CSE 446
Gradient Descent Part II

Natasha Jaques

UNIVERSITY of WASHINGTON

$$\hat{w}_{\mathrm{p}} = \underset{w}{\operatorname{argmin}} \sum_i \left(y_i - x_i^\top w\right)^2 + \lambda r(w)$$

$$\hat{w}_{\mathrm{c}} = \underset{w}{\operatorname{argmin}} \sum_i \left(y_i - x_i^\top w\right)^2 \quad \text{subject to} \quad r(w) \le \mu$$

Note that $\hat{w}_{\mathrm{p}}$ is a function of λ and $\hat{w}_{\mathrm{c}}$ is a function of μ. We could have written $\hat{w}_p(\lambda)$ and $\hat{w}_{\mathrm{c}}(\mu)$ to make this explicit, but we'll leave it implicit to keep the notation cleaner.
Assume for convenience that the solutions to the above optimization problems are unique.
Then for any $\lambda \ge 0$, there exists μ such that $\hat{w}_{\mathrm{p}} = \hat{w}_{\mathrm{c}}$, and vice versa.

Proof:

Given λ, solve for $\hat{w}_{\mathrm{p}}$, then set $\mu = r\left(\hat{w}_{\mathrm{p}}\right)$. Denote $f(w) = \sum_i \left(y_i - x_i^\top w\right)^2$.

Assume for contradiction that $\hat{w}_{\mathrm{p}} \neq \hat{w}_{\mathrm{c}}$.

We know that $\hat{w}_{\mathrm{p}}$ is feasible (i.e., $r\left(\hat{w}_{\mathrm{p}}\right) \le \mu$) because μ was set to exactly make $r\left(\hat{w}_p\right) = \mu$.

Since $\hat{w}_{\mathrm{c}}$ is the solution of the second optimization problem, we also know that:

(1) $r\left(\hat{w}_{\mathrm{c}}\right) \le \mu = r\left(\hat{w}_{\mathrm{p}}\right)$ (because $\hat{w}_{\mathrm{c}}$ is feasible by definition)

(2) $\mathrm{f}\left(\hat{w}_{\mathrm{c}}\right) < \mathrm{f}\left(\hat{w}_{\mathrm{p}}\right)$ (because $\hat{w}_{\mathrm{c}}$ minimizes $f(w)$ out of all feasible points, and $\hat{w}_{\mathrm{p}}$ is feasible).

Putting these together, we get that $\mathrm{f}\left(\hat{w}_{\mathrm{c}}\right) + \lambda r\left(\hat{w}_{\mathrm{c}}\right) < \mathrm{f}\left(\hat{w}_{\mathrm{p}}\right) + \lambda r\left(\hat{w}_{\mathrm{p}}\right)$.

But this is a contradiction because by definition, $\hat{w}_{\mathrm{p}}$ is supposed to minimize $\mathrm{f}(w) + \lambda r(w)$.

Thus, $\hat{w}_{\mathrm{p}} = \hat{w}_{\mathrm{c}}$.

Lecture plan

- Gradient descent algorithm + examples

- Theoretical analysis ← we are here
  - When does it work?
  - How quickly does it converge?
  - How do we choose a step size?
  - Key idea: Convexity

- Not tested on proof details, but concepts are important & practical

Convexity

- Optimization problems are hard to solve in general
- The exception: convex optimization
  - Objective is a convex function
  - Constraints are convex sets

- Special class of problems that can be solved efficiently
- Surprisingly common in practice

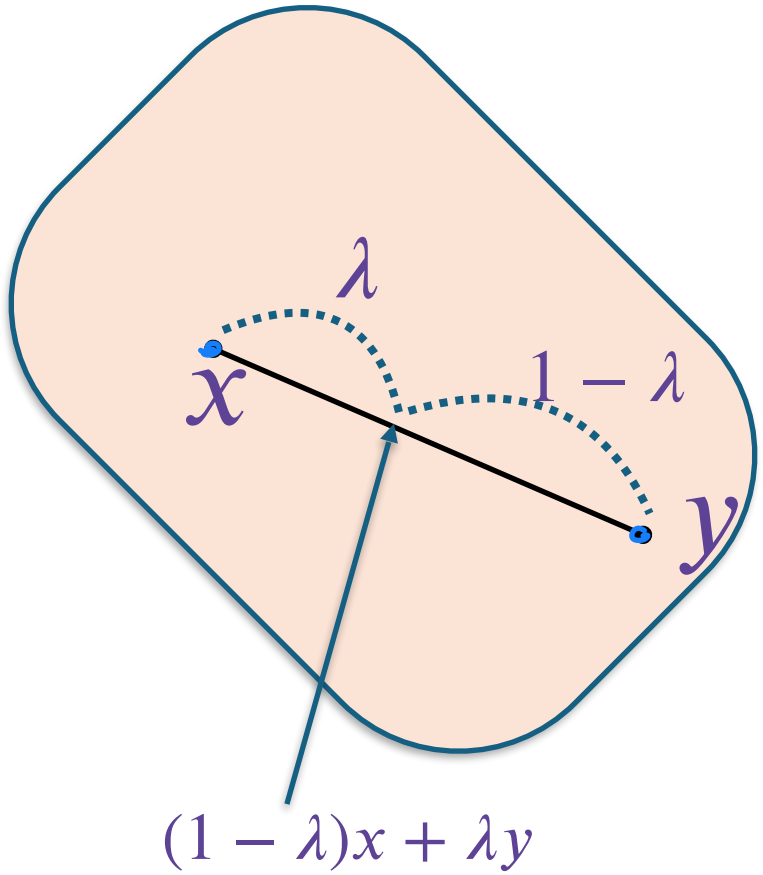
What is a convex set?

What is a convex set?

nothing to do w/ regularization

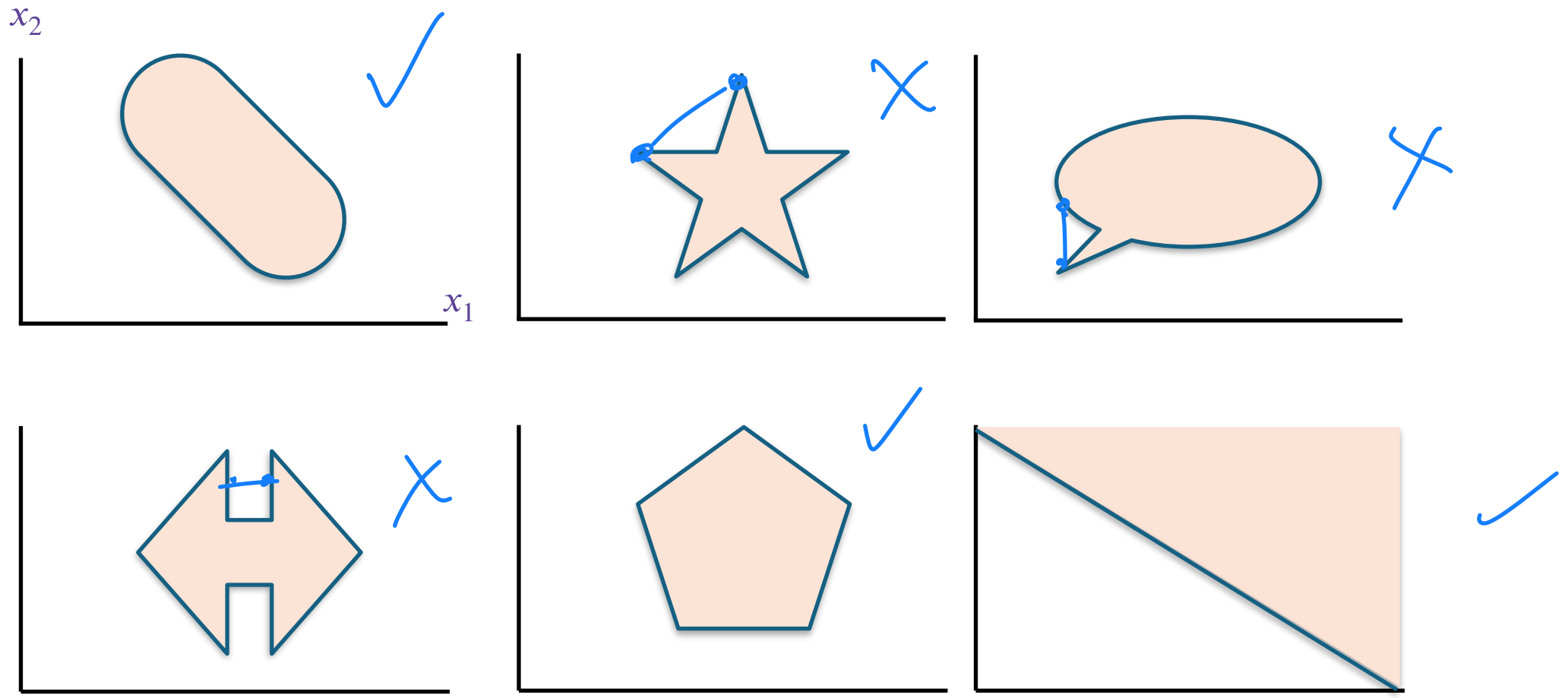
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

