Week 1 - L01
Convex Optimization and Gradient Descent: Basics

CS 295 Optimization for Machine Learning

Ioannis Panageas

Basics

Many machine learning problems involve learning parameters $\theta \in \Theta$ of a function, towards achieving an objective. Objectives are characterized by a loss function $L : \Theta \to \mathbb{R}$.

Example in supervised learning given n samples (x_i, y_i) where x is the input:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \overbrace{l(\underbrace{f(x_i, \theta)}_{\text{prediction}}, \underbrace{y_i}_{\text{label}})}^{\text{distance between } y_i \text{ and } f(x_i, \theta)}$$

Goal: $\min_{\theta \in \Theta} L(\theta)$

Typically solving $\min_{x \in \mathcal{X}} f(x)$ is NP-hard (computationally intractable).

Basics

Many machine learning problems involve learning parameters $\theta \in \Theta$ of a function, towards achieving an objective. Objectives are characterized by a loss function $L : \Theta \to \mathbb{R}$.

Example in supervised learning given n samples (x_i, y_i) where x is the input:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \overbrace{l(\underbrace{f(x_i, \theta)}_{\text{prediction}}, \underbrace{y_i}_{\text{label}})}^{\text{distance between } y_i \text{ and } f(x_i,\theta)}$$

Goal: $\min_{\theta \in \Theta} L(\theta)$

Typically solving $\min_{x \in \mathcal{X}} f(x)$ is NP-hard (computationally intractable).

Nevertheless, for certain classes of functions f, strong theoretical guarantees and efficient optimization algorithms exist!

- Classes of functions f: Convex!
- Algorithm: Gradient Descent!

Definitions

Definition (Convex combination)**.** *$z \in \mathbb{R}^d$ is a convex combination of $x1, x2, ..., x_n \in \mathbb{R}^d$ if*

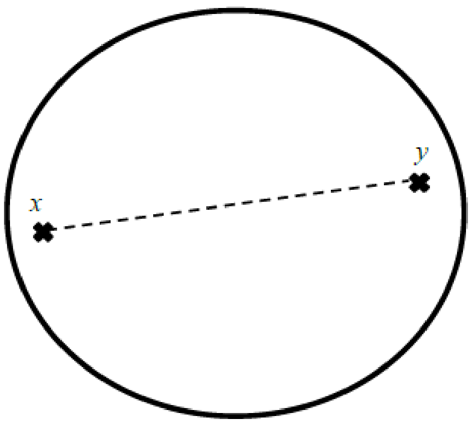
$$z = \sum \lambda_i x_i,\ \lambda_i \geq 0 \textit{ for all } i \textit{ and } \sum_i \lambda_i = 1.$$

Definitions

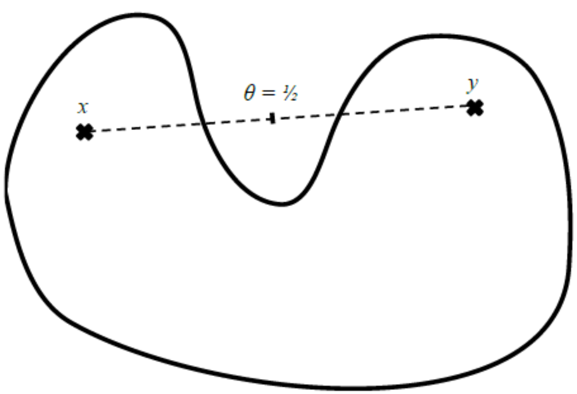
Definition (Convex combination). $z \in \mathbb{R}^d$ *is a convex combination of* $x1, x2, ..., x_n \in \mathbb{R}^d$ *if*

$$z = \sum \lambda_i x_i, \ \lambda_i \geq 0 \textit{ for all } i \textit{ and } \sum_i \lambda_i = 1.$$

Definition (Convex set). $\mathcal{X}$ *is a convex set if the convex combination of any two points in* $\mathcal{X}$ *belongs also in* $\mathcal{X}$.



