EE 546: Mathematics of High-dimensional Data

University of Southern California

Lecturer: Mahdi Soltanolkotabi August 22, 2018 Parts are adapted from student scribes

---

Lecture Note 1

1 When have we arrived at an optimal solution?

Consider the optimization problem

$$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}). \tag{1}$$

Definition 1 (optimal solution)

- *We say* $\boldsymbol{x}^*$ *is a solution of* (1) *if* $\boldsymbol{x}^* \in \Omega$ *and for all* $\boldsymbol{x} \in \Omega$ *we have* $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$.
- *We say* $\boldsymbol{x}^*$ *is a local solution of* (1) *if* $\boldsymbol{x}^* \in \Omega$ *and there is a neighborhood* $\mathcal{N}$ *around* $\boldsymbol{x}^*$ *such that for all* $\boldsymbol{x} \in \Omega \cap \mathcal{N}$ *we have* $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$.

Interestingly there exists infinitly differentiable function f (in fact of the form $f(\boldsymbol{x}) = \sum_{i,j} Q_{ij} x_i^2 x_j^2$) such that even checking local optimality is NP hard. This begs the question for which functions can we check optimality? One class of functions where we can check optimality is convex functions.

2 Convexity

Convex functions are functions for which the line between any two points lies above the graph. Formally

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad \theta \in [0,1].$$

One of the reasons we like convex functions in optimization is the following theorem.

Theorem 1 *For convex functions, local minima/maxima are global minima/maxima.*

Proof Assume that local minima are not global minima. If $\boldsymbol{x}_1, \boldsymbol{x}_2$ are two local minima such that $f(\boldsymbol{x}_1) < f(\boldsymbol{x}_2)$, and f is convex, then

$$f(\theta\boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2) \leq \theta f(\boldsymbol{x}_1) + (1-\theta) f(\boldsymbol{x}_2) < f(\boldsymbol{x}_2) \quad \forall \theta \in [0,1],$$

which is in contradiction with local optimality of $\boldsymbol{x}_1$. ■

3 Gradients, Hessians, and Taylor's theorem

For a function $f : \mathbb{R}^n \to \mathbb{R}$, we define

$$\text{Gradient} \quad \nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_i}\right]_{i=1,2,\ldots,n}$$

and

$$\text{Hessian} \quad \nabla^2 f(\boldsymbol{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1,2,\ldots,n}$$

The most fundamental theorem in all of optimization is one you learned in calculus.

Theorem 2 (Taylor's Theorem)

- *If f is continuously differentiable, then*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(t\boldsymbol{x} + (1-t)\boldsymbol{y})^T(\boldsymbol{y} - \boldsymbol{x}) \quad \textit{for some} \quad t \in [0,1].$$

- *If f is twice continuously differentiable, then*

$$\nabla f(\boldsymbol{y}) = \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(t\boldsymbol{y} + (1-t)\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})dt$$

  and

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f\,(t\boldsymbol{y} + (1-t)\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \quad \textit{for some} \quad t \in [0,1].$$

4 Equivalent conditions for convexity

We now state two equivalent conditions for convexity.

Theorem 3 (first order condition for convexity) *If f is continuously differentiable, then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}).$$

Stated differently, the first order Taylor approximation is a global lower bound for f.

Theorem 4 (second order condition for convexity) *If f is twice continuously differentiable on its domain ($dom(f)$), then f is convex if and only if for all $\boldsymbol{x} \in dom(f)$ we have $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$.*

4.1 Convexity through derivatives

A useful way to characterize convex function is through monotonicity of their gradients. This is the subject of the next definition.

Definition 2 *A mapping $g : \mathbb{R}^n \to \mathbb{R}^n$ is called monotone if for all $\boldsymbol{x}, \boldsymbol{y} \in dom(f)$ we have*

$$\langle g(\boldsymbol{x}) - g(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0.$$

It turns out that for differentiable functions convexity and monotonicity of gradients are equivalent.

Theorem 5 (monotonicity of gradients) *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the gradient mapping ∇f is monotone.*

Proof If f is convex by Theorem 3 for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

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

and

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

Combining these two identities we conclude that

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0,$$

which implies monotonicity of the gradients.

