Gradient methods for unconstrained problems

Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Quadratic minimization problems
- Strongly convex and smooth problems
- Convex and smooth problems
- Nonconvex problems

Differentiable unconstrained minimization

$$
\begin{aligned}
\text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\
\text{subject to} \quad & \boldsymbol{x} \in \mathbb{R}^n
\end{aligned}
$$

- f (objective or cost function) is differentiable

Iterative descent algorithms

Start with a point $\boldsymbol{x}^0$, and construct a sequence $\{\boldsymbol{x}^t\}$ s.t.

$$f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

- $\boldsymbol{d}$ is said to be a descent direction at $\boldsymbol{x}$ if

$$f'(\boldsymbol{x}; \boldsymbol{d}) := \underbrace{\lim_{\tau \downarrow 0} \frac{f(\boldsymbol{x} + \tau \boldsymbol{d}) - f(\boldsymbol{x})}{\tau}}_{\text{directional derivative}} = \nabla f(\boldsymbol{x})^\top \boldsymbol{d} < 0 \tag{2.1}$$

Iterative descent algorithms

Start with a point $\boldsymbol{x}^0$, and construct a sequence $\{\boldsymbol{x}^t\}$ s.t.

$$f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

- In each iteration, search in descent direction

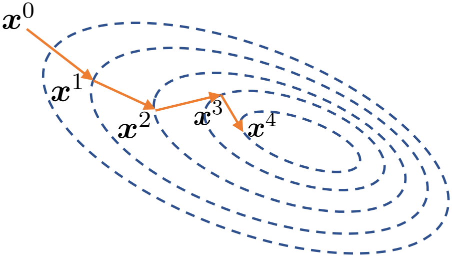
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \eta_t \boldsymbol{d}^t \tag{2.2}$$

where $\boldsymbol{d}^t$: descent direction at $\boldsymbol{x}^t$; $\eta_t > 0$: stepsize

Gradient descent (GD)

One of the most important examples of (2.2): **gradient descent**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) \tag{2.3}$$



- traced to Augustin Louis Cauchy '1847 ...

Gradient descent (GD)

One of the most important examples of (2.2): **gradient descent**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) \tag{2.3}$$

- descent direction: $\boldsymbol{d}^t = -\nabla f(\boldsymbol{x}^t)$
- a.k.a. steepest descent, since from (2.1) and Cauchy-Schwarz,

$$\underbrace{\arg\min_{\boldsymbol{d}:\|\boldsymbol{d}\|_2 \le 1} f'(\boldsymbol{x};\boldsymbol{d}) = \arg\min_{\boldsymbol{d}:\|\boldsymbol{d}\|_2 \le 1} \nabla f(\boldsymbol{x})^\top \boldsymbol{d} = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|_2}}_{\text{direction with the greatest rate of objective value improvement}}$$

Quadratic minimization problems

Quadratic minimization

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

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) := \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*)$$

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

Convergence for 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 $\boldsymbol{Q}$

- as we will see, η is chosen s.t. $\big|1 - \eta\lambda_n(\boldsymbol{Q})\big| = \big|1 - \eta\lambda_1(\boldsymbol{Q})\big|$
- the convergence rate is dictated by the condition number $\frac{\lambda_1(\boldsymbol{Q})}{\lambda_n(\boldsymbol{Q})}$ of $\boldsymbol{Q}$, or equivalently, $\frac{\max_{\boldsymbol{x}} \lambda_1(\nabla^2 f(\boldsymbol{x}))}{\min_{\boldsymbol{x}} \lambda_n(\nabla^2 f(\boldsymbol{x}))}$

Convergence for 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 $\boldsymbol{Q}$

- often called linear convergence or geometric convergence
  - since the error lies below a line on a log-linear plot of error vs. iteration count

Convergence for constant stepsizes

Proof: According to the GD update rule,

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

$$\Longrightarrow \qquad \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{I} - \eta_t \boldsymbol{Q}\| \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$$