↳ every point between 2 points in the set must also be in the set



What is a convex set?

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$



What is a convex function?

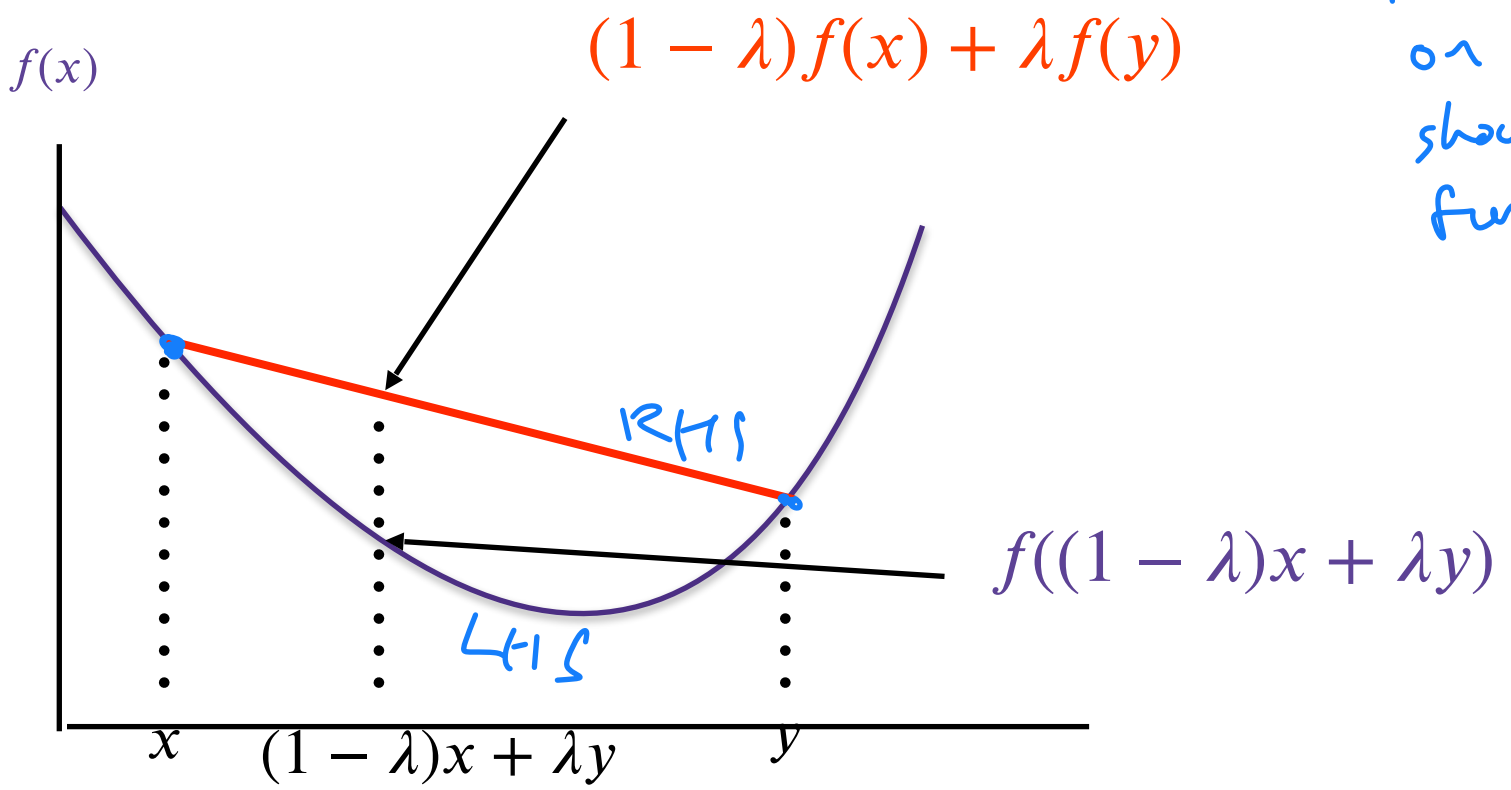
What is a convex function?

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$

LHS

RHS

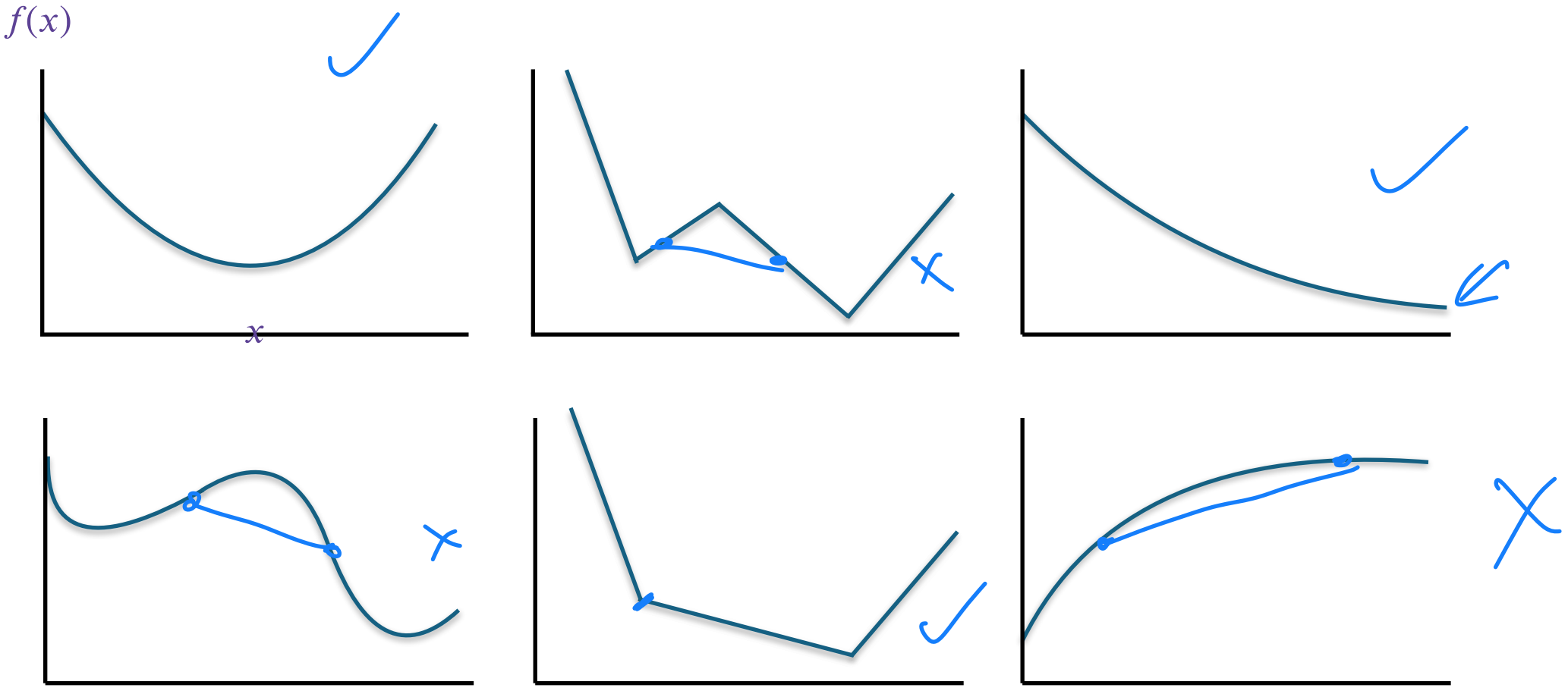
line b/w 2 points on the function should be above the function itself