Now to prove the other side assume that the mapping ∇f is monotone. Define $g(t) = f\left(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\right)$. Thus, $g'(t) = \langle \nabla f\left(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\right), \boldsymbol{y} - \boldsymbol{x} \rangle$. Note that monotonicity of ∇f implies that

$$t\left(g'(t) - g'(0)\right) = \langle \nabla f\left(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})\right) - \nabla f(\boldsymbol{x}), t(\boldsymbol{y} - \boldsymbol{x}) \rangle \geq 0.$$

The latter implies that for all $0 \leq t \leq 1$

$$g'(t) \geq g'(0)$$

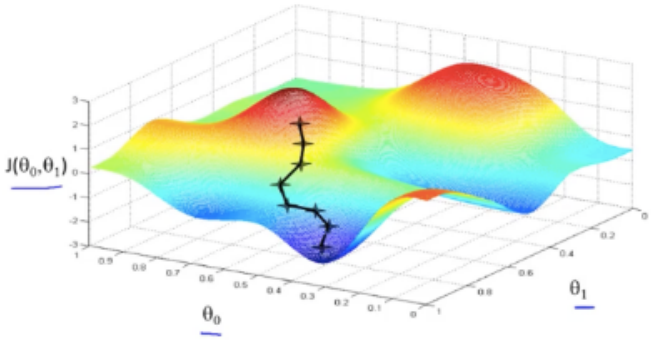
Thus,

$$f(\boldsymbol{y}) = g(1) = g(0) + \int_0^1 g'(t)dt \geq g(0) + g'(0) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Thus we have shown that for all $\boldsymbol{x}, \boldsymbol{y}$ monotonicity implies

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

The latter inequality together with Theorem 3 immediately implies that the function f is convex. ■



5 Gradient Descent

Gradient descent is perhaps the most widely used optimization algorithm. The strategy is simple: Go down hill, a.k.a. the descent direction.

Definition 3 (Descent direction) $\mathbf{v}$ *is a descent direction for* f *at* $\boldsymbol{x}$ *if*

$$f(\boldsymbol{x} + \mu \mathbf{v}) < f(\boldsymbol{x})$$

for some $\mu > 0$.

Lemma 6 *For a continuously differentiable function* f *in a neighborhood of* $\boldsymbol{x}$*, if* $\langle \mathbf{v}, \nabla f(\boldsymbol{x}) \rangle < 0$*, then* $\mathbf{v}$ *is a direction of descent.*

Proof First note that by continuity of gradient there exists a μ_0 such that

$$\langle \nabla f(\boldsymbol{x} + \mu \boldsymbol{v}), \mathbf{v} \rangle < 0, \quad \text{for all} \quad \mu \leq \mu_0.$$

By Taylor's theorem, we conclude that there exists $\tilde{\mu} \in [0, \mu]$ such that

$$f(\boldsymbol{x} + \mu \mathbf{v}) = f(\boldsymbol{x}) + \mu \langle \nabla f(\boldsymbol{x} + \tilde{\mu} \mathbf{v}), \mathbf{v} \rangle.$$

Thus using $\mu \langle \nabla f(\boldsymbol{x} + \tilde{\mu} \mathbf{v}), \mathbf{v} \rangle < 0$

$$f(\boldsymbol{x} + \mu \mathbf{v}) = f(\boldsymbol{x}) + \mu \langle \nabla f(\boldsymbol{x} + \tilde{\mu} \mathbf{v}), \mathbf{v} \rangle < f(\boldsymbol{x}).$$

Hence, $\boldsymbol{v}$ is a descent direction. ■

5.1 Heuristic development of gradient descent

Let's focus on the approximation

$$f(\boldsymbol{x} + \mu \boldsymbol{v}) \approx f(\boldsymbol{x}) + \mu \boldsymbol{v}^\top f(\boldsymbol{x}).$$

We want to make $\mu \boldsymbol{v}^\top f(\boldsymbol{x})$ as negative as possible, so we pick $\boldsymbol{v}$ such that

$$\inf_{||\boldsymbol{v}||_{\ell_2}=1} \boldsymbol{v}^\top \nabla f(\boldsymbol{x})$$

$$\boldsymbol{v} = -\frac{\nabla f(\boldsymbol{x})}{||\nabla f(\boldsymbol{x})||_{\ell_2}}$$

Gradient Descent $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu_t \nabla f(\boldsymbol{x}_t)$

- Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$
- $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu_t \nabla f(\boldsymbol{x}_t)$
- Repeat until convergence

5.2 Optimality conditions for gradient descent

Theorem 7 (Optimality Conditions) *The following are optimality conditions for the gradient descent algorithm:*

1. *Assume f is continuously differentiable and $\boldsymbol{x}^\star$ is a local minimum $\Rightarrow \nabla f(\boldsymbol{x}^\star) = \mathbf{0}$.*
2. *Assume $\nabla^2 f$ is continuous and $\boldsymbol{x}^\star$ is a local minimum $\Rightarrow \nabla^2 f(\boldsymbol{x}^\star) \succeq \mathbf{0}$.*
3. *If f is twice continuously differentiable, $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^\star) \succeq \mathbf{0} \Rightarrow \boldsymbol{x}^\star$ is a local minimum.*

