T = O(\frac{L}{\sigma} \log \frac{1}{\epsilon})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

Intuitively, this comes from the fact that when f has a large gradient, gradient descent takes bigger steps to approach the minimum faster. We omit the proof for this theorem, as it is just a variant of the proof for the L-smooth assumption.

3 Mirror Descent

3.1 Overview

Mirror descent is an iterative algorithm introduced by Nemirovski and Yudin in 1983 [3]. The algorithm is designed to minimize a convex function f with respect to an arbitrary norm $||\cdot||$. Note that to minimize f in some vector space $\mathcal{E}$, the mirror descent algorithm must work in the dual space $\mathcal{E}^*$ [4]. This is because any gradient $\nabla f(x)$ is defined in the dual space $\mathcal{E}^*$.

Note that previously, we did not have to work in the dual space as we used gradient descent to minimize with respect to the Hilbert space $\mathcal{H}$ (ℓ_2 norm), where $\mathcal{H}^* = \mathcal{H}$. Yet for an arbitrary vector space $\mathcal{E}$, the update step $x - \eta \nabla f(x)$ may not be defined as $x \in \mathcal{E}$ and $\nabla f(x) \in \mathcal{E}^*$.

Thus, in order to approximate x^* in the dual space, we must conduct the descent step in the dual space $\mathcal{E}^*$, and then map the result back into the primal space $\mathcal{E}$ with some map Φ. After mapping the result back into the primal space, the result may not be in the feasible region. Thus, we must project the result back into the feasible set K. We show a variant of Beck and Teboulle's version of the mirror descent method in Algorithm 2.

Algorithm 2 MD(x_0, η, Φ, T) [4]

Let $x_1 = x_0$
for $t = 1 \dots T$ **do**
    $\nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta \nabla f(x_t)$
    $x_{t+1} = \nabla \Phi^*(\nabla y_{t+1})$
end for
return $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$

Note that the mirror descent algorithm requires a map $\Phi(x) : K \to \mathbb{R}$ which is differentiable and strongly convex on K. We denote Φ^* as the conjugate of Φ, such that $\Phi^*(y) = \max_{x \in K} \{\langle x, y \rangle - \Phi(x)\}$. This conjugate function is needed in order to ensure x_{t+1} is projected back into the feasible region K after the transformation.

Mirror descent is also different from gradient descent in that we return $\bar{x}$ instead of x_T. This is because progress is not guaranteed at every mirror descent step. Nonetheless, convergence can still be ensured if we return the average of the previous T iterations.

3.2 Mirror Descent as Generalized Gradient Descent

Beck and Teboulle give an alternative characterization of the mirror descent algorithm. We first note that gradient descent can be generalized with the following update step with respect to the ℓ_2 norm:

$$x_{t+1} = argmin_{y \in K} \left\{ \nabla f(x_t)^T y + \frac{1}{2\eta} ||y - x_t||_2^2 \right\}$$

We can show that this generalization is equivalent to gradient descent by minimizing with respect to y:

Proof.

$$\begin{aligned} 0 &= \nabla f(x_t)^T y \frac{\partial}{\partial y} + \frac{1}{2\eta} ||y - x_t||_2^2 \frac{\partial}{\partial y} \\ 0 &= \nabla f(x_t) + \frac{1}{\eta}(y - x_t) \\ \frac{1}{\eta}(y - x_t) &= -\nabla f(x_t) \\ y &= x_t - \eta \nabla f(x_t) \end{aligned}$$

□

By replacing the ℓ_2-norm in the generalized gradient descent step with any proximity function, we can achieve a gradient descent algorithm for a different geometric manifold. Assume that we choose to use generalized gradient descent with the *Bregman divergence* proximity function.

Definition 9 (Bregman divergence [4])**.** *Let $\Phi : K \to \mathbb{R}$ be a map such that $\Phi(y) \geq \Phi(x)\langle \nabla\Phi(x), y - x\rangle + \frac{1}{2}||x - y||^2$. The Bregman divergence is a distance metric between two points x and y defined as follows*

$$B_\Phi(x, y) = \Phi(x) - \Phi(y) - \nabla\Phi(y)^T(x - y)$$

where $B_\Phi(x, x) = 0$ and $B_\Phi(x, y) \geq \frac{1}{2}||x - y||^2$. Note that $B(x, y)$ implies that B is defined for any valid Φ.

Beck and Teboulle show that mirror descent is equivalent to generalized gradient descent with the Bregman divergence metric, subject to a given map Φ [4].

Proof. Gradient descent with Bregman divergence is the following:

$$x_{t+1} = argmin_{y \in K} \left\{ \nabla f(x_t)^T y + \frac{1}{\eta} B_\Phi(y, x_t) \right\} \tag{1}$$

We take the derivative to find the optimality conditions:

$$0 \in \eta \nabla f(x_t) + \nabla\Phi(x_{t+1}) - \Phi(x_t) + N_K \tag{2}$$

where N_K is the normal cone of the closed convex set K. Rearranging terms, get the following:

$$x_{t+1} \in (\nabla\Phi + N_K)^{-1}(\nabla\Phi(x_t) - \eta\nabla(x_t)) \tag{3}$$

Because Φ is differentiable and strongly convex, Beck and Teboulle get that $(\Phi + N_K)^{-1}(z) = \Phi^*(z)$. Thus gradient descent with the Bregman divergence metric is exactly the same as mirror descent performed in one step:

$$x_{t+1} = \nabla\Phi^*(\nabla\Phi(x_t) - \eta\nabla f(x_t)) \tag{4}$$

Thus, the theorem is proved [4]. □

3.3 Mirror Descent as a Dual Method

In his 2014 paper, Zeyuan Allen-Zhu characterizes the mirror descent algorithm as a dual method to gradient descent [5]. This characterization stems from the analysis of mirror descent, which shows that the algorithm essentially uses lower-bounding hyper-planes to iteratively find the minimum of a convex function f.