What is a convex function?

smoothness ≠ convexity

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$



Convex functions and convex sets

Convex functions and convex sets

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

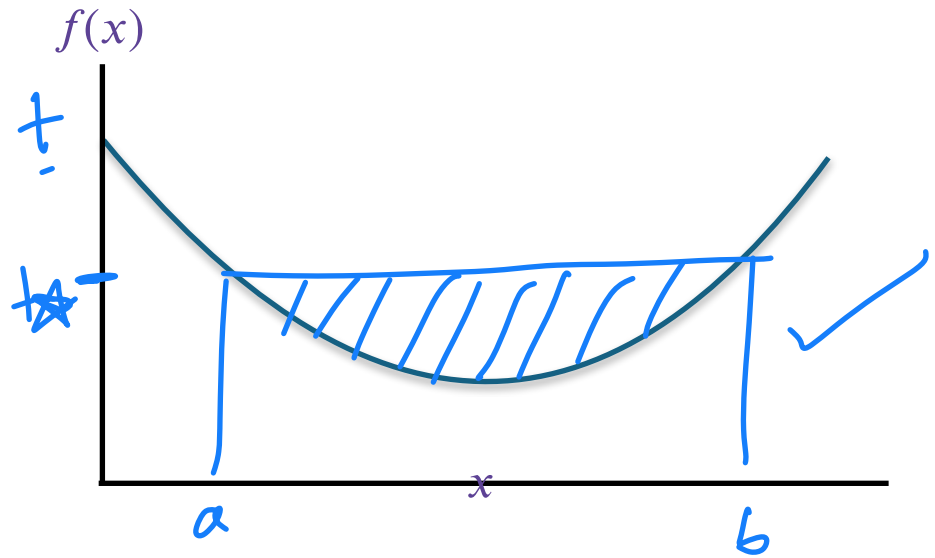
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$

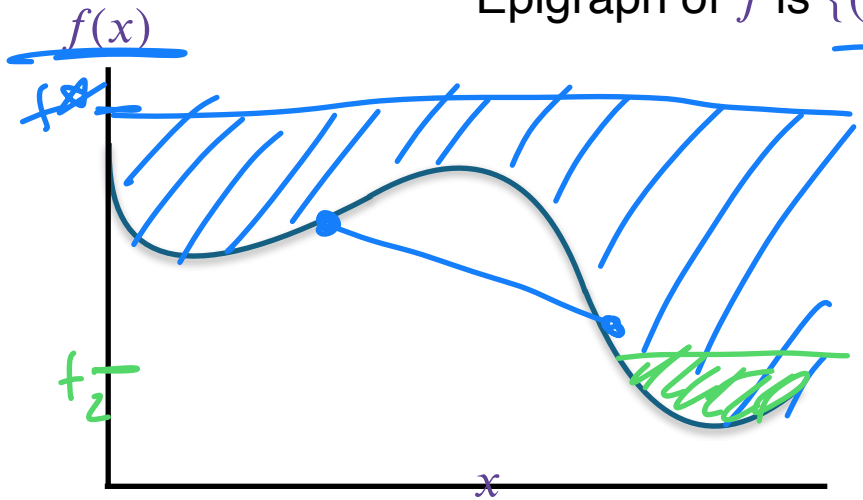
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

epigraph of

Graph of f is $\{(x,t) : f(x) = t\}$

Epigraph of f is $\{(x,t) : f(x) \leq t\}$





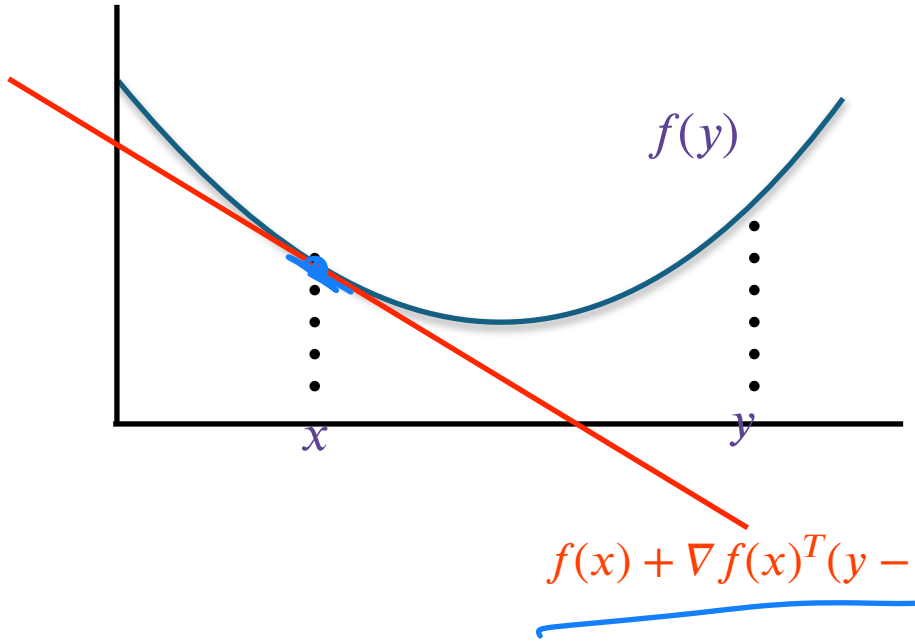
Convexity of differentiable functions

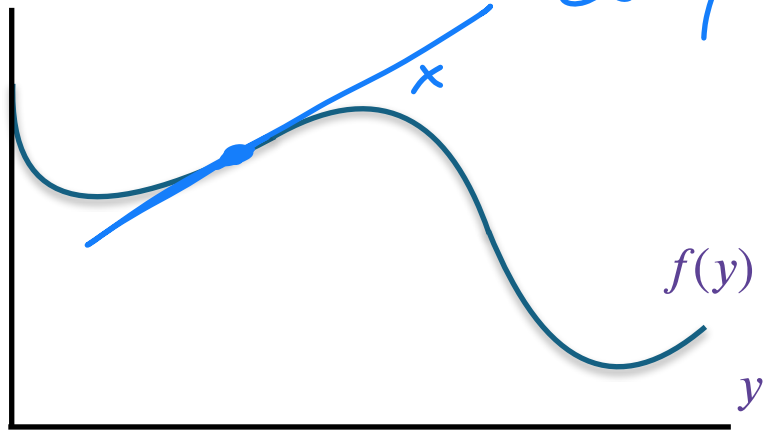
Convexity of differentiable functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x, t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