Proof

1. $-\nabla f(\mathbf{x}^\star)$ is a descent direction $\Rightarrow \nabla f(\boldsymbol{x}^\star) = 0$
2. $\boldsymbol{x}^\star$ is a local minimum $\Rightarrow f(\boldsymbol{x}^\star + t\boldsymbol{u}) \geq f(\boldsymbol{x}^\star)$ for all $\boldsymbol{u}$ and small enough t. By Taylor's theorem, $\exists \tilde{t} \in [0, t]$ such that

$$f(\boldsymbol{x}^\star + t\mathbf{u}) = f(\boldsymbol{x}^\star) + t\langle \nabla f(\boldsymbol{x}^\star), \mathbf{u} \rangle + \frac{1}{2} t^2 \mathbf{u}^\top \nabla^2 f(\boldsymbol{x}^\star + \tilde{t}\mathbf{u})\mathbf{u}$$

   By 1., we have

$$\begin{aligned} & f(\boldsymbol{x}^\star + t\mathbf{u}) - f(\boldsymbol{x}^\star) = \frac{1}{2} t^2 \mathbf{u}^\top \nabla^2 f(x^\star + \tilde{t}\mathbf{u})\mathbf{u} \\ \Rightarrow \quad & \forall \mathbf{u}, \quad \mathbf{u}^\top \nabla^2 f(\boldsymbol{x}^\star + \tilde{t}\mathbf{u})\mathbf{u} \geq 0 \\ \Rightarrow \quad & \forall \mathbf{u}, \quad \nabla^2 f(\boldsymbol{x}^\star + \tilde{t}\mathbf{u}) \succeq 0 \\ \Rightarrow \quad & \nabla^2 f(\boldsymbol{x}^\star) \succeq 0 \end{aligned}$$

3. $\exists r$ such that $\nabla^2 f(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in B(\boldsymbol{x}^\star, r)$ (ball with radius r). Pick $\mathbf{u}$ with $||\mathbf{u}||_{\ell_2} < r$, then we have $t \in [0, 1]$ such that

$$f(\boldsymbol{x}^\star + \mathbf{u}) = f(\boldsymbol{x}^\star) + \mathbf{u}^\top \nabla f(\boldsymbol{x}^\star) + \frac{1}{2} \mathbf{u}^\top \nabla^2 f(\boldsymbol{x} + t\mathbf{u})\mathbf{u} \geq f(\boldsymbol{x}^\star)$$

■

For the following, we will assume f to be convex.

Lemma 8 *For $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable, $\boldsymbol{x}$ is a global minimizer $\iff \nabla f(\boldsymbol{x}^\star) = 0$.*

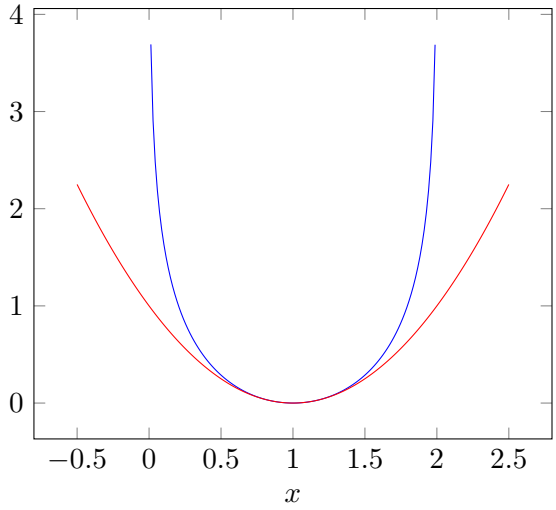


Figure 1: $-(\ln(x) + \ln(2-x))$ descends fast (blue), while $(x-1)^2$ descends slowly in comparison (red).

Proof We will assume $\nabla f(\mathbf{x}^\star) = 0$. Then,

$$f \text{ continuously differentiable} \iff f(\boldsymbol{x}) \geq f(\boldsymbol{x}^\star) + \langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \quad \forall \boldsymbol{x},$$

which implies that, if $\nabla f(\boldsymbol{x}^\star) = 0$, then

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^\star) \quad \forall \boldsymbol{x}.$$

If $\boldsymbol{x}$ is a global minimum, then $\nabla f(\boldsymbol{x}) = 0$. ■

6 Convergence of gradient descent

Gradient descent doesn't work well when functions change too rapidly. A way to quantify the rate of change of a function is its "Lipschitzness". Figure 1 shows a examples of functions that change and do not change too rapidly.

