Lecture 5. Gradient Descent

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah

Math 5750/6880, Fall 2023

Loss function

So far, we have formulated training machine learning models as

$$\min f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_i(\boldsymbol{x}) + R(\boldsymbol{x})$$

where $\boldsymbol{x}$ is the parameter of the machine learning model, $\mathcal{L}_i(\boldsymbol{x})$ is the loss of the ith training instance, and $R(\boldsymbol{x})$ is the regularization term.

How to find the optimal $\boldsymbol{x}^*$?

Gradient methods for unconstrained problems

Consider the following optimization problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \ \text{ subject to } \ \boldsymbol{x} \in \mathbb{R}^n,$$

where f (objective or cost function) is a differentiable. Starting with a point $\boldsymbol{x}^0$, gradient descent iterates as follows:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t). \tag{1}$$

GD is also known as steepest descent.

How fast gradient descent is?

Quadratic minimization problems

To get a sense of the convergence rate of GD, let's begin with quadratic objective functions

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*),$$

for some $n \times n$ matrix $Q \succ 0$ (positive definite), where $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*)$.

Quadratic minimization problems: constant stepsizes

Convergence rate: if $\eta_t \equiv \eta = \frac{2}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$$

where $\lambda_1(\boldsymbol{Q})$ (resp. $\lambda_n(\boldsymbol{Q})$) is the largest (resp. smallest) eigenvalue of $\boldsymbol{Q}$. Here η is chosen s.t. $|1 - \eta\lambda_n(\boldsymbol{Q})| = |1 - \eta\lambda_1(\boldsymbol{Q})|$.

Remark. The convergence rate is dictated by the condition number $\frac{\lambda_1(\boldsymbol{Q})}{\lambda_n(\boldsymbol{Q})}$ of $\boldsymbol{Q}$, or equivalently, $\frac{\max_{\boldsymbol{x}} \lambda_1(\nabla^2 f(\boldsymbol{x}))}{\min_{\boldsymbol{x}} \lambda_n(\nabla^2 f(\boldsymbol{x}))}$. This convergence rate is often called linear convergence or geometric convergence.

Quadratic minimization problems: constant stepsizes

Proof. According to the GD update rule,

$$\boldsymbol{x}^{t+1} - \boldsymbol{x}^* = \boldsymbol{x}^t - \boldsymbol{x}^* - \eta_t \nabla f(\boldsymbol{x}^t) = (\boldsymbol{I} - \eta_t \boldsymbol{Q})(\boldsymbol{x}^t - \boldsymbol{x}^*) \Rightarrow \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{I} - \eta_t \boldsymbol{Q}\|_2 \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2.$$

The claim then follows by observing that

$$\|\boldsymbol{I} - \eta_t \boldsymbol{Q}\|_2 = \underbrace{\max\{|1 - \eta_t \lambda_1(\boldsymbol{Q})|, |1 - \eta_t \lambda_n(\boldsymbol{Q})|\}}_{\text{optimal choice is } \eta_t = \frac{2}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}} = 1 - \frac{2\lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})} = \frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}.$$

Apply the above bound recursively to complete the proof.

We need to choose the step size to minimize $\|\boldsymbol{I} - \eta_t \boldsymbol{Q}\|_2$, resulting in the optimal choice of the step size η_t.

Quadratic minimization problems: Exact line search

Note that the stepsize rule $\eta_t \equiv \eta = \frac{2}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}$ relies on the spectrum of $\boldsymbol{Q}$, which requires preliminary experimentation. Another more practical strategy is the exact line search rule

$$\eta_t = \arg\min_{\eta \geq 0} f(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)). \tag{2}$$

Convergence rate: if $\eta_t = \arg\min_{\eta\geq 0} f(\boldsymbol{x}^t - \eta\nabla f(\boldsymbol{x}^t))$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \left(\frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}\right)^{2t} \Big(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\Big).$$

Note that the rate is stated in terms of the objective values, and the convergence rate is not faster than the constant stepsize rule.