Specifically, note that by the first-order property of convexity, any gradient $\nabla f(x)$ establishes a lower-bounding hyper-plane on f. Thus, each x_t in the mirror descent algorithm can be seen as a point establishing a lower-bounding hyper-plane $\nabla f(x_t)$ on f. Ultimately, mirror descent returns the average of these queries $\bar{x}$ in the hopes that the average of the hyper-planes has $\nabla f(\bar{x}) \approx \nabla f(x^*) = 0$.

Using this intuition, we note that mirror descent can be seen as a dual method to gradient descent. Specifically, while gradient descent improves with a larger gradient, the convergence analysis shows that mirror descent improves when ∇f is *smaller*. This is because when many of the lower-bounding hyper-planes have a gradient close to zero, we can establish a tighter bound on f.

3.4 Convergence

In order to prove the algorithms convergence, we will need to introduce the ρ-Lipschitz upper-bound on the gradient.

Definition 10 (ρ-Lipschitz [5])**.** *A function f is ρ-Lipschitz with respect to a norm $||\cdot||$ if $\forall x \in K$, $g \in \nabla f(x)$, $||g||_*^2 \leq \rho$.*

The norm $||\cdot||_*$ is the *dual norm* defined for $\nabla f \in \mathcal{E}^*$. The main proof relies on the following Mirror Descent Lemma, which Allen-Zhu proves using the properties Bregman divergence and the minimality of each mirror descent step. We omit the proof of the lemma for conciseness.

Lemma 11 (Mirror Descent Lemma [5])**.** *At each iteration t of mirror descent, we have that $\forall x \in K$:*

$$\eta(f(x_t) - f(x)) \leq \eta\langle\nabla f(x_t), x_t - x\rangle \leq \frac{\eta^2}{2}||\nabla f(x_t)||_*^2 + B(x_t, x) - B(x_{t+1}, x)$$

Note that letting $x = x^*$, the Mirror Descent Lemma says that our error from the optimum, $f(x_t) - f(x^*)$, is smaller than the decrease in the Bregman divergence at our current iteration, $B(x_t, x^*) - B(x_{t+1}, x^*)$, subject to some error proportional to the square of the size of the gradient $||\nabla f(x_t)||_*^2$.

Theorem 12 (Allen-Zhu [5]). *Assume f is convex and ρ-Lipschitz with $B(x_1, x^*) \leq D$, then mirror descent with $\eta = \frac{\sqrt{D}}{\rho\sqrt{T}}$ needs $T = O(\frac{D\rho^2}{\epsilon^2})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

Proof Sketch. From the Mirror Descent Lemma, we can telescope over all $1 \dots T$ iterations to show that $\eta T(f(\bar{x}) - f(x^*)) \leq T\frac{\rho^2\eta^2}{2} + D$. The proof is slightly intricate, and given in Appendix A for conciseness. This inequality then implies that at iteration t, the distance between the optimal x^* and $\bar{x}$ is bounded by the following:

$$f(\bar{x}) - f(x^*) \leq \frac{\rho^2\eta}{2} + \frac{D}{T\eta} \tag{1}$$

Letting $\eta = \frac{\sqrt{2D}}{L\sqrt{T}}$:

$$\frac{\rho^2\eta}{2} + \frac{D}{T\eta} = \frac{\rho\sqrt{2D}}{2\sqrt{T}} + \frac{D\sqrt{T}\rho}{T\sqrt{2D}} = \frac{\rho\sqrt{2D}}{\sqrt{T}} \tag{2}$$

Therefore, we get that

$$T \geq \frac{2D\rho^2}{\epsilon^2} \rightarrow f(\bar{x}) - f(x^*) \leq \epsilon \tag{3}$$

Thus, the theorem is proved [5]. □

4 Accelerated First-Order Methods

4.1 Overview

In the previous sections, we show a roughly $O(\frac{1}{\epsilon})$ convergence rate for gradient descent and a roughly $O(\frac{1}{\epsilon^2})$ convergence rate for mirror descent in the non-strongly convex case. The question remains whether there are first-order methods with accelerated convergence rates. Although such methods exist, the algorithms are more intricate and require much longer proofs.

This section describes two accelerated methods that achieve $O(\frac{1}{\sqrt{\epsilon}})$ convergence. The first is Nesterov's method, which was the first gradient descent method to achieve $O(\frac{1}{\sqrt{\epsilon}})$ convergence in 1983. The second method, introduced in 2014 by Zeyuan Allen-Zhu, is a more intuitive algorithm that achieves $O(\frac{1}{\sqrt{\epsilon}})$ convergence by combining mirror descent and gradient descent steps.

4.2 Nesterov's Method

It is possible to improve on gradient descent using *Nesterov's method* [6]. Nesterov's method can be conceptualized as using the *momentum* from the previous point x_t, when calculating the next point x_{t+1}.

