MS&E 213 / CS 269O : Chapter 4 - Acceleration*

By Aaron Sidford (sidford@stanford.edu)

October 31, 2020

In the last chapter we proved the following result about gradient descent for minimizing L-smooth μ-strongly convex functions.

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu \geq 0$. Let $x_0 \in \mathbb{R}^n$ and $x_* \in X_*(f)$ be arbitrary and let $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ for all $k \geq 0$. Then*

$$f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f_*] \, , \, \frac{L \cdot \|x_0 - x_*\|_2^2}{k+4}\right\}.$$

Consequently we can compute an ϵ-optimal point with

$$O\left(\min\left\{\frac{L}{\mu}\log\left(\frac{f(x_0) - f_*}{\epsilon}\right), \frac{L\|x_0 - x_*\|_2^2}{\epsilon}\right\}\right)$$

queries to a gradient oracle.

A natural question to ask, is is this optimal? If all we have is a gradient oracle, can we design an algorithm with improved query complexity? Here we address this question, showing how improved (optimal in some cases) query complexities through a technique often referred to as *acceleration*.

Acceleration is powerful and somewhat mysteriouss technique in optimization. Optimal rates for minimizing smooth convex functions were first achieved by Nesterov in 1983 [2] and has been shown to be quite powerful in obtaining faster methods in theory and in practice. Consequently, there are a number of perspectives on acceleration one could take, e.g. viewing it as a type of momentum based method, derviting it from continuous time dynamics, explaining the method as a careful combination of gradient descent and mirror descent [1] (an algorithm we will discuss later in the class). There are even interesting geometric views of the method, and varied potential based approaches to analyzing it, e.g. estimate sequences, and there have been attempts to explain the method as primal-dual.

Though there are many different approaches to obtaining improved, accelerated query complexities for minimizing smooth convex functions, they typically share a few common characterisitcs. They maintain more than a single point as the state of the algorithm and use more than just function error of the current point to analyze progress. In this chapter we will provide one particular principled derivation of accelerated smooth-convex minimization algorithms, simple potential function based analysis, and relate the methods to the acceleration.

This chapter only covers a few of the perspectives on acceleration; if you would like more references on any of the others, just let me know.

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material, there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future, and citations are often omitted. These notes are intended converge to a superset of the class material that is TODO-free with a more complete set of citations and pointers to the literature. Your feedback is welcome and highly encouraged. If anything is unclear, you find a bug or typo, or if you would find it particularly helpful for anything to be expanded upon, please do not hesitate to post a question on the Piazza or contact me directly at sidford@stanford.edu.

1 Lower Bounds for Faster Methods

So how should we improve upon gradient descent for smooth, convex function minimization? Recall that $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex if and only if all $x \in \mathbb{R}^n$ the induced upper bound, $L_x : \mathbb{R}^n \to \mathbb{R}$, and lower bound, $L_y : \mathbb{R}^n \to \mathbb{R}$, defined for all $y \in \mathbb{R}^n$ by

$$L_x(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2 \text{ and } U_x(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$$

satisfy

$$L_x(y) \leq f(y) \leq U_x(y) \text{ for all } y \in \mathbb{R}^n .$$

We saw that gradient descent was derived from just the upper bound implied by smoothness, U_x, as $x_{k+1} = \text{argmin}_x U_{x_k}(x) = x_k - \frac{1}{L}\nabla f(x_k)$. Although, strong convexity and the induced lower bounds, L_x, were used in the analysis of gradient descent, it was used in the derivatoin of the algorithm. Consequently, a natural idea to improve upon gradient descent is to leverage the lower-bounds induced by convexity.

Expanding upon this idea, we first focus on deriving an algorithm in case of strongly convex functions, $f : \mathbb{R}^n \to \mathbb{R}$. To leverage the lower-bounds induced by convexity, our algorithm will maintain both a point x_k with a function value we wish to decrease (as in gradient descent) and a lower bound function, $L_k : \mathbb{R}^n \to \mathbb{R}$, with the property that $L_k(x) \leq f(x)$ for all x. In every step of the algorithm our algorithm will seek to decrease the difference between the value of x_k and the minimum value of the lower bound L_k, i.e. $f(x_k) - \min_x L_k(x)$. Not that it suffices to decrease this quantity as $f(x_k) - \min_x L_k(x) \leq f(x_k) - f_*$ by the fact that $\min_x L_k(x) \leq L_k(x_*) \leq f(x_*) = f_*$.

With this approach in mind, what lower bound $L_k(x)$ should we use? We know that every query to a gradient oracle at a point x uses a lower bound L_x and one natural idea would be simply to let L_k be the maximum of all the lower bound functions we have computed, i.e. $L_k(x) \stackrel{\text{def}}{=} \max_{i=[k]} L_{x_i}(x)$ where $x_1, ..., x_k \in \mathbb{R}^n$ are the points at which the gradient oracle has been queried before. While this idea could possibly be made to work, the resulting lower bound function could be complex as k increases complicating both the derivation and analysis of such a method.

Instead, our accelerated algorithm will simply pick $L_k(x)$ to be simple quadratics, i.e. $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$. We pick such L_k as every L_x is of this form and such functions are closed under convex combinations, i.e. for any $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$, $L'_k(x) = \psi'_k + \frac{\mu}{2}\|x - v'_k\|_2^2$, and $\alpha \in [0, 1]$ it is the case that $\alpha L_k(x) + (1 - \alpha)L'_k(x) = \psi^\alpha_k + \frac{\mu}{2}\|x - v^\alpha_k\|_2^2$ for some ψ^α_k and v^α_k. Consequently, we can obtain and iterative combine such lower bounds with a gradient oracle. In the remainder of this section we prove some basic facts about such functions and then in the next section use them to more precisely derive an accelerated method for minimizing strongly convex functions.

First we give the a simple lemma characterizing quadratics.

