Optimization in Machine Learning: Gradient Descent Method

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Spring 2023)
http://cs182.sist.shanghaitech.edu.cn

Outline

Outline

The Gradient Descent Method

The optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}),$$

Gradient iteration:

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \gamma_r \cdot \nabla f(\mathbf{x}^r).$$

arguably, the most widely used optimization method in machine learning, which can be used for

- linear regression
- logistic regression
- neural networks
- etc.

Cost of Solving Linear Regression

- Problem: given $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, solve the linear regression problem with squared loss

$$f(\mathbf{w}) = \frac{1}{2} \sum_{t=1}^{n} \left(y^t - \mathbf{w}^T \mathbf{x}^t \right)^2 = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

- Two strategies for linear regression:
  - Closed-form solution by setting $\nabla f(\mathbf{w}) = \mathbf{0}$ (i.e., the normal equations):

$$\mathbf{w}^* = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{y}\right),$$

    which costs $O\left(nd^2 + d^3\right)$.
  - T iterations of gradient descent:

$$\mathbf{w}^{r+1} = \mathbf{w}^r - \gamma_r \mathbf{X}^\top \left(\mathbf{X}\mathbf{w}^r - \mathbf{y}\right), \ r = 0, 1, \ldots$$

    which costs $O(Tnd)$.
- GD is faster if total number of iterations T is not too big
  - If we only do $T < \max\left\{d, d^2/n\right\}$ iterations.

Cost of Solving Logistic Regression

- GD can also be applied to other models like logistic regression,
  - when labels $y^t \in \{0, 1\}$

$$f(\mathbf{w}) = -\sum_{t=1}^{n} \left[y^t \log p^t + \left(1 - y^t\right) \log \left(1 - p^t\right) \right]$$

    where $p^t = \frac{1}{1+\exp(\mathbf{w}^\top \mathbf{x}^t)}$
  - when labels $y^t \in \{-1, 1\}$

$$f(\mathbf{w}) = \sum_{t=1}^{n} \log \left(1 + \exp\left(-y^t \mathbf{w}^\top \mathbf{x}^t\right)\right)$$

  - Setting $\nabla f(\mathbf{w}) = \mathbf{0}$ gives a system of transcendental equations.
  - We cannot formulate it as a linear system or a linear program.
- But this objective function is convex and differentiable.
  - So GD converges to a global optimum.

The Gradient Descent Method

The optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),$$

Gradient iteration:

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \gamma_r \cdot \nabla f(\mathbf{x}^r).$$

Questions:

- why?
- where: $\mathbf{x}^r \to ?$
- when: rate of convergence $\mathbf{x}^r \to \mathbf{x}^\infty$ (how many iterations T of GD do we need?)

Outline

Functions

- continuity, Lipschitz continuity
- differentiability, continuous differentiability
- smoothness, Lipschitz smoothness
- convexity
- coercivity
- ...

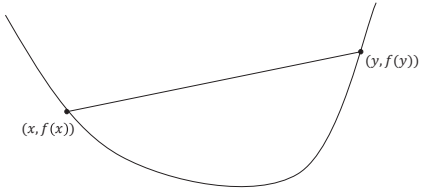
Two important classes of functions:

- convex functions
- smooth functions
- strongly convex functions
- ...

Convex functions

- A function $f : \mathrm{dom} f \to \mathbb{R}$ is convex if $\mathrm{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, and λ with $0 \leq \lambda \leq 1$, we have

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$



- f is convex iff its restriction to any line $\{\mathbf{x} + t\mathbf{v} : t \in \mathbb{R}\}$ is convex
- f is strictly convex if the inequality holds whenever $\mathbf{x} \neq \mathbf{y}$ and $0 < \lambda < 1$
- f is concave if $-f$ is convex

Convex functions

- Definition (zeroth order condition):

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- First order condition:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

- Second order condition:

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

Proof. Zeroth order condition $\Rightarrow$ First order condition
We first consider the case $x, y \in \mathbb{R}^1$. If $f : \mathbb{R} \to \mathbb{R}$ is convex and $x, y \in \mathrm{dom} f$, we have

$$f(y + \lambda(x - y)) \leq \lambda f(x) + (1 - \lambda) f(y),$$

and hence

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

If we divide both sides by λ and take the limit as $\lambda \to 0$, we obtain

$$\begin{aligned} f(x) &\geq f(y) + \lim_{\lambda \to 0} \frac{f(y + \lambda(x - y)) - f(y)}{\lambda} \\ &= f(y) + f'(y)(x - y). \end{aligned}$$

Proof. Zeroth order condition $\Rightarrow$ First order condition
Now we consider the case $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$ and consider f restricted to the line passing through them, i.e., the function defined by $g(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$, so

$$g'(t) = \nabla f(t\mathbf{x} + (1-t)\mathbf{y})^\top (\mathbf{x} - \mathbf{y}).$$

