

MS&E 213 / CS 269O : Chapter 5

Smooth Convex Generalizations *

By Aaron Sidford (sidford@stanford.edu)

October 20, 2020

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 obtain algorithms for minimizing a broad class of functions, even those which may be non-smooth. Second, these extensions will allow us to become familiar with some concepts we may use more extensively later in the course, e.g. dual norms and generalizations of strong convexity. Further, this chapter is meant to review many of the concepts we have seen so far and 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 our analysis of smooth convex function minimization to arbitrary norms. Formally, we define smoothness and strong convexity with respect to an arbitrary norm as follows. (See the appendix for further information about dual norms.)

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

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

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

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

In the remainder of this section we provide various equivalent characterizations of strong convexity and smoothness and bounds on different progress measures for these norms. In the next section we provide methods for provably optimizing a generalization of these functions.

1.1 Equivalent Characterizations of Strong Convexity

Here we give several equivalent characterizations of strong convexity.

*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 are intended 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 Piazza or contact me directly at sidford@stanford.edu.

Lemma 3. $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex for $\mu \geq 0$ with respect to 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. \square

Lemma 4 (Convexity of Differentiable Functions). *Differentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex with respect to norm $\|\cdot\|$ for $\mu \geq 0$ if and only if*

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

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

Lemma 5 (Convexity of Twice Differentiable Functions). *Twice differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex for $\mu \geq 0$ if and only if*

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

Proof. The proof is the same as in the case as for the case when $\|\cdot\| = \|\cdot\|_2$ and also follows from the smoothness chapter and Lemma 4. \square

1.2 Equivalent Characterizations of Smoothness

Here we give alternative characterizations of smoothness. Key to some of these proofs is the following fact regarding dual norms which is proved in the appendix.

Lemma 6. *For any norm $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$, $x \in \mathbb{R}^n$, and $\alpha > 0$ we have that*

$$\min_{y \in \mathbb{R}^n} x^\top y + \frac{\alpha}{2} \|y\|^2 = -\frac{1}{2\alpha} \|x\|_*^2$$

This lemma allows us to conclude generalizations of the fact that

$$\min_{y \in \mathbb{R}^n} f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|_2^2 = f(x) - \frac{1}{2\alpha} \|\nabla f(x)\|_2^2$$

which we used frequently in our analysis of smooth convex function minimization, i.e.

$$\min_{y \in \mathbb{R}^n} f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|^2 = \min_{d \in \mathbb{R}^n} f(x) + \nabla f(x)^\top d + \frac{\alpha}{2} \|d\|^2 = f(x) - \frac{1}{2\alpha} \|\nabla f(x)\|_*^2.$$

We first prove the following lemma relating smoothness to bounds on the quality of Taylor expansion based approximations.

Lemma 7. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth with respect to a norm $\|\cdot\|$ then for all $x, y \in \mathbb{R}^n$ we have*

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

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

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

□

Lemma 8. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable convex function such that for some norm $\|\cdot\|$ we have*

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

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

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

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

Combining, yields that

$$\begin{aligned} 0 &= \min_{z \in \mathbb{R}^n} g(z) \\ &\leq \min_{z \in \mathbb{R}^n} f(y) - [f(x) + \nabla f(x)^\top (y - x)] + (\nabla f(y) - \nabla f(x))^\top (z - y) + \frac{L}{2} \|z - y\|^2 \\ &= \min_{d \in \mathbb{R}^n} f(y) - [f(x) + \nabla f(x)^\top (y - x)] + (\nabla f(y) - \nabla f(x))^\top d + \frac{L}{2} \|d\|^2 \\ &= f(y) - [f(x) + \nabla f(x)^\top (y - x)] - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_*^2 \end{aligned}$$

where in the last step we used Lemma 6. Rearranging yields $\|\nabla f(y) - \nabla f(x)\|_*^2 \leq 2L \cdot [f(y) - [f(x) + \nabla f(x)^\top (y - x)]]$ as desired. Further, since by assumption

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

we have that

$$0 \leq \frac{L}{2} \|y - x\|^2 - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_*^2.$$

Rearranging and taking a square root yields that $\|\nabla f(y) - \nabla f(x)\|_* \leq L \|y - x\|$ and since y and x were arbitrary we see that f is L -smooth. □