The claim then follows by observing that

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

$$= 1 - \frac{2\lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})} = \frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}$$

Apply the above bound recursively to complete the proof

$\square$

Exact line search

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

$$\eta_t = \arg\min_{\eta \geq 0} f\big(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)\big) \tag{2.4}$$

Convergence for exact line search

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

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

- stated in terms of the objective values
- convergence rate not faster than the constant stepsize rule

Convergence for exact line search

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

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

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

Convergence for exact line search

Proof (cont.): From Kantorovich's inequality

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

we arrive at

$$\begin{aligned} f(\boldsymbol{x}^{t+1}) &\leq \left(1 - \frac{4\lambda_1(\boldsymbol{Q})\lambda_n(\boldsymbol{Q})}{\left(\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})\right)^2}\right) f(\boldsymbol{x}^t) \\ &= \left(\frac{\lambda_1(\boldsymbol{Q})-\lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}\right)^2 f(\boldsymbol{x}^t) \end{aligned}$$

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

Strongly convex and smooth problems

Strongly convex and smooth problems

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

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x})$$

where $f(\cdot)$ is strongly convex and smooth

- a twice-differentiable function f is said to be μ-strongly convex and L-smooth if

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

Convergence rate for strongly convex and smooth problems

Theorem 2.1 (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

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

where $\kappa := L/\mu$ is condition number; $\boldsymbol{x}^*$ is the minimizer

- 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. $\frac{\lambda_1(\boldsymbol{Q})-\lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q})+\lambda_n(\boldsymbol{Q})}$)

Convergence rate for strongly convex and smooth problems

Theorem 2.1 (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

- dimension-free: iteration complexity is $O\left(\frac{\log \frac{1}{\varepsilon}}{\log \frac{\kappa+1}{\kappa-1}}\right)$, which is independent of the problem size n *if κ does not depend on n*

Convergence rate for strongly convex and smooth problems

Theorem 2.1 (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

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

where $\kappa := L/\mu$ is condition number; $\boldsymbol{x}^*$ is the minimizer

- a direct consequence of Theorem 2.1 (using smoothness):

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

Proof of Theorem 2.1

It is seen from the fundamental theorem of calculus that

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

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

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

Repeat this argument for all iterations to conclude the proof

More on strong convexity

$f(\cdot)$ is said to be μ-strongly convex if

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

Equivalent first-order characterizations

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

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

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

More on strong convexity

$f(\cdot)$ is said to be μ-strongly convex if

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

Equivalent second-order characterization

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

More on smoothness

A convex function $f(\cdot)$ is said to be L-smooth if

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

Equivalent first-order characterizations (for *convex* functions)

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

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \ge \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \tfrac{L}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

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

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

More on smoothness

A convex function $f(\cdot)$ is said to be L-smooth if

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

Equivalent second-order characterization

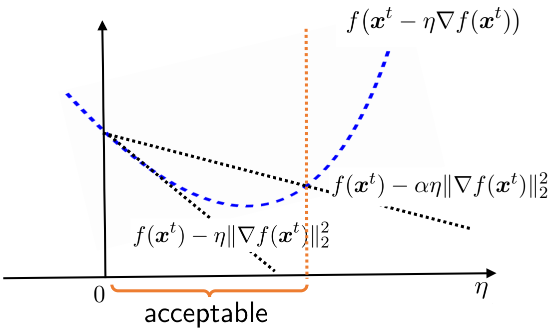
(v) $\|\nabla^2 f(\boldsymbol{x})\| \leq L, \quad \forall \boldsymbol{x}$ (for twice differentiable functions)

Backtracking line search

Practically, one often performs line searches rather than adopting constant stepsizes. Most line searches in practice are, however, *inexact*

