CSE 446
Gradient Descent Part II

Natasha Jaques

UNIVERSITY of WASHINGTON

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

When can we guarantee that gradient descent will work?

Convexity

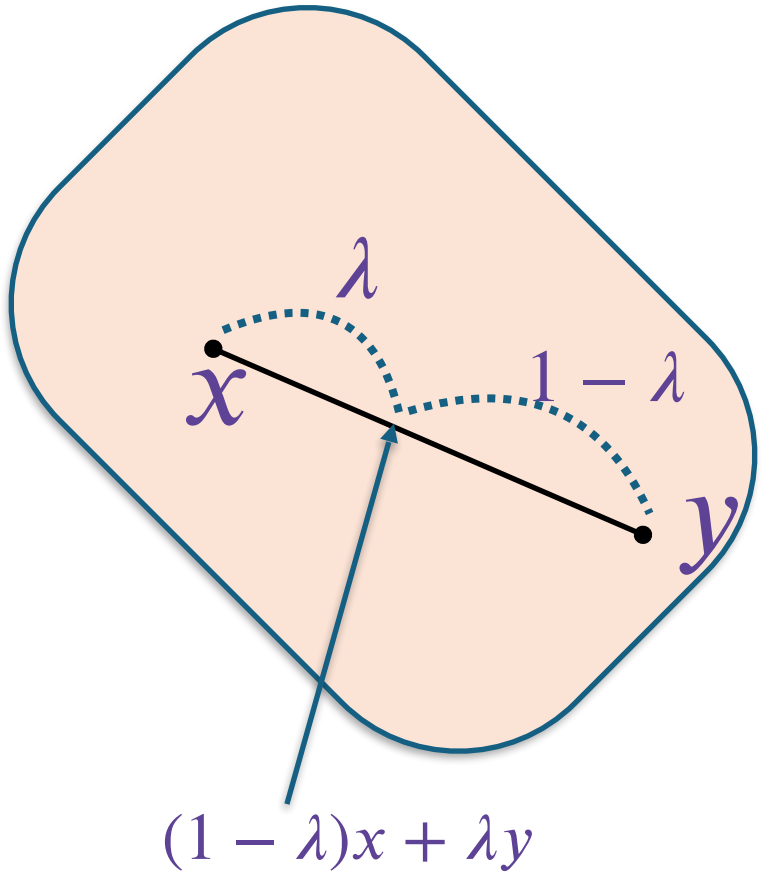
- Optimization problems are hard to solve in general
- The exception: convex optimization
  - Objective is a convex function
  - Constraints are convex sets

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

What is a convex set?