At each iteration, Nesterov's method first calculates the gradient descent step y_{t+1} as an intermediary. It then sets x_{t+1} to be a linear combination of y_{t+1} with the gradient

descent step y_t from the previous iteration. This linear combination effectively incorporates the gradient from the previous iteration in calculating x_t. In fact, because this is done at each iteration, Nesterov's method is essentially factoring in the entire gradient history every update! In this way, the algorithm is able to accelerate to a convex function's minimum with much fewer iterations.

The cleverness of Nesterov's method lies in the choice of λ, which specifies how y_{t+1} and y_t are combined to form x_{t+1}. We show the entire algorithm in Algorithm 3. Note that as $t \to \infty$, $\gamma_t = 1$. Thus, as we approach the minimum x^*, Nesterov's method turns into normal gradient descent and begins to ignore the gradient history.

Algorithm 3 Nesterov(x_0, L, T) [6]

Let $\lambda_1 = 1$
Let $x_1 = y_1 = x_0$
for $t = 1 \dots T$ **do**
    $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$
    $\gamma_t = \frac{\lambda_t - 1}{\lambda_{t+1}}$
    $y_{t+1} = x_t - \frac{1}{L}\nabla(x_t)$
    $x_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t$
end for
return x_T

Although the idea behind Nesterov's method is relatively intuitive, it is not immediately obvious why his algorithm converges at a faster rate than normal gradient descent. Yet improving on gradient descent, Nesterov's method gives us a roughly $O(\frac{1}{\sqrt{\epsilon}})$ convergence rate!

Theorem 13 (Nesterov [6]). *Assuming f is convex and L-smooth, Nesterov's accelerated gradient descent needs $T = O(\frac{\sqrt{L}D}{\sqrt{\epsilon}})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

We omit the proof of Nesterov's accelerated gradient descent method as the analysis is not enlightening and involves at least a page of algebra. For this reason, Nesterov's accelerated gradient descent method is often seen as an analytical trick, which lacks an intuitive geometric interpretation. Nonetheless, Nesterov's method is still crucial to the development of more intuitive accelerated first-order methods. In particular, while Nesterov's accelerated gradient descent was the first method to achieve $O(\frac{1}{\sqrt{\epsilon}})$ convergence for L-smooth convex functions, more recent papers have used similar ideas to achieve the same convergence with a more intuitive algorithm.

4.3 Linear Coupling

One of the weakest characteristics of gradient descent is that the algorithm always approaches the minimum of a function from above. This is because gradient descent is a primal-only optimization method, which by ignoring the dual problem, never establishes a lower bound on $f(x^*)$. On the other hand, we know that mirror descent is essentially a dual method, and thus approaches $f(x^*)$ from the other direction.

In 2014, Zeyuan Allen-Zhu combined these strategies to create an intuitive algorithm that matches Nesterov's method [5]. The main idea is that when the size of the gradient $||\nabla f(x)||_*$ is large, then gradient descent can make large steps to quickly approach the optimum x^*, and when $||\nabla f(x)||_*$ is small, mirror descent's $\bar{x}$ is more stable an accurate. Thus, combining them yields a faster algorithm overall.

Consider the case when $f(x_0) - f(x^*)$ is small, f is L-smooth, and $||\nabla f(x_t)||_*$ is always $\geq G$ or $\leq G$. We remember from Vishnoi's gradient descent proof, if $\eta = \frac{1}{2L}$, we get the following bound for the progress at each step:

$$f(x_t) - f(x_{t+1}) \geq \frac{1}{4L}||\nabla f(x_t)||_*^2$$

Thus, gradient descent makes $\frac{G^2}{4L}$ progress at each step, and needs only $T \geq \Omega(\frac{L\epsilon}{G^2})$ steps to converge. Additionally, from the mirror descent theorem, we know that when f is ρ-Lipschitz, mirror descent only needs $T \geq \frac{D\rho^2}{\epsilon^2}$ steps to converge. Since we assume the D is small and we know $||\nabla f(x)||_* \leq \rho$, we have that $T \geq \Omega(\frac{G^2}{\epsilon^2})$ gives us convergence. Thus, in either case

$$T \geq \Omega(\max\left\{\frac{L\epsilon}{G^2}, \frac{G^2}{\epsilon^2}\right\})$$

If $G = (L\epsilon^3)^{\frac{1}{4}}$, we get the following:

$$T \geq \Omega(\max\left\{\frac{L\epsilon}{\sqrt{L}\epsilon^{\frac{3}{2}}}, \frac{\sqrt{L}\epsilon^{\frac{3}{2}}}{\epsilon^2}\right\}) = \Omega(\frac{L}{\sqrt{\epsilon}})$$

This matches the convergence rate of Nesterov's accelerated gradient descent!

In order for this strategy to work, Allen-Zhu borrows Nesterov's idea of linear coupling in order to perform mirror and gradient descent steps at the same time. In particular, he uses a linear combination of both steps, with a parameter $\tau \in [0, 1]$ to control how the two steps are linearly combined. The rough algorithm is given in Algorithm 4.

