MS&E 213 / CS 269O : Chapter 5
Smooth Convex Generalizations *

By Aaron Sidford (sidford@stanford.edu)

November 26, 2019

Here we consider various generalization of the proofs we have seen in the previous chapters on minimizing smooth convex functions. We consider this for several reasons. First, the generalizations are useful to get fast algorithms for minimizing a broad class of functions, even those which a times are non-smooth. Second, these extensions will allow us to become familiar with some concepts we will use more extensively later in the course, i.e. dual norms and generalizations of strong convexity. Lastly, this chapter is meant to serve as a stepping stone towards research in the area as some of the results we will discuss are fairly recent.

1 Extension #1 - Smoothness and Strong Convexity In Other Norms

Our first extension is to generalize what we have done so far to arbitrary norms. Here we formally define smoothness and strong convexity in arbitrary norms and provide various equivalent characterizations. See the appendix for further information about dual norms.

Definition 1. We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to a norm $\|\cdot\|$ if and only if

$$\| \bigtriangledown f(x) - \bigtriangledown f(y)\|_* \leq L \cdot \|x - y\| \text{ for all } x, y \in \mathbb{R}^n .$$

Definition 2. We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex for $\mu \geq 0$ with respect to a norm $\|\cdot\|$ if and only if

$$f(t \cdot y + (1-t) \cdot x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} \cdot t \cdot (1-t) \cdot \|x - y\|^2 \text{ for all } x, y \in \mathbb{R}^n, t \in [0, 1]$$

1.1 Equivalent Characterizations of Strong Convexity

Here we give several equivalent characterizations of strong convexity.

Lemma 3. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex for $\mu \geq 0$ with respect to a norm $\|\cdot\|$ if and only if for all $x, h \in \mathbb{R}^n$ and $t \in [0, 1]$ we have that*

$$f(x + t \cdot h) \leq f(x) + t \cdot [f(x + h) - f(x)] - \frac{\mu}{2} \cdot t \cdot (1-t)\|h\|^2 .$$

Proof. Letting $h = y - x$ this statement can be seen to be equivalent to the Definition 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.

Lemma 4 (Convexity of Differentiable Functions)**.** *A differentiable function f is μ-strongly convex with respect to a norm $\|\cdot\|$ for $\mu \geq 0$ if and only if*

$$f(y) \geq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2 \text{ for all } x, y \in \mathbb{R}^n . \tag{1.1}$$

Proof. The proof is the same as in the case as for the case when $\|\cdot\| = \|\cdot\|_2$. □

Lemma 5 (Convexity of Twice Differentiable Functions)**.** *A twice differentiable function f is μ-strongly convex for $\mu \geq 0$ if and only if*

$$z^\top \bigtriangledown^2 f(x) z \geq \mu \|z\|^2 \text{ for all } x, z \in \mathbb{R}^n \tag{1.2}$$

Proof. The proof is the same as in the case as for the case when $\|\cdot\| = \|\cdot\|_2$. □

1.2 Equivalent Characterizations of Smoothness

Here we give alternative characterizations of smoothness.

Lemma 6. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to a norm $\|\cdot\|$ then for all $x, y \in \mathbb{R}^n$ we have*

$$\left|f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right| \leq \frac{L}{2}\|y-x\|^2 .$$

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

$$\begin{aligned}
\left|f(y) - [f(x) + \bigtriangledown f(x)^\top (y-x)]\right| &= \left|\int_0^1 \left[\bigtriangledown f(x_t) - \bigtriangledown f(x)\right]^\top (y-x) \cdot dt\right| \\
&\leq \int_0^1 \| \bigtriangledown f(x_t) - \bigtriangledown f(x)\|_* \cdot \|y-x\| \cdot dt \\
&\leq \int_0^1 L \cdot \|x_t - x\| \cdot \|y-x\| \cdot dt = \int_0^1 t \cdot L \cdot \|y-x\|^2 \cdot dt = \frac{L}{2}\|y-x\|^2 .
\end{aligned}$$

□

Lemma 7. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function such that for some norm $\|\cdot\|$ we have*

$$f(y) \leq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$$

for all $x, y \in \mathbb{R}^n$. Then for all $x, y \in \mathbb{R}^n$ we have $\| \bigtriangledown f(y) - \bigtriangledown f(x)\|_*^2 \leq 2L \cdot \left[f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right]$ and f is L-smooth with respect to $\|\cdot\|$.