A simple and effective scheme: *backtracking* line search

Backtracking line search

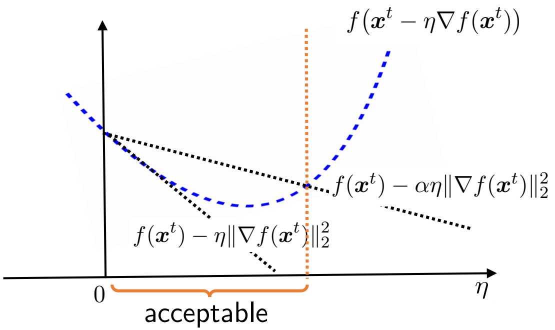


Armijo condition: for some $0 < \alpha < 1$

$$f(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)) < f(\boldsymbol{x}^t) - \alpha\eta\|\nabla f(\boldsymbol{x}^t)\|_2^2 \qquad (2.5)$$

- $f(\boldsymbol{x}^t) - \alpha\eta\|\nabla f(\boldsymbol{x}^t)\|_2^2$ lies above $f(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t))$ for small η
- ensures sufficient decrease of objective values

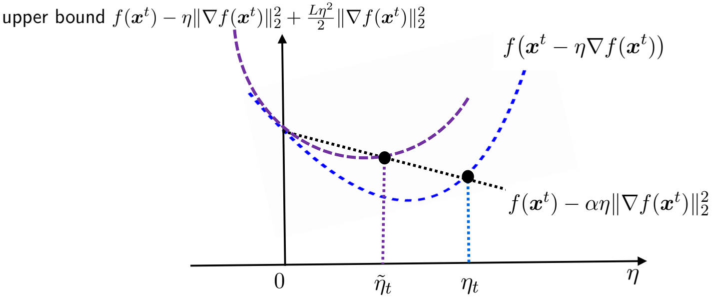
Backtracking line search



Algorithm 2.2 Backtracking line search for GD

1: Initialize $\eta = 1$, $0 < \alpha \le 1/2$, $0 < \beta < 1$
2: **while** $f\big(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)\big) > f(\boldsymbol{x}^t) - \alpha\eta \|\nabla f(\boldsymbol{x}^t)\|_2^2$ **do**
3: $\quad \eta \leftarrow \beta\eta$

Backtracking line search



Practically, backtracking line search often (but not always) provides good estimates on the local Lipschitz constants of gradients

Convergence for backtracking line search

Theorem 2.2 (Boyd, Vandenberghe '04)

Let f be μ-strongly convex and L-smooth. With backtracking line search,

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \left(1 - \min\left\{2\mu\alpha, \tfrac{2\beta\alpha\mu}{L}\right\}\right)^t \left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\right)$$

where $\boldsymbol{x}^*$ is the minimizer

Is strong convexity necessary for linear convergence?

So far we have established linear convergence under strong convexity and smoothness

Strong convexity requirement can often be relaxed

- local strong convexity
- regularity condition
- Polyak-Lojasiewicz condition

Example: logistic regression

Suppose we obtain m independent binary samples

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

where $\{\boldsymbol{a}_i\}$: known design vectors; $\boldsymbol{x}^\natural \in \mathbb{R}^n$: unknown parameters

Example: logistic regression

The maximum likelihood estimate (MLE) is given by (after a little manipulation)

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

- $\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\frac{\exp(-y_i\boldsymbol{a}_i^\top\boldsymbol{x})}{\big(1+\exp(-y_i\boldsymbol{a}_i^\top\boldsymbol{x})\big)^2}}_{\to 0 \text{ if } \boldsymbol{x}\to\infty} \boldsymbol{a}_i\boldsymbol{a}_i^\top \overset{\boldsymbol{x}\to\infty}{\longrightarrow} \boldsymbol{0}$

$$\Longrightarrow \qquad f \text{ is } 0\text{-strongly convex}$$

- Does it mean we no longer have linear convergence?