Convex set



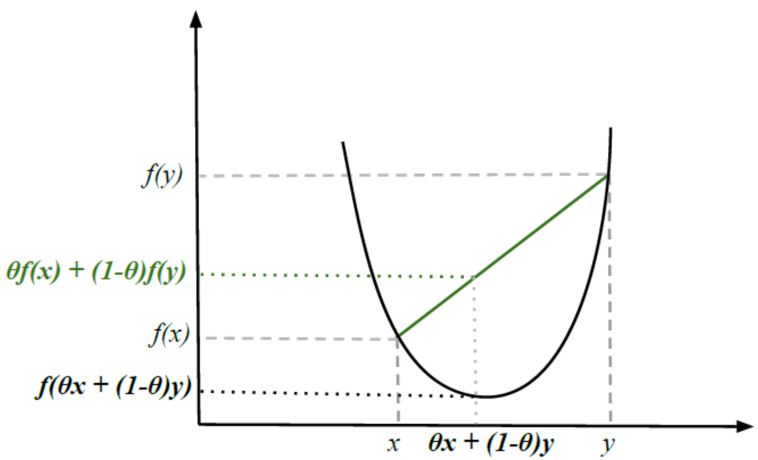
Non-convex set

Definitions cont.

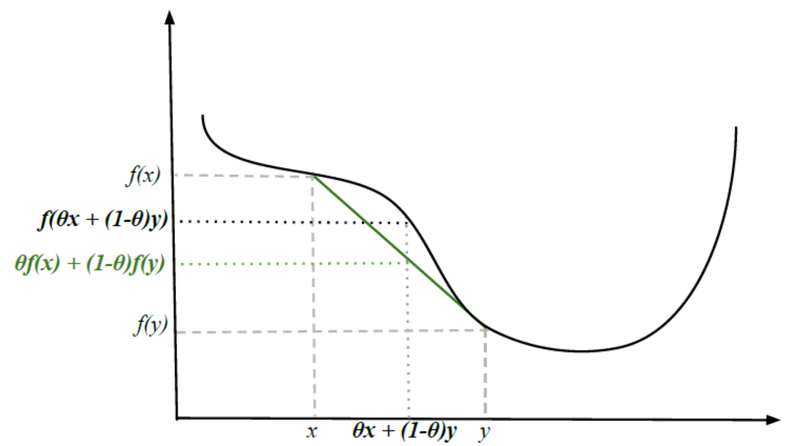
Definition (Convex function). *A function $f(x)$ is convex if and only if the domain $dom(f)$ is a convex set and $\forall x, y \in dom(f), \theta \in [0,1]$*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

Concave function f: $-f$ is convex, i.e., inequality above is reversed!
Moreover, if the inequality is strict, f is called strictly convex.



Convex function



Non-convex function

Basic Facts

Lemma (First order condtion for convexity)**.** *A differentiable function $f(x)$ is convex if and only if the domain $dom(f)$ is a convex set and $\forall x, y \in dom(f)$*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

Proof. ($\Rightarrow$)By convexity we have that (for all $t > 0$)

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x).$$

Rearranging a bit follows

$$f(x + t(y-x)) \leq t(f(y) - f(x)) + f(x).$$

Dividing by t we conclude:

$$f(y) - f(x) \geq \frac{f(x + t(y-x)) - f(x)}{t}.$$

Basic Facts

Proof (⇒) *cont.* Hence

$$f(y) - f(x) \geq \underbrace{\lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t}}_{\text{directional derivative}} = \nabla f(x)^\top (y - x).$$

Basic Facts

Proof (⇒) cont. Hence

$$f(y) - f(x) \geq \underbrace{\lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t}}_{\text{directional derivative}} = \nabla f(x)^\top (y - x).$$

Proof. (⇐) Choose first $z = tx + (1 - t)y$ for $t \in (0, 1)$ and moreover it holds that

- $f(x) \geq f(z) + \nabla f(z)^\top (x - z)$.
- $f(y) \geq f(z) + \nabla f(z)^\top (y - z)$.

Multiply first by t, second by $(1 - t)$ and add them up.

Basic Facts cont.

Lemma (Second order condtion for convexity)**.** *A twice differentiable function* $f(x)$ *is convex if and only if the domain* $dom(f)$ *is a convex set and* $\forall x \in dom(f)$

$$\nabla^2 f(x) \succeq 0.$$

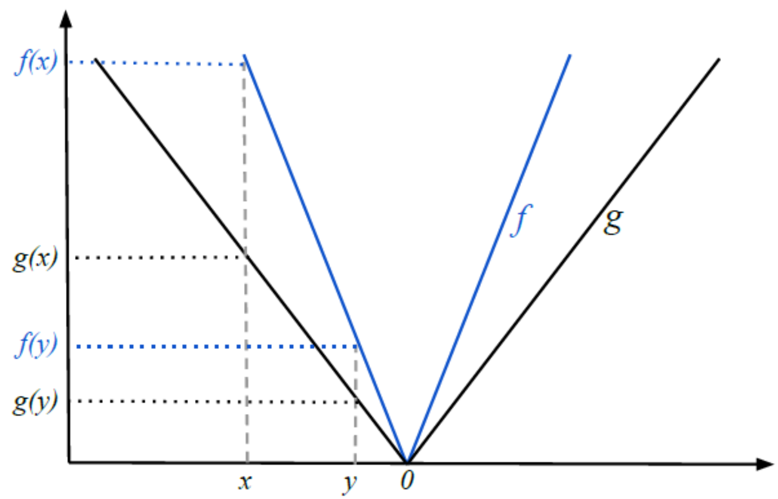
In words, the Hessian of f should be positive semi-definite.