Lemma 2 (Characterizing Quadratics)**.** *Twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ satisfies $\nabla^2 f(x) = \mathbf{A} \in \mathbb{R}^{n \times n}$ for all $x \in \mathbb{R}^n$ if and only if*

$$f(x) = f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2}(y - x)^\top \mathbf{A}(y - x) \text{ for all } x, y \in \mathbb{R}^n .$$

Consequently, for such functions if $x_* \in \mathbb{R}^n$ satisfies $\nabla f(x_*) = 0$ and $\|z\|_{\mathbf{A}} \stackrel{\text{def}}{=} \sqrt{z^\top \mathbf{A} z}$ for all z then

$$f(x) = f(x_*) + \frac{1}{2}\|x - x_*\|_{\mathbf{A}}^2 \text{ for all } x \in \mathbb{R}^n$$

Proof. Suppose $\nabla^2 f(x) = \mathbf{A} \in \mathbb{R}^{n \times n}$ for all $x \in \mathbb{R}^n$. Let $x_t = x + t(y - x)$ for all $t \in [0, 1]$. We have shown that

$$f(x) - f(y) - \nabla f(x)^\top (y - x) = \int_0^1 \int_0^t (y - x)\nabla^2 f(x_\alpha)(y - x) d\alpha dt = \int_0^1 \int_0^t (y - x)^\top \mathbf{A}(y - x) d\alpha dt$$

Since $\int_0^1 \int_0^t d\alpha dt = \int_0^1 t dt = \frac{1}{2}t^2|_0^1 = \frac{1}{2}$ the remaining results follow immediately. □

This lemma shows that we can re-write our lower bounds as quadratics centered around a particular point and obtain simple formulas for combining quadratics. (Note that these lemmas, and that fact, could also be proven by direct calculation.)

Lemma 3. *For differentiable $f : \mathbb{R}^n \to \mathbb{R}$, $L \geq 0$, and $y \in \mathbb{R}^n$ define $L_y : \mathbb{R}^n \to \mathbb{R}$ for all $x \in \mathbb{R}^n$ by*

$$L_y(x) = f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2 .$$

Then for all $x \in \mathbb{R}^n$ we have

$$L_y(x) = \psi_x + \frac{\mu}{2}\|y - v_x\|_2^2$$

where

$$\psi_x = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \text{ and } v_x = x - \frac{1}{\mu}\nabla f(x) .$$

Proof. Since,

$$L_x\left(x - \frac{1}{\mu}\nabla f(x)\right) = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2 ,$$
$$\nabla L_x\left(x - \frac{1}{\mu}\nabla f(x)\right) = \nabla f(x) - \frac{\mu}{\mu}\nabla f(x) = \vec{0} , \text{ and}$$
$$\nabla^2 L_x(z) = \mu \mathbf{I} \text{ for all } z \in \mathbb{R}^n$$

the result follows from Lemma 2. □

Next, we analyze combining quadratics. This lemma shows that when adding an α multiple of one quadratic to a $1 - \alpha$ multiple of another (with the same hessian), then we obtain a new quadratic with a center that interpolates linearly and with a minimum value that interpolates quadratically (e.g. the farther away the centers are, the more that combining can increases the lower bound).

Lemma 4. *Let $f_0, f_1 : \mathbb{R}^n \to \mathbb{R}$ be defined for all $x \in \mathbb{R}^n$ by*

$$f_0(x) \stackrel{\text{def}}{=} \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2 \quad \text{and} \quad f_1(x) = \psi_1 + \frac{\mu}{2}\|x - v_1\|_2^2$$

For $\psi_0, \psi_1 \in \mathbb{R}$, $v_0, v_1 \in \mathbb{R}^n$, and $\mu \geq 0$. Then for all $\alpha \in [0, 1]$ and $x \in \mathbb{R}^n$ we have

$$f_\alpha(x) \stackrel{\text{def}}{=} \alpha \cdot f_0(x) + (1 - \alpha) \cdot f_1(x) = \psi_\alpha + \frac{\mu}{2}\|x - v_\alpha\|_2^2$$

where

$$v_\alpha = \alpha v_0 + (1 - \alpha)v_1 \text{ and } \psi_\alpha = \alpha\psi_0 + (1 - \alpha)\psi_1 + \frac{\mu}{2}\alpha(1 - \alpha)\|v_1 - v_0\|_2^2 .$$

Proof. Note that

$$\nabla f_\alpha(x) = \alpha \cdot \mu \cdot (x - v_0) + (1 - \alpha) \cdot \mu(x - v_1) .$$

Since $v_\alpha - v_0 = (1 - \alpha)(v_1 - v_0)$ and $v_\alpha - v_1 = \alpha(v_0 - v_1)$ we have that

$$\nabla f_\alpha(v_\alpha) = \alpha \cdot \mu \cdot (1 - \alpha) \cdot (v_1 - v_0) + (1 - \alpha) \cdot \mu \cdot \alpha \cdot (v_0 - v_1) = 0 .$$

Further, this implies that

$$\begin{aligned} f_\alpha(v_\alpha) &= \alpha\left[\psi_0 + \frac{\mu}{2}\|v_\alpha - v_0\|_2^2\right] + (1-\alpha)\left[\psi_1 + \frac{\mu}{2}\|v_\alpha - v_1\|_2^2\right] \\ &= \alpha\psi_0 + (1-\alpha)\psi_1 + \frac{\mu}{2}\left[\alpha(1-\alpha)^2\|v_1 - v_0\|_2^2 + (1-\alpha)\alpha^2\|v_1 - v_0\|_2^2\right] \\ &= \alpha\psi_0 + (1-\alpha)\psi_1 + \frac{\mu}{2}\alpha(1-\alpha)\|v_1 - v_0\|_2^2 \end{aligned}$$

Since $\nabla^2 f_\alpha(x) = \alpha \cdot \mu \cdot \mathbf{I} + (1-\alpha)\cdot\mu\cdot\mathbf{I} = \mu\cdot\mathbf{I}$ for all $x \in \mathbb{R}^n$ the result then follows from Lemma 2. □

2 Accelerating Smooth Strongly Convex Minimization

As discussed in the previous section, our first acceleration approach is as follows. In every iteration k we maintain some $x_k \in \mathbb{R}^n$ and some lower bound function $L_k : \mathbb{R}^n \to \mathbb{R}$ such that for all $x \in \mathbb{R}^n$ we have $L_k(x) \le f(x)$. We restrict ourselves to L_k of the form, $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$ for some $\psi_k \in \mathbb{R}$ and some $v_k \in \mathbb{R}^n$. Now, since clearly $\min_{x\in\mathbb{R}^n} L_k(x) = \psi_k$ we know that

$$f(x_k) - \psi_k \ge f(x_k) - \min_x f(x) = f(x_k) - f_* \,.$$

Consequently, it suffices to show we can decrease $f(x_k) - \psi_k$ with few calls to a gradient oracle.