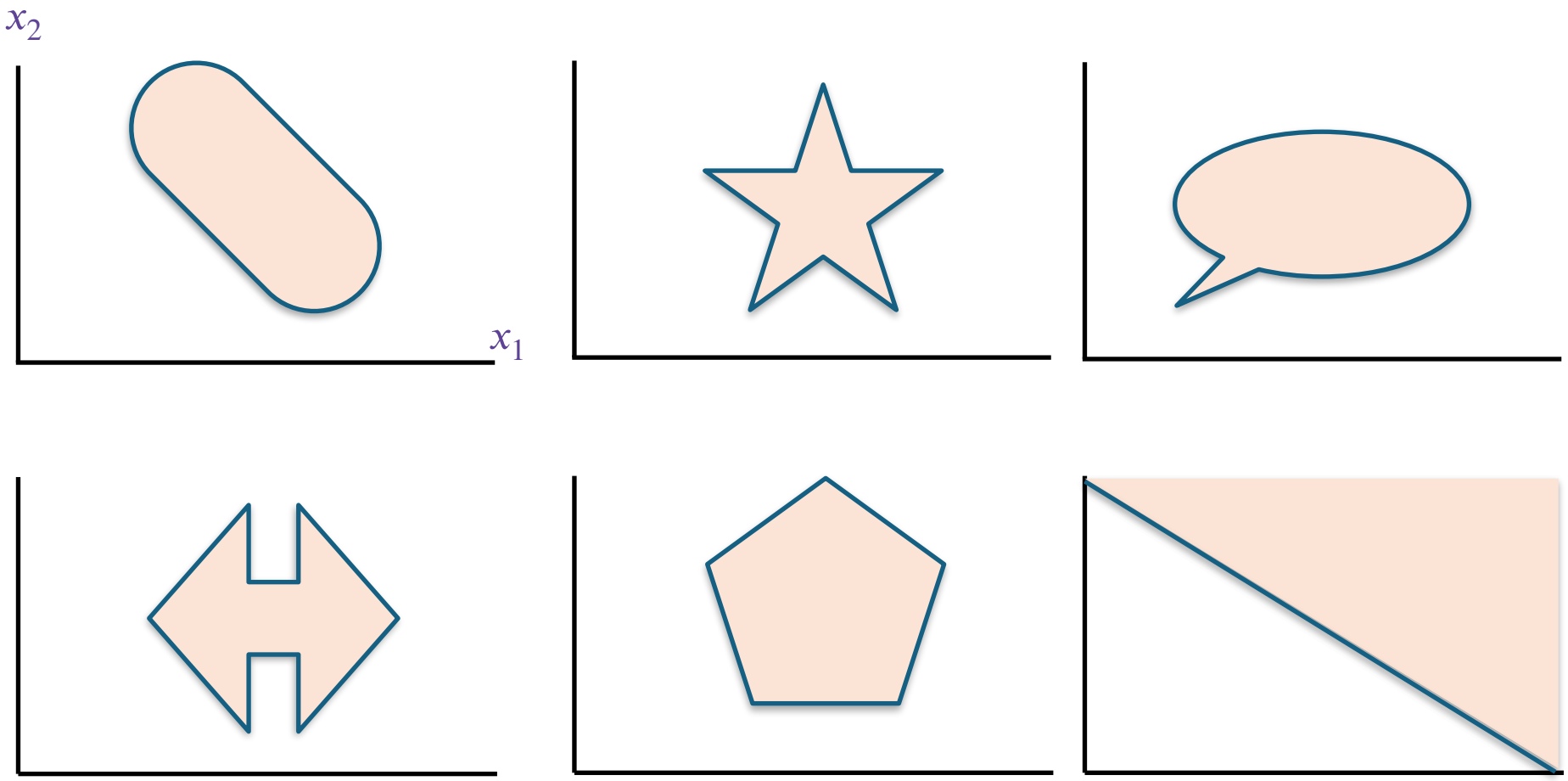
What is a convex set?

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 set?

