

# Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

## L4: Optimization Primal.

**Note:** all figures re-done on board in lecture.

## Lecture outline.

- 1 Review.
- 2 Convexity.
- 3 Descent methods.

## Reading.

- **Convexity and optimization:** Boyd and Vandenberghe “Convex Optimization”, Chapters 2-4.
- **Convexity** (*very optional! beyond this course.*): Hiriart-Urruty and Lemaréchal, “Fundamentals of Convex Analysis”; Borwein and Lewis, “Convex Analysis and Nonlinear Optimization”.

## Linear classification.

Suppose  $y^{(i)} \in \{-1, +1\}$ .

**ERM (Empirical Risk Minimization)** for linear classification:

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell \left( y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)} \right) \quad \text{or} \quad \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell \left( y^{(i)} \mathbf{w}^\top \phi(\mathbf{x}^{(i)}) \right).$$

Some choices for loss  $\ell$ :

$z \mapsto \mathbf{1}[z \leq 0]$  zero-one/classification,

$z \mapsto \frac{1}{2}(1 - z)^2$  least squares/linear regression,

$z \mapsto \ln(1 + \exp(-z))$  logistic.

## Linear classification.

Some choices for loss  $\ell$ :

$$z \mapsto \mathbf{1}[z \leq 0]$$

zero-one/classification,

$$z \mapsto \frac{1}{2}(1 - z)^2$$

least squares/linear regression,

$$z \mapsto \ln(1 + \exp(-z))$$

logistic.

## Linear classification.

Some choices for loss  $\ell$ :

$$z \mapsto \mathbf{1}[z \leq 0]$$

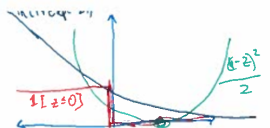
zero-one/classification,

$$z \mapsto \frac{1}{2}(1 - z)^2$$

least squares/linear regression,

$$z \mapsto \ln(1 + \exp(-z))$$

logistic.



## Descent methods.

How to solve for **weights**  $\mathbf{w} \in \mathbb{R}^d$  in

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) ?$$

Can use **gradient descent**. But why should it work?

## Descent methods.

How to solve for **weights**  $\mathbf{w} \in \mathbb{R}^d$  in

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) ?$$

Can use **gradient descent**. But why should it work?



**Today:** linear and logistic regression have no “bumps”.



## Descent methods and neural networks.

Neural nets are not linear models, but still have **weights**:

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)})$$

linear predictor,

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} f_{\mathbf{w}}(\mathbf{x}^{(i)}))$$

neural net.

Encounter “bumps” with neural nets?

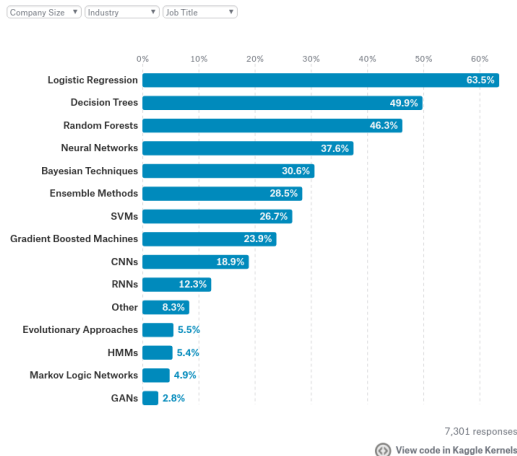
## Kaggle survey.

Lastly: is logistic regression relevant?

## Lastly: is logistic regression relevant?

### What data science methods are used at work?

Logistic regression is the most commonly reported data science method used at work for all industries *except* [Military and Security](#) where Neural Networks are used slightly more frequently.



# Convexity.

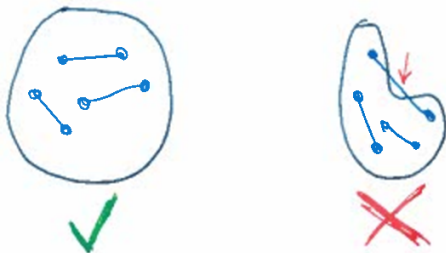
## Convexity.