The way we do this is simple. We pick some $\alpha \in [0,1]$ and let

$$y_k = \alpha \cdot x_k + (1-\alpha)\cdot v_k \,.$$

We se this point y_k to improve both our lower bound and upper bound. To improve the upper bound we take a gradient descent step and let $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$ and to update the lower bound, we take a convex combination of the old lower bound L_k and the lower bound from the point y_k, i.e. L_{y_k}. Formally we pick $\beta \in [0,1]$ and let $L_{k+1}(y) \stackrel{\text{def}}{=} \beta \cdot L_k(y) + (1-\beta)\cdot L_{y_k}(y)$ for all y.

That is the entire algorithm and the ultimate pseudocode for it is fairly short. We pick our lower bounds $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$ as this update rule keeps the L_k of this form and we can store these L_k compactly and consequently, this method is easy to implement.

What is tricky about this algorithm and the analysis of it, is reasoning about exactly what happens when we combine lower bounds. In the following lemma we analyze the change in L_k through a self contained helper lemma.

Lemma 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable μ-strongly convex function and let $L_k(x) \stackrel{\text{def}}{=} \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$ be such that $f(x) \ge L_k(x)$ for all x. Then for all $\beta \in [0,1]$ and $y_k \in \mathbb{R}^n$ we have that*

$$L_{k+1}(x) = \beta \cdot L_k(x) + (1-\beta)\cdot L_{y_k}(x)$$

where $L_{y_k}(x) = f(y_k) + \nabla f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$, satisfies $f(x) \ge L_{k+1}(x)$ for all $x \in \mathbb{R}^n$ and

$$L_{k+1}(x) = \psi_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|_2^2$$

where

$$\begin{aligned} v_{k+1} &= \beta \cdot v_k + (1-\beta)\cdot\left[y_k - \frac{1}{\mu}\nabla f(y_k)\right] \text{ and} \\ \psi_{k+1} &= \beta\cdot\psi_k + (1-\beta)\cdot\left[f(y_k) - \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2\right] + \frac{\mu}{2}\beta(1-\beta)\cdot\left\|v_k - \left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]\right\|_2^2 . \end{aligned}$$

Proof. Since f is μ-strongly convex, for all $x \in \mathbb{R}^n$ we know that $f(x) \geq L_{y_k}(x)$ and therefore

$$L_{k+1}(x) = \beta \cdot L_k(x) + (1-\beta) \cdot L_{y_k}(x) \leq \beta \cdot f(x) + (1-\beta) \cdot f(x) = f(x)\,.$$

Further, by Lemma 3 we know that

$$L_{y_k}(x) = \psi_{y_k} + \frac{\mu}{2}\|x - v_{y_k}\|_2^2 \text{ for all } x \in \mathbb{R}^n$$

where

$$\psi_{y_k} = f(y_k) - \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2 \text{ and } v_{y_k} = y_k - \frac{1}{\mu}\nabla f(y_k)\,.$$

Consequently, the result follows from Lemma .4. □

Using this we can analyze the combination of a gradient descent step from y_k and the combining of lower bounds.

Lemma 6. *Under the same assumptions of Lemma 5 if $y_k = \alpha \cdot x_k + (1-\alpha) \cdot v_k$ for some $\alpha \in (0,1)$ and $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$ then we have*

$$f(x_{k+1}) - \psi_{k+1} \leq \beta\left[f(x_k) - \psi_k\right] + \beta \cdot \left[1 - \alpha \cdot \frac{1-\beta}{1-\alpha}\right](f(y_k) - f(x_k)) + \left[\frac{(1-\beta)^2}{2\mu} - \frac{1}{2L}\right]\|\nabla f(y_k)\|_2^2\,.$$

Consequently if $\kappa = \frac{L}{\mu}$, $\beta = 1 - \sqrt{\frac{1}{\kappa}}$, and $\alpha = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$ then

$$f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left[f(x_k) - \psi_k\right]$$

Proof. Now by our assumption on y_k we have

$$v_k - y_k = \frac{1}{1-\alpha}\left[y_k - \alpha \cdot x_k\right] - y_k = \frac{\alpha}{1-\alpha}\left[y_k - x_k\right]\,.$$

Since $f(x_k) \geq f(y_k) + \nabla f(y_k)^\top (x_k - y_k)$ by convexity and $\|v_k - y_k\|_2^2 \geq 0$ trivially we have

$$\begin{aligned}
\frac{\mu}{2}\left\|v_k - \left[y_k - \frac{1}{\mu}\nabla f(y)\right]\right\|_2^2 &= \frac{\mu}{2}\left[\|v_k - y_k\|_2^2 + \frac{2}{\mu}\nabla f(y_k)^\top (v_k - y_k) + \frac{1}{\mu^2}\|\nabla f(y_k)\|_2^2\right] \\
&\geq \frac{\alpha}{1-\alpha} \cdot \nabla f(y_k)^\top (y_k - x_k) + \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2 \\
&\geq \frac{\alpha}{1-\alpha} \cdot [f(y_k) - f(x_k)] + \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2\,.
\end{aligned}$$

Consequently, by Lemma 5 we have

$$\psi_{k+1} \geq \beta \cdot \psi_k + (1-\beta) \cdot \left[f(y_k) - \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2\right] + \beta(1-\beta) \cdot \left[\frac{\alpha}{1-\alpha} \cdot [f(y_k) - f(x_k)] + \frac{1}{2\mu}\|\nabla f(y_k)\|_2^2\right]$$