6.1 Lipschitzness

Definition 4 (L-Lipschitz/Bounded Gradient) *The function f is L-Lipschitz if $\forall x, y \in dom(f)$, $|f(x) - f(y)| \leq L\|x - y\|_{\ell_2}$. Furthermore, if f is also differentiable, then*

$$||\nabla f(\boldsymbol{x})|| \leq L.$$

Lemma 9 *We have*

(1) If $|f(\boldsymbol{x}) - f(\mathbf{y})| \leq L\|\boldsymbol{x} - \mathbf{y}\|_{\ell_2}$ and f is differentiable then $\|\nabla f(\boldsymbol{x})\|_{\ell_2} \leq L$.

(2) If $\|\nabla f(\boldsymbol{x})\|_{\ell_2} \leq L$ then $|f(\boldsymbol{x}) - f(\mathbf{y})| \leq L\|\boldsymbol{x} - \mathbf{y}\|_{\ell_2}$.

Proof The first part follows from the mean value theorem. For the second part, by the Fundamental Theorem of Calculus, we have

$$f(\boldsymbol{x}) - f(\mathbf{y}) = \int_0^1 \langle \nabla f(t\boldsymbol{x} + (1-t)\mathbf{y}), (\boldsymbol{x} - \mathbf{y}) \rangle dt$$

Therefore,

$$\begin{aligned}
\|f(\boldsymbol{x}) - f(\mathbf{y})\| &= \left| \int_0^1 \langle \nabla f(t\boldsymbol{x} + (1-t)\mathbf{y}), (\boldsymbol{x} - \mathbf{y}) \rangle dt \right| \\
&\leq \int_0^1 \left| \langle \nabla f(t\boldsymbol{x} + (1-t)\mathbf{y}), (\boldsymbol{x} - \mathbf{y}) \rangle \right| dt \\
&\leq \int_0^1 \|\nabla f(t\boldsymbol{x} + (1-t)\mathbf{y})\| \cdot \|\boldsymbol{x} - \mathbf{y}\| dt \\
&= \left(\int_0^1 \|\nabla f(t\boldsymbol{x} + (1-t)\mathbf{y})\| dt \right) \cdot \|\boldsymbol{x} - \mathbf{y}\| \\
&\leq L\|\boldsymbol{x} - \mathbf{y}\|
\end{aligned}$$

■

Theorem 10 *Let f be a convex and L-Lipschitz function. Furthermore, assume the function is L-Lipschitz and the initial estimate obeys $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| \leq R$. Then, for a step-size obeying $\mu = \frac{R}{L\sqrt{t}}$ we have*

$$f\left(\frac{1}{t} \sum_{s=1}^{t} \boldsymbol{x}_s \right) - f(\boldsymbol{x}^\star) \leq \frac{RL}{\sqrt{t}}$$

Proof

$$\begin{aligned}
f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star) &\leq \langle \nabla f(\boldsymbol{x}_s), \boldsymbol{x}_s - \boldsymbol{x}^\star \rangle \\
&= \frac{1}{\mu} \langle \boldsymbol{x}_s - \boldsymbol{x}_{s+1}, \boldsymbol{x}_s - \boldsymbol{x}^\star \rangle \\
&\overset{(a)}{=} \frac{1}{2\mu} (\|\boldsymbol{x}_s - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{x}_s - \boldsymbol{x}_{s+1}\|^2 - \|\boldsymbol{x}_{s+1} - \boldsymbol{x}^\star\|^2) \\
&\overset{(b)}{=} \frac{1}{2\mu} \left(\|\boldsymbol{x}_s - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{s+1} - \boldsymbol{x}^\star\|^2 \right) + \frac{\mu}{2} \|\nabla f(\boldsymbol{x}_s)\|^2
\end{aligned}$$

Where (a) holds from the fact that

$$2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2,$$

and by taking $\mathbf{a} = \boldsymbol{x}_s - \boldsymbol{x}^\star$ and $\mathbf{b} = \boldsymbol{x}_s - \boldsymbol{x}_{s+1}$, and (b) holds because $\|\boldsymbol{x}_s - \boldsymbol{x}_{s+1}\|^2 = \mu^2 \|\nabla f(\boldsymbol{x}_s)\|$. Therefore, summing over all s, and canceling the telescopic sum values, we have

$$\begin{aligned}
\sum_{s=1}^{t} (f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)) &\leq \frac{1}{2\mu} \left(\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \right) + \frac{\mu}{2} \sum_{s=1}^{t} \|\nabla f(\boldsymbol{x}_s)\|^2 \\
&\leq \frac{1}{2\mu} \|x_1 - x^\star\|^2 + \frac{\mu}{2} L^2 t \\
&\leq \frac{R^2}{2\mu} + \frac{\mu}{2} L^2 t.
\end{aligned}$$