Algorithm 4 RoughAGM$(x_0, \eta, \tau, \Phi, T)$ [5]

Let $x_1 = y_1 = z_1 = x_0$
for $t = 1 \dots T$ **do**
    $x_{t+1} = \tau z_t + (1 - \tau)y_t$
    $y_{t+1} = GD(\frac{1}{L}, x_{t+1}, T)$
    $z_{t+1} = MD(\eta, x_{t+1}, \Phi, T)$
end for

Using the Mirror Descent Lemma, Allen-Zhu shows that for all $x \in K$

$$\eta\langle\nabla f(x_{t+1}), z_t - x\rangle \leq \eta^2 L(f(x_{t+1}) - f(y_{t+1})) + B(z_t, x) - B(z_{t+1}, x) \tag{1}$$

Assuming $x = x^*$, every iteration either the mirror descent step z_t brings us closer to x or the objective decreases from the gradient descent step. Unfortunately, this series does not telescope over all $t = 1 \dots T$ due to the presence of non-canceling x_{t+1} and y_{t+1} terms. Yet using the definition of convexity and the fact $\tau(x_{t+1} - z_t) = (1 - \tau)(y_t - x_{t+1})$, we can get a second inequality:

$$\eta\langle\nabla f(x_{t+1}), x_{t+1} - x\rangle - \eta\langle\nabla f(x_{t+1}), z_t - x\rangle \leq \frac{(1-\tau)\eta}{\tau}(f(y_t) - f(x_{t+1})) \tag{2}$$

Letting $\frac{1-\tau}{\tau} = \eta L$, we can add inequalities 1 and 2, which cancel to get the following lemma:

Lemma 14 (Allen-Zhu [5]). *Letting $\tau \in (0,1)$ satisfy $\frac{1-\tau}{\tau} = \eta L$, we get that $\forall x \in K$:*

$$\eta\langle\nabla f(x_{t+1}), x_{t+1} - x\rangle \leq L\eta^2(f(y_t) - f(y_{t+1})) + B(z_t, x) - B(z_{t+1}, x) \tag{3}$$

Notice that the use of τ allows us to bound the progress at a single step with a telescoping series. Telescoping over the series and letting $\bar{x} = \frac{1}{T}\sum_{i=1}^{T} x_t$, we can then bound the approximate error after T iterations. We then $f(y_0) - f(x^*) \leq \Theta$ and $B(x_0, x^*) \leq D$ to achieve the following result:

$$f(\bar{x}) - f(x^*) \leq \frac{1}{T}(\eta L\Theta + \frac{D}{\eta}) \tag{4}$$

Finally, letting $\eta = \sqrt{\frac{D}{L\Theta}}$, we get the following bound:

$$f(\bar{x}) - f(x^*) \leq \frac{2\sqrt{LD\Theta}}{T} \tag{5}$$

This means that every $T = 4\sqrt{\frac{LD}{\Theta}}$ steps, $f(\bar{x}) - f(x^*) \leq \frac{\Theta}{2}$, meaning the distance to the optimal x^* is halved. Thus, if we restart this procedure halving the distance with each run, we get $T = O(\sqrt{\frac{LD}{\epsilon}})$, as is desired. Unfortunately, because Θ and D are not always known in practice, this rough strategy doesn't work. However, the authors show that by letting η and τ change across iterations, this potential problem can be avoided. Algorithm 5 gives the full algorithm.

Algorithm 5 AGM(x_0, η, T) [5]

Assume f is L-Smooth with respect to $||\cdot||$
$x_1 = y_1 = z_1 = x_0$
for $t = 1 \dots T$ **do**
    $\eta_{t+1} = \frac{t+2}{2L}$
    $\tau_t = \frac{1}{\eta_{t+1}L}$
    $x_{t+1} = \tau z_t + (1-\tau)y_t$
    $y_{t+1} = argmin_{y\in K}\{\frac{L}{2}||y - x_{t+1}|| + \langle\nabla f(x_{t+1}), y - x_{t+1}\rangle\}$
    $z_{t+1} = argmin_{z\in K}\{B(z_k, z) + \langle\eta_{t+1}\nabla f(x_{t+1}), z - z_t\rangle\}$
end for
return y_T

Similar to Nesterov's method, as $T \to \infty$, $\tau_t = 0$ and the linear coupling algorithm becomes gradient descent. This makes sense as the gradient descent ensures progress and is more stable than mirror descent as we approach the optimum. Additionally, we note that this algorithm uses Nemirovski's mirror descent steps.

4.4 Convergence

Theorem 15 (Allen-Zhu [5]). *Assume f is L-smooth w.r.t. $||\cdot||$ on K, and Φ is 1-strongly convex w.r.t. $||\cdot||$ on K. Let D be any upper bound on $B(x_0, x^*)$. Then the linear coupling algorithm needs $T = O(\frac{\sqrt{DL}}{\sqrt{\epsilon}})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

Proof Sketch. Following a similar proof strategy, we achieves the following error bound after a single iteration:

$$\begin{aligned} \eta_{t+1}(f(x_{t+1}) - f(x)) &\leq \frac{(1-\tau_t)\eta_{t+1}}{\tau_t}(f(y_t) - f(x_{t+1})) && (1) \\ &- \eta_{t+1}^2 L(f(x_{t+1}) - f(y_{t+1})) + B(z_t, x) - B(z_{t+1}, x) && (2) \end{aligned}$$

for any $x \in K$. Letting $\tau_t = \frac{1}{\eta_{t+1}L}$:

$$\begin{aligned} \eta_{t+1}(f(x_{t+1}) - f(x)) &\leq (\eta_{t+1}^2 L - \eta_{t+1})f(y_t) - \eta_{t+1}^2 L f(y_{t+1}) && (3) \\ &+ \eta_{t+1} f(x_{t+1}) + B(z_t, x) - B(z_{t+1}, x) && (4) \end{aligned}$$