Assume f is convex, which implies g is convex, so by the argument above we have

$$g(1) \geq g(0) + g'(0)(1-0),$$

which means

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

□

Proof. First order condition $\Rightarrow$ Zeroth order condition
Suppose $\mathbf{z} = \lambda \mathbf{x} + (1-\lambda)\,\mathbf{y}$, we have

$$\begin{cases} f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^{\top} (\mathbf{x} - \mathbf{z}) & (a) \\ f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^{\top} (\mathbf{y} - \mathbf{z}) & (b) \end{cases}$$

Consider $\lambda \cdot (a) + (1-\lambda) \cdot (b)$, we have

$$\begin{aligned} \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) &\geq f(\mathbf{z}) + \nabla f(\mathbf{z})^{\top} (\lambda \mathbf{x} + (1-\lambda)\,\mathbf{y} - \mathbf{z}) \\ &= f(\mathbf{z}) \end{aligned}$$

□

Proof. First order condition $\Rightarrow$ Second order condition
With a $\tau > 0$, we have

$$f(\mathbf{x} + \tau\mathbf{d}) = f(\mathbf{x}) + \tau\nabla f(\mathbf{x})^\top \mathbf{d} + \frac{\tau^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x})\,\mathbf{d} + o\left(\|\tau\mathbf{d}\|^2\right).$$

By the first order condition, we have

$$f(\mathbf{x} + \tau\mathbf{d}) \geq f(\mathbf{x}) + \tau\nabla f(\mathbf{x})^\top \mathbf{d}.$$

Thus we obtain

$$\frac{\tau^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x})^\top \mathbf{d} + o\left(\|\tau\mathbf{d}\|^2\right) \geq 0.$$

Dividing both sides by τ^2 and take the limit as $\tau \to 0^+$, we obtain

$$\lim_{\tau\to 0^+}\left[\frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x})\,\mathbf{d} + \frac{o\left(\|\tau\mathbf{d}\|^2\right)}{\tau^2}\right] = \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x})\,\mathbf{d} \geq 0$$

□

Proof. Second order condition $\Rightarrow$ First order condition
For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, from second-order Taylor expansion we know that there exists a $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$ such that

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})^\top \nabla^2 f(\mathbf{z}) (\mathbf{x} - \mathbf{y})$$

We know $\nabla^2 f(\mathbf{z}) \succeq 0$, so we have

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

□

Smooth functions

- A function $f : \mathrm{dom} f \to \mathbb{R}$ is L-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, there exists a constant $L < +\infty$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

- "L-smooth" is also called "L-Lipschitz gradient", i.e., the gradient is Lipschitz continuous with Lipschitz constant L.
  - This is a fairly weak assumption, which is true in almost all ML models.
- The definition implicitly assumes f is differentiable.

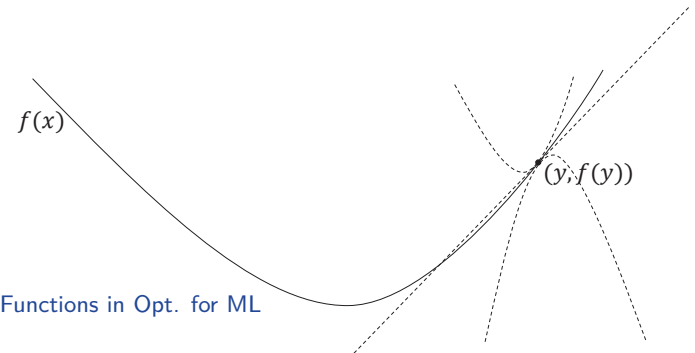
- Descent Lemma:

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

or, equivalently,

$$\begin{cases} f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, & \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \\ f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, & \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \end{cases}$$

meaning that f is bounded above and below by a quadratic function.



$$\boxed{|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.}$$

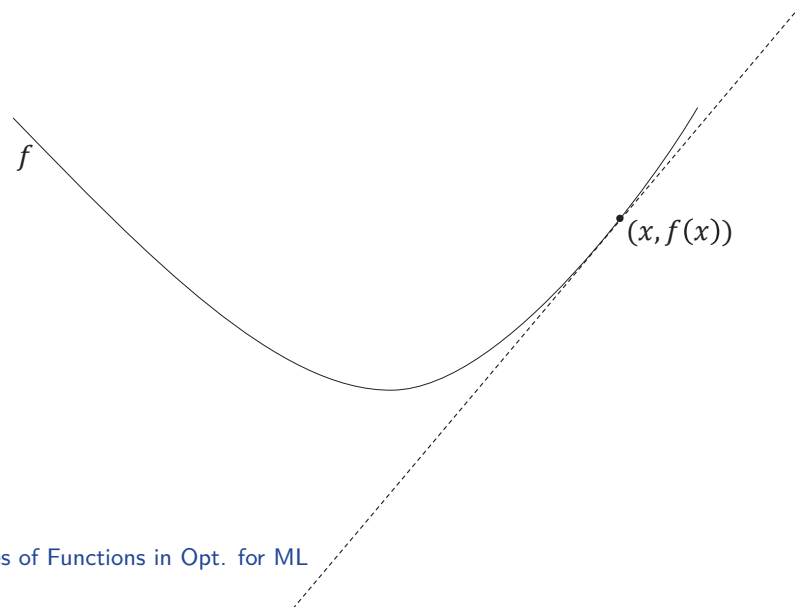
Proof. Using Taylor expansion:

$$f(\mathbf{x}) = f(\mathbf{y}) + \int_0^1 \nabla f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right)^\top (\mathbf{x} - \mathbf{y})\, dt.$$