Quadratic minimization problems: Exact line search – Proof

Proof. For notational simplicity, let $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}^t) = \boldsymbol{Q}(\boldsymbol{x}^t - \boldsymbol{x}^*)$. It can be verified that exact line search gives

$$\eta_t = \frac{\boldsymbol{g}^{t\top}\boldsymbol{g}^t}{\boldsymbol{g}^{t\top}\boldsymbol{Q}\boldsymbol{g}^t}.$$

This gives

$$\begin{aligned}
f(\boldsymbol{x}^{t+1}) &= \frac{1}{2}(\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t - \boldsymbol{x}^*) \\
&= \frac{1}{2}(\boldsymbol{x}^t - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^t - \boldsymbol{x}^*) - \eta_t \|\boldsymbol{g}^t\|_2^2 + \frac{\eta_t^2}{2}\boldsymbol{g}^{t\top}\boldsymbol{Q}\boldsymbol{g}^t \\
&= \frac{1}{2}(\boldsymbol{x}^t - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^t - \boldsymbol{x}^*) - \frac{\|\boldsymbol{g}^t\|_2^4}{2\boldsymbol{g}^{t\top}\boldsymbol{Q}\boldsymbol{g}^t} \\
&= \left(1 - \frac{\|\boldsymbol{g}^t\|_2^4}{(\boldsymbol{g}^{t\top}\boldsymbol{Q}\boldsymbol{g}^t)(\boldsymbol{g}^{t\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^t)}\right) f(\boldsymbol{x}^t)
\end{aligned}$$

where the last line uses $f(\boldsymbol{x}^t) = \frac{1}{2}(\boldsymbol{x}^t - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}^t - \boldsymbol{x}^*) = \frac{1}{2}\boldsymbol{g}^{t\top}\boldsymbol{Q}^{-1}\boldsymbol{g}^t$.

Quadratic minimization problems: Exact line search – Proof Cont'd

From Kantorovich's inequality

$$\frac{\|\boldsymbol{y}\|_2^4}{(\boldsymbol{y}^\top \boldsymbol{Q}\boldsymbol{y})(\boldsymbol{y}^\top \boldsymbol{Q}^{-1}\boldsymbol{y})} \geq \frac{4\lambda_1(\boldsymbol{Q})\lambda_n(\boldsymbol{Q})}{(\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q}))^2},$$

we arrive at

$$f(\boldsymbol{x}^{t+1}) \leq \Big(1 - \frac{4\lambda_1(\boldsymbol{Q})\lambda_n(\boldsymbol{Q})}{(\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q}))^2}\Big) f(\boldsymbol{x}^t) = \Big(\frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}\Big)^2 f(\boldsymbol{x}^t)$$

This concludes the proof since $f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x}) = 0$.

Strongly convex and smooth problems

Let's now generalize quadratic minimization to a broader class of problems

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}),$$

where $f(\cdot)$ is strongly convex and smooth. A twice-differentiable function f is said to be μ-strongly convex and L-smooth if

$$0 \preceq \mu \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}, \quad \forall \boldsymbol{x}.$$

Intuitively, the μ-strongly convex function is bounded below by a quadratic function; the L-smooth function is bounded above by another quadratic function.

Strongly convex and smooth problems: convergence rate

Theorem 1. [*GD for strongly convex and smooth functions*] *Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{2}{\mu+L}$, then*

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2,$$

where $\kappa := L/\mu$ is condition number; $\boldsymbol{x}^*$ is the minimizer. By smoothness, we further have

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t} \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2.$$

Remark. *Generalization of quadratic minimization problems: stepsize ($\eta = \frac{2}{\mu+L}$ vs. $\eta = \frac{2}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}$); contraction rate ($\frac{\kappa-1}{\kappa+1}$ vs. $\frac{\lambda_1(\boldsymbol{Q})-\lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}$).*

Remark. *Note that the convergence rate is dimension-free if κ does not depend on n.*

Strongly convex and smooth problems: convergence rate

Proof. By the fundamental theorem of calculus that

$$\nabla f(\boldsymbol{x}^t) = \nabla f(\boldsymbol{x}^t) - \underbrace{\nabla f(\boldsymbol{x}^*)}_{=0} = \Big(\int_0^1 \nabla^2 f(\boldsymbol{x}_\tau) d\tau \Big)(\boldsymbol{x}^t - \boldsymbol{x}^*),$$

where $\boldsymbol{x}_\tau := \boldsymbol{x}^t + \tau(\boldsymbol{x}^* - \boldsymbol{x}^t)$. Here, $\{\boldsymbol{x}_\tau\}_{0 \leq \tau \leq 1}$ forms a line segment between $\boldsymbol{x}^t$ and $\boldsymbol{x}^*$. Therefore,

$$\begin{aligned} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 &= \|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta \nabla f(\boldsymbol{x}^t)\|_2 = \|\Big(\boldsymbol{I} - \eta \int_0^1 \nabla^2 f(\boldsymbol{x}_\tau) d\tau\Big)(\boldsymbol{x}^t - \boldsymbol{x}^*)\|_2 \\ &\leq \sup_{0 \leq \tau \leq 1} \|\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{x}_\tau)\| \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \frac{L - \mu}{L + \mu} \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2. \end{aligned}$$

Repeat this argument for all iterations to conclude the proof.