Proof. Let $x \in \mathbb{R}^n$ be arbitrary and let $g : \mathbb{R}^n \to \mathbb{R}$ be defined for all $z \in \mathbb{R}^n$ by $g(z) = f(z) - \left[f(x) + \bigtriangledown f(x)^\top (z-x)\right]$. Now clearly g is convex (as f is) and therefore since $\bigtriangledown g(x) = \vec{0}$ we have

$$\min_{z \in \mathbb{R}^n} g(z) = g(x) = 0 .$$

However, by assumption, we have that for all $z, y \in \mathbb{R}^n$ it is the case that

$$\begin{aligned}
g(z) &\leq \left[f(y) + \bigtriangledown f(y)^\top (z-y) + \frac{L}{2}\|z-y\|^2\right] - \left[f(x) + \bigtriangledown f(x)^\top (z-x)\right] \\
&= f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right] + \left(\bigtriangledown f(y) - \bigtriangledown f(x)\right)^\top (z-y) + \frac{L}{2}\|z-y\|^2 \\
&\leq f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right] - \frac{1}{2L}\| \bigtriangledown f(y) - \bigtriangledown f(x)\|_*^2
\end{aligned}$$

However, by assumption

$$f(y) \leq f(x) + \bigtriangledown f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$$

and therefore

$$0 \leq \min_{z\in\mathbb{R}^n} g(z) \leq \frac{L}{2}\|y-x\|^2 - \frac{1}{2L}\| \bigtriangledown f(y) - \bigtriangledown f(x)\|_*^2 .$$

Rearranging and taking a square root yields the result. □

Lemma 8. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function and $\|\cdot\|$ is a norm then the condition*

$$\left|f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right| \leq \frac{L}{2}\|x-y\|^2 \text{ for all } x, y \in \mathbb{R}^n \tag{1.3}$$

is equivalent to the condition

$$\left|z^\top \bigtriangledown^2 f(x) z\right| \leq L \cdot \|z\|^2 \text{ for all } x, z \in \mathbb{R}^n . \tag{1.4}$$

Proof. First suppose (1.4) then for all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$ it is the case that if $x_t = x + t(y-x)$ then

$$\begin{aligned}
\left|f(y) - \left[f(x) + \bigtriangledown f(x)^\top (y-x)\right]\right| &= \left|\int_0^1 \int_0^t (y-x)^\top \bigtriangledown^2 f(x_\alpha)(y-x) d\alpha dt\right| \\
&\leq \int_0^1 \int_0^t \left|(y-x)^\top \bigtriangledown^2 f(x_\alpha)(y-x)\right| d\alpha dt \\
&\leq L \cdot \|y-x\|^2 \cdot \int_0^1 \int_0^t d\alpha dt = \frac{L}{2}\|y-x\|^2 ,
\end{aligned}$$

On the other hand suppose that (1.3) holds. Let $x, z \in \mathbb{R}^n$ be arbitrary and define $x_t \stackrel{\text{def}}{=} x + tz$ for all $t \in \mathbb{R}$ and $g(t) \stackrel{\text{def}}{=} f(x_t)$. We have $g'(t) = \bigtriangledown f(x_t)^\top z$, $g''(t) = z^\top \bigtriangledown^2 f(x_t) z$, and

$$g''(0) = \lim_{t\to 0} \frac{g'(t) - g'(0)}{t} = \lim_{t\to 0} \frac{(\bigtriangledown f(x_t) - \bigtriangledown f(x)^\top z}{t} = \lim_{t\to 0} \frac{(\bigtriangledown f(x_t) - \bigtriangledown f(x))^\top (x_t - x)}{t^2} .$$