The choice of τ_t gets us closer to a telescoping inequality. Simplifying this inequality leads to the following lemma:

Lemma 16 (Allen-Zhu [5]). *If $\tau_t = \frac{1}{\eta_{t+1}L}$, then $\forall x \in K$:*

$$\begin{aligned} \eta_{t+1}(f(x) - f(x_{t+1})) &\geq \eta_{t+1}^2 L f(y_{t+1}) - (\eta_{t+1}^2 L - \eta_{t+1})f(y_t) && (5) \\ &+ B(z_{t+1}, x) - B(z_t, x) && (6) \end{aligned}$$

We set $\eta_t^2 L = \eta_{t+1}^2 L - \eta_{t+1} + \frac{1}{4L}$ and then telescope over all t:

$$\eta_T^2 L f(y_T) + \sum_{t=1}^{T-1} \frac{1}{4L} f(y_t) + B(z_T, x) - B(z_0, x) \leq \sum_{t=1}^{T} \eta_t f(x) \tag{7}$$

We are now ready to substitute $x = x^*$. After T iterations, we have $\sum_{t=1}^{T} \eta_t = \frac{T(T+3)}{4L}$. Additionally, we know that $f(y_t) \geq f(x^*)$, $B(z_T, x^*) \geq 0$, and $B(z_0, x^*) = D$. Thus:

$$\frac{(T+1)^2 f(y^T) L}{4L^2} \leq (\frac{T(T+3)}{4L} - \frac{T-1}{4L}) f(x^*) + D \tag{8}$$

Simplifying, we get

$$f(y_T) - f(x^*) \leq \frac{4DL}{(T+1)^2} \tag{9}$$

Proving the theorem [5]. □

5 Stochastic Gradient Descent

5.1 Overview

Many convex minimization functions in machine learning have the following structure:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \phi(x)$$

Here, $f_i(x)$ is a convex function which is associated with the i-th observation in the data set and $\phi(x)$ is a convex proximal function. These functions often arise in least squares, maximum likelihood estimation, and empirical risk minimization.

Note that the standard gradient descent method would iterate over the gradient of all summand functions f_i. When the data set is enormous, this often becomes impractical. *Stochastic gradient descent* (SGD) has been shown to be faster, more reliable, less likely to reach a local minimum of the function [7]. With SGD, we sample $I \in \{1, 2, \dots n\}$, and calculate the new point according to the gradient at this f_I only. Algorithm 6 gives the full algorithm.

Algorithm 6 SGD(x_0, η, T) [7]

$x_1 = x_0$
for $t = 1 \dots T$ **do**
    Sample $I \in \{1, 2, \dots n\}$
    $x_{t+1} = x_t - \eta \nabla f_I(x_t)$
end for
return x_T

In this case, Algorithm 6 assumes that the proximal function $\phi(x)$ is zero. When the proximal function $\phi(x)$ is present, the SGD update step is generalized as the following [8]:

$$x_{t+1} = argmin_y \left\{ \frac{1}{2\eta}||y - x_t||_2^2 + \langle \nabla f_I(x), y \rangle + \phi(y) \right\}$$

Additionally, we note that on expectation, SGD is an unbiased estimator of gradient descent:

$$\nabla F(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) = \mathbb{E}[\nabla f_I(x)]$$

Thus, $\nabla \phi_I(x)$ is an unbiased estimator of the gradient of F at x, and has a cost of computing independent of n. SGD has many other advantages over gradient descent. While gradient descent can get stuck at a local minimum for non-convex functions, SGD can escape from a local minimum due to its random nature. Additionally, it is known to converge even when the objective function is not differentiable everywhere [7].

It is natural to ask whether SGD can also be accelerated. Unfortunately, convergence is much more difficult than for gradient descent due to the algorithms large variance. It is known that SGD has a slow standard convergence rate of roughly $O(\frac{1}{\epsilon^2})$ [9]. Even when F is strongly convex, the convergence rate is at most $O(\frac{1}{\epsilon})$ [8]. This is largely because SGD requires a decaying learning rate in order to reduce the variance and ensure convergence.

5.2 Variance Reduction

Many methods for accelerating SGD focus on reducing the variance of the algorithm. In 2013, Johnson and Zhang published the *stochastic variance reduced gradient* (SVRG) method in order to improve convergence rate [9]. In the σ-strongly convex setting, they achieve an ϵ-approximation in $T = O((n + \frac{L}{\sigma}) \log \frac{1}{\epsilon})$ iterations.

Let $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ be the function we wish to minimize. The method keeps a snapshot vector $\widetilde{x}$, updated every m iterations of stochastic gradient descent. It then computes a full gradient descent step with this vector as follows:

$$\nabla f(\widetilde{x}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\widetilde{x})$$

The update step is then modified to be the following:

$$x_{t+1} = x_t - \eta(\nabla f_I(x_t) - \nabla f_I(\widetilde{x}) + \nabla f(\widetilde{x}))$$

Note that $\mathbb{E}[\nabla f_I(x_t) - \nabla f_I(\widetilde{x}) + \nabla f(\widetilde{x})] = \nabla f(x_t)$ as before. By adding in a $\nabla f(\widetilde{x}) - \nabla f_I(\widetilde{x})$ term at every iteration, we reduce the variance at each iteration and stabilize the algorithm. In particular, as $x_t \to x^*$:

$$||X - E[X]||^2 = ||(\nabla f_I(x_t) - \nabla f_I(\widetilde{x}) + \nabla f(\widetilde{x})) - \nabla f(x_t)||^2 \to 0$$

Algorithm 7 gives the full SVRG descent algorithm [9].