Compare terms

$$\begin{aligned}
&|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \\
&= \left| \int_0^1 \nabla f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right)^\top (\mathbf{x} - \mathbf{y})\, dt - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \right| \\
&\leq \int_0^1 \left| \left(\nabla f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right) - \nabla f(\mathbf{y})\right)^\top (\mathbf{x} - \mathbf{y}) \right| dt \quad \text{(why?)} \\
&\leq \int_0^1 (1 - t)\, L\|\mathbf{x} - \mathbf{y}\|^2 dt = \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2.
\end{aligned}$$

First order expansion provides a good local approximation of f.



- If f is twice differentiable, f is L-smooth is equivalent to $\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$, or all the eigenvalues of $\nabla^2 f(\mathbf{x})$ are bounded above by L, i.e., $\mathbf{d}^T \nabla^2 f(\mathbf{x})\mathbf{d} \leq L\|\mathbf{d}\|^2$.

Proof.

Sufficiency: Suppose that f is L-smooth. Then by the fundamental theorem of calculus, for any $\mathbf{x} \in \mathbb{R}^n$ and $\alpha > 0$,

$$\nabla f(\mathbf{x} + \alpha\mathbf{d}) - \nabla f(\mathbf{x}) = \int_0^\alpha \nabla^2 f(\mathbf{x} + t\mathbf{d})\,\mathbf{d}\,dt.$$

Thus,

$$\left\| \left(\int_0^\alpha \nabla^2 f(\mathbf{x} + t\mathbf{d})\,\mathbf{d}\,dt \right) \mathbf{d} \right\|_2 = \|\nabla f(\mathbf{x} + \alpha\mathbf{d}) - \nabla f(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2 .$$

Dividing by α and taking the limit $\alpha \to 0^+$, we obtain

$$\left\| \nabla^2 f(\mathbf{x})\,\mathbf{d} \right\|_2 \leq L \|\mathbf{d}\|_2 \text{ for any } \mathbf{d} \in \mathbb{R}^n.$$

Necessity: Suppose that $\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$, which means $\left\|\nabla^2 f(x)\right\|_2 \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$. Then by the fundamental theorem of calculus, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\begin{aligned}\nabla f(\mathbf{y}) &= \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\, dt \\ &= \nabla f(\mathbf{x}) + \left(\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\, dt\right) \cdot (\mathbf{y} - \mathbf{x}).\end{aligned}$$

Then

$$\begin{aligned}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 &= \left\|\left(\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\, dt\right) \cdot (\mathbf{y} - \mathbf{x})\right\|_2 \\ &\leq \left(\int_0^1 \left\|\nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\right\|_2 dt\right) \|(\mathbf{y} - \mathbf{x})\|_2 \\ &\leq L \|(\mathbf{y} - \mathbf{x})\|_2.\end{aligned}$$

□

Outline

Unconstrained optimization

The optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}),$$

$f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

- Global minimizer: $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$.
- Local minimizer: $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for $\mathbf{x} \in \mathcal{N}(\mathbf{x}^*)$.

First order necessary condition:

If $\mathbf{x}^*$ is a local minimizer and f is continuously differentiable in an open neighborhood of $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$. [proof by contradiction]

Convexity: Local $\leftrightarrow$ Global

Convergence analysis

Let $\{\mathbf{x}^r\}_{r\in\mathbb{N}}$ be a sequence generated by an "algorithm".
Optimality (stationarity) measures $M(\mathbf{x}^r)$:

- convex: $\|\mathbf{x}^r - \mathbf{x}^*\|$ (for policy convergence), $f(\mathbf{x}^r) - f^*$ (for value convergence)
- non-convex: $\|\nabla f(\mathbf{x}^r)\|$.

Asymptotic convergence:

- $\lim_{r\to\infty} M(\mathbf{x}^r) = 0$.

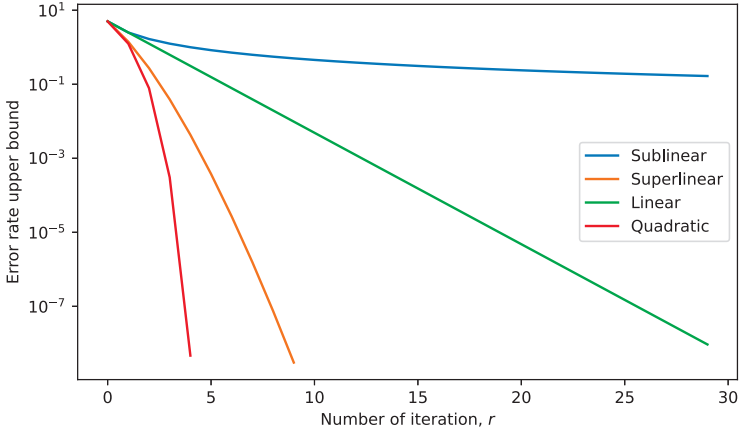
Non-asymptotic convergence, i.e., convergence rate: how fast?

