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

By Aaron Sidford (sidford@stanford.edu)

October 30, 2019

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 $x_{k+1} = x_k - \frac{1}{L} \bigtriangledown f(x_k)$ 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

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

oracle calls.

A natural question to ask, is is this optimal? If all we have is a gradient oracle, can we design an algorithm with improved bounds? Here we address this question, showing how better running times can be achieved through a technique typically referred to as *acceleration*.

Acceleration is one of the most mysterious techniques in optimization. Optimal rates for minimizing smooth convex functions were first achieved by Nesterov in 1983 [3] 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, i.e. viewing it as a type of momentum in continuous time, or a primal dual method or 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 [2] and varied potential based approaches to analyzing it, e.g. estimate sequences.

In these notes we present just one particular perspective on acceleration; if you would like more references on any of the others, just let me know.

1 Acceleration

So how should we accelerate? The idea here we use is pretty simple. If f is L-smooth μ-strongly convex we know that for all x the functions

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

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material and there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future. These notes will converge to a superset of the class material that is TODO-free. 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 discussion board or contact me directly at sidford@stanford.edu.

lower and upper bound f, i.e.

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

Whereas previously, our algorithms worked simply by greedily decreasing our function value using the upper bound, here we try to do better by using the lower bound as well. Whereas the gradient descent algorithm we analyzed before used a fixed step size that only depended upon the smoothness of the function, here we use strong convexity in designing our steps as well.

2 Acceleration Approach

The algorithm we use for our analysis is fairly straightforward. 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 $L_k(x) \leq f(x)$. 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

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

Consequently, it suffices to show we can decrease $f(x_k) - \psi_k$ at a fast rate.

The way we do this is simple. We let

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

and use 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 $x_{k+1} = y_k - \frac{1}{L} \bigtriangledown f(y_k)$ 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}, i.e. $L_{k+1}(y) = \beta \cdot L_k(y) + (1 - \beta) \cdot L_{y_k}(y)$ 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 the 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 next section we analyze this through a self contained helper lemma.

3 Quadratics and Combining Lower Bounds

Here we give a self contained lemma about the effect of taking convex combinations of quadratics. Proving these will be done in homework.

Lemma 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function where $\bigtriangledown^2 f(x) = \mathbf{A} \in \mathbb{R}^{n \times n}$ for all $x \in \mathbb{R}^n$ then for all x, y we have that*

$$f(x) = f(y) + \bigtriangledown f(y)^\top (x - y) + \frac{1}{2}(y - x)^\top \mathbf{A}(y - x) .$$

Proof. Let $x_t = x + t(y - x)$ for all $t \in [0, 1]$. Now we have shown that

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

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 result follows. □

This lemma shows that we can re-write our lower bounds as quadratics centered around a particular point.

Corollary 3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable μ-strongly convex function then for all $x, y \in \mathbb{R}^n$ we have*

$$f(y) \geq \psi_x + \frac{\mu}{2} \|y - v_x\|_2^2$$

where

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

Proof. We have already seen that

$$f(y) \geq L_x(y) \stackrel{\text{def}}{=} f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2\,.$$

However, we know that

$$L_x\left(x - \frac{1}{\mu} \bigtriangledown f(x)\right) = f(x) - \frac{1}{2\mu} \| \bigtriangledown f(x)\|_2^2$$

and

$$\bigtriangledown L_x\left(x - \frac{1}{\mu} \bigtriangledown f(x)\right) = \bigtriangledown f(x) - \frac{\mu}{\mu} \bigtriangledown f(x) = \vec{0}\,.$$

Since $\bigtriangledown^2 L_x(y) = \mu \mathbf{I}$ for all y the result follows from Lemma 2. □

Next, using Lemma 2 we show how to combine quadratic lower bounds.

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 $\alpha \in [0, 1]$ 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

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

and

$$\bigtriangledown^2 f_\alpha(x) = \alpha \cdot \mu \cdot \mathbf{I} + (1 - \alpha) \cdot \mu \cdot \mathbf{I} = \mu \cdot \mathbf{I}$$

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

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

and

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

The result then follows from Lemma 2 with $x = v_\alpha$ and $f = f_\alpha$. □

Intuitively, the above lemma says that the farther away the centers are, the more that combining them increases the lower bound.

4 Building An Accelerated Gradient Step

Using the analysis in the previous section yields the following bound for improving our lower bounds.