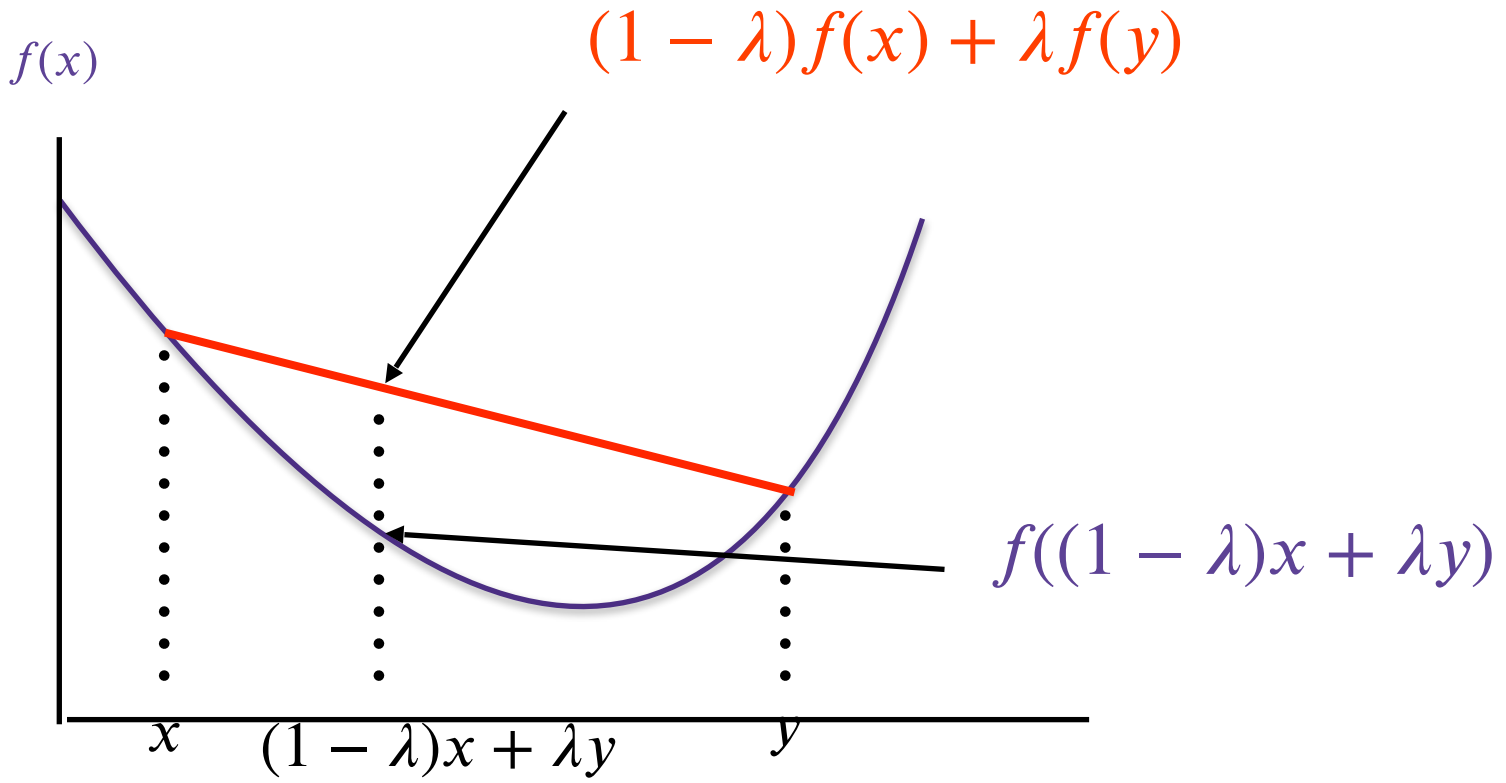
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$



What is a convex function?