- Order of convergence: q such that $\sup\left\{q \mid \lim_{r\to\infty} \frac{M(\mathbf{x}^{r+1})}{M(\mathbf{x}^r)^q} < \infty\right\}$
  - $q = 1$: linear convergence; $q = 2$: quadratic convergence
  - Most (non-finite) algorithms to solve mathematical programs are between linear and quadratic.
- Rate of convergence: the limiting ratio $\lim_{r\to\infty} \frac{M(\mathbf{x}^{r+1})}{M(\mathbf{x}^r)^q} = \eta$ given the order is q
  - Linear rate: order is 1 and rate is in $(0, 1)$, i.e., $\lim_{r\to\infty} \frac{M(\mathbf{x}^{r+1})}{M(\mathbf{x}^r)} = \eta < 1$. (also called exponential or geometric rate)

$$M(\mathbf{x}^{r+1}) \le \eta M(\mathbf{x}^r) \quad \Leftrightarrow \quad \log M(\mathbf{x}^{r+1}) \le \log M(\mathbf{x}^r) + \log \eta.$$

  - Sublinear rate: order is 1 and rate is 1, i.e., $\lim_{r\to\infty} \frac{M(\mathbf{x}^{r+1})}{M(\mathbf{x}^r)} = 1$.
  - Superlinear rate: order is 1 and rate is 0, i.e., $\lim_{r\to\infty} \frac{M(\mathbf{x}^{r+1})}{M(\mathbf{x}^r)} = 0$.

Figure for convergence rate



Outline

Convergence analysis - convex functions

Implication of convexity:

$$\nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^*) \geq f(\mathbf{x}) - f^* \geq 0$$

- $-\nabla f(\mathbf{x})$ is positively correlated to $\mathbf{x}^* - \mathbf{x}$
  - moving along $-\nabla f(\mathbf{x})$ direction gets closer to $\mathbf{x}^*$

Compute distance to $\mathbf{x}^*$:

$$\begin{aligned}
\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^r - \gamma_r \nabla f(\mathbf{x}^r) - \mathbf{x}^*\|^2 \\
&= \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma_r \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*) + \gamma_r^2 \|\nabla f(\mathbf{x}^r)\|^2 \\
&\leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma_r (f(\mathbf{x}^r) - f^*) + \gamma_r^2 \|\nabla f(\mathbf{x}^r)\|^2
\end{aligned}$$

Polyak's step size: $\gamma_r = \frac{f(\mathbf{x}^r) - f^*}{\|\nabla f(\mathbf{x}^r)\|^2}$

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - \frac{(f(\mathbf{x}^r) - f^*)^2}{\|\nabla f(\mathbf{x}^r)\|^2}.$$

Theorem

Let f be convex with bounded gradient $\|\nabla f(\mathbf{x}^r)\| \leq B$, then the sequence $(\mathbf{x}^r)_{r\in\mathbb{N}}$ generated by GD with step size $\gamma_r = \frac{f(\mathbf{x}^r)-f^*}{\|\nabla f(\mathbf{x}^r)\|^2}$ satisfies

$$\min_{r=0,\dots,T-1} f(\mathbf{x}^r) - f^* \leq \frac{B\|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{T}}.$$

Proof. Since $\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - \frac{(f(\mathbf{x}^r)-f^*)^2}{B^2}$, we have

$$(f(\mathbf{x}^r) - f^*)^2 \leq B^2 \left(\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2\right).$$

Then, using the telescoping sum

$$\begin{aligned} T \min_{r=0,\dots,T-1}(f(\mathbf{x}^r) - f^*)^2 &\leq \sum_{r=0}^{T-1}(f(\mathbf{x}^r) - f^*)^2 \\ &\leq B^2 \left(\|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^T - \mathbf{x}^*\|^2\right) \leq B^2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 \end{aligned}$$

Alternative proof

Fixed step size γ:

$$\begin{aligned}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma(f(\mathbf{x}^r) - f^*) + \gamma^2\|\nabla f(\mathbf{x}^r)\|^2 \\ &\leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma(f(\mathbf{x}^r) - f^*) + \gamma^2 B^2\end{aligned}$$

Regret interpretation: $f(\mathbf{x}^r) - f^*$ is large $\rightarrow$ $\mathbf{x}^{r+1}$ gets closer to $\mathbf{x}^*$

Rearranging terms

$$f(\mathbf{x}^r) - f^* \leq \frac{1}{2\gamma}\left(\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 + \gamma^2 B^2\right)$$

Arrive at

$$\min_{r=0,\ldots,T-1} f(\mathbf{x}^r) - f^* \leq \frac{1}{2T}\left(\frac{1}{\gamma}\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \gamma T B^2\right)$$

Optimal $\gamma^* = \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{T+1}B}$.