Why

$$\sup_{0\le\tau\le 1} \|\boldsymbol{I} - \eta\nabla^2 f(\boldsymbol{x}_\tau)\| \le \frac{L-\mu}{L+\mu}?$$

Choose η to minimize $\|\boldsymbol{I} - \eta\nabla^2 f(\boldsymbol{x})\|$.

Equivalent characterizations of strongly convex functions

$f(\cdot)$ is said to be μ-strongly convex if any of the following holds:

(i) $f(\boldsymbol{y}) \geq \underbrace{f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{\text{first-order Taylor expansion}} + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{y}.$

(ii) For all $\boldsymbol{x}, \boldsymbol{y}$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

(iii) $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{y}.$

(iv) $\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}, \ \forall \boldsymbol{x}$ (for twice differentiable functions)

Equivalent characterizations of smooth functions

$f(\cdot)$ is said to be L-smooth if any of the following holds:

(i) $f(\boldsymbol{y}) \leq \underbrace{f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{\text{first-order Taylor expansion}} + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{y}$.

(ii) For all $\boldsymbol{x}$ and $\boldsymbol{y}$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{L}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

(iii) $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{y}$.

(iv) $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \ \forall \boldsymbol{x}, \boldsymbol{y}$ (L-Lipschitz gradient).

(vi) $\|\nabla^2 f(\boldsymbol{x})\|_2 \leq L, \ \forall \boldsymbol{x}$ (for twice differentiable functions)

Is strong convexity necessary for linear convergence?

So far we have established linear convergence under strong convexity and smoothness, while the strong convexity requirement can be relaxed.

Is strong convexity necessary for linear convergence? Local strong convexity

Theorem 2. [GD for locally strongly convex and smooth functions] Let f be locally μ-strongly convex and L-smooth such that

$$\mu \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathcal{B}_0,$$

where $\mathcal{B}_0 := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2\}$ and $\boldsymbol{x}^*$ is the minimizer. Then the linear convergence still holds.

Proof. Suppose $\boldsymbol{x}^t \in \mathcal{B}_0$. Then repeating our previous analysis yields $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \leq \frac{\kappa-1}{\kappa+1}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$. This also means that $\boldsymbol{x}^{t+1} \in \mathcal{B}_0$, so the above bound continues to hold for the next iteration ...

Only deteriorates the global convergence constant.

Example. Consider the logistic regression problem, suppose we obtain m independent binary samples

$$y_i = \begin{cases} 1 & \text{with prob. } \frac{1}{1+\exp(-\boldsymbol{a}_i^\top \boldsymbol{x}^\dagger)} \\ -1 & \text{with prob. } \frac{1}{1+\exp(\boldsymbol{a}_i^\top \boldsymbol{x}^\dagger)} \end{cases}$$

where $\{\boldsymbol{a}_i\}$ are the design vectors and $\boldsymbol{x}^\dagger \in \mathbb{R}^n$ are the unknown parameters. The maximum likelihood estimate (MLE) is given by

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \log\Big(1 + \exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x})\Big)$$

Local strong convexity

Note that $\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \underbrace{\frac{\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x})}{(1+\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x}))^2}}_{\to 0 \ \ as \ \ \boldsymbol{x}\to\infty} \boldsymbol{a}_i \boldsymbol{a}_i^\top \to 0$, indicating that f is 0-strongly convex. However, the local strong convexity parameter is given by

$$\inf_{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{x}^*\|_2 \le \|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2} \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \frac{\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x})}{(1+\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x}))^2} \boldsymbol{a}_i \boldsymbol{a}_i^\top\right) \tag{3}$$

which is often strictly bounded away from 0, thus enabling linear convergence.

Implement it and see if you can observe linear convergence?

Regularity condition

We can also replace strong convexity and smoothness by the regularity condition:

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{1}{2L} \|\nabla f(\boldsymbol{x})\|_2^2, \ \forall \boldsymbol{x}, \tag{4}$$

which is equivalent to

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{1}{2L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*)\|_2^2, \ \forall \boldsymbol{x}.$$

Compared to strong convexity (which involves any pair $(\boldsymbol{x}, \boldsymbol{y})$), we only restrict ourselves to $(\boldsymbol{x}, \boldsymbol{x}^*)$.