1st order Taylor expansion → needs to below the curve every where

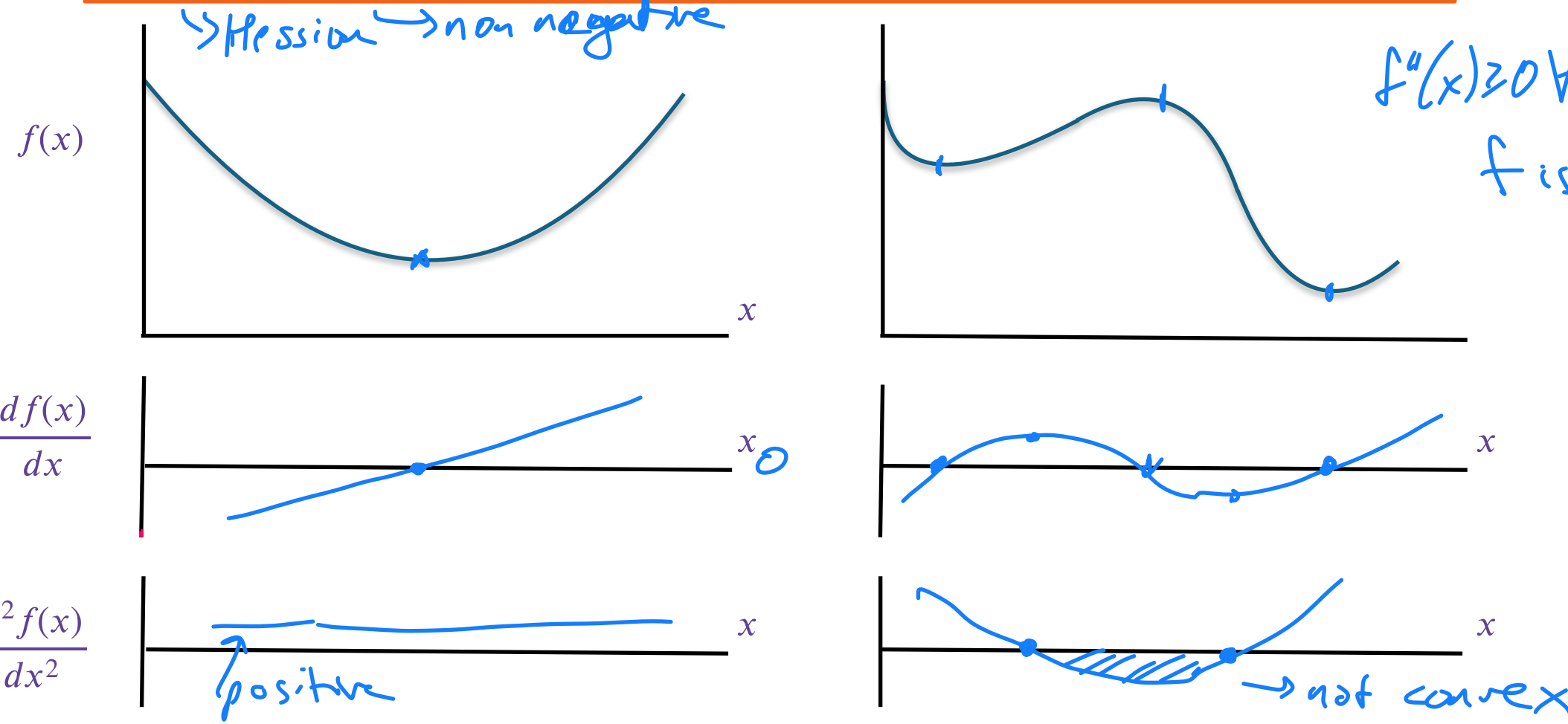




Convexity of twice-differentiable functions

Convexity of twice-differentiable functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$



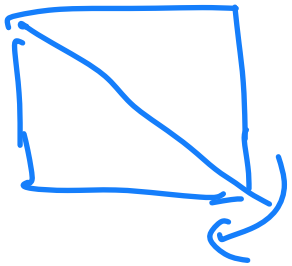
Convexity of twice-differentiable functions

Hessian

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$

$\nabla^2 f(x)_{ij} = \frac{d^2 f(x)}{d x_i d x_j} := A \in \mathbb{R}^{d \times d}$

symmetric $A_{ij} = A_{ji}$



always square & symmetric

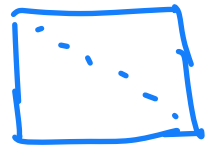
A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$

Positive semi-definite

$A \succeq 0 \iff \forall v \in \mathbb{R}^d, \; v^T A v \geq 0$

$\iff D_{ii} \geq 0$ where $A = Q^T D Q$

eigendecomposition $A = Q^T D Q$



eigenvalues

$D_{ii} \geq 0$ all eigenvalues are positive

Assume eigenvalues non-negative

$D_{ii} \geq 0 \; \forall i$

$v^T A v = v^T Q^T D Q v$ let $u = Qv$

$= u^T D u$

$= \sum_i u_i^2 D_{ii} \geq 0$

u_i^2: positive ≥ 0 bc squared

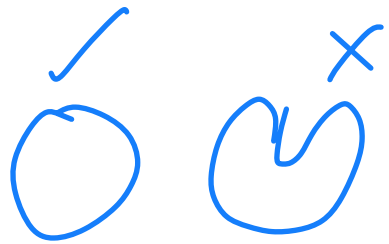
D_{ii}: positive ≥ 0 by assumption

→ Check if a loss that you're given is convex

Recap: Definitions of convexity

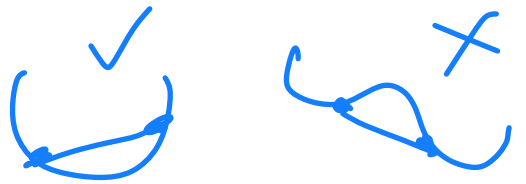
sets

A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$



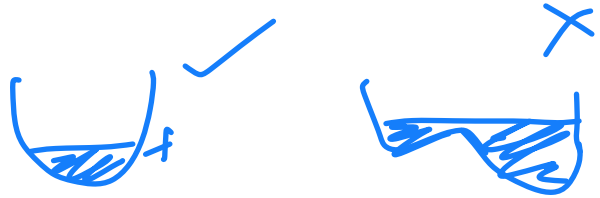
line above func

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$



epigraph

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the set $\{(x,t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex



1st order Taylor expansion

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable everywhere is convex if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in dom(f)$

slope line should be below func

Hessian

A function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice-differentiable everywhere is convex if $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$

gradient decreases goes up

Example: Ridge regression

How to prove convex?