Outline

Smooth functions

- Definition:

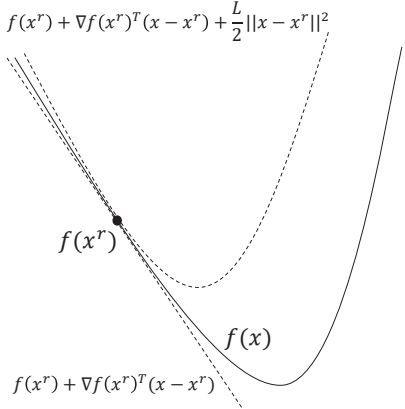
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

- Descent Lemma:

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

Smooth functions

- The descent lemma gives us a convex quadratic upper bound on f.



- This bound is minimized by a GD step from $\mathbf{x}^r$ with $\gamma = 1/L$.

Smooth functions

- GD iteration:

$$\begin{aligned}\mathbf{x}^{r+1} &= \mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r) \\ &= \arg\min_{\mathbf{x}} \underbrace{\left\{ f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{x}^r\|^2 \right\}}_{=L(\mathbf{x} \,|\, \mathbf{x}^r)} \text{ [verify it]}\end{aligned}$$

- Choose $\gamma \leq 1/L$: $L(\mathbf{x} \,|\, \mathbf{x}^r) \geq f(\mathbf{x})$

Proof of descent

- by descent lemma and $\gamma \leq 1/L$

$$\begin{aligned} f(\mathbf{x}) &\leq f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^r\|^2 \\ &\leq f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^r\|^2 \end{aligned}$$

let $\mathbf{x}^{r+1} = \mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r)$

$$\begin{aligned} f(\mathbf{x}^{r+1}) &\leq f(\mathbf{x}^r) - \gamma\|\nabla f(\mathbf{x}^r)\|^2 + \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2 \\ &= f(\mathbf{x}^r) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2. \end{aligned}$$

- If gradient is non-zero, $\gamma \leq 1/L$ is guaranteed to decrease objective.
- Amount we decrease grows with the size of the gradient.

Proof of descent

In fact, we can prove decay of $f(\mathbf{x}^r)$ for $\gamma < 2/L$:

$$f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r) - \gamma(1 - \frac{\gamma L}{2})\|\nabla f(\mathbf{x}^r)\|^2$$

[DIY]

Theorem

Let f be L-smooth, then the sequence $(\mathbf{x}^r)_{r\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 1/L$ satisfies

$$\min_{r=0,\dots,T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{\frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)}{T}.$$

Proof. Since $f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2$, we have

$$\|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{2}{\gamma}\left(f(\mathbf{x}^r) - f(\mathbf{x}^{r+1})\right).$$

Let's sum up the squared norms of all the gradients up to iteration T,

$$\sum_{r=0}^{T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{2}{\gamma}\sum_{r=0}^{T-1}\left(f(\mathbf{x}^r) - f(\mathbf{x}^{r+1})\right).$$

Using the telescoping sum and $f(\mathbf{x}^T) \geq f^*$,

$$T \min_{r=0,\ldots,T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \sum_{r=0}^{T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^T)\right) \leq \frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right),$$

and finally we get the convergence rate

$$\min_{r=0,\ldots,T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{\frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)}{T} = O(\frac{1}{T}).$$

□

- If GD runs for T iterations, we will find at least one r with $\|\nabla f(\mathbf{x}^r)\|^2 = O(\frac{1}{T})$ (or $\|\nabla f(\mathbf{x}^r)\| = O(\frac{1}{\sqrt{T}})$).
- This is a non-asymptotic result:
  - It holds on iteration 1, there is no "limit as $T \to \infty$" as in classic results.
  - But if T goes to ∞, argument can be modified to show that $\|\nabla f(\mathbf{x}^r)\|$ goes to 0.
- This convergence rate is dimension-independent.
  - It does not directly depend on dimension d.

- Instead, we're usually happy with $\|\nabla f(\mathbf{x}^r)\| \leq \epsilon$ for some small ϵ.
  - Given an ϵ, how many iterations does GD take for this to happen?

Since

$$\min_{r=0,\ldots,T-1} \|\nabla f(\mathbf{x}^r)\|^2 \leq \frac{\frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)}{T},$$

setting

$$\frac{\frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)}{T} \leq \epsilon$$

leads to

$$T \geq \frac{\frac{2}{\gamma}\left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)}{\epsilon}.$$

- GD requires $T = O(\frac{1}{\epsilon})$ iterations to achieve $\|\nabla f(\mathbf{x}^r)\|^2 \leq \epsilon$.
  - So if computing gradient costs $O(nd)$, total cost of GD is $O(nd/\epsilon)$.