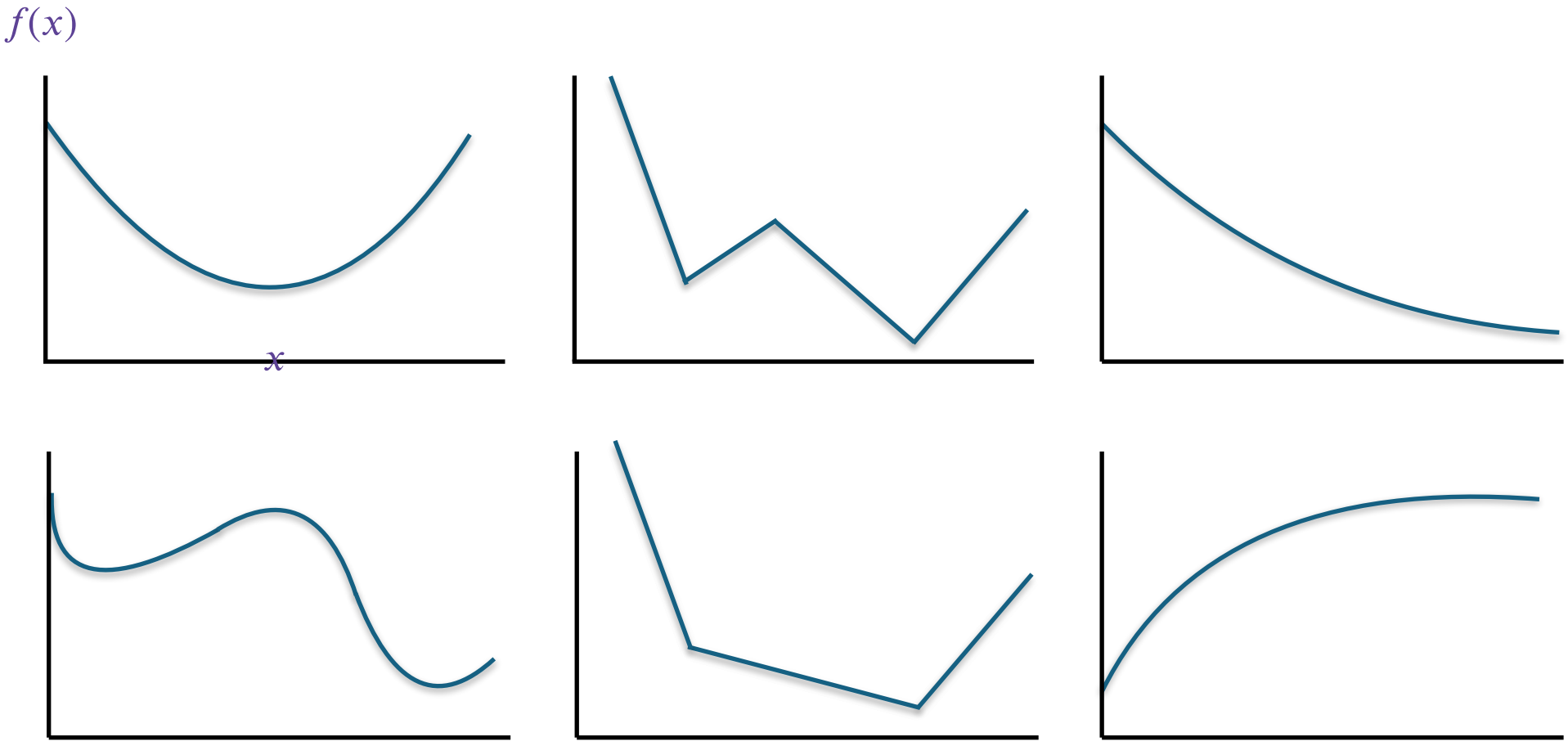
What is a convex function?