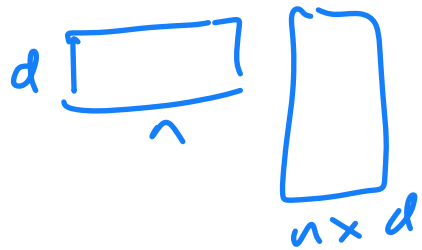
$$\underbrace{\underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_2^2}_{f(w)}$$

$$\underset{\|w\|_2^2 \le \rho}{\operatorname{argmin}} \|y - Xw\|_2^2$$

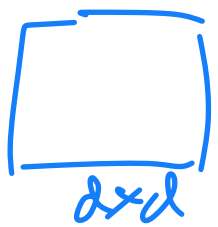
Let's use Hessian (easiest)

$$\nabla_w f(w) = 2X^T(Xw - y) + 2\lambda w$$

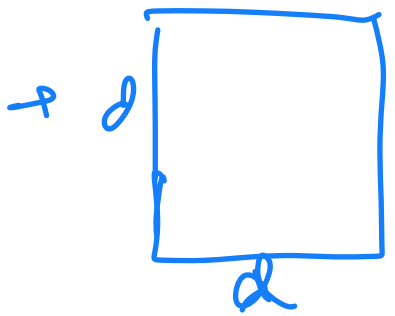
$$\nabla_w^2 f(w) = 2X^TX + 2\lambda I$$



PSD



$\lambda \ge 0$ positive



PD

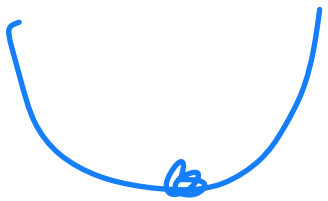
square & symmetric

Why $X^T X$ PSD?

$v \in \mathbb{R}^d$ $\quad v^T X^T X v = u^T u = \|u\|^2 \geq 0$

u

≥ 0

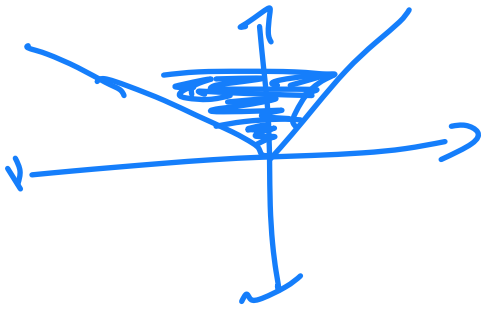


why convex good?

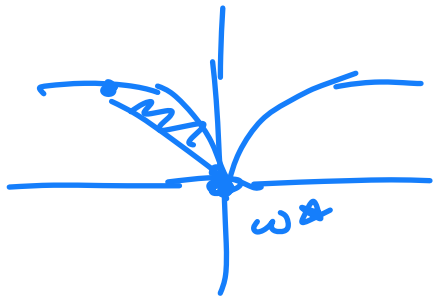
local minima = global minimum

Example: Lasso

$$\underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_1$$



Why not directly solve $\underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \operatorname{card}(w)$?



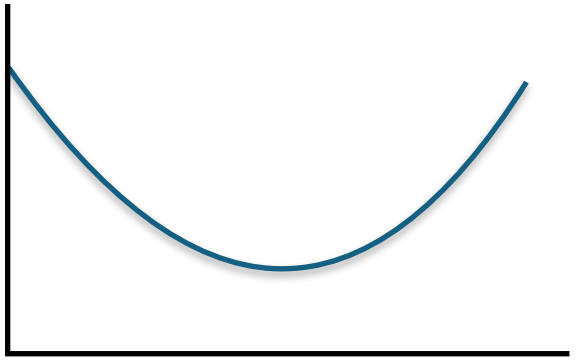
$\operatorname{card}(w)$ = count (w) where $w_i > 0$

Can interpret lasso as convex relaxation of cardinality objective

Convexity and gradient descent

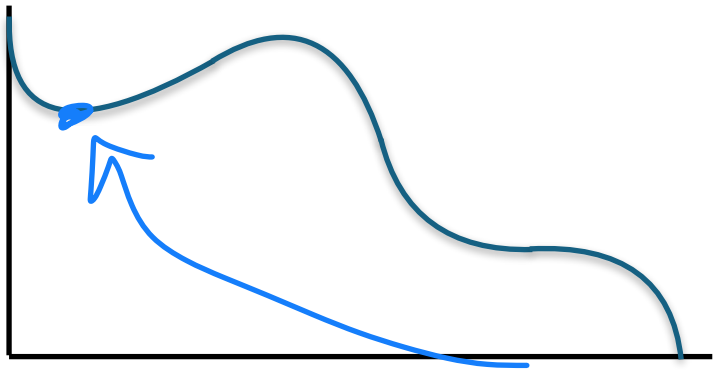
- All local minima are global minima

Convex function



Stationary points with $\nabla f(x) = 0$ are global minima

Non-convex function

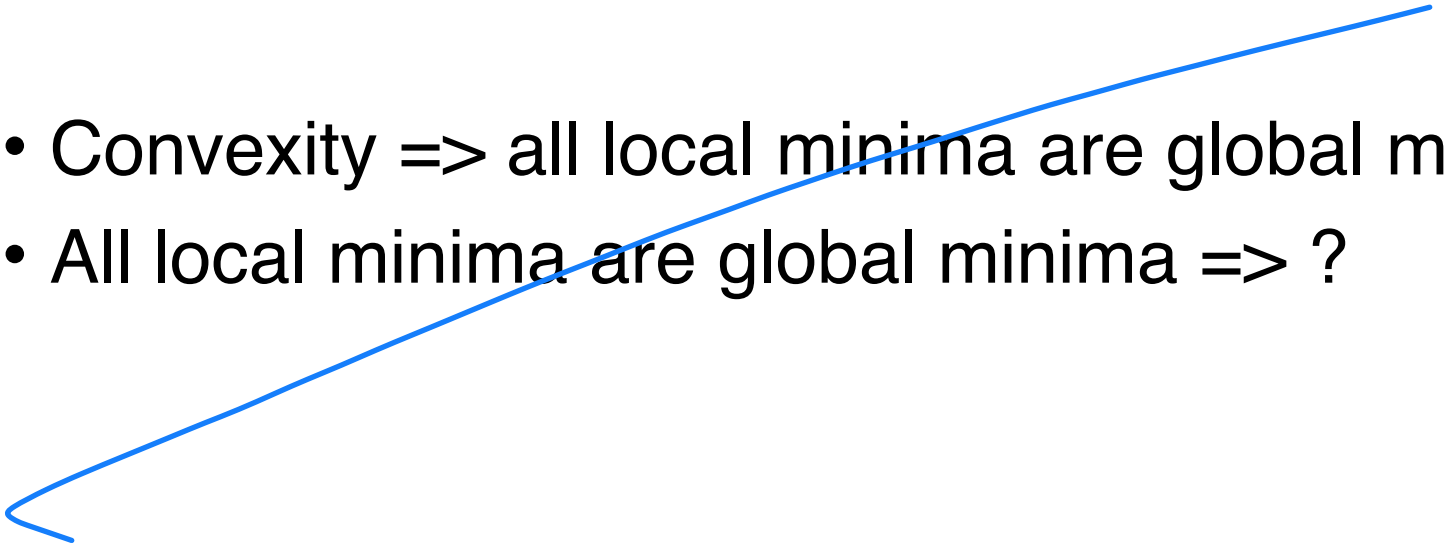


Stationary points with $\nabla f(x) = 0$ could be a local minima, a local maxima, or a saddle point

- Won't get stuck navigating the parameter constraints

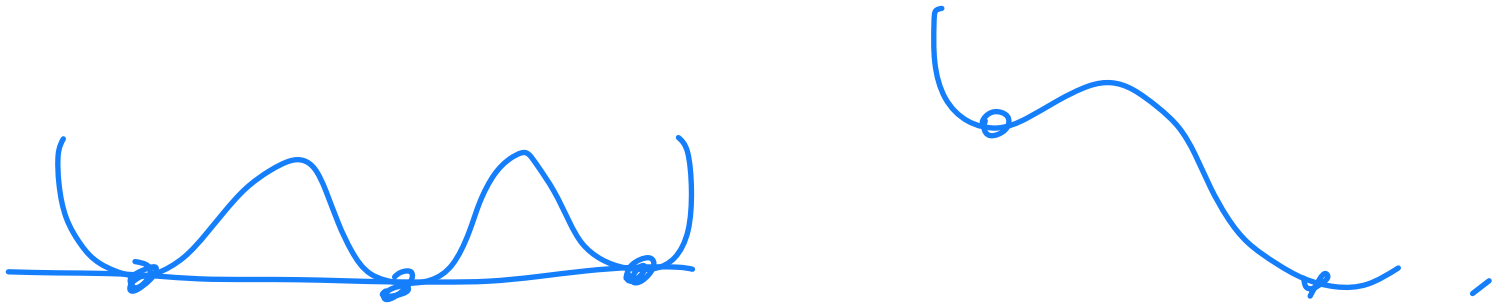
Convexity and gradient descent

- Convexity => all local minima are global minima
- All local minima are global minima => ?



