

## **Gradient methods for unconstrained problems**



Yuxin Chen

Princeton University, Fall 2019

# Outline

---

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

# Differentiable unconstrained minimization

---

$$\begin{array}{ll} \text{minimize}_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathbb{R}^n \end{array}$$

- $f$  (objective or cost function) is differentiable

# Iterative descent algorithms

---

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

$$f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t), \quad t = 0, 1, \dots$$

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

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

# Iterative descent algorithms

---

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

$$f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t), \quad t = 0, 1, \dots$$

- In each iteration, search in descent direction

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

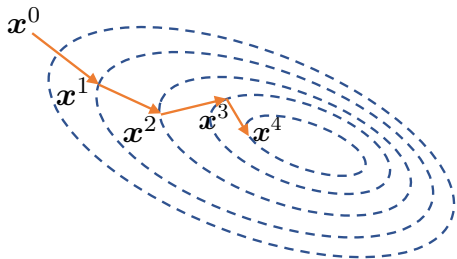
where  $\mathbf{d}^t$ : descent direction at  $\mathbf{x}^t$ ;  $\eta_t > 0$ : stepsize

# Gradient descent (GD)

---

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

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



- traced to Augustin Louis Cauchy '1847 ...

# Gradient descent (GD)

---

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

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

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

$$\arg \min_{\mathbf{d}: \|\mathbf{d}\|_2 \leq 1} f'(\mathbf{x}; \mathbf{d}) = \arg \min_{\mathbf{d}: \|\mathbf{d}\|_2 \leq 1} \nabla f(\mathbf{x})^\top \mathbf{d} = -\|\nabla f(\mathbf{x})\|_2$$

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

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

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

# Convergence for constant stepsizes

---

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

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

where  $\lambda_1(\mathbf{Q})$  (resp.  $\lambda_n(\mathbf{Q})$ ) is the largest (resp. smallest) eigenvalue of  $\mathbf{Q}$

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

# Convergence for constant stepsizes

---

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

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

where  $\lambda_1(Q)$  (resp.  $\lambda_n(Q)$ ) is the largest (resp. smallest) eigenvalue of  $Q$

- 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,

$$\begin{aligned}\mathbf{x}^{t+1} - \mathbf{x}^* &= \mathbf{x}^t - \mathbf{x}^* - \eta_t \nabla f(\mathbf{x}^t) = (\mathbf{I} - \eta_t \mathbf{Q})(\mathbf{x}^t - \mathbf{x}^*) \\ \implies \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2 &\leq \|\mathbf{I} - \eta_t \mathbf{Q}\| \|\mathbf{x}^t - \mathbf{x}^*\|_2\end{aligned}$$

The claim then follows by observing that

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

Apply the above bound recursively to complete the proof

□

## Exact line search

---

The stepsize rule  $\eta_t \equiv \eta = \frac{2}{\lambda_1(\mathbf{Q}) + \lambda_n(\mathbf{Q})}$  relies on the spectrum of  $\mathbf{Q}$ , which requires preliminary experimentation

Another more practical strategy is the **exact line search** rule

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

# Convergence for exact line search

---

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

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

- 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  $\mathbf{g}^t = \nabla f(\mathbf{x}^t) = \mathbf{Q}(\mathbf{x}^t - \mathbf{x}^*)$ . It can be verified that exact line search gives

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

This gives

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

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

# Convergence for exact line search

---

**Proof (cont.):** From Kantorovich's inequality

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

we arrive at

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

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

□



**Strongly convex and smooth problems**

# Strongly convex and smooth problems

---

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

$$\text{minimize}_x \quad f(x)$$

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

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

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

# Convergence rate for strongly convex and smooth problems

---

## Theorem 2.1 (GD for strongly convex and smooth functions)

Let  $f$  be  $\mu$ -strongly convex and  $L$ -smooth. If  $\eta_t \equiv \eta = \frac{2}{\mu+L}$ , then

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

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

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

# Convergence rate for strongly convex and smooth problems

---

## Theorem 2.1 (GD for strongly convex and smooth functions)

Let  $f$  be  $\mu$ -strongly convex and  $L$ -smooth. If  $\eta_t \equiv \eta = \frac{2}{\mu+L}$ , then

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

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

- dimension-free: iteration complexity is  $O\left(\frac{\log \frac{1}{\epsilon}}{\log \frac{\kappa+1}{\kappa-1}}\right)$ , which is independent of the problem size  $n$  if  $\kappa$  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  $\mu$ -strongly convex and  $L$ -smooth. If  $\eta_t \equiv \eta = \frac{2}{\mu+L}$ , then