Local strong convexity

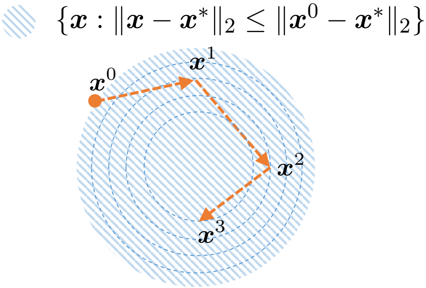
Theorem 2.3 (GD for locally strongly convex and smooth functions)

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 Theorem 2.1 continues to hold

Local strong convexity



- 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 $\boldsymbol{x}^{t+1} \in \mathcal{B}_0$, so the above bound continues to hold for the next iteration ...

Local strong convexity

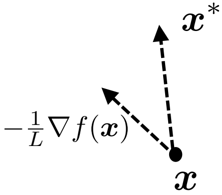
Back to the logistic regression example, the local strong convexity parameter is given by

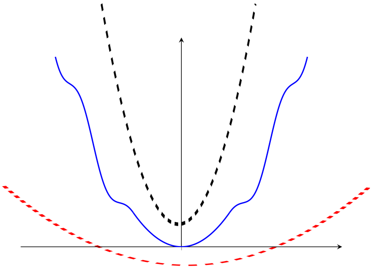
$$\inf_{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^0-\boldsymbol{x}^*\|_2} \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^{m} \frac{\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x})}{\big(1+\exp(-y_i \boldsymbol{a}_i^\top \boldsymbol{x})\big)^2} \boldsymbol{a}_i \boldsymbol{a}_i^\top\right) \tag{2.6}$$

which is often stricly bounded away from 0,[1] thus enabling linear convergence

[1]For example, when $\boldsymbol{x}^\star = \boldsymbol{0}$ and $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, one often has (2.6) $\geq c_0$ for some universal constant $c_0 > 0$ with high prob if $m/n > 2$ (Sur et al. '17)

Regularity condition

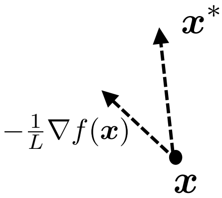


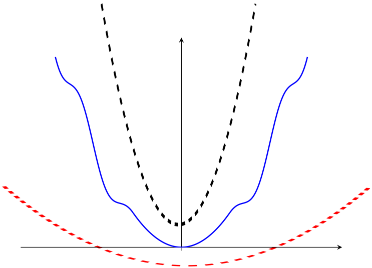


Another way is to replace strong convexity and smoothness by the following regularity condition:

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2, \quad \forall \boldsymbol{x} \tag{2.7}$$

Regularity condition

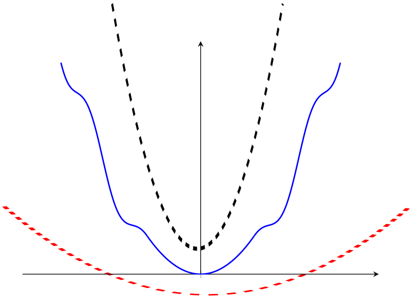




$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*)\|_2^2, \quad \forall \boldsymbol{x}$$

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

Convergence under regularity condition



Theorem 2.4

Suppose f satisfies (2.7). 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 of Theorem 2.4

It follows that

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

where (i) comes from (2.7)

Apply it recursively to complete the proof

Polyak-Lojasiewicz condition

Another alternative is the Polyak-Lojasiewicz (PL) condition

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu(f(\boldsymbol{x}) - f(\underbrace{\boldsymbol{x}^*}_{\text{minimizer}})), \quad \forall \boldsymbol{x} \tag{2.8}$$

- guarantees that gradient grows fast as we move away from the optimal objective value
- guarantees that every stationary point is a global minimum

Convergence under PL condition

Theorem 2.5

Suppose f *satisfies* (2.8) *and is* L*-smooth. If* $\eta_t \equiv \eta = \frac{1}{L}$*, then*