Convexity and gradient descent

- You can always run gradient descent whether $f(w)$ is convex or not!
- But if $f(w)$ is convex, we have guarantees on converging to the global minimum
- Linear regression, ridge regression, Lasso → all convex!



Lecture plan

- Gradient descent algorithm + examples

- Theoretical analysis
  - When does it work?
  - How quickly does it converge? ← we are here
  - How do we choose a step size?
  - Key idea: Convexity

- Not tested on proof details, but concepts are important & practical

Convergence analysis steps

1. Study single iteration: $f(w_{t+1})$ vs. $f(w_t)$
2. Piece iterations together to study how they converge

Single-iteration progress bound

Don't need convexity (yet).

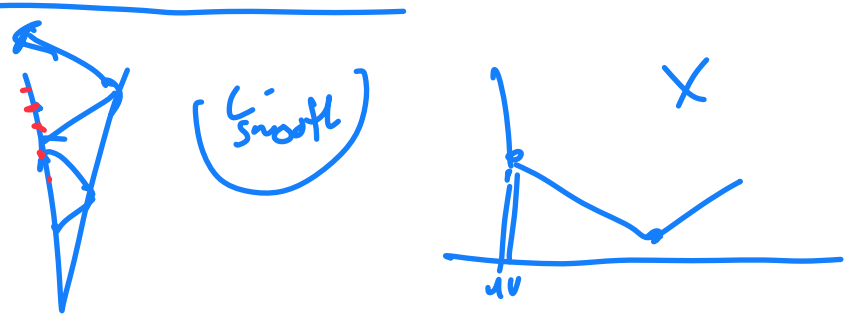
Assume f is C^2, and gradient of f is Lipschitz continuous.

There exists L such that:

For all u, v, $\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|$.

For all w, $\nabla^2 f(w) \preccurlyeq LI$.

For any $u,\ v$, $v^T \nabla^2 f(u) v \leq L\|v\|^2$.



Single-iteration progress bound

For any $u,\ v,\qquad v^T \nabla^2 f(u) v \le L\|v\|^2.$

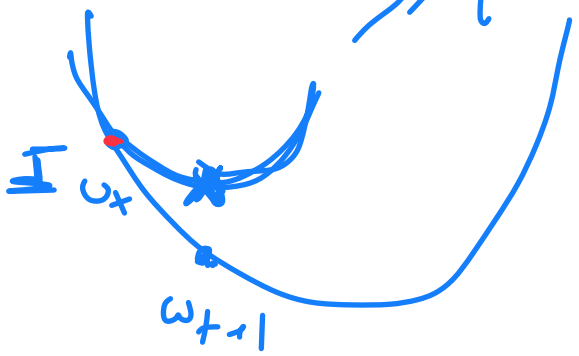
Take Taylor expansion:

$f(w_{t+1}) \le f(w_t)$

$w_t,\ f(w_t),\ \nabla f(w_t),\ f(w_{t+1})$

Recall: 2nd order Taylor expansion around point x_0

$f(x) \le f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$



$f(w_{t+1}) \le$ Taylor expansion

Single-iteration progress bound

For any $u,\ v,$ $\quad v^T \nabla^2 f(u) v \le L\|v\|^2.$ (1)

$u \in [w_{t+1}, w_t]$

$u \in \alpha w_t + (1-\alpha) w_{t+1}$

Take Taylor expansion:

$$f(w_{t+1}) = f(w_t) + (w_{t+1} - w_t)^T \nabla f(w_t) + \tfrac{1}{2}(w_{t+1} - w_t)^T \nabla^2 f(u)(w_{t+1} - w_t)$$

$$\le f(w_t) + (w_{t+1} - w_t)^T \nabla f(w_t) + \frac{L}{2}\|w_{t+1} - w_t\|^2$$

by assumption (1)

Please fill out course feedback!
https://uw.iasystem.org/survey/306248

Let $\Delta = w_{t+1} - w_t$ → the step we take

recall: gradient descent

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Pick Δ to minimize

$$\nabla f(w_t)^T \Delta + \frac{L}{2} \|\Delta\|_2^2$$

→ take deriv, set to zero: $\nabla f(w_t) + L\Delta = 0$

$$\Delta = \frac{-1}{L} \nabla f(w_t)$$

$$w_{t+1} = w_t - \frac{1}{L} \nabla f(w_t)$$

$$\eta = \frac{1}{L}$$ ←

Single-iteration progress bound

$$w_{t+1} = w_t - \frac{1}{L}\nabla f(w_t)$$

→ $\eta = \frac{1}{L}$ G.D.

recall: $w_{t+1} - w_t = -\eta \nabla f(w_t)$ ①

$$f(w_{t+1}) \le f(w_t) + \nabla f(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2}\|w_{t+1} - w_t\|_2^2$$

plug ①

$$\le f(w_t) - \eta \|\nabla f(w_t)\|^2 + \frac{L}{2}\eta^2 \|\nabla f(w_t)\|^2$$

$$f(w_{t+1}) \le f(w_t) - \frac{1}{2L}\|\nabla f(w_t)\|^2$$

→ one step progress bound

⇒ amount of guaranteed progress every step

more progress when? * large gradients

⇒ smooth function (small L)

Single-iteration progress bound

When $\eta = \frac{1}{L}$,

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

for any η in $0 < \eta < \frac{2}{L}$ ✓ valid

still make progress

Same argument shows any $\eta < \frac{2}{L}$ will decrease f.

Convergence analysis steps

1. Study single iteration: $f(w_{t+1})$ vs. $f(w_t)$
2. Piece iterations together to study how they converge

Convergence rate of gradient descent

converge to a minima or saddle point

$$\left\| \nabla f(w_t) \right\| \to 0$$

assume convexity

$$f(w_t) - f(w^*) \to 0$$

the minimum you converge to is the global minimum

Convergence rate of gradient descent

For some ϵ, how many iterations before $\left\| \nabla f(w_t) \right\|^2 \leq \epsilon$?