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 $\lambda \in [0, 1]$



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 $\lambda \in [0, 1]$



Convex functions and convex sets

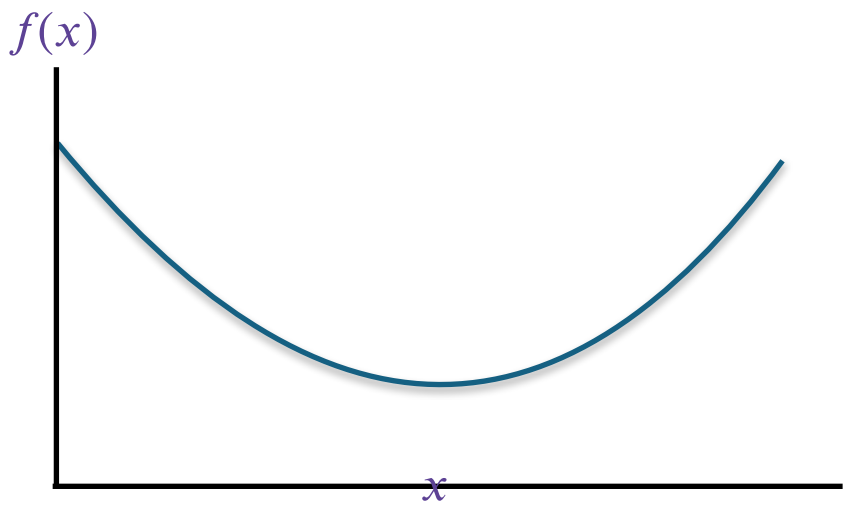
Convex functions and convex sets

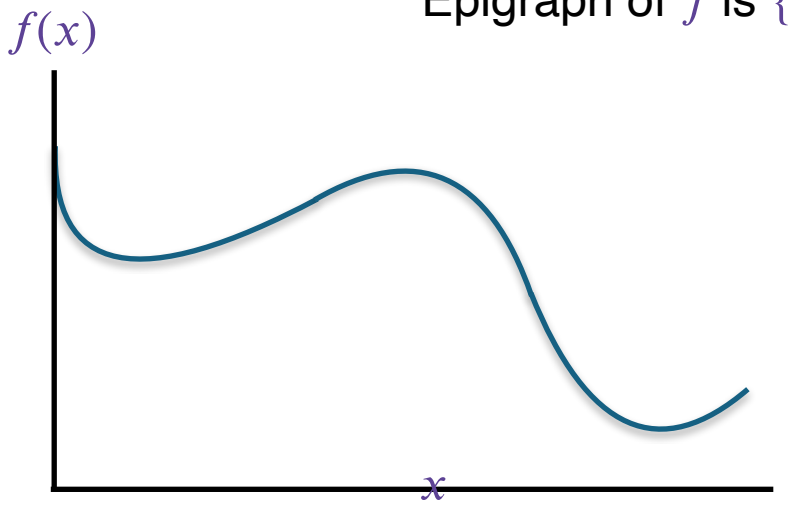
A set $K \subset \mathbb{R}^d$ is convex if $(1-\lambda)x + \lambda y \in K$ for all $x, y \in K$ and $\lambda \in [0,1]$