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

- guarantees linear convergence to the optimal objective value
- does NOT imply the uniqueness of global minima
- proof deferred to Page 2-45

Example: over-parameterized linear regression

- m data samples $\{\boldsymbol{a}_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}_{1 \le i \le m}$
- linear regression: find a linear model that best fits the data

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} \quad f(\boldsymbol{x}) \triangleq \frac{1}{2} \sum_{i=1}^{m} (\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2$$

Over-parameterization: model dimension $>$ sample size (i.e. $n > m$)

— *a regime of particular importance in deep learning*

Example: over-parametrized linear regression

While this is a convex problem, it is not strongly convex, since

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

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

Example: over-parametrized linear regression

Fact 2.6

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

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

- very mild assumption on $\{\boldsymbol{a}_i\}$
- no assumption on $\{y_i\}$

Example: over-parametrized linear regression

Fact 2.6

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

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

- **(aside)** while there are many global minima for this over-parametrized problem, GD has implicit bias
  - GD converges to a global min closest to initialization $\boldsymbol{x}^0$!

Proof of Fact 2.6

Everything boils down to showing the PL condition

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)\, f(\boldsymbol{x}) \tag{2.9}$$

If this holds, then the claim follows immediately from Theorem 2.5 and the fact $f(\boldsymbol{x}^*) = 0$

To prove (2.9), 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

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

which satisfies the PL condition (2.9) with $\mu = \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)$

Convex and smooth problems

Dropping strong convexity

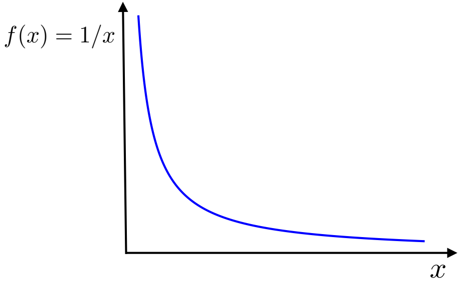
What happens if we completely drop (local) strong convexity?

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x})$$

- $f(\boldsymbol{x})$ is convex and smooth

Dropping strong convexity

Without strong convexity, it may often be better to focus on 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

Objective improvement and stepsize

Question:

- can we ensure reduction of the objective value (i.e. $f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t)$) without strong convexity?
- what stepsizes guarantee sufficient decrease?

Key idea: majorization-minimization

- find a *simple* majorizing function of $f(\boldsymbol{x})$ and optimize it instead

Objective improvement and stepsize

From the smoothness assumption,

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

(pick $\eta_t = 1/L$ to minimize the majorizing function)

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

Objective improvement

Fact 2.7

Suppose f is L-smooth. Then GD with $\eta_t = 1/L$ obeys

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

- for η_t sufficiently small, GD results in improvement in the objective value
- *does NOT rely on convexity!*

A byproduct: proof of Theorem 2.5

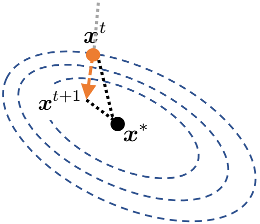
$$
\begin{aligned}
f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*) &\overset{\text{(i)}}{\leq} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^t)\|_2^2 \\
&\overset{\text{(ii)}}{\leq} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) - \frac{\mu}{L}\big(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\big)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

where (i) follows from Fact 2.7, and (ii) comes from the PL condition (2.8)

Apply it recursively to complete the proof

Improvement in estimation accuracy

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



$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$ is monotonically nonincreasing in t

Treating f as 0-strongly convex, we can see from our previous analysis for strongly convex problems that

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \le \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$$

Improvement in estimation accuracy

One can further show that $\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$ is strictly decreasing unless $\boldsymbol{x}^t$ is already the minimizer