Lemma 9. If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a twice differentiable function and $\|\cdot\|$ is a norm then the condition

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

is equivalent to the condition

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

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

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

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 $g'(t) = \nabla f(x_t)^\top z$, $g''(t) = z^\top \nabla^2 f(x_t) z$, and

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

However, since (1.3) holds we know that

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

and consequently

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

□

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

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

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

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

Proof. The proof follows directly from Lemma 6 as discussed in the previous section. □

Lemma 11. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth with respect to some $\|\cdot\|$ then for all $x \in \mathbb{R}^n$ if we let*

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

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

Proof. The proof of this is immediate from Lemma 10 and Lemma 7. \square

Lemma 12. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth with respect to some norm then for all $x_* \in X_*(f)$ we have that $\nabla f(x_*) = 0$ and for all $x \in \mathbb{R}^n$ it is the case that*

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

Proof. Note if $\nabla f(x_*) \neq 0$ then this would imply $\|\nabla f(x_*)\|_* \neq 0$ and consequently by Lemma 11 there is a point z with $f(z) \leq f(x_*) - \frac{1}{2L} \|\nabla f(x_*)\|_*^2 < f(x_*)$. Furthermore, this same lemma implies that implies that

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

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

Lemma 13. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is smooth and μ -strongly convex for $\mu > 0$ then for $x_* \in X_*(f)$ we have*

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

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

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

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

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

again by Lemma 10. \square

1.4 Algorithm Examples

The analysis of the preceding sections is enough to immediately obtain and analyze methods for provably minimizing smooth convex functions in various norms. Given $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that is L -smooth with respect to norm $\|\cdot\|$ a natural minimization is to simply repeatedly minimize the upper bound implied by smoothness, i.e. repeat

$$x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{L}{2} \|x - x_k\|^2. \quad (1.5)$$

By the analysis in the preceding sections we know that this implies that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_*^2.$$

Further, we know that $\frac{1}{2\mu} \|\nabla f(x_k)\|_*^2 \geq f(x_k) - f(x_*)$ and consequently

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

Therefore analogous convergence rates can be obtained for this method as were shown in the chapter on convex functions. This even generalizes to convex functions (i.e. $\mu = 0$) and in the next section we prove this for a more general class of functions.

In the remainder of this we look at what instances of the method given by (1.5) look like in different norms. This method is sometimes known as steepest descent or non-Euclidean gradient descent and in the remainder of the section we derive a closed form expression for a step of steepest descent in ℓ_∞ and ℓ_1 .

Lemma 14 (Gradient Descent in ℓ_∞). *For differentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if we define for $x \in \mathbb{R}^n$*

$$U(y) \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_\infty^2 \text{ for all } y \in \mathbb{R}^n$$

then

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

is minimizer of $U(y)$.

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

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

Note that

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

and

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

Combining these two inequalities yields that

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

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

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

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

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

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

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

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

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

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

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

Note that

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

and

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

Combining these two inequalities yields that

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

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

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

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

2 Composite Function Minimization

Here we generalize the setup in the previous section further. We consider the problem of minimizing a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ where $f(x) = g(x) + \psi(x)$ for all $x \in \mathbb{R}^n$, g is convex and L -smooth with respect to norm $\|\cdot\|$, and f is μ -strongly convex with respect to $\|\cdot\|$.

This problem is known as *composite function minimization* and it arises in many settings. We have already seen a few instances of these problems in this course. For example, if we want to solve the constrained minimization problem $\min_{x \in S} g(x)$ for some set S , we can write this problem as $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$ and $\psi(x) = \psi_S(x) = 0$ if $x \in S$ and $\psi_S(x) = \infty$ if $x \notin S$. Further, in the previous chapter we considered this problem in the setting where g is convex and L -smooth and $\psi(x) = \frac{\alpha}{2} \|x - x_0\|_2^2$ is chosen to improve the strong convexity of the problem so certain accelerated methods can be applied. More broadly, the technique of adding a simple function ψ to a given function f , e.g. $\psi(x) = \alpha \|x\|_1$ or $\psi(x) = \frac{\alpha}{2} \|x - x_0\|_2^2$, is a technique known as *regularization* and can be applied for a variety of reasons, e.g. $\psi(x) = \alpha \|x\|_1$ can help make solutions sparse.