Outline

Upper and lower bounds

Linear lowerbounds by convexity:

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

- Implication: $\nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*) \geq f(\mathbf{x}^r) - f^*$.

Quadratic upperbound by smoothness:

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

- Implications: $f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2$.

Convergence analysis - convex smooth functions

Distance to optimal point $\mathbf{x}^*$:

$$\begin{aligned}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r) - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^r - \mathbf{x}^*\|^2 \underbrace{-2\gamma \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*)}_{\text{convexity}} + \underbrace{\gamma^2 \|\nabla f(\mathbf{x}^r)\|^2}_{\text{smoothness}}\end{aligned}$$

Convexity:

$$-2\gamma \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*) \leq -2\gamma \left(f(\mathbf{x}^r) - f^*\right).$$

Smoothness:

$$\gamma^2 \|\nabla f(\mathbf{x}^r)\|^2 \leq 2\gamma \left(f(\mathbf{x}^r) - f(\mathbf{x}^{r+1})\right).$$

Combining

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma (f(\mathbf{x}^{r+1}) - f^*).$$

Theorem

Let f be convex and L-smooth, then the sequence $(\mathbf{x}^r)_{r\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 1/L$ satisfies

$$f(\mathbf{x}^T) - f^* \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|}{2\gamma T}.$$

Proof.

$$2\gamma T\left(f(\mathbf{x}^T) - f^*\right) \leq 2\gamma \sum_{r=0}^{T-1} f(\mathbf{x}^{r+1}) - f^* \leq \sum_{r=0}^{T-1} \left(\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2\right)$$
$$= \|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

□

Strong convexity

- Definition (zeroth order condition):

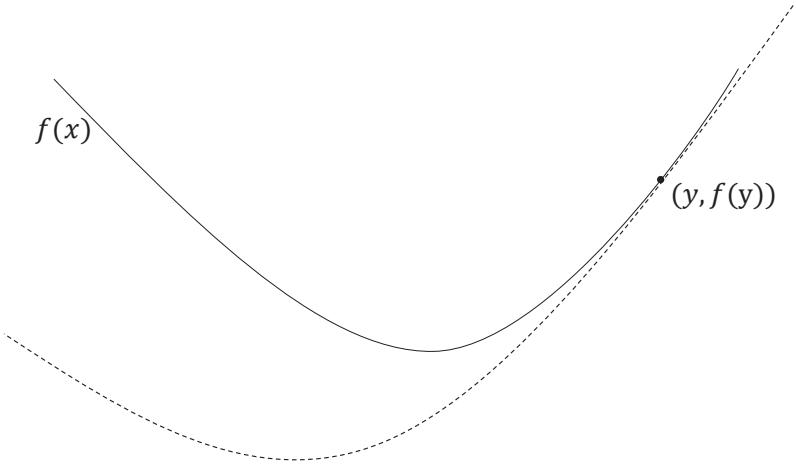
$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$$

- First order condition:

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

- Second order condition:

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}.$$


$f(x)$
$(y, f(y))$

Proof. Zeroth order condition $\Rightarrow$ First order condition
We can first introduce a function $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ and prove that $g(\mathbf{x})$ is convex if $f(\mathbf{x})$ is μ-strongly convex.
Since the function $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex if and only if its domain $\operatorname{dom}(g) = \operatorname{dom}(f)$ is convex and for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f)$ and $\lambda \in [0,1]$,

$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda) g(\mathbf{y}).$$

The latter inequality is the same as

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})+\frac{\mu}{2}\left[\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\|^2 - \lambda\|\mathbf{x}\|^2 - (1-\lambda)\|\mathbf{y}\|^2\right].$$

Now, using the identity (which holds since the norm is assumed to be Euclidean)

$$\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\|^2 - \lambda\|\mathbf{x}\|^2 - (1-\lambda)\|\mathbf{y}\|^2 = -\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2,$$

We can conclude that the convexity of g is equivalent to the convexity of $\mathrm{dom} f$ and the validity of the inequality

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$ and $\lambda \in [0,1]$, namely, to the μ-strong convexity of f.
Thus $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex. We obtain

$$g(\mathbf{x}) \geq g(\mathbf{y}) + \nabla g(\mathbf{y})^\top (\mathbf{x}-\mathbf{y}),$$

and further

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

□

Proof. First order condition $\Rightarrow$ Zeroth order condition
Suppose $\mathbf{z} = \lambda \mathbf{x} + (1-\lambda)\,\mathbf{y}$, we have

$$\begin{cases} f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2 & (a) \\ f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{z}\|^2 & (b) \end{cases}$$

Consider $\lambda \cdot (a) + (1-\lambda) \cdot (b)$, we have

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\lambda \mathbf{x} + (1-\lambda)\,\mathbf{y} - \mathbf{z}) + \frac{\mu}{2} \lambda (1-\lambda) \|\mathbf{x} - \mathbf{y}\|^2$$

$\Longleftrightarrow$

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2} \lambda (1-\lambda) \|\mathbf{x} - \mathbf{y}\|^2$$