Algorithm 7 SVRG(x_0, η, S, m) [9]

$\widetilde{x} = x_0$
for $s = 1 \dots S$ **do**
  $\widetilde{x} = \widetilde{x}_{s-1}$
  $\nabla f(\widetilde{x}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\widetilde{x_s})$
  $x_1 = \widetilde{x}$
  for $t = 1 \dots m$ **do**
    Sample $I \in \{1, 2, \dots n\}$
    $x_{t+1} = x_t - \eta(\nabla f_I(x_t) - \nabla f_I(\widetilde{x}) + \nabla f(\widetilde{x}))$
  end for
  $\widetilde{x_s} = x_t$ for $t \in \{0 \dots m-1\}$
end for
return $\widetilde{x_S}$

Theorem 17 (Johnson and Zhang [9])**.** *Assume all f_i are convex, L-smooth, and F is σ-strongly convex. Let x^* be where F is minimum and let $F(x_0) - F(x^*) = \Theta$. Assume m is sufficiently large such that*

$$\alpha = \frac{1}{m\sigma\eta(1 - 2L\eta)} + \frac{2L\eta}{(1 - 2L\eta)} < 1$$

Then on expectation we get exponential convergence,

$$\mathbb{E}[F(\widetilde{x_s}) - F(x^*)] \leq \alpha^s \Theta$$

Proof Sketch. We provide a proof sketch of Johnson and Zhang's theorem. Consider $g_I(x) = f_I(x) - f_I(x^*) - \nabla f_I(x^*)^T(x - x^*)$. Using the fact that all f_i are L-smooth, Johnson and Zhang show the following gradient bound:

$$||\nabla f_I(x) - \nabla f_I(x^*)||_2^2 \leq 2Lg_I(x) \tag{1}$$

We can then use this inequality to show that the gradient moves us closer to $F(x^*)$ on expectation:

$$\mathbb{E}[\nabla f_I(x_t) - \nabla f_I(\widetilde{x}) + \nabla f(\widetilde{x})]_2^2 \leq 4L[F(x_t) + F(\widetilde{x}) - 2F(x^*)] \tag{2}$$

We can use this to show the distance between x_{t+1} and x^* decreases at each iteration:

$$\begin{aligned} \mathbb{E}[||x_{t+1} - x^*||_2^2] &\leq \mathbb{E}[||x_t - x^*||_2^2] - 2\eta(1 - 2L\eta)[F(x_t) - F(x^*)] &(3)\\ &+ 4L\eta^2[F(\widetilde{x}) - F(x^*)] &(4)\end{aligned}$$

We notice that this series can be telescoped over all $t = 1 \dots m$ to obtain the following inequality:

$$\begin{aligned} 2\eta(1 - 2L\eta)m\mathbb{E}[F(\widetilde{x_s}) - F(x^*)] &\leq 4Lm\eta^2 E[F(\widetilde{x}_{s-1}) - F(x^*)] &(5)\\ &+ \mathbb{E}[||x_0 - x^*||]_2^2 - \mathbb{E}[||x_m - x^*||_2^2] &(6)\end{aligned}$$

We can discard the term $\mathbb{E}[||x_m - x^*||_2^2] \geq 0$. We can then use the strong convexity property to bound the expected distance between x_0 and x^*. We can then bound the expected difference $\mathbb{E}[F(\widetilde{x_s}) - F(x^*)]$ as follows:

$$\mathbb{E}[F(\widetilde{x_s}) - F(x^*)] \leq [\frac{1}{m\sigma\eta(1 - 2L\eta)} + \frac{2L\eta}{(1 - 2L\eta)}]\mathbb{E}[F(\widetilde{x}_{s-1}) - F(x^*)] \tag{7}$$

Over all s, we get that $\mathbb{E}[F(\widetilde{x_s}) - F(x^*)] \leq \alpha^s\mathbb{E}[F(\widetilde{x_0}) - F(x^*)]$ and the theorem is proven. Additionally, we note that this is equivalent to achieving an ϵ-approximation in $T = O((n + \frac{L}{\sigma})\log\frac{1}{\epsilon})$ iterations [9].

□

6 Accelerated Stochastic Gradient Descent

6.1 Overview

In 2016, Zeyuan Allen-Zhu introduced a stochastic gradient method that allowed for accelerated, $O(\frac{1}{\sqrt{\epsilon}})$ convergence in general and improved convergence when F is strongly convex (by a factor of $\sqrt{\frac{L}{\sigma}}$) [8]. The method, called *Katyusha*, has the fastest known convergence rate and can be applied to a variety of convex objectives.

Katyusha combines linear coupling and variance reduction in order to efficiently descend toward the minimum x^*. To choose x_{t+1}, katyusha uses a convex function of three variables: $\widetilde{x}, y_t$, and z_t.