Lemma 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable μ-strongly convex function and let $L_k(x) = \psi_k + \frac{\mu}{2}\|x - v_k\|_2^2$ by such that $f(x) \geq L_k(x)$ 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 $L_{y_k}(x) = f(y_k) + \bigtriangledown f(y_k)^\top (x - y_k) + \frac{\mu}{2}\|x - y_k\|_2^2$, satisfies $f(x) \geq L_{k+1}(x)$ 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

$$v_{k+1} = \beta \cdot v_k + (1-\beta) \cdot \left[y_k - \frac{1}{\mu} \bigtriangledown f(y_k) \right]$$

and

$$\psi_{k+1} = \beta \cdot \psi_k + (1-\beta) \cdot \left[f(y_k) - \frac{1}{2\mu}\| \bigtriangledown f(y_k)\|_2^2 \right] + \frac{\mu}{2}\beta(1-\beta) \cdot \left\| v_k - \left[y_k - \frac{1}{\mu} \bigtriangledown f(y_k) \right] \right\|_2^2 .$$

Proof. First, note that $\bigtriangledown L_{y_k}(y_k - \frac{1}{\mu} \bigtriangledown f(y_k)) = \vec{0}$ and $L_{y_k}(y_k - \frac{1}{\mu} \bigtriangledown f(y_k)) = L_{y_k}(y_k) - \frac{1}{2\mu}\| \bigtriangledown f(y_k)\|_2^2$ and consequently, by Lemma 2 we have that

$$L_{y_k}(x) = f(y_k) - \frac{1}{2\mu}\| \bigtriangledown f(y_k)\|_2^2 + \frac{\mu}{2}\|x - (y_k - \frac{1}{\mu} \bigtriangledown f(y_k))\|_2^2$$

The result then follows by applying Lemma .4. □

Using this we can analyze a gradient descent step.

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 $x_{k+1} = y_k - \frac{1}{L} \bigtriangledown f(y_k)$ then we have*

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

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

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

Consequently, by Lemma 5 we have

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

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

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

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 $\alpha = \frac{1}{1+\frac{1}{\sqrt{\kappa}}} = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$. □

5 Accelerated Gradient Descent Guarantees

In the last section we showed 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*

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

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 $\| \bigtriangledown f(x_0) \|_2^2 \leq 2L \cdot [f(x_0) - f_*]$ we obtain the desired upper bound on $f(x_0) - \psi_0$. □

Putting this all together yields the following.

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

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

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 chose 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. □

6 Improving the Analysis

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 gradient oracle.

However, the κ in the logarithmic term can be improved just by slightly improving the analysis. Rather than tracking $f(x_k) - \psi_k$ there is another natural potential function can be used, the sum of the function error of $f(x_k)$ and the appropriately scaled distance of v_k to the optimal point, i.e. $\frac{\mu}{2}\|x_* - v_k\|_2^2$. In the homework, you will show that the same elements of the above proof yield the following.

Lemma 9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-smooth μ-strongly convex function and let $\kappa = \frac{L}{\mu}$ and for any $x_k, v_k \in \mathbb{R}^n$ let*

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

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

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

Proof. Let $z_{k+1} = \beta \cdot v_k + (1-\beta) \cdot y_k$. We have

$$\begin{aligned} r_{k+1} &= \frac{\mu}{2}\left\|z_{k+1} - x_* - \frac{(1-\beta)}{\mu} \bigtriangledown f(y_k)\right\|_2^2 \\ &= \frac{\mu}{2}\|z_{k+1} - x_*\|_2^2 + (1-\beta) \bigtriangledown f(y_k)^\top (x_* - z_{k+1}) + \frac{(1-\beta)^2}{2\mu}\| \bigtriangledown f(y_k)\|_2^2 \,. \end{aligned}$$

Now, since $\|\cdot\|_2^2$ is convex and since f is μ strongly convex, we have

$$\begin{aligned} \frac{\mu}{2}\|z_{k+1} - x_*\|_2^2 &= \frac{\mu}{2}\|\beta \cdot (v_k - x_*) + (1-\beta) \cdot (y_k - x_*)\|_2^2 \\ &\leq \frac{\mu}{2} \cdot \beta \cdot \|v_k - x_*\|_2^2 + \frac{(1-\beta)\mu}{2}\|y_k - x_*\|_2^2 \\ &\leq \beta \cdot r_k - (1-\beta) \cdot \left[f(y_k) - f(x_*) + \bigtriangledown f(y_k)^\top (x_* - y_k)\right] \end{aligned}$$

where in the last step we used that $f(x_*) \geq f(y_k) + \bigtriangledown f(y_k)^\top (x_* - y_k) + \frac{\mu}{2}\|y_k - x_*\|_2^2$. Combining yields that

$$r_{k+1} \leq \beta \cdot r_k - (1-\beta) \cdot [f(y_k) - f(x_*)] + (1-\beta) \cdot \bigtriangledown f(y_k)^\top (y_k - z_{k+1}) + \frac{(1-\beta)^2}{2\mu}\| \bigtriangledown f(y_k)\|_2^2$$