Can you think the geometry of this condition?

Regularity condition

Theorem 3. Suppose f satisfies (4). If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2.$$

Proof. It follows that

$$\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 &= \|\boldsymbol{x}^t - \boldsymbol{x}^* - \frac{1}{L}\nabla f(\boldsymbol{x}^t)\|_2^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^t)\|_2^2 - \frac{2}{L}\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^t)\rangle \\
&\underbrace{\leq}_{(4)} \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\mu}{L}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 \\
&= \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2.
\end{aligned}$$

Apply it recursively to complete the proof.

Polyak-Lojasiewicz condition

Another alternative is the PL condition

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu\Big(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\Big), \ \forall \boldsymbol{x}. \tag{5}$$

It guarantees that gradient grows fast as we move away from the optimal value, and also guarantees that every stationary point is a global minimum (may not be unique).

Theorem 4. Suppose f satisfies (5) and is L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^t \Big(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\Big).$$

We will prove this theorem later.

Example. (over-parameterized linear regression) Suppose we have data $\{\boldsymbol{a}_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}_{1\le i\le m}$. We want to find a linear model that best fits the data

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) := \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2 = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]^\top \in \mathbb{R}^{m\times n}.$$

Over-parameterization: model dim > sample size i.e. $n > m$, *(an important regime in deep learning).*
The problem is convex but not strongly convex, since

$$\nabla^2 f(\boldsymbol{x}) = \sum_{i=1}^{m} \boldsymbol{a}_i \boldsymbol{a}_i^\top \text{ is rank-deficient if } n > m. \ \ rank \nabla^2 f(\boldsymbol{x}) \le m < n, \ \text{0 is an eigenevalue}$$

But for most "non-degenerate" cases ($\nabla^2 f(\boldsymbol{x})$ has rank m), one has $f(\boldsymbol{x}^*) = 0$ and the PL condition is met, and hence GD converges linearly.

Polyak-Lojasiewicz condition

Corollary. Suppose that $A = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]^\top \in \mathbb{R}^{m \times n}$ has rank m, and that $\eta_t \equiv \eta_t = \frac{1}{\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^\top)}$. Then GD obeys

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \le \left(1 - \frac{\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)}{\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^\top)}\right)^t \Big(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\Big), \ \forall t.$$

Remark. Note that while there are many global minima for this over-parameterized problem, GD has implicit bias. GD converges to a global min closest to initialization $\boldsymbol{x}^0$!

An active research area! How over-parameterization helps training and generalization?

Polyak-Lojasiewicz condition

Proof. Everything boils down to showing the PL condition

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top) f(\boldsymbol{x}). \tag{6}$$

If this holds, then the claim follows immediately from the Theorem above and the fact $f(\boldsymbol{x}^*) = 0$. To prove (6), let $\boldsymbol{y} = [y_i]_{1 \leq i \leq m}$, and observe $\nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$. Then

$$\|\nabla f(\boldsymbol{x})\|_2^2 = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{A}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \geq \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 = 2\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top) f(\boldsymbol{x})$$

which satisfies the PL condition (6) with $\mu = \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)$.

Convex and smooth problems

Consider $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ is convex and smooth.

Without strong convexity, it may often be better to focus on the objective improvement (rather than improvement on estimation error).

Example. Consider $f(x) = 1/x(x > 0)$. GD iterates $\{x^t\}$ might never converge to $x^* = \infty$. In comparison, $f(x^t)$ might approach $f(x^*) = 0$ rapidly.

Convex and smooth problems

Consider the objective improvement, from the smoothness assumption,

$$\begin{aligned} f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^t) &\leq \nabla f(\boldsymbol{x}^t)^\top (\boldsymbol{x}^{t+1} - \boldsymbol{x}^t) + \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 \\ &= \underbrace{-\eta_t \|\nabla f(\boldsymbol{x}^t)\|_2^2 + \frac{\eta_t^2 L}{2}\|\nabla f(\boldsymbol{x}^t)\|_2^2}_{\text{majorizing function of objective reduction due to smoothness}} \end{aligned} \tag{7}$$

Let $\eta_t = 1/L$, the majorizing function is minimized, which gives

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

Lemma 1. [Objective improvement] Suppose f is L-smooth. Then GD with $\eta_t = 1/L$ obeys

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

Note that the above result does not rely on convexity.