However, since (1.3) holds we know that

$$\begin{aligned}
\left|(\bigtriangledown f(x_t) - \bigtriangledown f(x))^\top (x_t - x)\right| &= \left|\left(f(x_t) - \left[f(x) + \bigtriangledown f(x)^\top (x_t - x)\right]\right) + \left(f(x) - \left[f(x_t) + \bigtriangledown f(x_t)^\top (x - x_t)\right]\right)\right| \\
&\leq 2 \cdot \frac{L}{2} \cdot \|x_t - x\|^2 = L \cdot t^2 \cdot \|z\|^2
\end{aligned}$$

and consequently

$$\left|z^\top \bigtriangledown^2 f(x) z\right| = |g''(t)| \leq \lim_{t\to 0} \left|\frac{(\bigtriangledown f(x_t) - \bigtriangledown f(x))^\top (x_t - x)}{t^2}\right| \leq L \cdot \|z\|^2 .$$

□

1.3 Optimality Bounds

Here we show how to use the above analysis to bound the progress by gradient descent and the distance to optimality under various norms. This gives the techniques needed to analyze gradient descent, however we will not do this formally here as we give even more general analysis in the next section.

Lemma 9. *For differentiable $f \in \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ let $Q : \mathbb{R}^n \to \mathbb{R}$ be defined for all x by*

$$Q(x) = f(x_0) + \bigtriangledown f(x_0)^\top (x - x_0) + \frac{\alpha}{2}\|x - x_0\|^2$$

where $\|\cdot\|$ is an arbitrary norm and $\alpha \geq 0$. Then we have that

$$\min_x Q(x) = f(x_0) - \frac{1}{2\alpha}\| \bigtriangledown f(x_0)\|_*^2 .$$

Proof. The proof follows directly from Lemma 6 in the appendix regarding norms. □

Lemma 10. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to some $\|\cdot\|$ then for all $x \in \mathbb{R}^n$ if we let*

$$y_* = \text{argmin}_{y \in \mathbb{R}^n} f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2$$

then $f(y_*) \leq f(x) - \frac{1}{2L}\| \bigtriangledown f(x)\|_*^2$.

Proof. The proof of this is immediate from Lemma 9 and Lemma 6. □

Lemma 11. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to some norm then for all $x_* \in X_*(f)$ we have that $\bigtriangledown f(x_*) = 0$ and for all $x \in \mathbb{R}^n$ it is the case that*

$$\frac{1}{2L}\| \bigtriangledown f(x)\|_*^2 \leq f(x) - f(x_*) \leq \frac{L}{2} \cdot \|x - x_*\|^2 .$$

Proof. Note if $\bigtriangledown f(x_*) \neq 0$ then this would imply $\| \bigtriangledown f(x_*)\|_* \neq 0$ and consequently by Lemma 10 there is a point z with $f(z) \leq f(x_*) - \frac{1}{2L}\| \bigtriangledown f(x_*)\|_*^2 < f(x_*)$. Furthermore, this same lemma implies that implies that

$$f_* \leq f(x) - \frac{1}{2L}\| \bigtriangledown f(x)\|_*^2$$

giving the left hand side of the desired identity. The right hand side follows from the fact that $f(x) \leq f(x_*) + \bigtriangledown f(x_*)^\top (x - x_*) + \frac{L}{2}\|x - x_*\|^2$ by smoothness. □

Lemma 12. *If $f : \mathbb{R}^n \to \mathbb{R}$ is smooth and μ-strongly convex for $\mu > 0$ then for $x_* \in X_*(f)$ we have*

$$\frac{1}{2\mu}\| \bigtriangledown f(x)\|_*^2 \geq f(x) - f(x_*) \geq \frac{\mu}{2}\|x - x_*\|^2 .$$