Combining this with the fact that $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$ yields

$$\begin{aligned}
f(x_{k+1}) - \psi_{k+1} \leq \beta\alpha \cdot \frac{1-\beta}{1-\alpha} \cdot f(x_k) - \beta\psi_k + \left[1 - (1-\beta) - \alpha\beta \cdot \frac{1-\beta}{1-\alpha}\right] f(y_k) \\
+ \left[\frac{1-\beta}{2\mu} - \beta \cdot (1-\beta) \cdot \frac{1}{2\mu} - \frac{1}{2L}\right]\|\nabla f(y_k)\|_2^2
\end{aligned}$$

This yields the first formula. The values for β and α were chosen by solving for $(1-\beta)^2 = \frac{\mu}{L} = \frac{1}{\kappa}$ yielding the first formula and then solving for $\alpha \cdot \frac{1-\beta}{1-\alpha} = 1$ which yields that $\frac{\alpha}{\sqrt{\kappa}} = 1 - \alpha$ which then yields $\alpha = \frac{1}{1+\frac{1}{\sqrt{\kappa}}} = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$. □

The preceding lemma shows how to construct a step that decreased an upper bound on $f(x_k) - f_*$ by a multiplicative $1 - \sqrt{\frac{\mu}{L}}$ in every iteration. To turn this into a full algorithm, all that remains is to show how to bound the initial error, i.e. how to get an initial quadratic lower bound on our function. However, by Lemma 3 we already know that how to get a lower bound, so all that remains is to analyze the initial error with this lower bound.

Lemma 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function then for any x_0 we have that for*

$$\psi_0 = f(x_0) - \frac{1}{2\mu}\|\nabla f(x_0)\|_2^2 \text{ and } v_0 = x_0 - \frac{1}{\mu}\nabla f(x_0)$$

it is the case that $f(x) \geq L_0(x) \stackrel{\text{def}}{=} \psi_0 + \frac{\mu}{2}\|x - v_0\|_2^2$ and $f(x_0) - \psi_0 \leq \frac{L}{\mu} \cdot [f(x_0) - f_*]$.

Proof. The fact that $f(x) \geq L_0(x)$ is immediate from Lemma 3. Since $\|\nabla f(x_0)\|_2^2 \leq 2L \cdot [f(x_0) - f_*]$ we obtain the desired upper bound on $f(x_0) - \psi_0$. □

Putting this all together yields the following. Note that although our algorithm was motivated by maintaining L_k, we in fact only need to maintain v_k in the algorithm.

Theorem 8 (An Accelerated Gradient Descent Method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and let $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$. For arbitrary $x_0 \in \mathbb{R}^n$ compute $v_0 = x_0 - \frac{1}{\mu}\nabla f(x_0)$ and for all $k \geq 0$ let*

- $y_k = \alpha \cdot x_k + (1 - \alpha) \cdot v_k$ *for* $\alpha = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$
- $v_{k+1} = \beta \cdot v_k + (1 - \beta) \cdot \left[y_k - \frac{1}{\mu}\nabla f(y_k)\right]$ *for* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Then we have that $f(x_k) - f_* \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \cdot \kappa \cdot [f(x_0) - f_*]$ and consequently we can compute an ϵ-optimal point for f with $1 + \lceil \sqrt{\kappa} \log(\kappa \cdot [f(x_0) - f_*]/\epsilon) \rceil$ queries to a gradient oracle.

Proof. For all k if we let $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$ then we have by previous lemmas that there is a way to choose the ψ_k such that $f(x) \geq L_k(x)$ for all x and therefore $f(x_k) - f_* \leq f(x_k) - \psi_k$. We have also proven that this can be done so that $f(x_{k+1}) - \psi_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)[f(x_k) - \psi_k]$ and $f(x_0) - \psi_0 \leq \kappa \cdot [f(x_0) - f_*]$ yielding the result. □

A natural question to ask is can accelerated gradient descent algorithm be further improved? It can be shown that the dependence on κ in the asymptotic rate cannot, in general, be improved if the function can be accessed only through a first-order oracle and there is no dependence on dimension.

However, the κ in the logarithmic term can be improved just by slightly improving the analysis. Careful inspection of the analysis in this section shows that in fact it suffices for $L_k(x_*) \leq f(x_*)$, i.e. $L_k(x) \leq f(x)$ for all $x \in \mathbb{R}^n$ is not required. More formally, it can be shown that if $L_0(x_*) \leq f(x_*)$ then for all k we have $L_k(x_*) \leq f(x_*)$ and again

$$f(x_k) - f_* \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k [f(x_0) - \psi_0] .$$

Further, since $L_0(x) \stackrel{\text{def}}{=} \psi_0 + \frac{\mu}{2}\|v_0 - x\|_2^2$ for $v_0 = x_0$ and $\psi_0 = f_* - \frac{\mu}{2}\|x_0 - x_*\|_2^2$ trivially has the property that $L_0(x_*) \leq f(x_*)$ and $f(x_0) - \psi_0 = f(x_0) - f_* + \frac{\mu}{2}\|x_0 - x_*\|_2^2 \leq 2[f(x_0) - f_*]$ by strong convexity we see that if in Theorem 8 we simply set $v_0 = x_0$ then the κ factor can be changed to a 2.

This analysis can be strengthened and shown more directly simply by changing the potential function used to analyze progress. Rather than analyzing $f(x_k) - \psi_k$ it can be shown that $f(x_k) - f_* + \frac{\mu}{2}\|v_0 - x_*\|_2^2$ also decreases by a multiplicative $1 - \kappa^{-1/2}$ in each iteration. We how this in Section 4.

3 Non-strongly Convex Functions

How can we use the above result to minimize non-strongly convex functions? There is a fairly general trick to reduce non-strongly convex function minimization to strongly convex function minimization and that is *regularization*. This is a commonly used term with multiple applications and interpretations. When we use this term in the class, we will simply use it to refer to the idea of adding a simple function we understand to improve the behavior of our iterative methods.