Proof. Exercise 1 for homework...

More Definitions

Definition (Lipschitz function). *A function* $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ *is L-Lipschitz continuous iff for* $L > 0$ *and* $\forall x, y \in dom(f)$

$$\|f(x) - f(y)\|_2 \leq L \|x - y\|_2 .$$



L_f-Lipschitz continuous function f and a L_g-Lipschitz continuous function g with $L_f > L_g$.

More Definitions cont.

Definition (Smoothness). *A continuously differentiable function* $f(x)$ *is L-smooth if its gradient is L-Lipschitz, i.e., there exists a* $L > 0$ *and* $\forall x, y \in dom(f)$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 .$$

Definition (Strongly convex). *A function* $f(x)$ *is* α*-strongly convex if for* $\alpha > 0$ *and* $\forall x \in dom(f)$

$$f(x) - \frac{\alpha}{2} \|x\|_2^2 \text{ is convex.}$$

Exercise 2. *Suppose* $f(x)$ *is differentiable and* α*-strongly convex. Then* $\forall x, y \in dom(f)$

$$f(y) - f(x) \geq \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|_2^2 .$$

Minimizing convex functions

- We examine this class of functions because are easier to minimize.

Lemma (Gradient zero). *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be differentiable and convex.* x^* *is a minimizer if and only if* $\nabla f(x^*) = 0$. *Hence all minimizers give same* f*-value.*

Proof. ($\Leftarrow$)By FOC for convexity we have that $\forall x \in \mathrm{dom}(f)$

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) = f(x^*).$$

Minimizing convex functions

- We examine this class of functions because are easier to minimize.

Lemma (Gradient zero). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex. x^* is a minimizer if and only if $\nabla f(x^*) = 0$. Hence all minimizers give same f-value.*

Proof. ($\Leftarrow$)By FOC for convexity we have that $\forall x \in \mathrm{dom}(f)$

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) = f(x^*).$$

Proof. ($\Rightarrow$) Choose $t > 0$ small enough such that $y := x^* - t\nabla f(x^*)$ is in $\mathrm{dom}(f)$. By Taylor we have

$$\begin{aligned} f(y) - f(x^*) &= \nabla f(x^*)^\top (y - x^*) + o(\|y - x^*\|_2) \\ &= -t\,\|\nabla f(x^*)\|_2^2 + o(\|t\nabla f(x^*)\|_2). \end{aligned}$$

For t small enough $f(y) - f(x^*) < 0$ if $\nabla f(x^*) \neq 0$ (contradiction).

Gradient Descent (GD) (for differentiable functions)

Definition (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable (want to minimize). The algorithm below is called gradient descent*

$$x_{k+1} = x_k - \alpha \nabla f(x_k).$$

Remarks

- α is called the stepsize. Intuitively the smaller, the slower the algorithm.
- α may or may not depend on k.
- If GD converges, it means that $\nabla f(x) \to 0$, so we should have "convergence" to the minimizer (for f convex)!
- The minimizers of f are fixed points of GD.

Analysis of GD for L-Lipschitz

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex (want to minimize) and L-Lipschitz. Let $R = \|x_1 - x^*\|_2$, the distance between the initial point x_1 and minimizer x^*. It holds for $T = \frac{R^2 L^2}{\epsilon^2}$*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{\epsilon}{L^2}$.

Remarks

- The speed of convergence is independent of dimension d.
- This result gives a rate of $O\left(\frac{1}{\epsilon^2}\right)$. With smoothness assumptions we can do $O\left(\frac{1}{\epsilon}\right)$.
- There is Nesterov's accelerated method that can achieve $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ (under smoothness).
- With smoothness and strong-convexity assumptions we can do $O\left(\ln\frac{1}{\epsilon}\right)$.
- The theorem does not imply pointwise convergence $f(x_T) \to f(x^*)$.