Proof. First we note that since f is smooth we have $\bigtriangledown f(x_*) = 0$ and therefore

$$f(x) \geq f(x_*) + \bigtriangledown f(x_*)^\top (x - x_*) + \frac{\mu}{2}\|x - x_*\|^2$$

gives the desired bounds on the right hand side. Next we note that

$$f(x_*) \geq \min_y f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2 \geq f(x) - \frac{1}{2\mu}\| \bigtriangledown f(x)\|_*^2$$

again by Lemma 9. □

2 Composite Function Minimization

Before we formally show how to minimize functions that are smooth or strongly convex with respect to other norms, we wish to generalize things a little further. Here we consider how to do constrained function minimization, regularized function minimization, or minimization that is the sum of a smooth part and a non-smooth part. For example, in problems we may want to restrict our attention to the domain when coordinates are not too large, i.e. $\min_{\|x\|_\infty \leq 1} f(x)$, or where we want to constrain that the total weight of the coordinates is not too large in the hopes of getting sparse solutions, i.e. $\min_x f(x) + \alpha\|x\|_1$.

The abstraction we consider for all these problems is as follows. We try to minimize $f(x) = g(x) + \psi(x)$ where g is L-smooth with respect to some norm and g is convex and f is convex. We may even consider the case where $\psi \to \mathbb{R} \cup \infty$ so we can use ψ as the indicator function of a set, i.e. to solve $\min_{x \in S} g(x)$ we instead solve $f(x) = g(x) + \psi_S(x)$ where $\psi_S(x) = 0$ if $x \in S$ and $\psi_S(x) = \infty$ if $x \notin S$. This problem is known as *composite function minimization.*

The question we wish to address is how should we make progress on the problem and what should the oracle be? One hope is that perhaps both $\psi(x)$ and the norm are simple enough that we can minimizer quadratic $+ \psi(x)$ in other words we could hope to minimize

$$\min_y g(x) + \bigtriangledown g(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 + \psi(y)$$

Formally, we assume that we can compute gradients of g and that the structure of ψ is simple enough that the above can be computed efficiently given the gradient for ψ.

Lemma 13. *Let $f(x) = g(x) + \psi(x)$ where $f, g, \psi : \mathbb{R}^n \to \mathbb{R}$ and g is smooth with respect to some norm $\|\cdot\|$ and g is convex then for all $x \in \mathbb{R}^n$ if we let*

$$z = \text{argmin}_y U_x(y) \stackrel{\text{def}}{=} g(x) + \bigtriangledown g(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 + \psi(y)$$

then we have that

$$f(z) \leq \min_{y \in \mathbb{R}^n} f(y) + \frac{L}{2}\|y - x\|^2 .$$

Proof. The proof is immediate from the facts that $U_x(y) \geq f(y)$ due to smoothness and that $g(x) + \bigtriangledown g(x)^\top (y - x) \leq g(y)$ by convexity □

Thus to turn this sort of thing into an algorithm, we just need to bound the value of

$$\min_x f(x) + \frac{L}{2}\|y - x\|^2$$

however, below we show how to do this generally by convexity.

2.1 Strongly Convex Composite Function Minimization

Lemma 14. *If f is μ-strongly convex then*

$$\min_y f(y) + \frac{L}{2}\|x - y\|^2 \leq f(x) - \frac{\mu}{L + \mu}\left[f(x) - f_*\right]$$

Proof. Let x_* denote a minimizer of f and let $x_t = x + t(x_* - x)$. then by convexity we have that

$$\begin{aligned}
\min_y f(y) + \frac{L}{2}\|x-y\|^2 &\leq \min_{t\in[0,1]} f(x_t) + \frac{L}{2}\|x_t - x\|^2 \\
&\leq \min_{t\in[0,1]} f((1-t)x + t\cdot x_*) + \frac{L}{2}\cdot t^2 \cdot \|x - x_*\|^2 \\
&\leq \min_{t\in[0,1]} (1-t)\cdot f(x) + t\cdot f(x_*) - \frac{\mu}{2}\cdot t\cdot(1-t)\cdot\|x-x_*\|^2 + \frac{L}{2}\cdot t^2\cdot\|x-x_*\|^2 \\
&= f(x) + \min_{t\in[0,1]} -[f(x) - f_*]\cdot t + \left(\frac{L+\mu}{2}\cdot t^2 - \frac{\mu}{2}t\right)\cdot\|x-x_*\|^2
\end{aligned}$$