Conceptually, y_t and z_t ressemble variables in methods we have seen before. y_t is set to the variance reduced gradient for the current iteration while z_t is a variance reduced mirror descent step. Note that from the linear coupling algorithm, combining gradient descent and mirror descent achieves the same convergence as Nesterov's method. Additionally, the

algorithm uses $\widetilde{x}$ as the snapshot vector in a similar manner to variance reduction. The use of $\widetilde{x}$ reduces the variance of the stochastic updates, and ensures that the momentum never carries us too far in the wrong direction. The $\widetilde{x}$ vector is referred to as the *Katyusha momentum*. Algorithm 8 gives the full algorithm with the L-smooth assumption.

Algorithm 8 Katyusha(x_0, σ, S) [8]

$m = 2n$
$\tau_2 = \frac{1}{2}$
Initialize $\widetilde{x_0}, y_0, z_0 = x_0$
for $s = 1 \dots S$ **do**
    $\tau_{1,s} = \frac{2}{s+4}$
    $\eta_s = \frac{1}{3\tau_{1,s}L}$
    $\mu_s = \nabla f(\widetilde{x}_s)$
    Initialize $y_1 \dots y_m$
    for $t = 1 \dots m$ **do**
        Sample $I \in [n]$
        $x_t = \tau_{1,s} z_{t-1} + \tau_2 \widetilde{x}_s + (1 - \tau_{1,s} - \tau_2) y_{t-1}$
        $\hat{\nabla}_t = \mu_s + \nabla f_I(x_t) - \nabla f_I(\widetilde{x}_s)$
        $y_t = argmin_y \{\frac{3L}{2}||y - x_t||^2 + \langle \hat{\nabla}_t, y \rangle + \phi(y)\}$
        $z_t = argmin_z \{\frac{1}{2\eta_s}||z - z_t||^2 + \langle \hat{\nabla}_t, z \rangle + \phi(z)\}$
    end for
    $\widetilde{x}_{s+1} = \frac{1}{m} \sum_{t=1}^{m} y_t$
end for
return $\widetilde{x}_S$

The algorithm reflects many of the similarities of the linear coupling algorithm in how $\tau_{1,s}$ and η_s decay, allowing the steps to become variance reduced gradient descents overtime. Also note that when the proximal function $\phi(x)$ is zero, the update steps for z_t and y_t become variance reduced mirror descent and gradient descent steps respectively. In the strongly convex setting, the only difference is that τ_1 and η are set in the beginning and do not converge over time. Additionally, the strongly convex setting, takes into account σ when calculating snapshot vector $\widetilde{x}_{s+1}$.

6.2 Convergence

Theorem 18 (Allen-Zhu [8])**.** *Assume each $f_i(x)$ is convex and L-smooth. Let $F(x_0) - F(x^*) = \Theta$ and $||x_0 - x^*|| = D$. Then Katyusha needs $T = O(\frac{n\sqrt{\Theta} + \sqrt{nLD}}{\sqrt{\epsilon}})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

Proof Sketch. We can think of $\widetilde{x}$ as the snapshot parameter, y_t as the variance reduced gradient step, and z_t as the variant reduced mirror step. In proving the theorem, Allen-Zhu first proves five lemmas, and then ends with the final proof. The entire proof is several pages long, so we will provide a conceptual proof sketch for conciseness.

- We first use the L-Smooth property to show that, on expectation, y_t will always decrease $F(x)$ during a given iteration. The proof resembles any proof of gradient descent convergence, but now includes the proximal function term.
- We then upper-bound the variance $E[||\hat{\nabla}_t - \nabla f(x_t)||]$ as tightly as possible. This is to show the variance at any given step is never too large.
- From the mirror descent step for z_t, we derive a variant of the Mirror Descent Lemma for any $x \in K$:

$$\begin{aligned} \langle \hat{\nabla}_t, z_t - x \rangle + \eta\phi(z_t) + \eta\phi(x) &\leq \frac{-1}{2}||z_{t-1} - z_t||^2 && (1) \\ &+ \frac{1}{2}||z_{t-1} - x||^2 - \frac{1}{2}||z_t - x||^2 && (2) \end{aligned}$$

- Combine the previous three lemmas and the definition of x_t to bound the term $\eta_s \langle \nabla f(x_t), z_{t-1} - x \rangle - \eta_s \phi(x)$ from above.
- Using the parameters $\tau_{1,s}$ and τ_2, we can then establish the following inequality for a given iteration of Katyusha (see Appendix B for why this differs from Allen-Zhu's paper):

$$\begin{aligned} 0 &\leq \frac{\eta_s(1 - \tau_{1,s} - \tau_2)}{\tau_{1,s}} D_{t-1} - \frac{\eta_s}{\tau_{1,s}} E[D_t] + \frac{\eta_s \tau_2}{\tau_{1,s}} \widetilde{D}_s && (3) \\ &+ \frac{1}{2}||z_{t-1} - x^*||^2 - \frac{1}{2}\mathbb{E}[||z_t - x^*||^2] && (4) \end{aligned}$$

  where $D_t = F(y_t) - F(x^*)$ and $\widetilde{D}_s = F(\widetilde{x}_s) - F(x^*)$. This inequality gives us a relationship between the progress toward $F(x^*)$ in the objective and our distance from x^*. We see that every iteration, either the mirror descent step z_t brings us closer to x^* or the objective $F(x^*)$ decreases, subject to error proportional to $\widetilde{D}_s$ (which is always decreasing).
- Telescoping this inequality across all iterations $t = 1 \dots T$ and $s = 1 \dots S$ and simplifying, we get the following inequality:

$$\begin{aligned} \mathbb{E}[F(\widetilde{x}_S) - F(x^*)] &\leq O(\frac{\tau_{1,S}}{m})(\frac{1 - \tau_{1,0} - \tau_2}{\tau_{1,0}^2}(F(x_0) - F(x^*)) && (5) \\ &+ \frac{n\tau_2}{\tau_{1,0}^2}(F(\widetilde{x}_0) - F(x^*)) + \frac{3L}{2}||z_0 - x^*||^2) && (6) \\ &= O(\frac{1}{m^2 S^2})(m^2\Theta + Lm||z_0 - x^*||^2) && (7) \\ &= O(\frac{1}{T^2})(n^2\Theta + LnD^2) && (8) \end{aligned}$$