□

Proof. First order condition $\Rightarrow$ Second order condition
With a $\tau > 0$, we have

$$f(\mathbf{x} + \tau d) = f(\mathbf{x}) + \tau \nabla^\top f(\mathbf{x})\, \mathbf{d} + \frac{\tau^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x})\, \mathbf{d} + o\left(\|\tau \mathbf{d}\|^2\right).$$

By the first order condition, we have

$$f(\mathbf{x} + \tau \mathbf{d}) \geq f(\mathbf{x}) + \tau \nabla f(\mathbf{x})^\top \mathbf{d} + \frac{\mu}{2} \|\tau \mathbf{d}\|^2.$$

Thus we obtain

$$\frac{\tau^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x})\, \mathbf{d} + o\left(\|\tau \mathbf{d}\|^2\right) \geq \frac{\mu}{2} \|\tau \mathbf{d}\|^2.$$

Dividing both sides by τ^2 and take the limit as $\tau \to 0$, we obtain

$$\begin{aligned} \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x})^\top \mathbf{d} + \lim_{\tau \to 0} \frac{o\left(\|\tau \mathbf{d}\|^2\right)}{\tau^2} \geq \frac{\mu}{2} \|\mathbf{d}\|^2 &\iff \mathbf{d}^\top \nabla^2 f(\mathbf{x})\, \mathbf{d} \geq \mu \|\mathbf{d}\|^2 \\ &\iff \nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}. \end{aligned}$$

□

Proof. Second order condition $\Rightarrow$ First order condition
For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$, from second-order Taylor expansion we know that there exists a $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$ such that

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})^\top \nabla^2 f(\mathbf{z}) (\mathbf{x} - \mathbf{y})$$

We know $\nabla^2 f(\mathbf{z}) \succeq \mu \mathbf{I}$, so we have

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

□

Upper and lower bounds

Quadratic lowerbound by convexity:

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

- Implication: Improved lower bound:

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

Quadratic upperbound by smoothness:

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

- Implication: Same descent inequality: $f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2$.

Convergence analysis - strongly convex smooth functions

Distance to optimal point $\mathbf{x}^*$:

$$\begin{aligned}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r) - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^r - \mathbf{x}^*\|^2 \underbrace{-2\gamma \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*)}_{\text{s-cvx}} + \underbrace{\gamma^2 \|\nabla f(\mathbf{x}^r)\|^2}_{\text{smoothness}}\end{aligned}$$

Strong convexity:

$$-2\gamma \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*) \leq -2\gamma \left(f(\mathbf{x}^r) - f^*\right) - \mu\gamma \|\mathbf{x}^r - \mathbf{x}^*\|^2.$$

Smoothness:

$$\gamma^2 \|\nabla f(\mathbf{x}^r)\|^2 \leq 2\gamma \left(f(\mathbf{x}^r) - f(\mathbf{x}^{r+1})\right).$$

Combining

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq (1 - \mu\gamma)\|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma(f(\mathbf{x}^{r+1}) - f^*).$$

Convergence analysis - strongly convex smooth functions

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\mathbf{x}^r)_{r\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 1/L$ satisfies

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq \frac{1-\mu\gamma}{1+\mu\gamma}\|\mathbf{x}^r - \mathbf{x}^*\|^2.$$

Proof. To complete the proof we lowerbound $f(\mathbf{x}^{r+1}) - f^*$ using strong convexity

$$f(\mathbf{x}^{r+1}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^\top(\mathbf{x}^{r+1} - \mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2.$$

Note $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Hence

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq (1-\mu\gamma)\|\mathbf{x}^r - \mathbf{x}^*\|^2 - \mu\gamma\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2$$

Improved rate using co-coercivity

Monotonicity: an operator T is monotone if

$$(T(\mathbf{x}) - T(\mathbf{x}))^\top (\mathbf{x} - \mathbf{y}) \geq 0.$$

- f is convex iff

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq 0.$$

  Proof. [DIY]
  "$\Rightarrow$" follows from definition; "$\Leftarrow$" use Taylor expansion.

- f is strongly convex iff

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \mu \|\mathbf{x} - \mathbf{y}\|^2$$

A sanity check: think of $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x}$

Improved rate using co-coercivity

- If f is μ-strongly convex and L-smooth, then

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

- Distance to optimal point $\mathbf{x}^*$:

$$\begin{aligned}
&\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \\
&= \|\mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r) - \mathbf{x}^*\|^2 \\
&= \|\mathbf{x}^r - \mathbf{x}^*\|^2 \underbrace{-2\gamma \nabla f(\mathbf{x}^r)^\top (\mathbf{x}^r - \mathbf{x}^*)}_{\text{s-cvx}} + \gamma^2 \|\nabla f(\mathbf{x}^r)\|^2 \\
&= \|\mathbf{x}^r - \mathbf{x}^*\|^2 - 2\gamma (\nabla f(\mathbf{x}^r) - \nabla f(\mathbf{x}^*))^\top (\mathbf{x}^r - \mathbf{x}^*) + \gamma^2 \|\nabla f(\mathbf{x}^r)\|^2 \\
&\leq \left(1 - 2\gamma \frac{\mu L}{\mu + L}\right) \|\mathbf{x}^r - \mathbf{x}^*\|^2 - \gamma \left(\frac{2}{\mu + L} - \gamma\right) \|\nabla f(\mathbf{x}^r)\|^2
\end{aligned}$$