Fact 2.8

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

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

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

Proof of Fact 2.8

It follows that

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

Monotonicity of gradient sizes

When $\eta_t = 1/L$, gradient sizes are also monotonically non-increasing

Lemma 2.9

Let f be convex and smooth. If $\eta_t \equiv \eta = 1/L$, then GD obeys

$$\left\|\nabla f(x^{t+1})\right\|_2 \leq \left\|\nabla f(x^t)\right\|_2$$

As a result, GD enjoys at least 3 types of monotonicity as t grows:

- objective value $f(x^t) \searrow$
- estimation error $\|x^t - x^*\|_2 \searrow$
- gradient size $\|f(x^t)\|_2 \searrow$

Proof of Lemma 2.9

Recall that the fundamental theorem of calculus gives

$$\begin{aligned}\nabla f(x^{t+1}) &= \nabla f(x^t) + \int_0^1 \nabla^2 f(x_\tau)\big(x^{t+1} - x^t\big)\mathrm{d}\tau \\ &= \underbrace{\left(I - \eta \int_0^1 \nabla^2 f(x_\tau)\mathrm{d}\tau\right)}_{=:\,B} \nabla f(x^t),\end{aligned}$$

where $x_\tau := x^t + \tau(x^{t+1} - x^t)$. When $\eta \leq 1/L$, it is easily seen that

$$0 \preceq B \preceq I \qquad \Longrightarrow \qquad 0 \preceq B^2 \preceq I$$

We can thus derive

$$\big\|\nabla f(x^{t+1})\big\|_2^2 - \big\|\nabla f(x^t)\big\|_2^2 = \nabla f(x^t)^\top (B^2 - I) \nabla f(x^t) \leq 0$$

Convergence rate for convex and smooth problems

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

Theorem 2.10 (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 $f(\cdot)$

- attains ε-accuracy within $O(1/\varepsilon)$ iterations (vs. $O(\log \frac{1}{\varepsilon})$ iterations for linear convergence)

Proof of Theorem 2.10 (cont.)

From Fact 2.7,

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

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

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

$$\Longrightarrow \qquad \|\nabla f(\boldsymbol{x}^t)\|_2 \geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2} \overset{\text{Fact 2.8}}{\geq} \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}$$

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

$$\Delta_{t+1} - \Delta_t \leq -\frac{1}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}\Delta_t^2 =: -\frac{1}{w_0}\Delta_t^2 \tag{2.10}$$

Proof of Theorem 2.10 (cont.)

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{w_0}\Delta_t^2$$

Dividing both sides by $\Delta_t \Delta_{t+1}$ and rearranging terms give

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

$$\Longrightarrow \qquad \frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{w_0} \qquad \text{(since } \Delta_t \geq \Delta_{t+1} \text{ (Fact 2.7))}$$

$$\Longrightarrow \qquad \frac{1}{\Delta_t} \geq \frac{1}{\Delta_0} + \frac{t}{w_0} \geq \frac{t}{w_0}$$

$$\Longrightarrow \qquad \Delta_t \leq \frac{w_0}{t} = \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}$$

as claimed

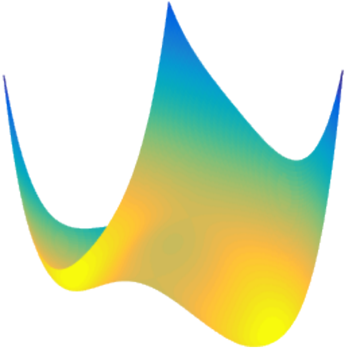
Nonconvex problems

Nonconvex problems are everywhere

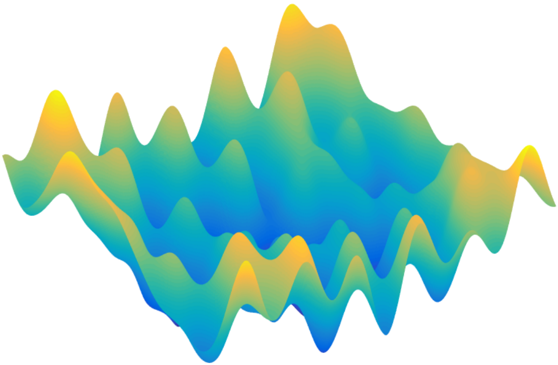
Many empirical risk minimization tasks are nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \text{data})$$

- low-rank matrix completion
- blind deconvolution
- dictionary learning
- mixture models
- **learning deep neural nets**
- ...