**Convexity** will formalize “no bumps”.

**Convexity** is pervasive in mathematics, not just optimization.

## Convex sets.

A set is **convex** if it contains all line segments:



In symbols:  $C \subseteq \mathbb{R}^d$  is convex when

$$\{\mathbf{x}, \mathbf{y}\} \subseteq C \quad \implies \quad [\mathbf{x}, \mathbf{y}] \subseteq C$$

where  $[\mathbf{x}, \mathbf{y}] = \{\alpha \mathbf{x} + (1 - \alpha) \mathbf{y} : \alpha \in [0, 1]\}$ .

## Convex set operations.

**Convex hull** is similar to putting a rubber band around data:



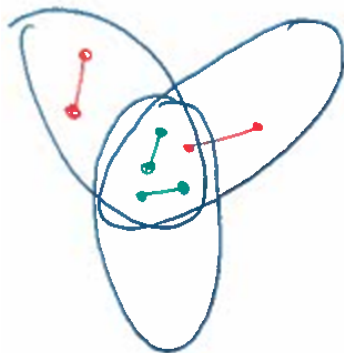
Rigorously: intersection of all convex supersets.

Alternatively (for finite set  $S = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ ):  
all **convex combinations**:

$$\text{conv}(S) := \left\{ \sum_{i=1}^k \alpha_i \mathbf{x}_i : \alpha_i \geq 0, \sum_{i=1}^k \alpha_i = 1 \right\}.$$

## Convex set intersections.

**Convex hull** is the intersection of all convex supersets. . .  
. . . why is convexity preserved under intersection?



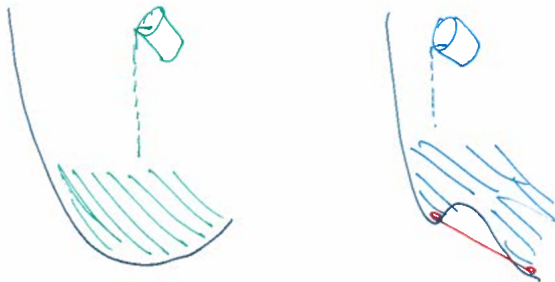
**Example. Polyhedron**  $\left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{Ax} \leq \mathbf{b} \right\}$ .



## Convex functions.

Bucket fill a function from above: this is the **epigraph**

$$\text{epi}(f) := \left\{ (\mathbf{x}, r) : \mathbf{x} \in \mathbb{R}^d, r \in \mathbb{R}, f(\mathbf{x}) \leq r \right\}.$$

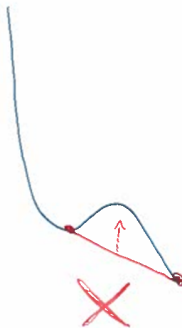
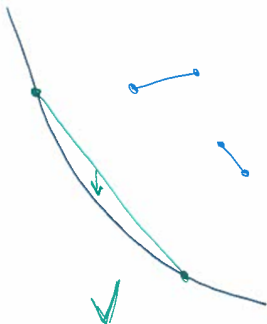


$f$  is a **convex function** when  $\text{epi}(f)$  is a convex set.

## Convex functions — algebraic form.

Equivalently: for any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ ,  $\alpha \in [0, 1]$ ,

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



## Convex functions — no bumps!

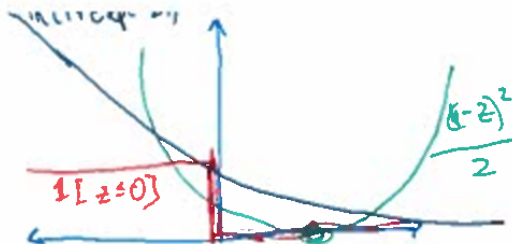
Gradient descent works when no “bumps”.



Convexity implies no bumps!

## Convex losses.

The logistic and least squares losses *look* convex:



### Question.

(a) How to prove this?

(b) What about convexity of  $\text{risk } \mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell \left( y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)} \right)$  ?