Now, setting $t = \frac{\mu}{L+\mu}$ yields the desired result. □

Theorem 15. *Let f be a μ-strongly convex function with respect to some norm $\|\cdot\|$ and let x_i be any sequence such that*

$$f(x_{i+1}) \leq \min_y f(y) + \frac{L}{2}\|y - x_i\|^2$$

then we have that

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

2.2 Non-strongly Convex Composite Function Minimization

Lemma 16. *If f is convex and $x_* \in X_*(f)$ then*

$$\min_y f(y) + \frac{L}{2}\|x-y\|^2 \leq f(x) - \frac{f(x)-f_*}{2}\cdot\min\left\{\frac{f(x)-f_*}{L\|x-x_*\|^2}, 1\right\}$$

Proof. Let x_* denote a minimizer of f and let $x_t = x + t(x_* - x)$. then by convexity we have that

$$\begin{aligned}
\min_y f(y) + \frac{L}{2}\|x-y\|^2 &\leq \min_{t\in[0,1]} f(x_t) + \frac{L}{2}\|x_t - x\|^2 \\
&\leq \min_{t\in[0,1]} f((1-t)x + t\cdot x_*) + \frac{L}{2}\cdot t^2 \cdot \|x - x_*\|^2 \\
&\leq \min_{t\in[0,1]} (1-t)\cdot f(x) + t\cdot f(x_*) + \frac{L}{2}\cdot t^2\cdot\|x-x_*\|^2 \\
&= f(x) + \min_{t\in[0,1]} -[f(x) - f_*]\cdot t + \frac{L}{2}\cdot t^2\cdot\|x-x_*\|^2
\end{aligned}$$

Now the derivative of the above with respect to t is $-[f(x) - f_*] + L\cdot t\cdot\|x - x_*\|^2$. Consequently, the minimum (ignoring the $t \in [0,1]$ constraint) is at $t = \frac{1}{L}\cdot\frac{f(x)-f_*}{\|x-x_*\|^2}$ which would yield

$$\min_y f(y) + \frac{L}{2}\|x-y\|^2 \leq f(x) - \frac{1}{2L}\cdot\frac{[f(x)-f_*]^2}{\|x-x_*\|^2} \,.$$

Since $f(x) - f_* \geq 0$ and $\|x - x_*\|^2 \geq 0$ the only reason why might not get to use this value of t is that $\frac{1}{L}\cdot\frac{f(x)-f_*}{\|x-x_*\|^2} \geq 1$ However in this case we have that $\|x - x_*\|^2 \leq \frac{1}{L}\cdot[f(x) - f_*]$ and consequently

$$\min_y f(y) + \frac{L}{2}\|x-y\|^2 \leq f(x_*) + \frac{1}{2}[f(x) - f_*] = f(x) - \frac{1}{2}[f(x) - f_*] \,.$$

Taking the worse of these two bounds yields the result. □

Theorem 17. *Let f be a μ-strongly convex function with respect to some norm $\|\cdot\|$ and let x_i be any sequence such that*

$$f(x_{i+1}) \leq \min_y f(y) + \frac{L}{2}\|y - x_i\|^2$$

then for $D = \max_{x \in \mathbb{R}^n : f(x) \leq f(x_0)} \min_{x_* \in X_*(f)} \|x - x_*\|$ we have that for all $k \geq 1$

$$f(x_k) - f_* \leq \frac{2 \cdot L \cdot D^2}{k+3}.$$

Proof. Let $\epsilon_k = f(x_k) - f_*$. Then by Lemma 16 we have that

$$\epsilon_{k+1} \leq \epsilon_k - \frac{\epsilon_k}{2} \cdot \min\left\{\frac{\epsilon_k}{L \cdot D^2}, 1\right\}$$

Consequently, $\epsilon_{k+1} \leq \epsilon_k$ for all k. We also clearly have that $\epsilon_{k+1} \leq \frac{L}{2} \cdot D^2$ and therefore

$$\epsilon_k - \epsilon_{k+1} \geq \frac{\epsilon_k}{2} \cdot \min\left\{\frac{\epsilon_{k+1}}{L \cdot D^2}, \frac{2\epsilon_{k+1}}{L \cdot D^2}\right\} = \frac{\epsilon_k \cdot \epsilon_{k+1}}{2 \cdot L \cdot D^2}$$