This implies that

$$\frac{1}{t}\sum_{s=1}^{t} f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star) \leq \frac{R^2}{2\mu t} + \frac{\mu}{2}L^2$$

Therefore, using Jensen's inequality, we conclude that

$$f\left(\frac{1}{t}\sum_{s=1}^{t}\mathbf{x_s}\right) - f(\boldsymbol{x}^\star) \leq \frac{1}{t}\sum_{s=1}^{t}\left(f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star)\right) \leq \frac{R^2}{2\mu t} + \frac{\mu}{2}L^2 \leq \frac{RL}{\sqrt{t}}$$

with $\mu = \frac{R}{L\sqrt{t}}$ and convexity. ■

The previous result implies that to achieve an error $\leq \epsilon$, we need $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations. A few remarks are in order

- This result cannot be improved,
- It is independent of n (the dimension of the data!)
- We can vary the step size in each iteration: $\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - \mu_s \nabla f(\boldsymbol{x}_s)$, $\mu_s = \frac{R}{L\sqrt{s}}$. This convergence rate is optimal for L-Lipschitz functions.

6.2 Projection onto Sets

We have looked at optimization without constraints. But, what happens if we do have them?

Proposition 11 *Assume Ω is convex and closed, we have the following optimization problem with constraints*

$$\min_x f(\boldsymbol{x})$$
$$s.t.\ \boldsymbol{x} \in \Omega$$

Definition 5 (Projection onto Ω) *The projection onto Ω is defined as*

$$\Pi_\Omega(\boldsymbol{x}) = \arg\min_{\mathbf{y}\in\Omega} \|\boldsymbol{x} - \mathbf{y}\|$$

Examples of projections are:

- $\Omega = \{\boldsymbol{x} : \boldsymbol{x} \geq 0\} \Rightarrow$ All $\boldsymbol{x} < \mathbf{0}$ are set to $\mathbf{0}$.
- $\Omega = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq 1\} \Rightarrow \Pi_\Omega(y) = \begin{cases} \mathbf{y}, & \|\mathbf{y}\| \leq 1 \\ \frac{\mathbf{y}}{\|\mathbf{y}\|}, & \|\mathbf{y}\| > 1 \end{cases}$

Lemma 12 *Let $\boldsymbol{x} \in \Omega$ (convex) and $\mathbf{y} \in \mathbb{R}^n$. Then*

$$\langle \Pi_\Omega(\mathbf{y}) - \boldsymbol{x}, \Pi_\Omega(\mathbf{y}) - \mathbf{y}\rangle \leq 0.$$

Corollary 13 *Following from Lemma 12,*

$$\|\Pi_\Omega(\mathbf{y}) - \boldsymbol{x}\|^2 + \|\mathbf{y} - \Pi_\Omega(\mathbf{y})\|^2 \leq \|\mathbf{y} - \boldsymbol{x}\|^2$$

Corollary 14 *For $x \in \Omega$, $\forall y$,*

$$\|\Pi_\Omega(\mathbf{y}) - \boldsymbol{x}\| \leq \|\mathbf{y} - \boldsymbol{x}\|$$

For gradient descent, the iterations look like:

$$\boldsymbol{x}_{t+1} = \Pi_\Omega(\boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t))$$

Let $\mathbf{y}_{t+1} = \boldsymbol{x}_t - \mu \nabla f(\mathbf{x_t})$, and $\boldsymbol{x}^\star \in \Omega$. Then,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \leq \|\mathbf{y}_{t+1} - \boldsymbol{x}^\star\|$$

Out starting point for Proof of Theorem 10 was $f(\boldsymbol{x}_s) - f(\boldsymbol{x}^\star) \leq \langle \nabla f(\boldsymbol{x}_s), \boldsymbol{x}_s - \boldsymbol{x}^\star \rangle^2$. The exact same proof works with projections.

Lemma 15 *All of the following are equivalent ($\iff$):*

- *f is convex and L-smooth*
- *$g(\boldsymbol{x}) = \frac{L}{2}\boldsymbol{x}^\top \boldsymbol{x} - f(\boldsymbol{x})$ is convex*
- *$f(\mathbf{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \mathbf{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\mathbf{y} - \boldsymbol{x}\|^2$*
- *$\nabla^2 f(\boldsymbol{x}) \leq LI$, where I is the identity matrix*