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

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



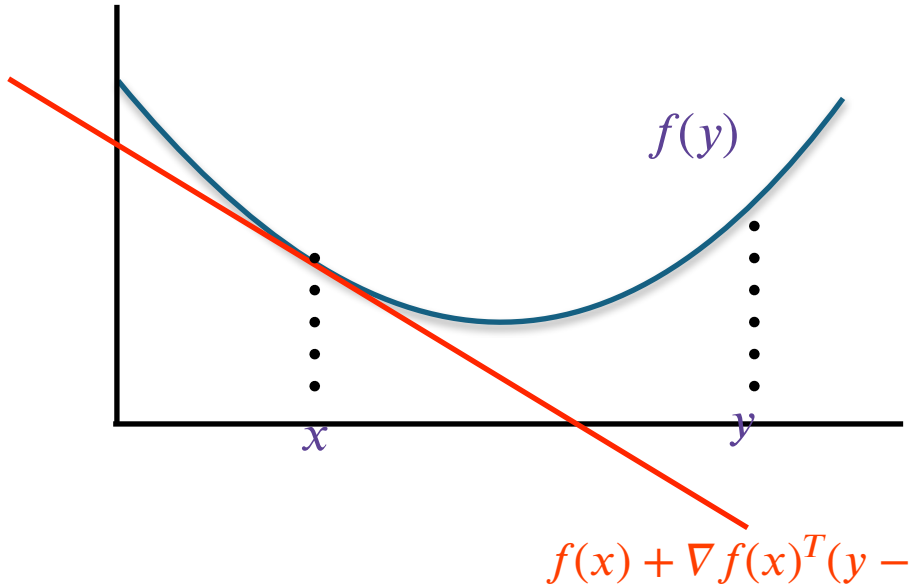


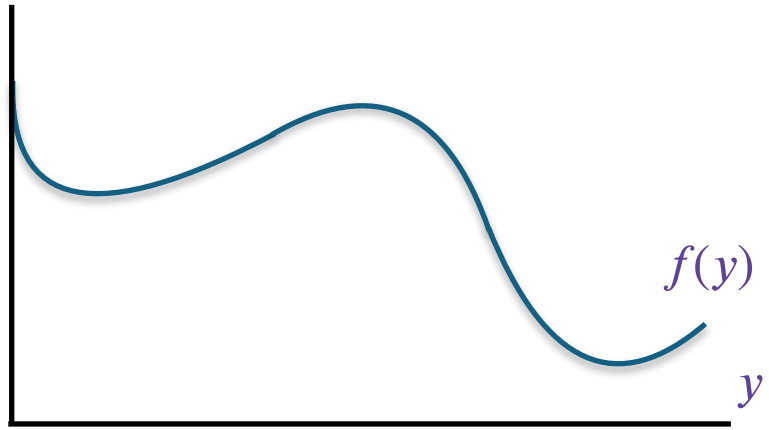
Convexity of differentiable functions

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 $x, y \in dom(f)$

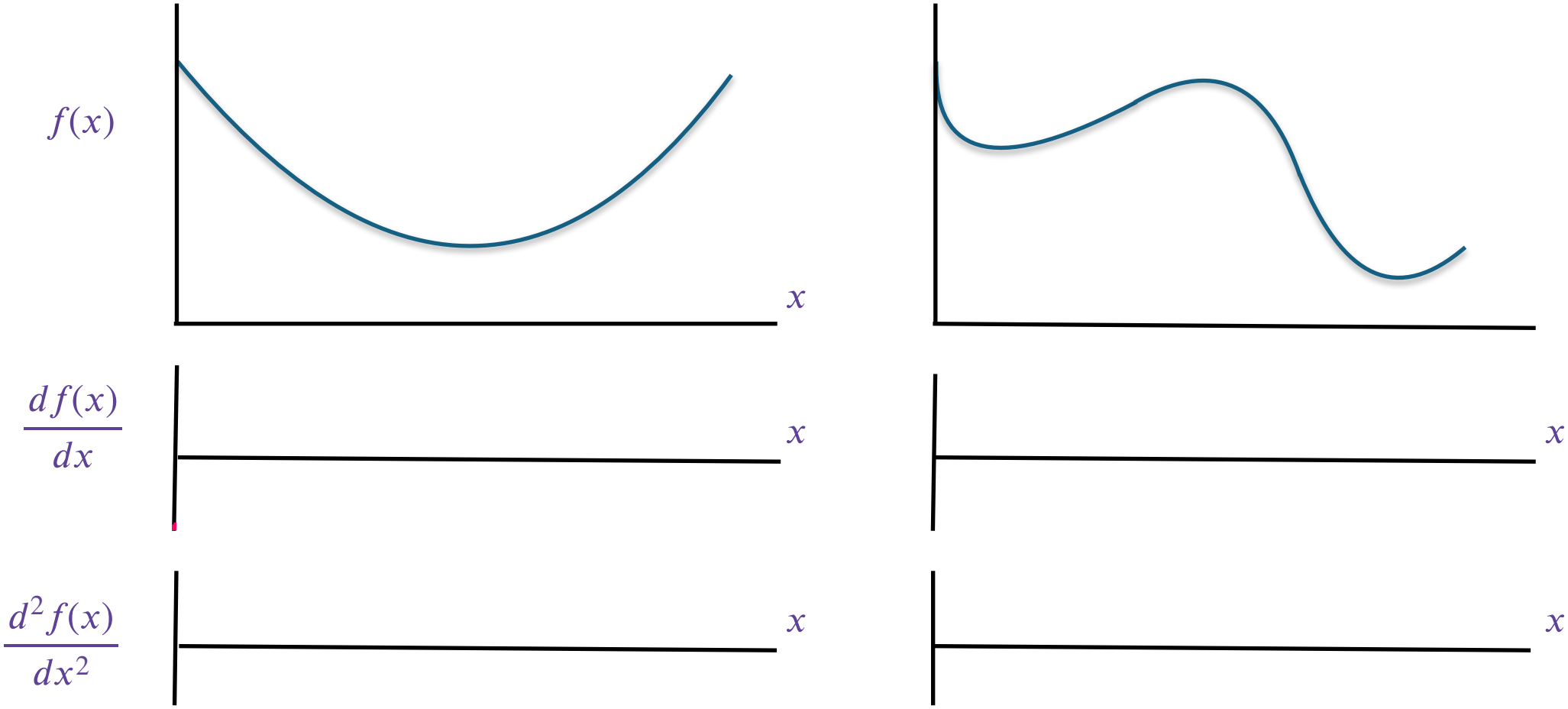




Convexity of twice-differentiable functions

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

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 $x \in dom(f)$

Recap: Definitions of convexity

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

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 $x, y \in dom(f)$

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 $x \in dom(f)$