Consequently,

$$\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} \geq \frac{1}{2 \cdot L \cdot D^2}$$

and since $\epsilon_1 \leq \frac{L}{2} \cdot D^2$ we have

$$\frac{1}{\epsilon_k} - \frac{4}{2 \cdot L \cdot D^2} \geq \frac{1}{\epsilon_k} - \frac{1}{\epsilon_1} \geq \frac{k-1}{2 \cdot L \cdot D^2}$$

yielding the result. (Note have to check $k = 1$ case separately.) □

2.3 Acceleration

This can be done as well in many cases (but not necessarily all). The idea is that you just need to be able to build lower bounds as well and then you can do similar analysis. For example, if you can exactly minimize $f(x) + \frac{\mu}{2}\|x - x_k\|_2^2$ for values of x_k then you can accelerate. I'll either add the proof of this here or give it as homework

3 What do these algorithms look like?

The results of the last section are quite versatile. Applying them in different settings can yield a variety of variants of gradient descent. Here we give two quick examples of gradient descent in other norms. First, in the following lemma we derive a closed form expression for a step of gradient descent in ℓ_∞.

Lemma 18 (Gradient Descent in ℓ_∞)**.** *For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ if we define for $x \in \mathbb{R}^n$*

$$U(y) \stackrel{\text{def}}{=} f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_\infty^2$$

then

$$y_* = x - \frac{\| \bigtriangledown f(x)\|_1}{L} \text{sign}(\bigtriangledown f(x)) \text{ where } \text{sign}(z)_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } z_i > 0 \\ -1 & \text{if } z_i < 0 \\ 0 & \text{if } z_i = 0 \end{cases}$$

is minimizer of $U(y)$.

Proof. For arbitrary $z \in \mathbb{R}^n$ let

$$a \stackrel{\text{def}}{=} x - \|z - x\|_\infty \text{sign}(\bigtriangledown f(x))$$

Note that

$$\|a - x\|_\infty = \|z - x\|_\infty$$

and

$$\bigtriangledown f(x)^\top (a - x) = -\| \bigtriangledown f(x)\|_1 \cdot \|z - x\|_\infty \leq \bigtriangledown f(x)^\top (z - x)\,.$$

Combining these two inequalities yields that

$$U(a) \leq f(x) + \bigtriangledown f(x)^\top (z - x) + \frac{L}{2}\|z - x\|_\infty^2 = U(z)\,.$$

Since z was arbitrary, this implies that there is always a minimizer of $U(y)$ of the form $x - \alpha \cdot \text{sign}(\bigtriangledown f(x))$. However, since

$$U(x - \alpha \cdot \text{sign}(\bigtriangledown f(x)) = f(x) - \alpha \cdot \| \bigtriangledown f(x)\|_1 + \frac{L}{2}\alpha^2$$

and this is minimized when $\alpha = \| \bigtriangledown f(x)\|_1 / L$, i.e. the derivative of the above expression is 0, the result follows. □

This lemma shows that gradient descent steps in ℓ_∞ are simply gradient descent steps where the amount of movement in each coordinate is changed to be the same. In the other extreme, in the next lemma we show that gradient descent in ℓ_1 is essentially a type of coordinate descent. Here, the steps can be made so that at most one coordinate changes in each iteration.

Lemma 19 (Gradient Descent in ℓ_1)**.** *For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ if we define for $x \in \mathbb{R}^n$*

$$U(y) \stackrel{\text{def}}{=} f(x) + \bigtriangledown f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_1^2$$

then for $i_* \in \text{argmax}_{i \in [n]} | \bigtriangledown f(x)_j |$ and

$$y_* = x - \frac{\| \bigtriangledown f(x)\|_\infty}{L} \text{sign}(\bigtriangledown f(x))_{i_*} \vec{1}_{i_*} \text{ where } \text{sign}\,(z)_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } z_i > 0 \\ -1 & \text{if } z_i < 0 \\ 0 & \text{if } z_i = 0 \end{cases}$$

we have that y_* is minimizer of $U(y)$.