Furthermore, since

$$y_k - z_{k+1} = y_k - \beta \cdot v_k - (1-\beta) \cdot y_k = \beta \cdot (y_k - v_k) = \frac{\beta\alpha}{1-\alpha} \cdot (x_k - y_k)$$

and $f(x_k) \geq f(y_k) + \bigtriangledown f(y_k)^\top (x_k - y_k)$ we have that

$$r_{k+1} \leq \beta \cdot r_k - (1-\beta) \cdot [f(y_k) - f_*] + \beta\alpha \cdot \frac{1-\beta}{1-\alpha} \cdot [f(x_k) - f(y_k)] + \frac{(1-\beta)^2}{2\mu}\| \bigtriangledown f(y_k)\|_2^2$$

Since $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\| \bigtriangledown f(y_k)\|_2^2$ and we chose α and β so that $\frac{(1-\beta)^2}{2\mu} = \frac{1}{2L}$ and $\alpha \cdot \frac{1-\beta}{1-\alpha} = 1$ we have

$$\epsilon_{k+1} + r_{k+1} \leq f(y_k) + \beta \cdot r_k - (1-\beta) \cdot [f(y_k) - f_*] + \beta \cdot (f(x_k) - f(y_k) = \beta \cdot [\epsilon_k + r_k] \ .$$

□

This gives the following improved analysis of accelerated gradient descent

Theorem 10 (Accelerated Gradient Descent (Improved))**.** *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 $v_0 = x_0$ and for all $k \geq 0$ let*

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

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. By the previous theorem we have that for $x_* \in X_*(f)$ it is the case that

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

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

7 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. This gives another popular statement of the accelerated gradient descent algorithm.

Theorem 11. *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 $x_1 = x_0 - \frac{1}{L} \bigtriangledown f(x_0)$ and for all $k \geq 1$ let*

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

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 obtain this by massaging the previous algorithm and rewriting the iterates in terms of x_{k-1} instead of v_k. Note that in Theorem 10 we have that

$$v_{k-1} = \frac{1}{1-\alpha}\left[y_{k-1} - \alpha \cdot x_{k-1}\right] = y_{k-1} + \frac{\alpha}{1-\alpha}\left[y_{k-1} - x_{k-1}\right] \text{ and } \bigtriangledown f(y_{k-1}) = L(y_{k-1} - x_k)$$

Consequently, we have that

$$\begin{aligned} v_k &= \beta v_{k-1} + (1-\beta)\cdot[y_{k-1} - \frac{1}{\mu} \bigtriangledown f(y_{k-1})] \\ &= y_{k-1} + \frac{\alpha}{1-\alpha}\beta\left[y_{k-1} - x_{k-1}\right] - (1-\beta)\cdot\kappa\cdot\left[y_{k-1} - x_k\right] . \end{aligned}$$

Now since $\alpha(1-\beta)/(1-\alpha) = 1$ and $\kappa = (1-\beta)^{-2}$ we have that

$$v_k = y_{k-1} + \frac{\beta}{1-\beta}\left[y_{k-1} - x_{k-1}\right] - \frac{1}{1-\beta}\left[y_{k-1} - x_k\right] .$$

This in turn implies that

$$\begin{aligned} y_k &= \alpha\cdot x_k + (1-\alpha)\cdot v_k = x_k + (1-\alpha)\left[v_k - x_k\right] \\ &= x_k + \frac{1-\alpha}{1-\beta}\left[(1-\beta)\left[y_{k-1} - x_k\right] + \beta\left[y_{k-1} - x_{k-1}\right] - \left[y_{k-1} - x_k\right]\right] \\ &= x_k + \frac{1-\alpha}{1-\beta}\beta\left[x_k - x_{k-1}.\right] \end{aligned}$$

Now since $\alpha = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$ and $\beta = 1 - \frac{1}{\sqrt{\kappa}}$ we have that

$$1 - \alpha = \frac{1}{1+\sqrt{\kappa}} \text{ and } \frac{\beta}{1-\beta} = \sqrt{\kappa} - 1$$

yielding that

$$\frac{1-\alpha}{1-\beta}\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

and thus the result follows by Theorem 10. □

8 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 fairly overloaded term with all sorts of applications and interpretations, particularly in machine learning. When we use this term in the class though, 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 12. *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 to decay the value of μ used, 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 and they are similar in some sense. 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. We will talk more on this type of reduction in the next chapter.

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] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *CoRR*, abs/1506.08187, 2015.

[3] 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.