Thus, after $T = O(\frac{n\sqrt{\Theta} + \sqrt{nLD}}{\sqrt{\epsilon}})$ iterations, we have $\mathbb{E}[F(\widetilde{x}_S) - F(x^*)] \leq \epsilon$, and the theorem is proved [8].

□

There is also a proof for the strongly convex case, which with similar analysis proves the following bound:

Theorem 19 (Allen-Zhu [8]). *Assume each $f_i(x)$ is convex, L-smooth, and the proximal function $\phi(x)$ is σ-strongly convex. Let $F(x_0) - F(x^*) = \Theta$. Then Katyusha needs $T = O((n + \sqrt{\frac{nL}{\sigma}}) \log \frac{\Theta}{\epsilon})$ iterations to achieve $f(x_T) - f(x^*) \leq \epsilon$.*

7 Appendix

7.1 A.

A step-by-step proof using the Mirror Descent Lemma which shows $T\eta(f(\bar{x}) - f(x^*)) \leq T\frac{\rho^2\eta^2}{2} + D$. We start with the Mirror Descent Lemma:

$$\eta(f(x_t) - f(x)) \leq \eta\langle \nabla f(x_t), x_t - x\rangle \leq \frac{\eta^2}{2}||\nabla f(x_t)||^2 + B(x_t, x) - B(x_{t+1}, x)$$

Let $x = x^*$ and $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$. From $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$:

$$\eta T(f(\bar{x}) - f(x^*)) = \eta T(f(\frac{1}{T}\sum_{t=0}^{T-1} x_t) - f(x^*))$$

Definition of convexity:

$$\eta T(f(\frac{1}{T}\sum_{t=0}^{T-1} x_t) - f(x^*)) \leq \eta T(\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*))$$

First inequality from the lemma:

$$\begin{aligned}\eta T(\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*)) &\leq \eta T\langle \frac{1}{T}\sum_{t=0}^{T-1} \nabla f(x_t), x_t - x^*\rangle \\ &= \sum_{t=0}^{T-1} \eta\langle \nabla f(x_t), x_t - x^*\rangle\end{aligned}$$

Second inequality from the lemma:

$$\eta T(\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*)) \leq \sum_{t=0}^{T-1}(\frac{\eta^2}{2}||\nabla f(x_t)||^2 + B(x_t, x^*) - B(x_{t+1}))$$

Telescoping the right-hand side:

$$\eta T(\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*)) \leq \sum_{t=0}^{T-1} \frac{\eta^2}{2}||\nabla f(x_t)||^2 + B(x_0, x^*) - B(x_T, x^*)$$

From f being ρ-Lipschitz:

$$\eta T(f(\bar{x}) - f(x^*)) \leq T\frac{\rho^2\eta^2}{2} + B(x_0, x^*) - B(x_T, x^*)$$

Letting $D = B(x_0, x^*)$:

$$\eta T(f(\bar{x}) - f(x^*)) \leq T\frac{\rho^2\eta^2}{2} + D$$

7.2 B.

Zeyuan Allen-Zhu's paper currently shows the following inequality for one round of Katyusha:

$$\begin{aligned} 0 &\leq \frac{\eta_s(1-\tau_{1,s}-\tau_2)}{\tau_{1,s}} D_{t-1} - \frac{\eta_s}{\tau_{1,s}} E[D_t] + \frac{m\eta_s\tau_2}{\tau_{1,s}}\widetilde{D}_s \quad (1) \\ &+ \frac{1}{2}||z_{t-1} - x^*||^2 - \frac{1}{2}\mathbb{E}[||z_t - x^*||^2] \quad (2) \end{aligned}$$

Notice the extra coefficient m on the $\widetilde{D}_s$ term. This m should only be present after telescoping over $t = 1 \dots m$. Professor Allen-Zhu confirmed this was a typo after I emailed him.

References

[1] Nisheeth K. Vishnoi. A mini-course on convex optimization: with a view toward designing fast algorithms, 2014.

[2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

[3] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience series in discrete mathematics. Wiley,, 1983.

[4] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, May 2003.

[5] Zeyuan Allen-Zhu and Lorenzo Orecchia. A novel, simple interpretation of nesterov's accelerated method as a combination of gradient and mirror descent. *CoRR*, abs/1407.1537, 2014.

[6] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[7] Léon Bottou. Stochastic gradient learning in neural networks. In *In Proceedings of Neuro-Nîmes. EC2*, 1991.

[8] Zeyuan Allen-Zhu. Katyusha: The First Direct Acceleration of Stochastic Gradient Methods. In *STOC*, 2017.

[9] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 315–323, USA, 2013. Curran Associates Inc.