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

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

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

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

# Proof of Theorem 2.1

---

It is seen from the fundamental theorem of calculus that

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

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

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

Repeat this argument for all iterations to conclude the proof

## More on strong convexity

---

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

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

### Equivalent first-order characterizations

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

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

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

## More on strong convexity

---

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

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

### Equivalent second-order characterization

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



## More on smoothness

---

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

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

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

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

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

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

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

## More on smoothness

---

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

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

### Equivalent second-order characterization

$$(v) \quad \|\nabla^2 f(\mathbf{x})\|_2 \leq L, \quad \forall \mathbf{x} \quad (\text{for twice differentiable functions})$$

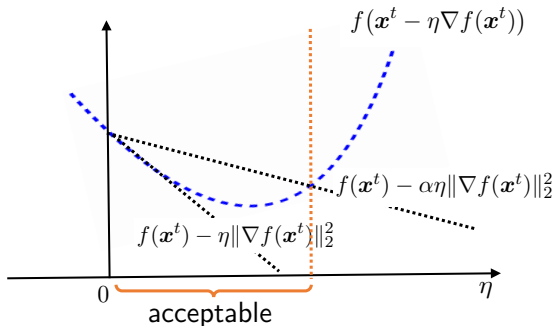
# Backtracking line search

---

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



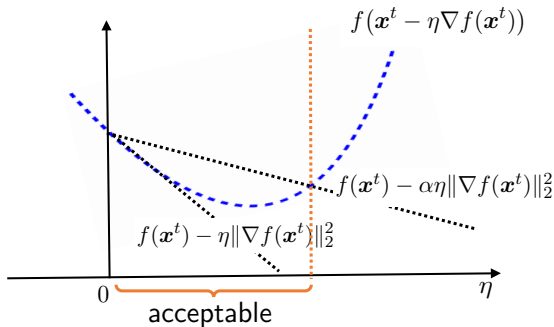
**Armijo condition:** for some  $0 < \alpha < 1$

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

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

# Backtracking line search

---



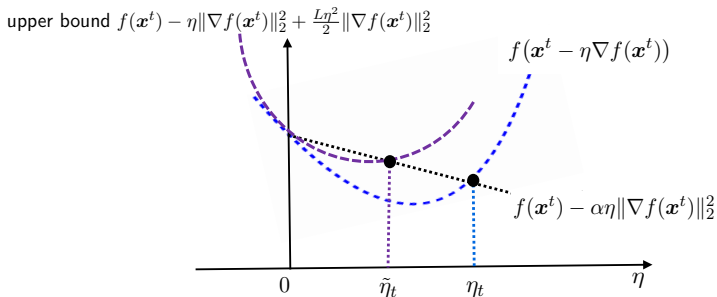
---

## Algorithm 2.2 Backtracking line search for GD

---

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

# 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  $\mu$ -strongly convex and  $L$ -smooth. With backtracking line search,

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

where  $\mathbf{x}^*$  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

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

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

## Example: logistic regression

---

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

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

$$\bullet \quad \nabla^2 f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \underbrace{\frac{\exp(-y_i \mathbf{a}_i^\top \mathbf{x})}{(1 + \exp(-y_i \mathbf{a}_i^\top \mathbf{x}))^2} \mathbf{a}_i \mathbf{a}_i^\top}_{\rightarrow 0 \text{ if } \mathbf{x} \rightarrow \infty} \xrightarrow{\mathbf{x} \rightarrow \infty} \mathbf{0}$$

$\implies$   $f$  is 0-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*  $\mu$ -strongly convex and  $L$ -smooth such that

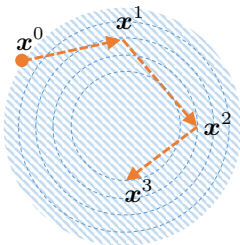
$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L \mathbf{I}, \quad \forall \mathbf{x} \in \mathcal{B}_0$$

where  $\mathcal{B}_0 := \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_2\}$  and  $\mathbf{x}^*$  is the minimizer.  
Then Theorem 2.1 continues to hold

# Local strong convexity

---

$$\{ \mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_2 \}$$



- Suppose  $\mathbf{x}^t \in \mathcal{B}_0$ . Then repeating our previous analysis yields  $\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2 \leq \frac{\kappa-1}{\kappa+1} \|\mathbf{x}^t - \mathbf{x}^*\|_2$
- This also means  $\mathbf{x}^{t+1} \in \mathcal{B}_0$ , 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

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

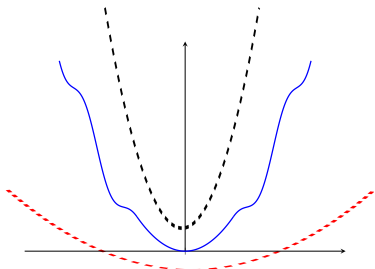
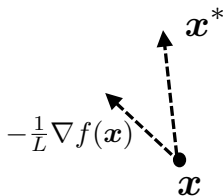
which is often strictly bounded away from 0,<sup>1</sup> thus enabling linear convergence