Our goal in this section is to obtain methods for provably minimizing f given access to a gradient oracle for g and access to the entire function ψ . How should we make progress on the problem and how exactly is the oracle access to ψ that is needed? One hope is that perhaps both $\psi(x)$ and the norm are simple enough that we can minimize the quadratic $+ \psi(x)$. In other words we could hope to minimize

$$\min_{y \in \mathbb{R}^n} g(x) + \nabla g(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2 + \psi(y)$$

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 . We analyze this setup in this section and in the following lemma bound the progress made by such a step.

Lemma 16. *Let $f(x) = g(x) + \psi(x)$ where $f, g, \psi : \mathbb{R}^n \rightarrow \mathbb{R}$ 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*

$$z = \operatorname{argmin}_y U_x(y) \stackrel{\text{def}}{=} g(x) + \nabla g(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2 + \psi(y) \quad (2.1)$$

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 $g(x) + \nabla g(x)^\top (y - x) \leq g(y)$ by convexity \square

Interestingly, this lemma shows that to analyze a method which repeatedly applies (2.1), we simply need to bound the value of

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

The method which iterates $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x) + \frac{L}{2} \|x - x_k\|^2$ is known as the *proximal point method* and consequently the preceding lemma shows that if we can simply analyze how much progress on function minimization is achieved by one step of proximal point we can analyze the composite function minimization method which repeatedly implies (2.1).

At first glance this analysis may seem somewhat difficult as we wish to do this without any assumptions on ψ (other than the convexity of f). It may be the case that f and ψ are non-differentiable. Nevertheless, we will show that convexity is enough to perform this analysis and in the following section we analyze these methods in the cases when $\mu > 0$ and $\mu = 0$.

2.1 Strongly Convex Composite Function Minimization

Lemma 17. *If arbitrarily (possibly non-differentiable) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex, $L > 0$, and $\|\cdot\|$ is a norm then*

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

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. \square

Theorem 18. *If arbitrarily (possibly non-differentiable) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex, $L > 0$, $\|\cdot\|$ is a norm, and $x_i \in \mathbb{R}^n$ for $i \geq 0$ is any sequence such that*

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

then

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

Corollary 19. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex with respect to norm $\|\cdot\|$ and $f(x) = g(x) + \psi(x)$ for all $x \in \mathbb{R}^n$, g is convex L -smooth with respect to $\|\cdot\|$ and for some $x_0 \in \mathbb{R}^n$ and all $k \geq 0$*

$$x_{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^n} g(x_k) + \nabla g(x_k)^\top (y - x_k) + \frac{L}{2} \|y - x_k\|^2 + \psi(y)$$

then for all $k \geq 0$

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

2.2 Non-strongly Convex Composite Function Minimization

Lemma 20. *If arbitrarily (possibly non-differentiable) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, $L > 0$, $\|\cdot\|$ is a norm, 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. \square

Theorem 21. *If arbitrarily (possibly non-differentiable) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, $L > 0$, $\|\cdot\|$ is a norm, and $x_i \in \mathbb{R}^n$ for $i \geq 0$ is 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 20 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.) □

Corollary 22. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex with respect to norm $\|\cdot\|$ and $f(x) = g(x) + \psi(x)$ for all $x \in \mathbb{R}^n$, g is convex L -smooth with respect to $\|\cdot\|$ and for some $x_0 \in \mathbb{R}^n$ and all $k \geq 0$*

$$x_{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^n} g(x_k) + \nabla g(x_k)^\top (y - x_k) + \frac{L}{2} \|y - x_k\|^2 + \psi(y)$$

then for all $k \geq 1$

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

2.3 Acceleration

A natural question is, are the methods above optimal? Can they be accelerated? Interestingly, acceleration is possible in some (but not necessarily all). The idea to achieve this is to build lower bounds as well and then perform 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.