Challenges



- there may be bumps and local minima everywhere
    - e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)
- no algorithm can solve nonconvex problems efficiently in all cases

Typical convergence guarantees

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

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

Making gradients small

Suppose we are content with any (approximate) stationary point ...

This means that our goal is merely to find a point $\boldsymbol{x}$ with

$$\|\nabla f(\boldsymbol{x})\|_2 \le \varepsilon \quad \text{(called } \varepsilon\text{-approximate stationary point)}$$

Question: can GD achieve this goal? If so, how fast?

Making gradients small

Theorem 2.11

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

- *In general, GD obeys*

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

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

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

- GD finds an ε-approximate stationary point in $O(1/\varepsilon^2)$ iterations
- does not imply GD converges to stationary points; it only says that $\exists$ approximate stationary point in the GD trajectory

Proof of Theorem 2.11

From Fact 2.7, we know

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

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

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

$$\Longrightarrow \qquad \min_{t_0 \leq k < t}\|\nabla f(\boldsymbol{x}^k)\|_2 \leq \sqrt{\frac{2L\left(f(\boldsymbol{x}^{t_0}) - f(\boldsymbol{x}^*)\right)}{t - t_0}} \qquad (2.11)$$

Proof of Theorem 2.11 (cont.)

For a general $f(\cdot)$, taking $t_0 = 0$ immediately estalishes the claim

If $f(\cdot)$ is convex, invoke Theorem 2.10 to obtain

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

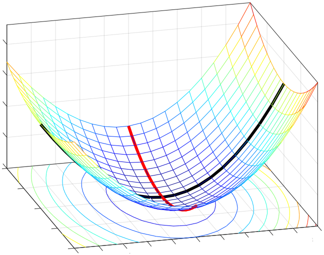
Taking $t_0 = t/2$ and combining it with (2.11) give

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

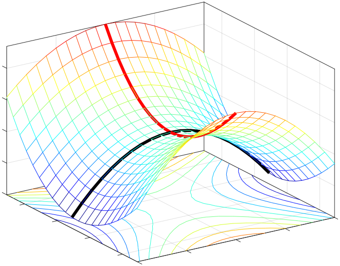
In view of Lemma 2.9, $\min_{t_0 \leq k < t} \|\nabla f(\boldsymbol{x}^k)\|_2 = \|\nabla f(\boldsymbol{x}^t)\|_2$, thus concluding the proof

Escaping saddles

There are at least two kinds of points with vanishing gradients



global and local minimum



saddle point

Saddle points look like "unstable" critical points; can we hope to at least avoid saddle points?

Escaping saddle points

GD cannnot always escape saddles

- e.g. if $\underbrace{\boldsymbol{x}^0 \text{ happens to be a saddle}}_{\text{can often be prevented by random initialization}}$, then GD gets trapped (since $\nabla f(\boldsymbol{x}^0) = \boldsymbol{0}$)

Fortunately, under mild conditions, randomly initialized GD converges to local (sometimes even global) minimum almost surely (Lee et al.)!

Example

Consider a simple *nonconvex* quadratic minimization problem

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$$

- $\boldsymbol{A} = \boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top$, where $\|\boldsymbol{u}_1\|_2 = \|\boldsymbol{u}_2\|_2 = 1$ and $\boldsymbol{u}_1^\top\boldsymbol{u}_2 = 0$

This problem has (at least) a saddle point: $\boldsymbol{x} = \boldsymbol{0}$ (why?)