---

<sup>1</sup>For example, when  $\mathbf{x}^* = \mathbf{0}$  and  $\mathbf{a}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ , 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

---

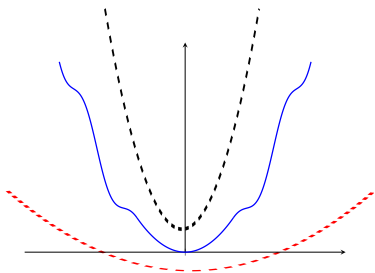
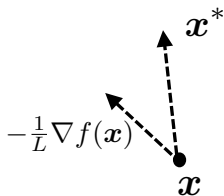


Another way is to replace strong convexity and smoothness by the following regularity condition:

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

# Regularity condition

---

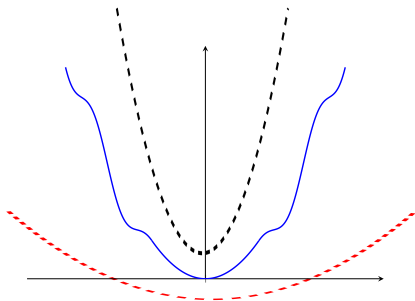


$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 + \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2, \quad \forall \mathbf{x}$$

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

# Convergence under regularity condition

---



## Theorem 2.4

Suppose  $f$  satisfies (2.7). If  $\eta_t \equiv \eta = \frac{1}{L}$ , then

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



## Proof of Theorem 2.4

---

It follows that

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

where (i) comes from (2.7)

Apply it recursively to complete the proof

# Polyak-Lojasiewicz condition

---

Another alternative is the Polyak-Lojasiewicz (PL) condition

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

- 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

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

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

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

**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

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

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

## Example: over-parametrized linear regression

---

### Fact 2.6

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

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

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

## Example: over-parametrized linear regression

### Fact 2.6

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

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

- **(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  $\mathbf{x}^0$ !

## Proof of Fact 2.6

---

Everything boils down to showing the PL condition

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2\lambda_{\min}(\mathbf{A}\mathbf{A}^\top) f(\mathbf{x}) \quad (2.9)$$

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

To prove (2.9), let  $\mathbf{y} = [y_i]_{1 \leq i \leq m}$ , and observe  $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})$ . Then

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

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



## **Convex and smooth problems**

# Dropping strong convexity

---

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

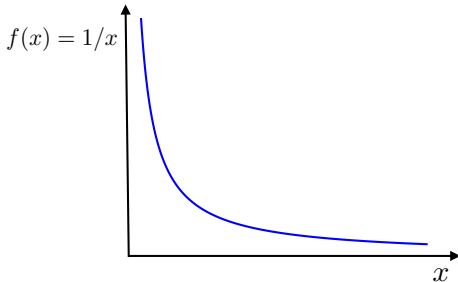
$$\text{minimize}_x \quad f(x)$$

- $f(x)$  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.  $f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t)$ ) without strong convexity?
- what stepsizes guarantee sufficient decrease?

## Key idea: **majorization-minimization**

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

# Objective improvement and stepsize

---

From the smoothness assumption,

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

majorizing function of objective reduction due to smoothness

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

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

# Objective improvement

---

## Fact 2.7

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

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

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

## A byproduct: proof of Theorem 2.5

---

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

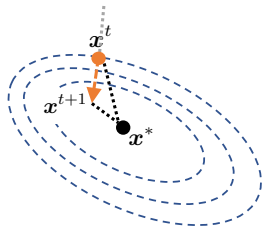
where (i) follows from Fact 2.7, and (ii) comes from the PL condition (2.8)

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  $\eta_t$  is not too large



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

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

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



## Improvement in estimation accuracy

---

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

### Fact 2.8

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

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

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

## Proof of Fact 2.8

---

It follows that

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

# Convergence rate for convex and smooth problems

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

## Theorem 2.9 (GD for convex and smooth problems)

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

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

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

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

## Proof of Theorem 2.9 (cont.)

---

From Fact 2.7,

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

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

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}^t) &\geq \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^* - \mathbf{x}^t) \geq -\|\nabla f(\mathbf{x}^t)\|_2 \|\mathbf{x}^t - \mathbf{x}^*\|_2 \\ \implies \|\nabla f(\mathbf{x}^t)\|_2 &\geq \frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|_2} \stackrel{\text{Fact 2.8}}{\geq} \frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{\|\mathbf{x}^0 - \mathbf{x}^*\|_2} \end{aligned}$$