## Other convex functions.

- Exponential:  $e^x$ .
- Negative logarithm:  $-\log(x)$  over  $\mathbb{R}_{>0}$ .
- Negative entropy:  $x \log(x)$  over  $\mathbb{R}_{>0}$ .
- Norms:  $\|\mathbf{x}\|_p$  for  $p \geq 1$ .
- Log-Sum-Exp:  $\ln(\exp(\mathbf{x}_1) + \cdots + \exp(\mathbf{x}_d))$ .

## Three checks for convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$

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

## Three checks for convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$

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

Derivatives:  $\forall \mathbf{x}, \mathbf{y}$ ,

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

(This implies *increasing slopes*:  $(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0$ .)

## Three checks for convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$

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

Derivatives:  $\forall \mathbf{x}, \mathbf{y}$ ,

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

(This implies *increasing slopes*:  $(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0$ .)

Hessians:  $\forall \mathbf{x}$ ,

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

(We'll use this for least squares and logistic losses.)



## Strict convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$ :

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

Derivatives:  $\forall \mathbf{x}, \mathbf{y}$ ,

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

Hessians:  $\forall \mathbf{x}$ ,

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

## Strict convexity.

Function values:  $\forall \mathbf{x} \neq \mathbf{y}, \forall \alpha \in (0, 1)$ :

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) < \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}).$$

Derivatives:  $\forall \mathbf{x} \neq \mathbf{y}$ ,

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

Hessians:  $\forall \mathbf{x}$ ,

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

## $\lambda$ -Strong-Convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$

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

Derivatives:  $\forall \mathbf{x}, \mathbf{y}$

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

Hessians:  $\forall \mathbf{x},$

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

## $\lambda$ -Strong-Convexity.

Function values:  $\forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]$

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

Derivatives:  $\forall \mathbf{x}, \mathbf{y}$

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

Hessians:  $\forall \mathbf{x},$

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

## Convexity of key losses.

Logistic loss  $z \mapsto \ln(1 + \exp(-z))$  is **strictly convex**.

Squared loss  $z \mapsto \frac{1}{2}(1 - z)^2$  is **1-strongly-convex**.

What about the **risk**  $\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)})$  ?

## Operations preserving convexity.

If  $(f_1, \dots, f_k)$  convex and  $(\alpha_1, \dots, \alpha_k)$  nonnegative,

$$\mathbf{w} \mapsto \alpha_1 f_1(\mathbf{w}) + \dots + \alpha_k f_k(\mathbf{w}) \quad \text{is convex.}$$

If  $f$  is convex, then for any matrix  $\mathbf{A}$  and vector  $\mathbf{b}$ ,

$$\mathbf{w} \mapsto f(\mathbf{A}\mathbf{w} + \mathbf{b}) \quad \text{is convex.}$$

If  $\mathcal{F}$  is a *set* of convex functions,

$$\mathbf{w} \mapsto \sup_{f \in \mathcal{F}} f(\mathbf{w}) \quad \text{is convex.}$$

## Convexity of linear prediction with convex losses.

Suppose loss  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  is convex.

Collect  $y^{(i)} \mathbf{x}^{(i)}$  into a single matrix  $\mathbf{A} \in \mathbb{R}^{n \times d}$ :

$$\mathbf{A} := \begin{bmatrix} \leftarrow y^{(1)} \mathbf{x}^{(1)} \rightarrow \\ \vdots \\ \leftarrow y^{(n)} \mathbf{x}^{(n)} \rightarrow \end{bmatrix}.$$

Then  $\mathbf{v} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{v}_i)$  is convex, as is

$$\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^n \frac{1}{n} \ell((\mathbf{A} \mathbf{w})_i),$$

Alternatively:

$$\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^n \frac{1}{n} \ell\left(\mathbf{w}^\top (y^{(i)} \mathbf{x}^{(i)})\right).$$

## Convexity of linear prediction with convex losses.

Suppose loss  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  is convex.

Alternatively:

$$\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^n \frac{1}{n} \ell \left( \mathbf{w}^\top (y^{(i)} \mathbf{x}^{(i)}) \right).$$



## Convexity of linear prediction with convex losses.

Suppose loss  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  is convex.