Proof. For arbitrary $z \in \mathbb{R}^n$ let an $j \in \text{argmax}_{i \in [n]} | \bigtriangledown f(x)_i |$ and

$$a \stackrel{\text{def}}{=} x - \|z - x\|_1 \text{sign}(\bigtriangledown f(x))_j \vec{1}_j$$

Note that

$$\|a - x\|_1 = \|z - x\|_1$$

and

$$\bigtriangledown f(x)^\top (a - x) = -\| \bigtriangledown f(x)\|_\infty \cdot \|z - x\|_1 \leq \bigtriangledown f(x)^\top (z - x)\,.$$

Combining these two inequalities yields that

$$U(a) \leq f(x) + \bigtriangledown f(x)^\top (z - x) + \frac{L}{2}\|z - x\|_1^2 = U(z)\,.$$

Since z was arbitrary, this implies that there is always a minimizer of $U(y)$ of the form $x - \alpha \cdot \text{sign}(\bigtriangledown f(x))_j \vec{1}_j$ for $j \in \text{argmax}_{i \in [n]} | \bigtriangledown f(x)_i |$. However, since

$$U(x - \alpha \cdot \text{sign}(\bigtriangledown f(x))_j \vec{1}_j) = f(x) - \alpha \cdot \| \bigtriangledown f(x)\|_\infty + \frac{L}{2}\alpha^2$$

and this is minimized when $\alpha = \| \bigtriangledown f(x)\|_\infty / L$, i.e. the derivative of the above expression is 0, the result follows. □

4 Extension #3 Coordinate Descent

Here we consider a different direction for generalizing our smooth convex minimization results. Rather than assuming a possibly more powerful oracle, i.e. minimizing the upper bounds considered in the last section, here instead we consider a weaker or more-fine grained oracle.

Definition 20 (Partial Derivative Oracle)**.** For differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ a *partial derivative oracle* is an oracle that given as query $x \in \mathbb{R}^n$ and $i \in [n]$ outputs $\frac{\partial}{\partial x_i} f(x)$.

Note that an evaluation of partial derivative oracle provides less information than a gradient oracle (since we can perform one partial derivative evaluation with 1 gradient evaluation). Furthermore, it can be a lot less information as in general implement 1 gradient evaluation takes n evaluations of a partial derivative oracle.

Given that partial derivative oracles are weaker, why are they a good idea? The hope and the reason they might be used is that in general, computing a partial derivative of a function could be cheaper than computing a gradient of a function. In some cases it might be that computing a partial derivative is a $1/n$ fraction of the cost of evaluating a gradient. Consequently, we might hope that the total cost of a partial derivative based algorithm is cheaper.

So what is a good assumption to make on our function to take advantage of this? One idea is that instead of assuming that the whole function is smooth, we could just assume that each partial derivative.

Definition 21. For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ we say that coordinate $i \in [n]$ for f is L_i-smooth if for all $x \in \mathbb{R}^n$ the function $g : \mathbb{R} \to \mathbb{R}$ defined as $g(h) = f(x + h \cdot \vec{1}_i)$ is L_i-smooth.

This definition is essentially says that f is L_i smooth a coordinate $i \in [n]$ if at any point restricting to the line where only coordinate i changes, yields a L_i smooth function. This assumption allows us to bound how much progress we possibly make when we just change coordinate i. How? We can just use our bounds from before and do gradient descent on g_x, i.e. only change coordinate. We call such an update a coordinate descent step.