The idea we use here is simple. Instead of minimizing $f(x)$ directly, given some point we just minimize $g(x) = f(x) + \frac{\mu}{2}\|x - x_0\|_2^2$. Clearly this function is μ strongly convex and thus we can apply accelerated gradient descent as analyzed above to it. Below we analyze the performance of this scheme.

Lemma 9. *If f is a L-smooth convex function then given any $x_0 \in \mathbb{R}^n$ we can compute an ϵ-optimal point with*

$$\left\lceil \sqrt{1 + \frac{L \cdot \|x_0 - x_*\|_2^2}{\epsilon}} \log\left(\frac{L \cdot \|x_0 - x_*\|_2^2}{\epsilon}\right) \right\rceil$$

queries for any $x_* \in X_*(f)$.

Proof. Given $x_0 \in \mathbb{R}^n$ we run accelerated gradient descent to minimize $g(x) = f(x) + \frac{\mu}{2}\|x - x_0\|_2^2$ for a value of μ we pick later. Since g is μ-strongly convex and $L + \mu$ smooth we know that by accelerated gradient descent we can compute an ϵ-sub-optimal point, denoted x_ϵ, for g with $\lceil \sqrt{\frac{L+\mu}{\mu}} \log\left(2 \cdot [g(x_0) - g_*]/\epsilon\right) \rceil$ gradient queries.

Now, since $g(x) \geq f(x)$ for all x we have $g_* \geq f_*$. Furthermore, since $g(x_0) = f(x_0)$ we have

$$g(x_0) - g_* \leq f(x_0) - f_* \leq \frac{L}{2} \cdot \|x_0 - x_*\|_2^2 .$$

Furthermore, by the definition of g and x_ϵ if we let x_f denote a minimizer of f we have that

$$f(x_\epsilon) \leq \epsilon + \min_{x \in \mathbb{R}^n} g(x) \leq \epsilon + g(x_f) = \epsilon + f_* + \frac{\mu}{2}\|x_0 - x_*\|_2^2 .$$

Consequently, computing $x_{\frac{1}{2}\epsilon}$ for $\mu = \frac{\epsilon}{\|x_0 - x_*\|_2^2}$ yields the desired result. □

There are two natural ways to remove the log factor in the above analysis. The first is to change the accelerated gradient descent algorithm itself and the second is to minimize $f(x) + \frac{\mu}{2}\|x - x_0\|_2^2$ in phases changing perhaps what the regularization is with respect to. Both can be used to remove the logarithmic factors. The first has the advantage of perhaps being a more natural way of running the algorithm, but the second has the virtue of being a fairly general reduction. In the next section we provide such a direct accelerated method.

4 Direct Potential Based Proofs of Optimal Dimension-Free Rates

Here we show how to strengthen then analysis of the preceding sections providing method which we can directly analyze to show they achieve the optimal dimension free rates for minimizing smooth strongly convex functions. The main result of this section is the following.

Theorem 10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu \geq 0$. Let $x_0 \in \mathbb{R}^n$ and $x_* \in X_*(f)$ be arbitrary. There is an algorithm which with k queries to a gradient oracle outputs a point x_k with*

$$f(x_k) - f_* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k 2\left[f(x_0) - f_*\right]$$

and method which outputs a point x_k with

$$f(x_k) - f_* \leq \frac{L\|x_0 - x_*\|_2^2}{2k^2}\,.$$

Consequently, it is possible to compute an ϵ-optimal point with

$$O\left(\min\left\{\sqrt{\frac{L}{\mu}}\log\left(\frac{f(x_0)-f_*}{\epsilon}\right),\sqrt{\frac{L\|x_0-x_*\|_2^2}{\epsilon}}\right\}\right)$$

queries to a gradient oracle.

We prove this theorem by analyzing a slightly generalized variant of the accelerated algorithms we have seen earlier in this section. The accelerated method we saw earlier maintained two points, $x_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^n$, combined them to obtain a new point $y_k = \alpha_k \cdot x_k + (1-\alpha_k)\cdot v_k$. Then let v_{k+1} be a combination of v_k, y_k, and the gradient at y_k, i.e. $v_{k+1} = \beta_k \cdot v_k + (1-\beta_k)\cdot y_k - \eta_k \nabla f(y_k)$, and let x_{k+1} be the result of a gradient descent step from y_k, i.e. $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$. To derive our accelerated methods we analyze one step of this method for arbitrary α_k, β_k, and γ_k. In particular we analyze the effect of this method on the function error of x_k, i.e. $\epsilon_k = f(x_k) - f_*$ and the squared, scaled distances from v_k to a minimizer x_*, i.e. $r_k \stackrel{\text{def}}{=} \frac{1}{2}\|v_k - x_*\|_2^2$.

Lemma 11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function for $\mu \geq 0$ and suppose that for some x_k and v_k we let*

- $y_k = \alpha_k \cdot x_k + (1-\alpha_k)\cdot v_k$
- $v_{k+1} = \beta_k \cdot v_k + (1-\beta_k)\cdot y_k - \eta_k \nabla f(y_k)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Then if x_* is a minimizer of f and we let $\epsilon_k \stackrel{\text{def}}{=} f(x_k) - f_*$ and $r_k \stackrel{\text{def}}{=} \frac{1}{2}\|v_k - x_*\|_2^2$ then we have

$$\eta_k^2 L \cdot \epsilon_{k+1} + r_{k+1} \leq \beta_k \cdot r_k + \frac{\alpha_k\beta_k}{1-\alpha_k}\cdot\eta_k\cdot\epsilon_k + E_k$$

where

$$E_k = \left[(1-\beta_k) - \mu\cdot\eta_k\right] r_k^y + \eta_k\left[\eta_k L - 1 - \frac{\alpha_k\beta_k}{1-\alpha_k}\right]\epsilon_k^y\,.$$

Consequently, if $\mu\cdot\eta_k = (1-\beta_k)$ and $\eta_k L = 1 + \frac{\alpha_k\beta_k}{1-\alpha_k} = \frac{1-(1-\beta_k)\alpha_k}{1-\alpha_k}$ then

$$\eta_k^2 L\cdot\epsilon_{k+1} + r_{k+1} \leq \beta_k\cdot r_k + (\eta_k L - 1)\cdot\eta_k\cdot\epsilon_k\,.$$