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

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

## Proof of Theorem 2.9 (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

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

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

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

$$\implies \Delta_t \leq \frac{w_0}{t} = \frac{2L \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{t}$$

as claimed

# **Nonconvex problems**

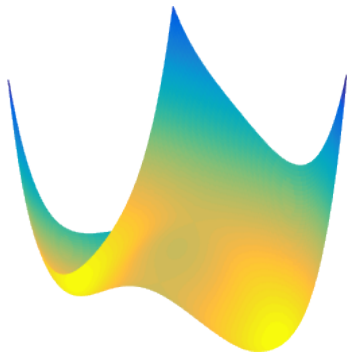
# Nonconvex problems are everywhere

---

Many empirical risk minimization tasks are nonconvex

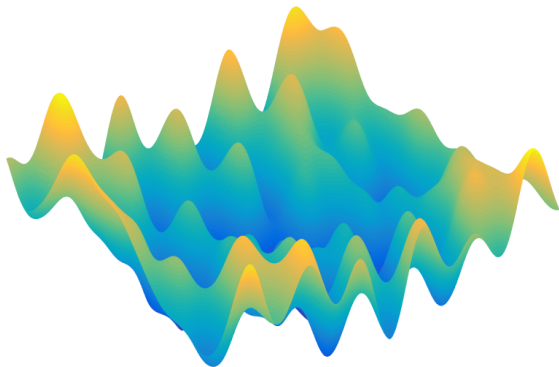
$$\text{minimize}_{\mathbf{x}} \quad f(\mathbf{x}; \text{data})$$

- 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.  $\nabla f(\mathbf{x}) = \mathbf{0}$ )
- 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  $\mathbf{x}$  with

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

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

# Making gradients small

## Theorem 2.10

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

- In general, GD obeys

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

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

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

- GD finds an  $\varepsilon$ -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 2.10

---

From Fact 2.7, we know

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

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

$$\begin{aligned} \frac{1}{2L} \sum_{k=t_0}^{t-1} \|\nabla f(\mathbf{x}^k)\|_2^2 &\leq \sum_{k=t_0}^{t-1} (f(\mathbf{x}^k) - f(\mathbf{x}^{k+1})) = f(\mathbf{x}^{t_0}) - f(\mathbf{x}^t) \\ &\leq f(\mathbf{x}^{t_0}) - f(\mathbf{x}^*) \\ \implies \min_{t_0 \leq k < t} \|\nabla f(\mathbf{x}^k)\|_2 &\leq \sqrt{\frac{2L(f(\mathbf{x}^{t_0}) - f(\mathbf{x}^*))}{t - t_0}} \end{aligned} \quad (2.11)$$

## Proof of Theorem 2.10 (cont.)

---

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

If  $f(\cdot)$  is convex, invoke Theorem 2.9 to obtain

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

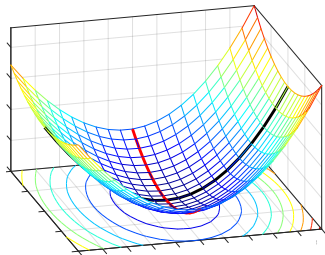
Taking  $t_0 = t/2$  and combining it with (2.11) give

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

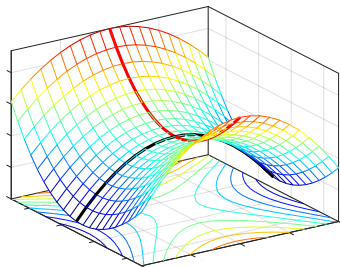
# Escaping saddles

---

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 cannot always escape saddles

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

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

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

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

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

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

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



## Example (cont.)

---

### Fact 2.11

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

$$\|\mathbf{x}^t\|_2 \rightarrow \infty \quad \text{as } t \rightarrow \infty$$

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

## Example (cont.)

---

**Proof of Fact 2.11:** Observe that

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

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

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

$$\begin{aligned} \implies \mathbf{x}^t &= (\mathbf{I} - \eta\mathbf{A})\mathbf{x}^{t-1} = \dots = (\mathbf{I} - \eta\mathbf{A})^t\mathbf{x}^0 \\ &= \mathbf{I}_\perp\mathbf{x}^0 + \underbrace{(1 - \eta)^t(\mathbf{u}_1^\top\mathbf{x}^0)}_{=:\alpha_t}\mathbf{u}_1 + \underbrace{(1 + \eta)^t(\mathbf{u}_2^\top\mathbf{x}^0)}_{=:\beta_t}\mathbf{u}_2 \end{aligned}$$

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

# Reference

---

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

# Reference

---

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