Alternatively:

$$\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^n \frac{1}{n} \ell \left( \mathbf{w}^\top (y^{(i)} \mathbf{x}^{(i)}) \right).$$

**Therefore** linear and logistic regression are convex minimization!

## Convexity of linear prediction with convex losses.

Suppose loss  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  is convex.

Alternatively:

$$\mathbf{w} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^n \frac{1}{n} \ell \left( \mathbf{w}^\top (y^{(i)} \mathbf{x}^{(i)}) \right).$$

**Therefore** linear and logistic regression are convex minimization!

**Therefore** gradient descent should work!

## **Convexity: subgradients.**

## Convexity and differentiability.

Many useful convex functions are *not differentiable*.



$$x \mapsto |x|.$$

**Question:** how can we do gradient descent?

## Subgradients.

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

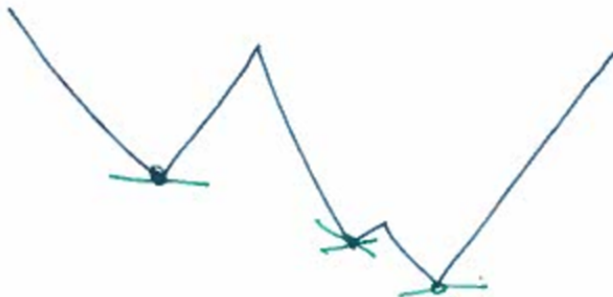


**Subdifferential set:**

$$\partial f(\mathbf{x}) = \left\{ \mathbf{s} \in \mathbb{R}^d : \forall \mathbf{y} . f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{s}^\top (\mathbf{y} - \mathbf{x}) \right\} .$$

## Aside: neural network subdifferentials.

Standard neural network packages (tensorflow, pytorch, etc.) give weird “descent directions” for things without even subgradients



## Subgradients: first order condition.

Suppose  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex.

**First order conditions:** For any  $\mathbf{y} \in \mathbb{R}^d$ ,

$$0 \in \partial f(\mathbf{y}) \quad \Longleftrightarrow \quad f(\mathbf{y}) = \inf_{\mathbf{x}} f(\mathbf{x}).$$

**Proof.** By definition of subgradient!

## Subgradients: first order condition.

Suppose  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex.

**First order conditions:** For any  $\mathbf{y} \in \mathbb{R}^d$ ,

$$0 \in \partial f(\mathbf{y}) \quad \Longleftrightarrow \quad f(\mathbf{y}) = \inf_{\mathbf{x}} f(\mathbf{x}).$$

**Proof.** By definition of subgradient!

**Magic of convexity:** local information gives global structure.



## Subgradients: Jensen's inequality.

If  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex, then  $\mathbb{E}f(\mathbf{X}) \geq f(\mathbb{E}\mathbf{X})$ .

**Proof.** Set  $\mathbf{y} := \mathbb{E}\mathbf{X}$ , and pick any  $\mathbf{s} \in \partial f(\mathbb{E}\mathbf{X})$ . Then

$$\mathbb{E}f(\mathbf{X}) \geq \mathbb{E} \left( f(\mathbf{y}) + \mathbf{s}^\top (\mathbf{X} - \mathbf{y}) \right) = f(\mathbf{y}) + \mathbf{s}^\top \mathbb{E}(\mathbf{X} - \mathbf{y}) = f(\mathbf{y}).$$

**Note.** This inequality comes up often!

## Further topics.

If you like this material,  
e.g., you'd like to see another reason why  $\frac{1}{2}\|\cdot\|^2$  has “1/2”, see

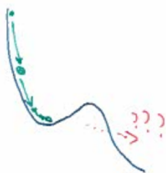
- Hiriart-Urruty and Lemaréchal, “Fundamentals of Convex Analysis”;
- Borwein and Lewis, “Convex Analysis and Nonlinear Optimization”.

## **Descent methods.**

todo: smooth gd, smooth gd convex, smooth+sc gd, vague sgd, descent directions and newton, maybe some others.

## Gradient descent.

- 1 Let  $\mathbf{w}_0 \in \mathbb{R}^d$  be given.
- 2 For  $i \in (0, 1, \dots, t)$ :
  - 1  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .



**Note.** Can relax  $\nabla f(\mathbf{w}_{i-1})$  in various ways.

## Smoothness.

$\lambda$ -**strong-convexity** was a Taylor lower bound:  $\forall \mathbf{x}, \mathbf{y}$ ,

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

Say  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\beta$ -**smooth** when reverse holds:  $\forall \mathbf{x}, \mathbf{y}$ ,

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

## Smooth, non-convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\alpha_i = 1/\beta$ ,

$$\min_{i \leq t} \|\nabla f(\mathbf{w}_{i-1})\|^2 \leq \frac{1}{t} \sum_{i=1}^t \|\nabla f(\mathbf{w}_{i-1})\|^2 \leq \frac{2\beta}{t} \left( f(\mathbf{w}_0) - \inf_{\mathbf{w}} f(\mathbf{w}) \right).$$

## Smooth, non-convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\alpha_i = 1/\beta$ ,

$$\min_{i \leq t} \|\nabla f(\mathbf{w}_{i-1})\|^2 \leq \frac{1}{t} \sum_{i=1}^t \|\nabla f(\mathbf{w}_{i-1})\|^2 \leq \frac{2\beta}{t} \left( f(\mathbf{w}_0) - \inf_{\mathbf{w}} f(\mathbf{w}) \right).$$

**Proof.** Averaging the inequalities (for each  $i \leq t$ )

$$\begin{aligned} f(\mathbf{w}_i) &\leq f(\mathbf{w}_{i-1}) - \nabla f(\mathbf{w}_{i-1})^\top (\mathbf{w}_i - \mathbf{w}_{i-1}) + \frac{\beta}{2} \|\mathbf{w}_i - \mathbf{w}_{i-1}\|^2 \\ &= f(\mathbf{w}_{i-1}) - \frac{1}{2\beta} \|\nabla f(\mathbf{w}_{i-1})\|^2 \end{aligned}$$

gives

$$\frac{1}{t} \sum_{i=1}^t \|\nabla f(\mathbf{w}_{i-1})\|^2 \leq \frac{2\beta}{t} (f(\mathbf{w}_0) - f(\mathbf{w}_t)).$$

## Smooth, convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If **convex**  $f$  is  $\beta$ -smooth and  $\alpha_i = 1/\beta$ , for every  $\mathbf{u} \in \mathbb{R}^d$

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \frac{1}{t} \sum_{i=1}^t (f(\mathbf{w}_i) - f(\mathbf{u})) \leq \frac{\beta}{2t} \left( \|\mathbf{w}_0 - \mathbf{u}\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2 \right).$$



## Smooth, convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If **convex**  $f$  is  $\beta$ -smooth and  $\alpha_i = 1/\beta$ , for every  $\mathbf{u} \in \mathbb{R}^d$

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \frac{1}{t} \sum_{i=1}^t (f(\mathbf{w}_i) - f(\mathbf{u})) \leq \frac{\beta}{2t} (\|\mathbf{w}_0 - \mathbf{u}\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2).$$

**Proof.** Average the inequalities (for each  $i \leq t$ )

$$\begin{aligned} \|\mathbf{w}_i - \mathbf{u}\|^2 &= \|\mathbf{w}_{i-1} - \mathbf{u}\|^2 - 2\alpha_i \nabla f(\mathbf{w}_{i-1})^\top (\mathbf{w}_{i-1} - \mathbf{u}) + \alpha_i^2 \|\nabla f(\mathbf{w}_{i-1})\|^2 \\ &\leq \|\mathbf{w}_{i-1} - \mathbf{u}\|^2 + 2\alpha_i (f(\mathbf{u}) - f(\mathbf{w}_{i-1})) + 2\alpha_i^2 \beta (f(\mathbf{w}_{i-1}) - f(\mathbf{w}_i)) \\ &= \|\mathbf{w}_{i-1} - \mathbf{u}\|^2 + \frac{2}{\beta} (f(\mathbf{u}) - f(\mathbf{w}_i)). \end{aligned}$$

## Smooth and strongly convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\lambda$ -strongly-convex,  $\alpha_i = 1/\beta$ , and  $\mathbf{u}$  is optimal,