Assumptions:

- Gradient is Lipschitz continuous (as before)
- Step size is small enough (assume $1/L$)
- f is bounded below by $f(w^*)$

Proof sketch:

- Each iteration decreases f by at least $\frac{1}{2L}\left\| \nabla f(w_t) \right\|^2$
- Can't decrease below $f(w^*)$
- So $\left\| \nabla f(w_t) \right\|^2$ must be decaying fast enough

Convergence rate of gradient descent

$$f(w_{t+1}) \le f(w_t) - \frac{1}{2L}\|\nabla f(w_t)\|_2^2$$

(*) previously shown

$$\|\nabla f(w_t)\|_2^2 \le 2L\left(f(w_t) - f(w_{t+1})\right)$$

$$\sum_{t=0}^{T} \|\nabla f(w_t)\|_2^2 \le 2L \sum_{t=0}^{T} \left[f(w_t) - f(w_{t+1})\right]$$

// telescoping sum

$$= 2L\left[f(w_0) - \cancel{f(w_1)} + \cancel{f(w_1)} - \cancel{f(w_2)} + \dots + \cancel{f(w_{T-1})} + f(w_T)\right]$$

$$= 2L\left[f(w_0) - f(w_T)\right]$$

$$\le 2L\left[f(w_0) - f(w^*)\right]$$

// if you have convexity

← initial gap (depends on w_0)

↓ analyze LHS in next slide

LHS?

$$\sum_{t=0}^{T} \|\nabla f(\omega_t)\|_2^2 \leq 2L\left[f(\omega_0) - f(\omega^*)\right]$$

$$\downarrow$$

$$T \cdot \min_{t=0}^{T} \|\nabla f(\omega_t)\|_2^2 \leq 2L\left[f(\omega_0) - f(\omega^*)\right]$$

$$\rightarrow \quad \min \|\nabla f(\omega_t)\|_2^2 \leq \frac{2L}{T}\left[f(\omega_0) - f(\omega^*)\right] \leq \varepsilon$$

$$\sum_{t=0}^{T} \|\nabla f(\omega_t)\|_2^2 \geq T \cdot \min \|\nabla f(\omega_t)\|_2^2$$

Convergence rate of gradient descent

L-smooth, smoother it is, faster convergence

$$T \geq \frac{2L\left(f(w_0) - f(w^*)\right)}{\epsilon}$$

$f(w_0) - f(w^*)$ → initial gap

ϵ ← closeness to $f(w^*)$

Gradient descent requires

$T = O(1/\epsilon)$ iterations

to achieve $\left\|\nabla f(w_t)\right\|^2 \leq \epsilon$

$\exists C$ for $\epsilon > \bar{\epsilon}$ s.t. $T_\epsilon \leq \frac{C}{\epsilon}$

Convergence rate of gradient descent

$$\left\| \nabla f(w_t) \right\| \leq \epsilon$$

$T_\epsilon = O(1/\epsilon)$

if L-smooth

$$f(w_t) - f(w^*) \leq \epsilon$$

if L-smooth & convex

$T_\epsilon = O(1/\epsilon)$

Strongly convex

$\nabla^2 f \geq \mu I$

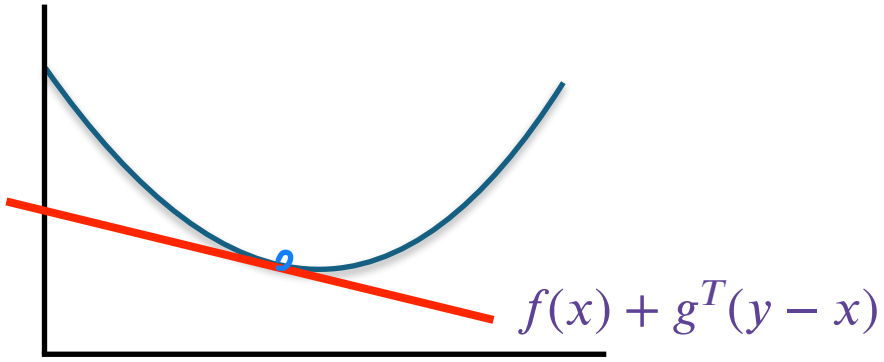
$\hookrightarrow T_\epsilon = O(\log \frac{1}{\epsilon})$

Lasso revisited

Subgradients

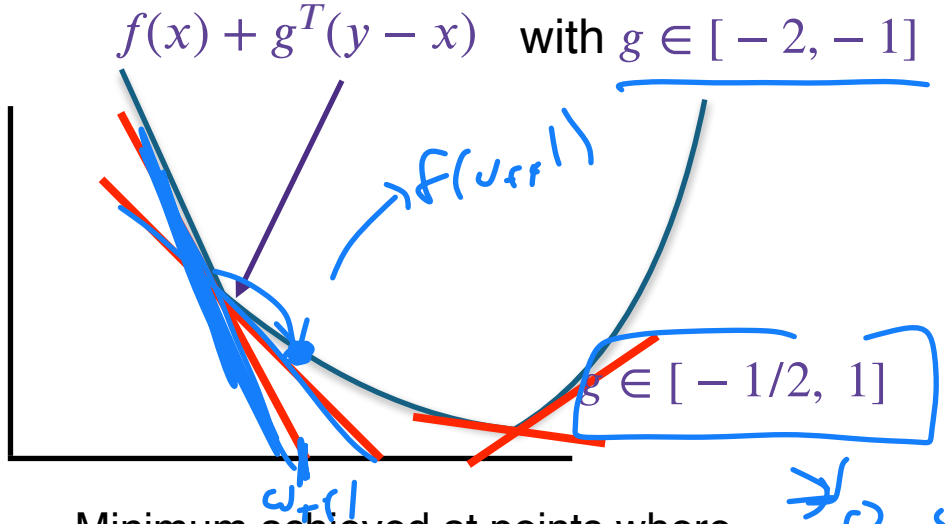
A vector $g \in \mathbb{R}^d$ is a **subgradient** at x if it satisfies
$f(y) \geq f(x) + g^T(y - x)$ for all $y \in \mathbb{R}^d$

Smooth convex function



Gradient is unique sub-gradient
Minimum at points where gradient is 0

Non-smooth convex function

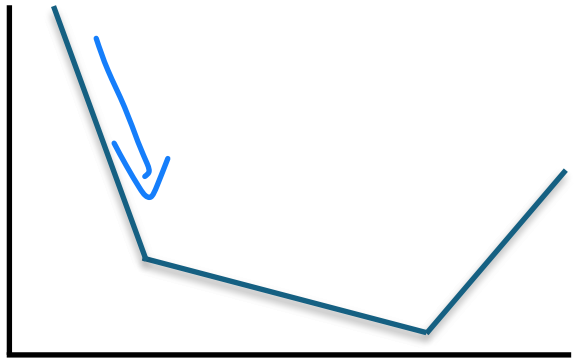


Minimum achieved at points where subgradient set includes 0 vector

Subgradient descent for non-smooth functions

For each t,

Find any subgradient g_t, then set $w_{t+1} \leftarrow w_t - \eta_t g_t$



Works on non-smooth convex functions

Slower compared to smooth convex functions

Gradients don't get smaller near global minima

- Instead of last iterate w_t, keep track of best one
- Step size needs to decrease with t

Stochastic gradient descent (SGD)

$$\hat{w} = \operatorname*{argmin}_w \frac{1}{n} \sum_{i=1}^{n} \ell_i(w)$$

MSE

Gradient descent: $w_{t+1} = w_t - \eta \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell_i(w) \right) \Big|_{w=w_t}$

all examples

Stochastic gradient descent: $w_{t+1} = w_t - \eta \nabla_w \ell_{I_t}(w) \Big|_{w=w_t}$