3 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 23 (Partial Derivative Oracle). For differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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, i.e. the gradient is Lipschitz, we could just assume that each partial derivative is

Definition 24. For differentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$ we say that coordinate $i \in [n]$ for f is L_i -smooth if for all $x \in \mathbb{R}^n$ the function $g : \mathbb{R} \rightarrow \mathbb{R}$ defined as $g(h) = f(x + h \cdot \vec{1}_i)$ is L_i -smooth.

This definition 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 25. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function that is L_i smooth in coordinate i . Then for all $x \in \mathbb{R}^n$ if we let $y = x - \frac{1}{L_i} \nabla f(x)_i \cdot \vec{1}_i$ where $\vec{1}_i(j)$ has value 1 if $j = i$ and 0 otherwise we have

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

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

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

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

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 26. If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $y = x - \frac{1}{L_j} [\nabla f(x)]_j \cdot \vec{1}_j$ then

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

Proof. We have by the Lemma 25 that

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

\square

Theorem 27 ((Randomized) Coordinate Descent). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $O\left(\sum_{i \in [n]} \frac{L_i}{\mu} \cdot \log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$ oracle calls.

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

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

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

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 and this may be proved in a later version of the note.

Theorem 28 (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.*

We conclude this chapter (and unit on smooth function minimization) by taking a closer look at randomized coordinate descent for solving linear systems. Suppose we wish to solve $\mathbf{A}x = b$ where $\mathbf{A} = \mathbf{A}^\top$ is symmetric and positive definite, meaning that $x^\top \mathbf{A}x > 0$ for all $x \neq 0$ or equivalently that $\lambda_{\min}(\mathbf{A}) > 0$. Solving such systems is equivalent to solving the following optimization problem

$$\min_x f(x) = \frac{1}{2} x^\top \mathbf{A}x - b^\top x$$

as $\nabla f(x) = \mathbf{A}x - b$ and $\nabla^2 f(x) = \mathbf{A}$, i.e. f is convex and the gradient is $\vec{0}$ only at solutions to the linear system. Note that f is $\lambda_{\max}(\mathbf{A})$ -smooth and $\lambda_{\min}(\mathbf{A})$ -strongly convex and consequently gradient descent computes an ϵ -optimal point for f with $O(\frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} \log(\epsilon_0/\epsilon))$ gradient evaluations where $\epsilon \leq \epsilon_0 = f(x_0) - f_*$.

How many iterations does coordinate descent require? Note that for all $i \in [n]$ it is the case that $g(h) = f(x + h \cdot \vec{1}_i)$ satisfies

$$g''(h) = \vec{1}_i^\top \nabla^2 f(x + h \vec{1}_i) \vec{1}_i = \vec{1}_i^\top \mathbf{A} \vec{1}_i = \mathbf{A}_{ii}$$

and consequently f is \mathbf{A}_{ii} -smooth in coordinate i . Consequently, coordinate descent computes an expected ϵ -optimal point with

$$O\left(\frac{\sum_{i \in [n]} \mathbf{A}_{ii}}{\lambda_{\min}(\mathbf{A})} \log(\epsilon_0/\epsilon)\right) = O\left(\frac{\sum_{i \in [n]} \lambda_i(\mathbf{A})}{\lambda_{\min}(\mathbf{A})} \log(\epsilon_0/\epsilon)\right)$$

partial derivative evaluations. Note that this is always more than the number of gradient evaluations as $\sum_{i \in [n]} \lambda_i(\mathbf{A}) \geq \lambda_{\max}(\mathbf{A})$.

However, note that the cost of computing a gradient of f involves applying \mathbf{A} to a vector, which could involve looking at n^2 entries of \mathbf{A} . However, computing a partial derivative of f just involves looking at a row of \mathbf{A} , i.e. n entries, and an entry of b_i . Consequently, a partial derivative computation could be a $1/n$ fraction of the cost of a gradient evaluation. When this happens the total complexity of coordinate descent is never worse than that of gradient descent and is better by however much $\frac{1}{n} \sum_{i \in [n]} \lambda_i(\mathbf{A})$ is smaller than $\lambda_{\max}(\mathbf{A})$, i.e. the average eigenvalue is smaller than the largest. Thus we see that coordinate descent can yield improvements over gradient descent when in some sense the average smoothness over the coordinates is better than the worst-case smoothness of the function.