Theorem 4 (Recap). Suppose f satisfies PL condition and is L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^t \Big(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\Big).$$

Proof of Theorem 4. [Linear convergence of GD under the PL condition]

$$\begin{aligned} f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*) &\underbrace{\leq}_{(i)} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^t)\|_2^2 \\ &\underbrace{\leq}_{(ii)} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) - \frac{\mu}{L}(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)) \\ &= \left(1 - \frac{\mu}{L}\right)(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)) \end{aligned}$$

where (i) follows from Lemma of the objective improvement, and (ii) comes from the PL condition.

Convex and smooth problems

GD is not only improving the objective value, but is also dragging the iterates towards minimizer(s), as long as η_t is not too large.

Lemma 2. Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^t)\|_2^2,$$

where $\boldsymbol{x}^*$ is any minimizer of $f(\cdot)$.

Convex and smooth problems

Proof. It follows that

$$
\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 &= \|\boldsymbol{x}^t - \boldsymbol{x}^* - \eta(\nabla f(\boldsymbol{x}^t) - \underbrace{\nabla f(\boldsymbol{x}^*)}_{=0})\|_2^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \underbrace{2\eta\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\rangle}_{\geq \frac{2\eta}{L}\|\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\|_2^2 \ (smooth+cvx)} + \eta^2\|\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\|_2^2 \\
&\leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{2\eta}{L}\|\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\|_2^2 + \eta^2\|\nabla f(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^*)\|_2^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^t) - \underbrace{\nabla f(\boldsymbol{x}^*)}_{=0}\|_2^2 \ (since \ \eta = 1/L).
\end{aligned}
$$

Convergence of GD for convex and smooth problems

However, without strong convexity, convergence is typically much slower than linear (or geometric) convergence.

Theorem 5. [GD for convex and smooth problems] Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then GD obeys

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}$$

where $\boldsymbol{x}^*$ is any minimizer of $f(\cdot)$. That is, GD attains ϵ-accuracy within $O(1/\epsilon)$ iterations.

Can be accelerated using Nesterov's accelerated gradient!

Proof of the convergence of GD for convex and smooth problems

From Lemma of objective improvement,

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

To infer $f(\boldsymbol{x}^t)$ recursively, it is often easier to replace $\|\nabla f(\boldsymbol{x}^t)\|_2$ with simpler functions of $f(\boldsymbol{x}^t)$. Use convexity and Cauchy-Schwarz to get

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) \geq \nabla f(\boldsymbol{x}^t)^\top(\boldsymbol{x}^* - \boldsymbol{x}^t) \geq -\|\nabla f(\boldsymbol{x}^t)\|_2\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$$

therefore

$$\|\nabla f(\boldsymbol{x}^t)\|_2 \geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2} \underbrace{\geq}_{Lemma\ 2} \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}.$$

Proof of the convergence of GD for convex and smooth problems

Setting $\Delta_t := f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)$ and combining the above bounds yield

$$\Delta_{t+1} - \Delta_t \leq -\frac{1}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}\Delta_t^2 := -\frac{1}{w_0}\Delta_t^2, \tag{8}$$

i.e. $\Delta_{t+1} \leq \Delta_t - \frac{1}{w_0}\Delta_t^2$. Dividing both sides by $\Delta_t \Delta_{t+1}$ and rearranging terms give

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{w_0}\frac{\Delta_t}{\Delta_{t+1}}.$$

Proof of the convergence of GD for convex and smooth problems

Since $\Delta_t \geq \Delta_{t+1}$, thus $\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{w_0}$, and therefore

$$\frac{1}{\Delta_t} \geq \frac{1}{\Delta_0} + \frac{t}{w_0} \geq \frac{t}{w_0} \Rightarrow \Delta_t \leq \frac{w_0}{t} = \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}.$$

Nonconvex problems

We cannot hope for efficient global convergence to global minima in general, but we may have:

- convergence to stationary points (i.e. $\nabla f(\boldsymbol{x}) = 0$)
- convergence to local minima
- local convergence to global minima (i.e., when initialized suitably)

Convergence of GD for nonconvex problems

Theorem 6. Let f be L-smooth and $\eta_k \equiv \eta = 1/L$. Assume t is even,

In general, GD obeys

$$\min_{0 \le k \le t} \|\nabla f(\boldsymbol{x}^k)\|_2 \le \sqrt{\frac{2L(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*))}{t}}$$

If $f(\cdot)$ is convex, then GD obeys

$$\min_{t/2 \le k < t} \|\nabla f(\boldsymbol{x}^k)\|_2 \le \frac{4L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}{t}$$

Remark. GD finds an ϵ-approximate stationary point in $O(1/\epsilon^2)$ iterations.