Analysis of GD for L-Lipschitz

Proof. It holds that

$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$ FOC for convexity,

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\le \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \text{ law of Cosines,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \text{ Def. of GD,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

$$\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \text{ Def. of GD,} \\
&\leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \text{ Exercise 3.}
\end{aligned}$$

Exercise 3. *Suppose $f(x)$ is L-Lipschitz continous. Then $\forall x \in dom(f)$*

$$\|\nabla f(x)\|_2 \leq L.$$

Analysis of GD for L-Lipschitz

Proof cont. Since

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2},$$

taking the telescopic sum we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) &\leq \frac{1}{2\alpha T}(\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\alpha L^2}{2}. \\ &\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \text{ by choosing appropriately } \alpha, T.\end{aligned}$$

The claim follows by convexity since $\frac{1}{T}\sum_{t=1}^{T} f(x_t) \geq f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)$ (Jensen's inequality).

Recap Lecture 1

- Introduction to Convex Optimization
  - Easy to minimize objectives (generally is NP-hard).
  - Focus on Gradient Descent.
  - GD has rate of convergence $O\left(\frac{L^2}{\epsilon^2}\right)$ for L-Lipschitz.
- Today
  - GD has rate of convergence $O\left(\frac{L}{\epsilon}\right)$ for L-smooth.
  - GD has rate of convergence $O\left(\frac{L}{\mu}\ln\frac{1}{\epsilon}\right)$ for L-smooth, μ-convex.
  - *Projected* GD (similar analysis) for constrained optimization.

Analysis of GD for L-smooth

Theorem (Gradient Descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex (want to minimize) and L-smooth. Let $R = \|x_1 - x^*\|_2$. It holds for $T = \frac{2R^2L}{\epsilon}$*

$$f(x_{T+1}) - f(x^*) \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Remarks

- Speed of convergence is independent of dimension d.
- This result gives a rate of $O\left(\frac{1}{\epsilon}\right)$, different choice of stepsize.
- The theorem implies convergence $f(x_T) \to f(x^*)$.

Analysis of GD for L-smooth

Before showing the proof, we show some important claims for L-smooth functions.

Claim 1. *Let f be a differentiable and L-smooth, then*

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

Proof. It holds that

$$f(y) - f(x) - \nabla f(y)^\top (x - y) = \int_0^1 \nabla f(y + t(x - y))^\top (x - y) dt - \nabla f(y)^\top (x - y)$$

Analysis of GD for L-smooth

Before showing the proof, we show some important claims for L-smooth functions.

Claim 1. *Let f be a differentiable and L-smooth, then*

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

Proof. It holds that

$$\begin{aligned} f(y) - f(x) - \nabla f(y)^\top (x - y) &= \int_0^1 \nabla f(y + t(x - y))^\top (x - y) dt - \nabla f(y)^\top (x - y) \\ &= \left(\int_0^1 \nabla f(y + t(x - y)) dt - \nabla f(y) \right)^\top (x - y) \end{aligned}$$

Analysis of GD for L-smooth

Before showing the proof, we show some important claims for L-smooth functions.

Claim 1. *Let f be a differentiable and L-smooth, then*

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

Proof. It holds that

$$\begin{aligned}
f(y) - f(x) - \nabla f(y)^\top (x - y) &= \int_0^1 \nabla f(y + t(x-y))^\top (x-y)dt - \nabla f(y)^\top (x-y) \\
&= \left(\int_0^1 \nabla f(y + t(x-y))dt - \nabla f(y) \right)^\top (x-y) \\
&= \left(\int_0^1 \{\nabla f(y + t(x-y)) - \nabla f(y)\} dt \right)^\top (x-y)
\end{aligned}$$

Analysis of GD for L-smooth

Before showing the proof, we show some important claims for L-smooth functions.

Claim 1. *Let f be a differentiable and L-smooth, then*

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

Proof. It holds that

$$\begin{aligned}
f(y) - f(x) - \nabla f(y)^\top (x - y) &= \int_0^1 \nabla f(y + t(x-y))^\top (x-y) dt - \nabla f(y)^\top (x - y) \\
&= \left(\int_0^1 \nabla f(y + t(x-y)) dt - \nabla f(y) \right)^\top (x - y) \\
&= \left(\int_0^1 \{\nabla f(y + t(x-y)) - \nabla f(y)\} dt \right)^\top (x - y) \\
\text{using } L\text{-smoothness} \qquad &\leq L \textstyle\int_0^1 t dt \, \|x - y\|_2^2 = \frac{L}{2} \|x - y\|_2^2 .
\end{aligned}$$