Example: Ridge regression

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

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

Example: Lasso

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

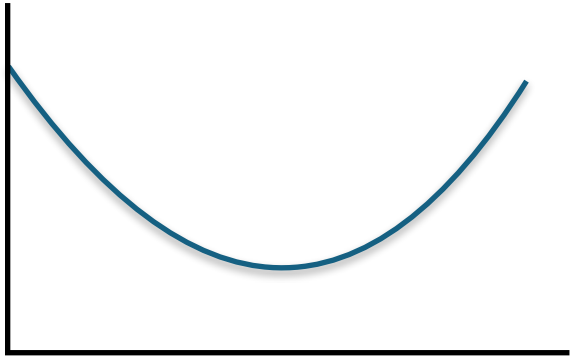
Why not directly solve $\operatorname*{argmin}_{w} \|y - Xw\|_2^2 + \lambda \operatorname{card}(w)$?

Can interpret lasso as convex relaxation of cardinality objective

Convexity and gradient descent

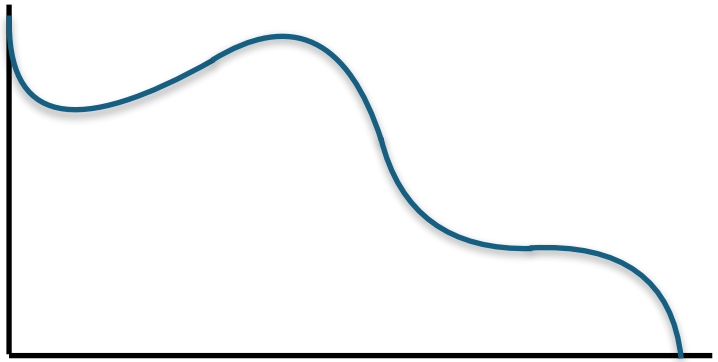
- All local minima are global minima

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, $\left\| \nabla f(u) - \nabla f(v) \right\| \leq L \left\| u - v \right\|$.

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

Take Taylor expansion:

Single-iteration progress bound

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

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

Single-iteration progress bound

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

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

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

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

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

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

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

Convergence rate of gradient descent

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

Gradient descent requires

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

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

Convergence rate of gradient descent

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

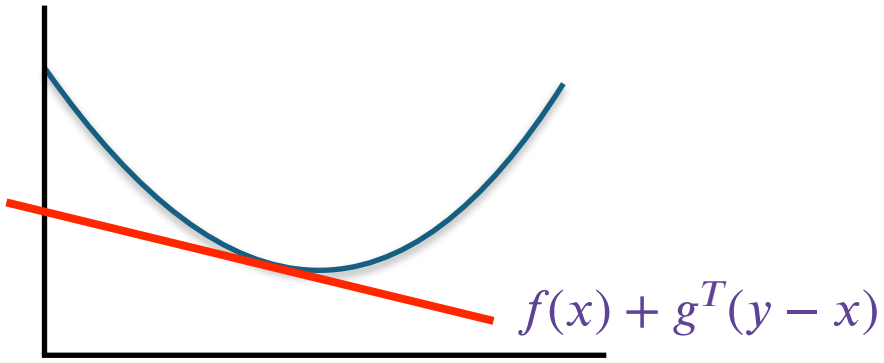
$$f(w_t) - f(w^*) \leq \epsilon$$

Lasso revisited

Subgradients

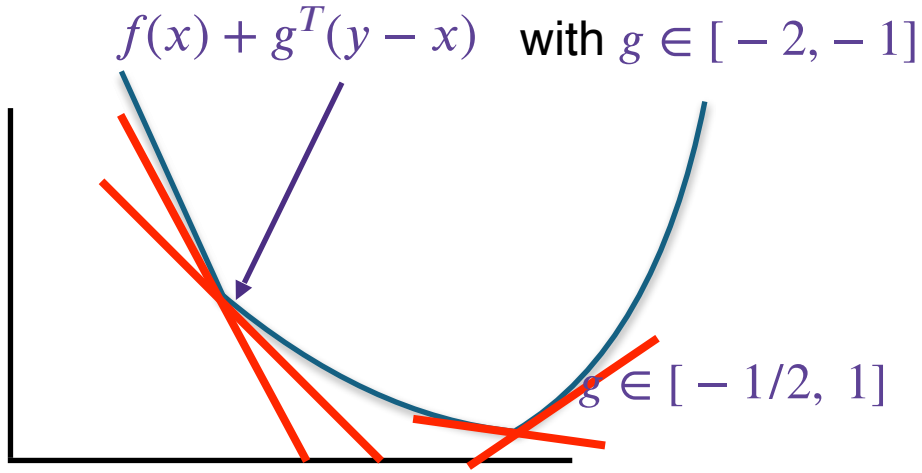
A vector $g \in \mathbb{R}^d$ is a **subgradient** at x if it satisfies
$f(y) \geq f(x) + g^T(y - x)$ for all $y \in \mathbb{R}^d$

Smooth convex function



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

Non-smooth convex function

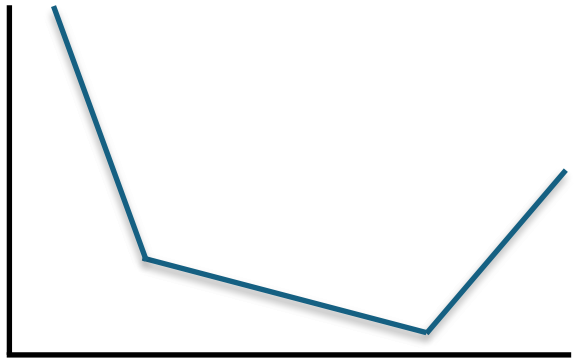


Minimum achieved at points where
subgradient set includes 0 vector

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)

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

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

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 $\{1, \ldots, n\}$

n times faster per iteration!

And can even be better minimizer.

Minibatch stochastic gradient descent

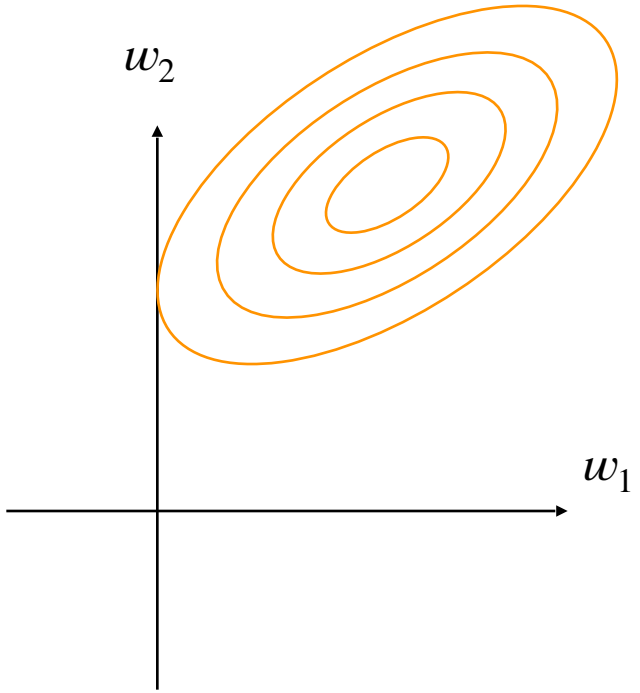
- Instead of one iterate, average B stochastic gradients together
- Advantages:
  - Smaller variance (by $1/B$)
  - Parallelization: Each gradient in the minibatch can be computed in parallel

- 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



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