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \exp(-t\lambda/\beta) .$$

## Smooth and strongly convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\lambda$ -strongly-convex,  $\alpha_i = 1/\beta$ , and  $\mathbf{u}$  is optimal,

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \exp(-t\lambda/\beta).$$

**Proof.** Since

$$\begin{aligned} f(\mathbf{u}) &= \inf_{\mathbf{v}} f(\mathbf{v}) = \inf_{\mathbf{v}} f(\mathbf{w}_i + \mathbf{v}) \\ &\geq \inf_{\mathbf{v}} \left( f(\mathbf{w}_i) + \nabla f(\mathbf{w}_i)^\top \mathbf{v} + \frac{\lambda}{2} \|\mathbf{v}\|^2 \right) = f(\mathbf{w}_i) - \frac{1}{2\lambda} \|\mathbf{w}_i\|^2, \end{aligned}$$

then

$$f(\mathbf{w}_i) \leq f(\mathbf{w}_{i-1}) - \frac{1}{2\beta} \|\mathbf{w}_{i-1}\|^2 \leq f(\mathbf{w}_{i-1}) - \frac{\lambda}{\beta} (f(\mathbf{w}_{i-1}) - f(\mathbf{u}));$$

recurse.

## Smooth and strongly convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\lambda$ -strongly-convex,  $\alpha_i = 1/\beta$ , and  $\mathbf{u}$  is optimal,

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \exp(-t\lambda/\beta) .$$

## Smooth and strongly convex.

**Gradient descent:**  $\mathbf{w}_0$  given;  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1})$ .

If  $f$  is  $\beta$ -smooth and  $\lambda$ -strongly-convex,  $\alpha_i = 1/\beta$ , and  $\mathbf{u}$  is optimal,

$$f(\mathbf{w}_t) - f(\mathbf{u}) \leq \exp(-t\lambda/\beta).$$

**Note.**  $\beta/\lambda$  is a **condition number**.

**Example.** Ridge regression  $w \mapsto \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$  is  $(\lambda_{\max}(\mathbf{X}^\top \mathbf{X}) + \lambda)$ -smooth and  $(\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) + \lambda)$ -strongly-convex. Increasing  $\lambda$  brings condition number closer to 1.

## Inexact gradients — descent directions.

We don't quite need  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \nabla f(\mathbf{w}_{i-1}) \dots$

$\dots$  can do  $\mathbf{w}_i := \mathbf{w}_{i-1} - \alpha_i \mathbf{v}_i$  with  $\mathbf{v}_i^\top \nabla(\mathbf{w}_{i-1}) > 0$ ;  
in this case,  $\mathbf{v}_i$  is a **descent direction**.

Can still prove rates.

## Inexact gradients — stochastic gradients.

Can replace gradient **stochastic gradient**  $\mathbf{v}_i$ ;  
namely,  $\mathbb{E}(\mathbf{v}_i) = \nabla f(\mathbf{w}_{i-1})$ .

**Example.** With linear prediction, we can use

$$\mathbf{v}_i := y^{(i)} \mathbf{x}^{(i)} \ell' \left( \mathbf{w}_{i-1}^\top (y^{(i)} \mathbf{x}^{(i)}) \right),$$

or a **minibatch**

$$\mathbf{v}_i := \frac{1}{|S_i|} \sum_{(\mathbf{x}, y) \in S_i} y \mathbf{x} \ell' \left( \mathbf{w}_{i-1}^\top (y \mathbf{x}) \right).$$

**Rates.**  $1/\sqrt{t}$  (Lipschitz  $f$ , bounded domain), or  $1/t$  (strong convexity).

**Why?** “Batch” gradient is expensive!

## Other methods.

- Momentum/Acceleration (rate  $1/t^2$  or  $\exp(-t\sqrt{\lambda/\beta})$ ).
- Line searches.
- SVRG.
- ADMM.
- ...



## **Key topics.**

- Logistic regression is glorious.
- Gradient descent hates bumps.
- Convex sets.
- Convex functions.
- ERM with convex losses of linear predictors.
- Gradient descent convergence rates.