Remark. Note that it does not imply GD converges to stationary points; it only says that $\exists$ approximate stationary point in the GD trajectory.

Proof of Theorem 6

From Lemma 1 (Depend on L-smooth only), we know

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|_2^2 \le f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1}), \quad \forall k.$$

This leads to a telescopic sum when summed over $k = t_0$ to $k = t - 1$:

$$\frac{1}{2L}\sum_{k=t_0}^{t-1}\|\nabla f(\boldsymbol{x}^k)\|_2^2 \le \sum_{k=t_0}^{t-1}\left(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1})\right) = f(\boldsymbol{x}^{t_0}) - f(\boldsymbol{x}^t) \le f(\boldsymbol{x}^{t_0}) - f(\boldsymbol{x}^*)$$

therefore,

$$\min_{t_0 \le k < t}\|\nabla f(\boldsymbol{x}^k)\|_2 \le \sqrt{\frac{2L(f(\boldsymbol{x}^{t_0}) - f(\boldsymbol{x}^*))}{t - t_0}}. \tag{9}$$

Proof of Theorem 6

For a general $f(\cdot)$, taking $t_0 = 0$ immediately establishes the claim.
If $f(\cdot)$ is convex, invoke Theorem 5 to obtain

$$f(\boldsymbol{x}^{t_0}) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t_0}$$

Taking $t_0 = t/2$ and combining it with (9) give

$$\min_{t_0 \leq k < t} \|\nabla f(\boldsymbol{x}^k)\|_2 \leq \frac{2L}{\sqrt{t_0(t - t_0)}} \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2 = \frac{4L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}{t}.$$

Gradient Descent For Solving Near-singular Linear Equations

- Consider solving the linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{b} \in \mathbb{R}^n$.

- Solving the above linear equations is equivalent to solving the following unconstrained optimization problem

$$\min_{\boldsymbol{u} \in \mathbb{R}^m} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2.$$

- Over-parameterization: $m \gg n$.

Near singular system

- Consider a nearly singular system: $\boldsymbol{A}_\epsilon \boldsymbol{u} = \boldsymbol{g}\ (\boldsymbol{A}_\epsilon = \boldsymbol{A}_0 + \epsilon \boldsymbol{I})$

$$\boldsymbol{A}_0 = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \quad \boldsymbol{g} = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \in R\left(\boldsymbol{A}_0\right).$$

- $\sigma(\boldsymbol{A}_0) = \{3, 1, 0\}$.

- Write $\boldsymbol{u} \in \mathbb{R}^3 = u_1 \boldsymbol{e}_1 + u_2 \boldsymbol{e}_2 + u_3 \boldsymbol{e}_3$ as

$$\boldsymbol{u} = \tilde{u_1} \boldsymbol{e}_1 + \tilde{u_2} \boldsymbol{e}_2 + \tilde{u_3} \boldsymbol{e}_3 + \tilde{u_4} \boldsymbol{p} = \boldsymbol{P} \tilde{\boldsymbol{u}}$$

where

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{p} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \in \ker\left(\boldsymbol{A}_0\right).$$

- Notice that solving $\boldsymbol{A}_\epsilon \boldsymbol{u} = \boldsymbol{g}$ is equivalent to solving

$$\boldsymbol{A}_\epsilon \boldsymbol{P} \tilde{\boldsymbol{u}} = \boldsymbol{g} \Longleftrightarrow \left(\boldsymbol{P}^\top \boldsymbol{A}_{\boldsymbol{\epsilon}} \boldsymbol{P} \right) \tilde{\boldsymbol{u}} = \boldsymbol{P}^\top \boldsymbol{g},$$

resulting in the following semi-definite system (a singular system):

$$\begin{pmatrix} 1+\epsilon & -1 & 0 & \epsilon \\ -1 & 2+\epsilon & -1 & \epsilon \\ 0 & -1 & 1+\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 3\epsilon \end{pmatrix} \tilde{\boldsymbol{u}} = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix}$$

- GD can solve singular problems easily but the near-singular problem

- GD with stepsize 0.7 till $\|\boldsymbol{A}\boldsymbol{u}^k - \boldsymbol{b}\| \leq 10^{-8}$

ϵ	Original	Expanded
1.	37	13
10^{-1}	236	14
10^{-2}	1,918	14
10^{-3}	16,115	16
10^{-4}	130,168	16
10^{-5}	>1M	16
10^{-9}	>1M	15
10^{-10}	21	15
0	20	