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\mathbf{x}^r)_{r\in\mathbb{N}}$ generated by GD with step size $\gamma \leq 2/(\mu + L)$ satisfies

$$\|\mathbf{x}^{r+1} - \mathbf{x}^*\|^2 \leq \left(1 - 2\gamma\frac{\mu L}{L + \mu\gamma}\right)\|\mathbf{x}^r - \mathbf{x}^*\|^2.$$

- Set $\gamma = \frac{2}{\mu+L}$ gives rate $r = 1 - \frac{4\mu L}{(\mu+L)^2}$
- Condition number $\kappa = L/\mu \Rightarrow r = \left(\frac{\kappa-1}{\kappa+1}\right)^2$
- Ill conditioning (κ large) leads to worse convergence rate

Outline

Multivariate Chain Rule

- If $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, then $h(\mathbf{x}) = f(\mathbf{g}(\mathbf{x}))$ has gradient

$$\nabla h(\mathbf{x}) = \nabla \mathbf{g}(\mathbf{x})^\top \nabla f(\mathbf{g}(\mathbf{x})),$$

  where $\nabla \mathbf{g}(\mathbf{x})$ is the Jacobian (since $\mathbf{g}$ is multi-output)
- If $\mathbf{g}$ is an affine map $\mathbf{x} \to \mathbf{A}\mathbf{x} + \mathbf{b}$ so that $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ then we obtain

$$\nabla h(\mathbf{x}) = \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

- Further, for the Hessian we have

$$\nabla^2 h(\mathbf{x}) = \mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{x} + \mathbf{b})\,\mathbf{A}.$$

Gradient and Hessian of Logistic Regression

- Logistic regression gradient is

$$\nabla f(\mathbf{w}) = \mathbf{X}^\top \mathbf{c},$$

  where $\mathbf{c}$ is a vector with $c_t = -y^t h\left(-y^t \mathbf{w}^\top \mathbf{x}^t\right)$ and h is the sigmoid function.
  - GD costs $O(nd)$ per iteration to compute $\mathbf{X}\mathbf{w}^r$ and $\mathbf{X}^T\mathbf{c}$.
- Logistic regression Hessian is

$$\nabla^2 f(\mathbf{w}) = \mathbf{X}^\top \mathbf{D}\mathbf{X},$$

  where $\mathbf{D}$ is a diagonal matrix with $d_{tt} = h\left(y^t \mathbf{w}^\top \mathbf{x}^t\right) h\left(-y^t \mathbf{w}^\top \mathbf{x}^t\right)$.

Convexity of Logistic Regression

- Since the sigmoid function is non-negative, we can compute $\mathbf{D}^{\frac{1}{2}}$, and

$$\mathbf{v}^\top \mathbf{X}^\top \mathbf{D}\mathbf{X}\mathbf{v} = \mathbf{v}^\top \mathbf{X}^\top \mathbf{D}^{\frac{1}{2}} \mathbf{D}^{\frac{1}{2}} \mathbf{X}\mathbf{v} = \left(\mathbf{D}^{\frac{1}{2}} \mathbf{X}\mathbf{v}\right)^\top \left(\mathbf{D}^{\frac{1}{2}} \mathbf{X}\mathbf{v}\right) = \left\| \mathbf{X}\mathbf{D}^{\frac{1}{2}} \mathbf{v} \right\|^2 \geq 0,$$

  so $\mathbf{X}^\top \mathbf{D}\mathbf{X}$ is positive semidefinite and logistic regression is convex.
  - It becomes strictly convex if you add ℓ_2-regularization, making solution unique.

Smoothness of Logistic Regression

- Logistic regression Hessian is

$$\begin{aligned}\nabla^2 f(\mathbf{w}) &= \sum_{t=1}^{n} \underbrace{h\left(y^t \mathbf{w}^\top \mathbf{x}^t\right) h\left(-y^t \mathbf{w}^\top \mathbf{x}^t\right)}_{d_{tt}} \mathbf{x}^t \left(\mathbf{x}^t\right)^\top \\ &\preceq \frac{1}{4} \sum_{t=1}^{n} \mathbf{x}^t \left(\mathbf{x}^t\right)^\top \\ &= \frac{1}{4} \mathbf{X}^\top \mathbf{X}.\end{aligned}$$

- In the second line we use that $h(\alpha) \in (0,1)$ and $h(-\alpha) = 1 - h(\alpha)$.
  - This means that $d_{tt} \leq \frac{1}{4}$.
- So for logistic regression, we can take $L = \frac{1}{4} \max\left\{\text{eig}\left(\mathbf{X}^\top \mathbf{X}\right)\right\}$.