Proof. Letting $z_k \stackrel{\text{def}}{=} \beta_k\cdot v_k + (1-\beta_k)\cdot y_k$ we see that

$$\begin{aligned} r_{k+1} &= \frac{1}{2}\|z_k - x_* - \eta_k\nabla f(y_k)\|_2^2 \\ &= \frac{1}{2}\|z_k - x_*\|_2^2 - \eta_k\nabla f(y_k)^\top(z_k - x_*) + \frac{\eta_k^2}{2}\|\nabla f(y_k)\|_2^2\,. \end{aligned}$$

Now, letting $\epsilon_k^y \stackrel{\text{def}}{=} f(y_k) - f_*$ and $r_k^y \stackrel{\text{def}}{=} \frac{1}{2}\|y_k - x_*\|_2^2$ we see that by smoothness

$$\epsilon_{k+1} = f(x_{k+1}) - f(x_*) \leq f(y_k) - f(x_*) - \frac{1}{2L}\|\nabla f(y_k)\|_2^2 = \epsilon_k^y - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$$

and by convexity of $\|\cdot\|_2^2$

$$\frac{1}{2}\|z_k - x_*\|_2^2 = \frac{1}{2}\|\beta_k(v_k - x_*) + (1-\beta_k)(y_k - x_*)\|_2^2 \le \beta_k \cdot r_k + (1-\beta_k)\cdot r_k^y\,.$$

Combining yields that

$$r_{k+1} \le \beta_k \cdot r_k + (1-\beta_k)\cdot r_k^y - \eta_k \nabla f(y_k)^\top (z_k - x_*) + \eta_k^2 L[\epsilon_k^y - \epsilon_{k+1}]\,.$$

Further, since

$$(1-\alpha_k)[v_k - y_k] = \alpha_k[y_k - x_k] \text{ and } v_k = y_k + \frac{\alpha_k}{1-\alpha_k}[y_k - x_k]$$

we have that

$$z_k - x_* = \beta_k \cdot v_k + (1-\beta_k)\cdot y_k = y_k - x_* + \frac{\alpha_k\beta_k}{1-\alpha_k}[y_k - x_k]\,.$$

By μ-strong-convexity we have that for all $z \in \mathbb{R}^n$

$$\begin{aligned} f(z) - f(x_*) &\ge f(y_k) - f(x_*) + \nabla f(y_k)^\top (z - y_k) + \frac{\mu}{2}\|z - x_k\|_2^2 \\ &= \epsilon_k^y - \nabla f(y_k)^\top (y_k - z) + \frac{\mu}{2}\|z - y_k\|_2^2\,. \end{aligned}$$

Consequently, picking $z = x_*$ implies

$$-\nabla f(y_k)^\top (y_k - z) \le -\epsilon_k^y - \mu r_k^y \text{ and } -\nabla f(y_k)^\top (y_k - x_k) \le -[\epsilon_k^y - \epsilon_k] - \mu r_k^y\,.$$

Combining yields

$$-\nabla f(y_k)^\top (z_k - x_*) \le -\epsilon_k^y - \mu r_k^y - \frac{\alpha_k\beta_k}{1-\alpha_k}\cdot[\epsilon_k^y - \epsilon_k]$$

and

$$r_{k+1} \le \beta_k \cdot r_k + (1-\beta_k)\cdot r_k^y - \eta_k\left[\epsilon_k^y + \mu r_k^y + \frac{\alpha_k\beta_k}{1-\alpha_k}\cdot[\epsilon_k^y - \epsilon_k]\right] + \eta_k^2 L[\epsilon_k^y - \epsilon_{k+1}]$$

which yields the result. □

This lemma turns the problem of designing an accelerated scheme to picking α_k, β_k, and η_k to certify as much progress as possible. We focus on choosing parameters where E_k in the preceding lemma its 0, i.e.

$$1 - \beta_k = \mu\eta_k \text{ and } L\eta_k = 1 + \frac{\alpha_k\beta_k}{1-\alpha_k}$$

though other schemes are possible.

In the case of strongly convex functions, $\mu > 0$ we wish to show that in

$$\eta_k^2 L \cdot \epsilon_{k+1} + r_{k+1} \le \beta_k \cdot r_k + (\eta_k L - 1)\cdot\eta_k\cdot\epsilon_k$$

we have a multiplicative decrease of β_k from r_k and ϵ_k to r_{k+1} and ϵ_{k+1}, i.e. that the decrease is balanced among the two error measures. Consequently, we wish to have for $\eta_k > 0$

$$(\eta_k L - 1)\eta_k = \beta_k \cdot \eta_k^2 L \text{ or equivalently } \eta_k(1-\beta_k)L = 1$$

Letting $\kappa = L/\mu$ and combining with $1 - \beta_k = \mu\eta_k$ this implies that $\mu\eta_k = \frac{1}{L\eta_k}$ and $\eta_k = \frac{1}{\sqrt{\mu L}}$ and $\beta_k = 1 - \kappa^{-1/2}$. Further, $L\eta_k = 1 + \frac{\alpha_k\beta_k}{1-\alpha_k} = \frac{1-\alpha_k(1-\beta_k)}{1-\alpha_k}$ implies $\sqrt{\kappa}(1-\alpha_k) = 1 - \frac{\alpha_k}{\sqrt{\kappa}}$. Thus, $(\kappa-1)\alpha = \kappa - \sqrt{\kappa}$ and $\alpha = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$. Thus, this choice of parameters recovers the strongly convex algorithm we analyzed earlier! We analyze this algorithm below.