I_t drawn uniformly at random from $\{1, \ldots, n\}$

1 example at a time

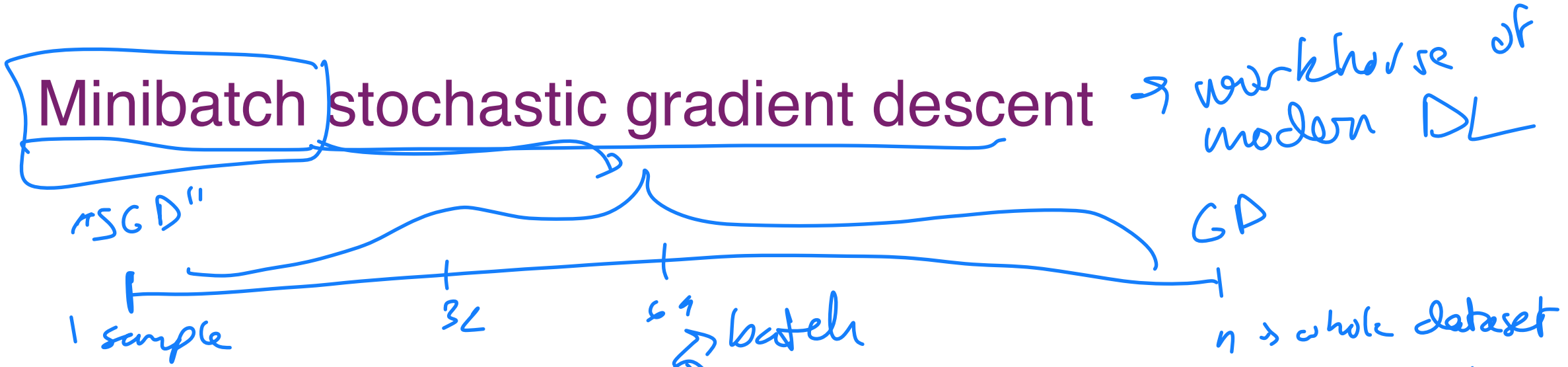
→ n times faster per iteration!

→ in practice parallelizable

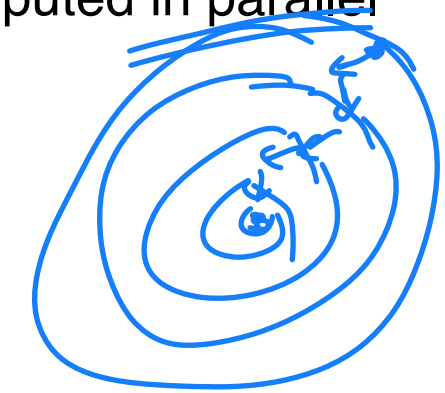
→ And can even be better minimizer.

→ function value decreases faster in terms of gradient computations

Minibatch stochastic gradient descent

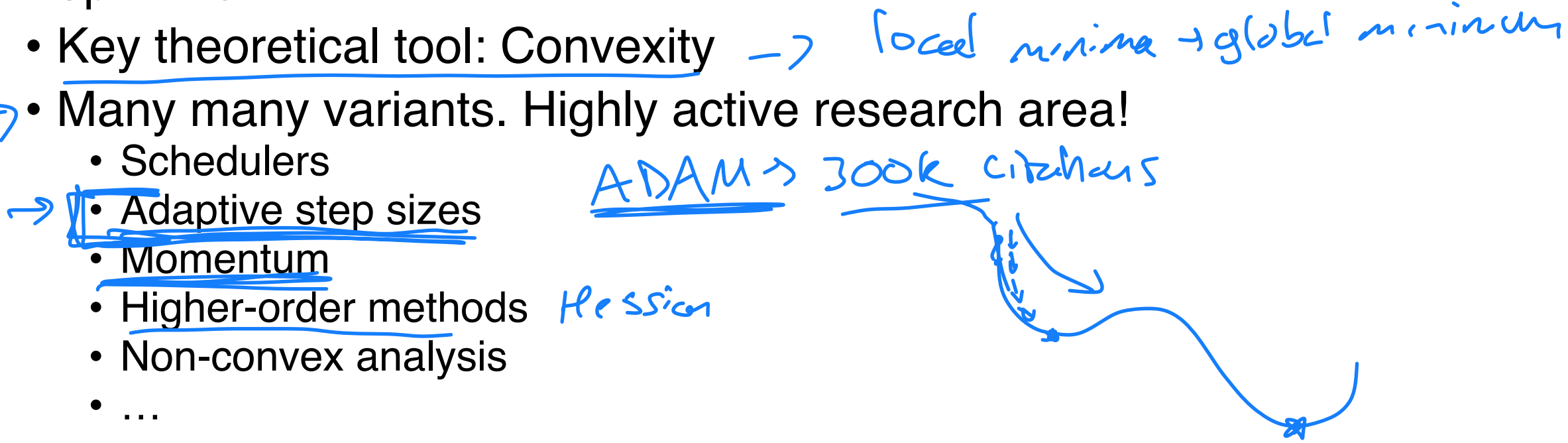


- Instead of one iterate, average B stochastic gradients together
- Advantages:
  - Smaller variance (by $1/B$) (vs 1-sample SGD)
  - Parallelization: Each gradient in the minibatch can be computed in parallel
- This is very widely used!



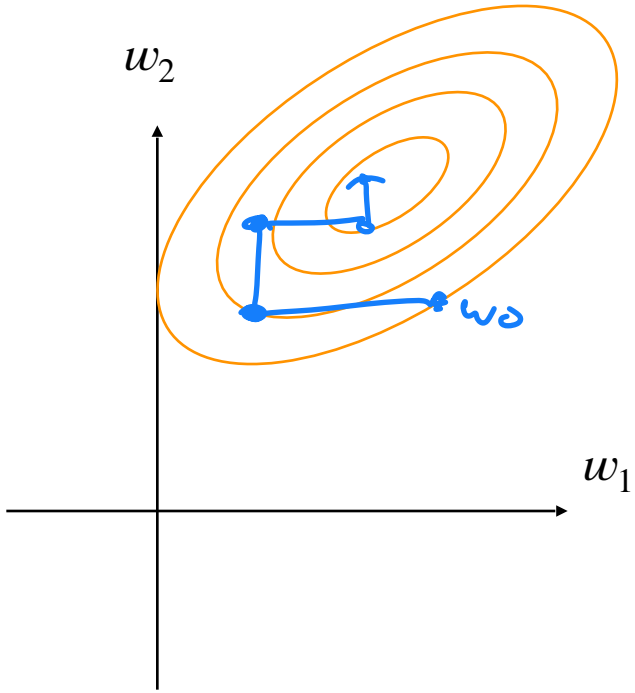
Summary

- Closed form -> iterative methods
- (Minibatch stochastic) gradient descent as a general-purpose optimizer
- Key theoretical tool: Convexity
- Many many variants. Highly active research area!
  - Schedulers
  - Adaptive step sizes
  - Momentum
  - Higher-order methods
  - Non-convex analysis
  - …



Bonus: Coordinate descent

↳ optimize 1 weight (coordinate) at a time
↳ solve 1D optimization each time



eventually reach w^* if convex & differentiable

Coordinate descent for lasso

$$\hat{w} = \underset{w}{\operatorname{argmin}} \ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - x_i^\top w\right)^2 + \lambda\|w\|_1$$

$$\frac{d}{dw_{k_n}} f(w)$$
$$= \sum_{i=1}\left(x_i^\top w - y_i\right)x_{ik} + \lambda\operatorname{sign}\left(w_k\right)$$
$$= \sum_{i=1}^{n}\left(\sum_{j\neq k} x_{ij}w_j + x_{ik}w_k - y_i\right)x_{ik} + \lambda\operatorname{sign}\left(w_k\right)$$
$$= \sum_{i=1}^{n}\left(\sum_{j\neq k} x_{ij}w_j - y_i\right)x_{ik} + w_k\sum_{i=1}^{n}x_{ik} + \lambda\operatorname{sign}\left(w_k\right) \ni 0$$
$$\dots$$

Further reading

- Example gradient descent code on class website
- Boyd and Vandenberghe, Convex Optimization https://stanford.edu/~boyd/cvxbook/
- Mark Schmidt's CPSC 540 notes: https://www.cs.ubc.ca/~schmidtm/Courses/540-W18/L4.pdf
- 3Blue1Brown