Analysis of GD for L-smooth

Claim 2. *Let f be a differentiable, convex and L-smooth, then*

$$f(x^*) - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\left\|\nabla f(x)\right\|_2^2.$$

Proof. Set $z = x - \frac{1}{L}\nabla f(x)$. First inequality is trivial (definition of minizer).

$$f(z) - f(x) \leq \nabla f(x)^\top (z - x) + \tfrac{L}{2}\left\|z - x\right\|_2^2 \text{ using Claim 1,}$$

Analysis of GD for L-smooth

Claim 2. *Let f be a differentiable, convex and L-smooth, then*

$$f(x^*) - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2 .$$

Proof. Set $z = x - \frac{1}{L}\nabla f(x)$. First inequality is trivial (definition of minizer).

$$f(z) - f(x) \leq \nabla f(x)^\top (z - x) + \frac{L}{2}\|z - x\|_2^2 \text{ using Claim 1,}$$

$$= -\frac{1}{L}\nabla f(x)^\top \nabla f(x) + \frac{L}{2}\frac{1}{L^2}\|\nabla f(x)\|_2^2 ,$$

Analysis of GD for L-smooth

Claim 2. *Let f be a differentiable, convex and L-smooth, then*

$$f(x^*) - f(x) \le f(x - \frac{1}{L}\nabla f(x)) - f(x) \le -\frac{1}{2L}\|\nabla f(x)\|_2^2 .$$

Proof. Set $z = x - \frac{1}{L}\nabla f(x)$. First inequality is trivial (definition of minizer).

$$\begin{aligned} f(z) - f(x) &\le \nabla f(x)^\top (z - x) + \tfrac{L}{2}\|z - x\|_2^2 \text{ using Claim 1,} \\ &= -\tfrac{1}{L}\nabla f(x)^\top \nabla f(x) + \tfrac{L}{2}\tfrac{1}{L^2}\|\nabla f(x)\|_2^2 , \\ &= -\tfrac{1}{2L}\|\nabla f(x)\|_2^2 . \end{aligned}$$

Analysis of GD for L-smooth

Proof of Theorem. Assume $\|x_t - x^*\|_2$ is decreasing in t (**Exercise 4 to prove**).

Using Claim 2,

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} \|\nabla f(x_t)\|_2^2 .$$

Analysis of GD for L-smooth

Proof of Theorem. Assume $\|x_t - x^*\|_2$ is decreasing in t (**Exercise 4 to prove**).

Using Claim 2,

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} \|\nabla f(x_t)\|_2^2 .$$

From convexity we get,

$$\begin{aligned} f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*) &\leq \|\nabla f(x_t)\|_2 \|x_t - x^*\|_2 \ \text{(C-S inequality} \\ &\leq \|\nabla f(x_t)\|_2 \|x_0 - x^*\|_2 \ \text{(Assumption).} \end{aligned}$$

Analysis of GD for L-smooth

Proof of Theorem. Assume $\|x_t - x^*\|_2$ is decreasing in t (**Exercise 4 to prove**).

Using Claim 2,

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} \|\nabla f(x_t)\|_2^2 .$$

From convexity we get,

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*) \leq \|\nabla f(x_t)\|_2 \|x_t - x^*\|_2 \text{ (C-S inequality}$$
$$\leq \|\nabla f(x_t)\|_2 \|x_0 - x^*\|_2 \text{ (Assumption).}$$

Combining the two

$$f(x_{t+1}) - f(x^*) - (f(x_t) - f(x^*)) \leq -\frac{1}{2L} \frac{(f(x_t) - f(x^*))^2}{R^2}.$$

Setting $\delta_t = f(x_t) - f(x^*)$, we get $\delta_{t+1} \leq \delta_t - \frac{\delta_t^2}{2LR^2}$

Analysis of GD for L-smooth

Proof of Theorem. Assume $\|x_t - x^*\|_2$ is decreasing in t (**Exercise 4 to prove**).

Using Claim 2,

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} \|\nabla f(x_t)\|_2^2 .$$

Easy to show (board) $\delta_t \leq \frac{2LR^2}{t-1}$.

QED

Combining the two

$$f(x_{t+1}) - f(x^*) - (f(x_t) - f(x^*)) \leq -\frac{1}{2L} \frac{(f(x_t) - f(x^*))^2}{R^2}.$$

Setting $\delta_t = f(x_t) - f(x^*)$, we get $\delta_{t+1} \leq \delta_t - \frac{\delta_t^2}{2LR^2}$