Lemma 12. *In the setting of Lemma 11 if $\mu > 0$, $\kappa = \frac{L}{\mu}$, $\alpha_k = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$, and $\beta_k = 1 - \frac{1}{\sqrt{\kappa}}$ then*

$$\epsilon_{k+1} + \frac{\mu}{2} r_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right) \left[\epsilon_k + \frac{\mu}{2} r_k\right] .$$

Proof. Let $\eta = (1-\beta)\frac{1}{\mu} = \frac{1}{\sqrt{\mu L}}$. Note that

$$1 - \beta_k = \mu \eta_k \text{ and } L\eta_k = 1 + \frac{\alpha_k \beta_k}{1 - \alpha_k}$$

as we argued above and therefore by Lemma 11

$$\eta_k^2 L \cdot \epsilon_{k+1} + r_{k+1} \leq \beta_k \cdot r_k + (\eta_k L - 1) \cdot \eta_k \cdot \epsilon_k .$$

Further, since $\eta_k L - 1 = \eta_k^2 L \cdot \beta_k$ as argued above and $\eta_k^2 L = \frac{1}{\mu}$ the result follows by multiplying each side by μ. □

Leveraging this lemma we obtain the optimal dimension-free rate for minimizing smooth, μ-strongly convex functions for $\mu > 0$.

Theorem 13 (Strongly Convex Accelerated Gradient Descent)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and let $\kappa \stackrel{\text{def}}{=} \frac{L}{\mu}$. For arbitrary $x_0 \in \mathbb{R}^n$ let $x_0 = v_0$ and for all $k \geq 0$ let*

- $y_k = \alpha \cdot x_k + (1 - \alpha) \cdot v_k$ *for* $\alpha = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$
- $v_{k+1} = \beta \cdot v_k + (1 - \beta) \cdot \left[y_k - \frac{1}{\mu} \nabla f(y_k)\right]$ *for* $\beta = 1 - \frac{1}{\sqrt{\kappa}}$
- $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$

Then we have that $f(x_k) - f_* \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \cdot 2 \cdot [f(x_0) - f_*]$ and consequently we can compute an ϵ-optimal point for f with $\lceil \sqrt{\kappa} \log(2 \cdot [f(x_0) - f_*]/\epsilon) \rceil$ queries to a gradient oracle.

Proof. Noting that $(1 - \beta) \cdot \frac{1}{\mu} = \frac{1}{\sqrt{\mu L}}$ and applying Lemma 12 we have that

$$f(x_k) - f_* + \frac{\mu}{2} \|v_k - x_*\|_2^2 \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left[f(x_0) - f_* + \frac{\mu}{2} \|v_0 - x_*\|_2^2\right] .$$

Since $\frac{\mu}{2} \|v_0 - x_*\|_2^2 \leq f(x_0) - f_*$ by strong convexity, the result follows. □

Next, we consider the setting when $\mu = 0$. In this case, to set $1 - \beta_k = \mu \eta_k$ we see that we should set $\beta_k = 1$. Further, to set $\eta_k = \frac{1}{L} \left[\frac{\alpha_k \beta_k}{1 - \alpha_k} + 1\right]$ this implies we wish to set $\eta_k = \frac{1}{L(1-\alpha_k)}$. This choice of parameters yields that

$$\eta_k^2 L \cdot \epsilon_{k+1} + r_{k+1} \leq (\eta_k L - 1) \cdot \eta_k \cdot \epsilon_k + r_k .$$

To obtain an algorithm, we simply imagine we have a bound on $A_k \cdot \epsilon_k + r_k$, solve for η_k such that $(\eta_k L - 1) \cdot \eta_k = A_k$ and consider the algorithm this induces. We analyze this formally in the following lemma.

Lemma 14. *In the setting of Lemma 11 if $A_k \geq 0$ and $\mu = 0$ if we let $\eta_k = \frac{1 + \sqrt{1 + 4LA_k}}{2L}$ (i.e. the positive solution to $(L\eta_k - 1)\eta_k = A_k$), $\alpha_k = 1 - \frac{1}{\eta_k L}$, and $\beta_k = 1$ then*

$$(A_k + \eta_k) \cdot \epsilon_{k+1} + r_{k+1} \leq A_k \cdot \epsilon_k + r_k .$$

Proof. Note that $1-\beta_k = 0 = \mu\eta_k$ and $\eta_k = \frac{1}{L}\left[\frac{\alpha_k\beta_k}{1-\alpha_k}+1\right] = \frac{1}{L(1-\alpha_k)}$. Thus by Lemma 11 we have

$$\eta_k^2 L \cdot \epsilon_{k+1} + r_{k+1} \leq \beta \cdot r_k + (L\eta_k - 1)\eta_k \cdot \epsilon_k \,.$$

Further, note that η_k is a solution to $(L\eta_k - 1)\eta_k = A_k$. Since $L\eta_k^2 = A_k + \eta_k$ the result follows. □

Leveraging this lemma we obtain the optimal dimension-free rate for minimizing smooth, convex functions.

Theorem 15 (Convex Accelerated Gradient Descent)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth convex function and suppose that for some $A_0 \geq 0$, $x_0, v_0 \in \mathbb{R}^n$, and all $k \geq 0$ we let*

- $\eta_k = \frac{1+\sqrt{1+4LA_k}}{2L}$, $\alpha_k = 1 - \frac{1}{\eta_k L}$, *and* $A_{k+1} = A_k + \eta_k$
- $y_k = \alpha_k \cdot x_k + (1-\alpha_k)\cdot v_k$
- $v_{k+1} = v_k - \eta_k \nabla f(y_k)$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Then if x_* is a minimizer of f and we let $\epsilon_k \stackrel{\text{def}}{=} f(x_k) - f_*$ and $r_k \stackrel{\text{def}}{=} \frac{1}{2}\|v_k - x_*\|_2^2$ then for all $k \geq 0$ we have

$$A_k \cdot \epsilon_k + r_k \leq A_0 \cdot \epsilon_0 + r_0 \,. \tag{4.1}$$

Further, $A_k \geq (\sqrt{A_0} + \frac{k}{\sqrt{L}})^2$. Consequently, if $v_0 = x_0$ and $A_0 = 0$ we have that for all $k \geq 0$

$$f(x_k) - f_* \leq \frac{\|x_0 - x_*\|_2^2}{2k^2} \,.$$

Proof. By Lemma 14 we have that $A_{k+1}\cdot\epsilon_{k+1} + r_k \leq A_k \cdot \epsilon_k + r_k$ for all $k \geq 0$. Applying this repeatedly yields (4.1). Further, since for all $k \geq 0$ we have $\eta_k \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}}$ we have that

$$A_{k+1} = A_k + \eta_k \geq A_k + \frac{1}{2L} + \sqrt{\frac{A_k}{L}} = \left(\sqrt{A_k} + \frac{1}{\sqrt{L}}\right)^2$$

Consequently, $\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{\sqrt{L}}$ and applying this repeatedly yields that $\sqrt{A_k} \geq \sqrt{A_0} + \frac{k}{\sqrt{L}}$ for all $k \geq 0$ yielding the result. □

We now have everything to prove the main theorem of this section.