Lemma 22. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function that is L_i smooth in coordinate i. Then for all $x \in \mathbb{R}^n$ if we let $y = x - \frac{1}{L_i} \bigtriangledown f(x)_i \cdot \vec{1}_i$ where $\vec{1}_i(j)$ has value 1 if $j = i$ and 0 otherwise we have*

$$f(y) \leq f(x) - \frac{1}{2L_i} \left[\bigtriangledown f(x)_i\right]^2 .$$

Proof. Let $g(h) = f(x + h \cdot \vec{1}_i)$ for all $h \in \mathbb{R}$. By assumption g is L_i smooth and therefore

$$g\left(\frac{1}{L_i} g'(0)\right) \leq g(0) - \frac{1}{2L_i} \left[g'(0)\right]^2$$

However, $g'(0) = \bigtriangledown f(x)^\top \cdot \vec{1}_i = [\bigtriangledown f(x)]_i$ yielding the desired result. □

So we have shown that just by changing a single coordinate we can provably decrease the value of the function. How can we turn this into an efficient algorithm? Well, we could always find the coordinate with the largest ratio of partial derivative squared to L_i and change it. This might be efficient, but it is expensive as it involves looking at all the coordinates. Instead we could do something simpler and pick a random coordinate. This is known as randomized coordinate descent and we analyze it as follows.

Lemma 23. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function that for each $i \in [n]$ is L_i smooth in coordinate i then if we pick $j \in [n]$ with $\Pr[j = i] = p_i$ and let $y = x - \frac{1}{L_j}[\bigtriangledown f(x)]_j \cdot \vec{1}_j$ then*

$$\mathbb{E} f(y) \leq f(x) - \sum_{i \in [n]} \frac{p_i}{2 \cdot L_i} \cdot [\bigtriangledown f(x)]_i^2 .$$

Proof. We have by the Lemma 22 that

$$\mathbb{E}f(y) = \sum_{i\in[n]} p_i \cdot f\left(x - \frac{1}{L_i}[\bigtriangledown f(x)]_i \cdot \vec{1}_i\right) \leq \sum_{i\in[n]} p_i \cdot \left[f(x) - \frac{1}{L_i}\left[\bigtriangledown f(x)\right]_i^2\right].$$

□

Theorem 24 ((Randomized) Coordinate Descent)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex function such that for all i coordinate i of f is L_i-smooth. Further suppose that starting from some x_0 in $\mathbb{R}^n$ for all $k \geq 0$ we pick i_k with probability L_i/S where $S = \sum_{i\in[n]} L_i$. Then we have that*

$$\mathbb{E}f(x_k) - f_* \leq \left(1 - \frac{\mu}{S}\right)^k [f(x_0) - f_*]$$

and consequently an ϵ-optimal point can be computed with an expected $O(\sum_{i\in[n]} \frac{L_i}{\mu} \cdot \log\left(\frac{f(x_0)-f_*}{\epsilon}\right))$ oracle calls.

Proof. Note that for any fixed x_k looking at the expected effect of picking a random i_k yields

$$\mathbb{E}_{i_k} f(x_{k+1}) \leq \sum_{i\in[n]} \frac{L_i}{S}\left[f(x_k) - \frac{1}{2L_i}[\bigtriangledown f(x_k)]_i^2\right] \leq f(x_k) - \frac{1}{2S}\| \bigtriangledown f(x_k)\|_2^2 .$$

Consequently, by strong convexity (as we saw previous for gradient descent) we have that

$$\mathbb{E}_{i_k} f(x_{k+1}) - f_* \leq \left(1 - \frac{\mu}{S}\right)[f(x_k) - f_*] .$$

Applying the low of total expectation to take into account the randomness in computing $f(x_k)$ and applying induction then yields the desired result. □

In the homework, we will possibly show that this rate is always worse than that of gradient descent. However, coordinate descent can in some cases lead to faster algorithms if the cost of computing partial derivatives is less than that of computing gradients.

Note that the rate can be improved by acceleration. We will likely prove this in a later homework assignment.

Theorem 25 (Accelerated Coordinate Descent)**.** *Let f be a μ-strongly convex function such that for all i coordinate i of f is L_i-smooth. Then given any x_0 in $\mathbb{R}^n$ there is a method that computes an expected ϵ-optimal point with only $O(\sum_{i\in[n]} \sqrt{\frac{L_i}{\mu}} \cdot \log(\frac{f(x_0)-f_*}{\epsilon}))$ oracle calls.*