- if $\boldsymbol{x}^0 = \boldsymbol{0}$, then GD gets stuck at $\boldsymbol{0}$ (i.e. $\boldsymbol{x}^t \equiv \boldsymbol{0}$)
- what if we initialize GD randomly? can we hope to avoid saddles?

Example (cont.)

Fact 2.12

If $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, *then with prob. approaching 1, GD with* $\eta < 1$ *obeys*

$$\|\boldsymbol{x}^t\|_2 \to \infty \qquad \text{as } t \to \infty$$

- Interestingly, GD (almost) never gets trapped in the saddle $\mathbf{0}$!

Example (cont.)

Proof of Fact 2.12: Observe that

$$\boldsymbol{I} - \eta \boldsymbol{A} = \boldsymbol{I}_{\perp} + (1-\eta)\boldsymbol{u}_1\boldsymbol{u}_1^{\top} + (1+\eta)\boldsymbol{u}_2\boldsymbol{u}_2^{\top}$$

where $\boldsymbol{I}_{\perp} := \boldsymbol{I} - \boldsymbol{u}_1\boldsymbol{u}_1^{\top} - \boldsymbol{u}_2\boldsymbol{u}_2^{\top}$. It can be easily verified that

$$(\boldsymbol{I} - \eta \boldsymbol{A})^t = \boldsymbol{I}_{\perp} + (1-\eta)^t\boldsymbol{u}_1\boldsymbol{u}_1^{\top} + (1+\eta)^t\boldsymbol{u}_2\boldsymbol{u}_2^{\top}$$

$$\begin{aligned}
\Longrightarrow \qquad \boldsymbol{x}^t &= (\boldsymbol{I} - \eta \boldsymbol{A})\boldsymbol{x}^{t-1} = \cdots = (\boldsymbol{I} - \eta \boldsymbol{A})^t \boldsymbol{x}^0 \\
&= \boldsymbol{I}_{\perp}\boldsymbol{x}^0 + \underbrace{(1-\eta)^t\left(\boldsymbol{u}_1^{\top}\boldsymbol{x}^0\right)}_{=:\,\alpha_t}\boldsymbol{u}_1 + \underbrace{(1+\eta)^t\left(\boldsymbol{u}_2^{\top}\boldsymbol{x}^0\right)}_{=:\,\beta_t}\boldsymbol{u}_2
\end{aligned}$$

Clearly, $\alpha_t \to 0$ as $t \to \infty$, and $\underbrace{|\beta_t| \to \infty}_{\text{and hence } \|\boldsymbol{x}^t\|_2 \to \infty}$ as long as $\underbrace{\beta_0 \neq 0}_{\text{happens with prob. 1}}$

Reference

- "*Convex optimization and algorithms*," D. Bertsekas, 2015.
- "*Convex optimization: algorithms and complexity*," S. Bubeck, *Foundations and trends in machine learning*, 2015.
- "*First-order methods in optimization*," A. Beck, Vol. 25, *SIAM*, 2017.
- "*Convex optimization*," S. Boyd, L. Vandenberghe, *Cambridge university press*, 2004.
- "*Introductory lectures on convex optimization: A basic course*," Y. Nesterov, *Springer Science & Business Media*, 2013.
- "*The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled shi-square*," P. Sur, Y. Chen, E. Candes, Probability Theory and Related Fields, 2017.
- "*How to make the gradients small*," Y. Nesterov, *Optima*, 2012.

Reference

- "*Gradient descent converges to minimizers*," J. Lee, M. Simchowitz, M. Jordan, B. Recht, *COLT*, 2016.
- "*Nonconvex optimization meets low-rank matrix factorization: an overview*," Y. Chi, Y. Lu, Y. Chen, *IEEE Transactions on Signal Processing*, 2019.