Proof of Theorem 10. This follows immediately from Theorem 13 and Theorem 15 □

5 Momentum

Another popular viewpoint or perspective on acceleration is that it can be viewed as gaining momentum in some sense, i.e. once you move in the direction of the gradient you keep moving in that direction for some time afterwards. This view can be confirmed by a rearranging of the variables in the method we derived and gives alternative statement of the accelerated gradient descent algorithms we have derived.

Lemma 16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and suppose that for all $k \geq 0$ it is the case that $x_k, v_k, y_k \in \mathbb{R}^n$ are given by Lemma 11 with $\eta_k L = 1 + \frac{\alpha_k\beta_k}{1-\alpha_k}$ then for all $k \geq 1$ we have*

- $y_k = x_k + \left(\left(\frac{1-\alpha_k}{1-\alpha_{k-1}}\right) \cdot \alpha_{k-1}\beta_{k-1}\right)(x_k - x_{k-1})$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Proof. For all $k \geq 0$ we have that

$$v_k = \frac{1}{1-\alpha_k}[y_k - \alpha_k \cdot x_k] = y_k + \frac{\alpha_k}{1-\alpha_k}[y_k - x_k] \text{ and } \nabla f(y_k) = L(y_k - x_{k+1})$$

Consequently, we have that

$$\begin{aligned} v_{k+1} &= \beta_k v_k + (1-\beta_k) \cdot y_k - \eta_k \cdot \nabla f(y_k) \\ &= y_k + \frac{\alpha_k \beta_k}{1-\alpha_k}[y_k - x_k] - L \cdot \eta_k \cdot [y_k - x_{k+1}] \\ &= x_{k+1} + \frac{\alpha_k \beta_k}{1-\alpha_k} \cdot [x_{k+1} - x_k] \end{aligned}$$

where we used that $\eta_k L = 1 + \frac{\alpha_k \beta_k}{1-\alpha_k}$ in the last line. Consequently,

$$\begin{aligned} y_{k+1} &= \alpha_{k+1} \cdot x_{k+1} + (1-\alpha_{k+1}) \cdot v_{k+1} \\ &= x_{k+1} + \left(\frac{1-\alpha_{k+1}}{1-\alpha_k}\right) \cdot \alpha_k \beta_k \cdot [x_{k+1} - x_k] \end{aligned}$$

yielding the desired result. □

Theorem 17. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and let $\kappa = \frac{L}{\mu}$. For arbitrary $x_0 \in \mathbb{R}^n$ let $x_1 = x_0 - \frac{1}{L}\nabla f(x_0)$ and for all $k \geq 1$ let*

- $y_k = x_k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)(x_k - x_{k-1})$
- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Then we have that $f(x_k) - f_* \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \cdot 2 \cdot [f(x_0) - f_*]$ and consequently we can compute an ϵ-optimal point for f with $\lceil \sqrt{\kappa} \log(2 \cdot [f(x_0) - f_*]/\epsilon) \rceil$ queries to a gradient oracle.

Proof. We will show that the sequence of x_k generated by this method are the same as those generated by Theorem 15 and therefore the result follows. First, note that as $v_0 = x_0$ it is the case that $y_0 = x_0$ and $x_1 = x_0 - \frac{1}{L}\nabla f(x_0)$. Further, as for $\alpha_k = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}$ and $\beta_k = 1 - \frac{1}{\sqrt{\kappa}} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}$ it is the case that

$$\left(\frac{1-\alpha_{k+1}}{1-\alpha_k}\right) \cdot \alpha_k \beta_k = \left(\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}\right) \cdot \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right) = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

the result follows from Lemma 16. □

Theorem 18. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and let $\kappa = \frac{L}{\mu}$. For arbitrary $x_0 \in \mathbb{R}^n$ let $x_1 = x_0 - \frac{1}{L}\nabla f(x_0)$, $A_1 = \frac{1}{L}$, $\eta_0 = \frac{1}{L}$, and for all $k \geq 1$ let*

- $\eta_k = \frac{1+\sqrt{1+4LA_k}}{2L}$, $\alpha_k = 1 - \frac{1}{\eta_k L}$, *and* $A_{k+1} = A_k + \eta_k$
- $y_k = x_k + \left(\frac{\eta_{k-1}}{\eta_k} \cdot \alpha_{k-1}\right) \cdot (x_k - x_{k-1})$

- $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Then if x_* is a minimizer of f we have that for all $k \geq 0$ that

$$f(x_k) - f_* \leq \frac{\|x_0 - x_*\|_2^2}{2k^2}\,.$$

Proof. We will show that the sequence of x_k generated by this method are the same as those generated by Theorem 13 and therefore the result follows. First, note that as $v_0 = x_0$ it is the case that $y_0 = x_0$ and $x_1 = x_0 - \frac{1}{L}\nabla f(x_0)$. Further, for $A_0 = 0$ we see that $\eta_0 = \frac{1}{L}$ and $A_1 = \frac{1}{L}$.Further, as for $\beta_k = 1$ we have

$$\left(\frac{1-\alpha_k}{1-\alpha_{k-1}}\right) \cdot \alpha_{k-1}\beta_k = \frac{\eta_{k-1}}{\eta_k} \cdot \alpha_{k-}$$

the result follows from Lemma 16. □

References

[1] Zeyuan Allen Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 3:1–3:22, 2017.

[2] Y. E. NESTEROV. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.