Application/data

Classified data?

yes: Categorical labels (classification) or quantitative (regression)
no: Similarity (clustering) or positioning (dimensionality reduction)

4 levels of abstraction

Model/method

CS 189

e.g.

- decision function: linear, polynomials, logistic, neural net...
- nearest neighbors, decision trees
- features: edge detectors, embeddings, raw pixels
- low vs. high capacity model: how sinuous/complicated?
  - ↳ affects overfitting/underfitting, inference

Optimization Problem

- variables/parameters, objective function, constraints
  e.g. unconstrained, convex programs, least squares linear regression, PCA

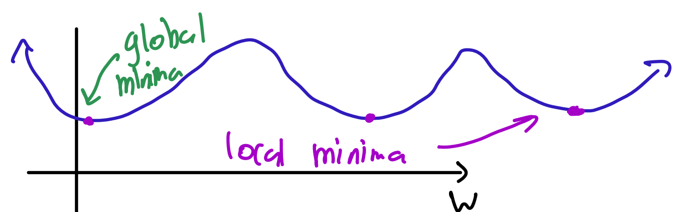
Optimization Algorithms

e.g. gradient descent, SGD, Simplex, SVD (for PCA)

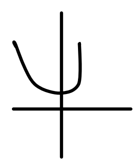
Optimization Problems

Unconstrained: given continuous objective function f, find $\vec{w}$ that minimizes/maximizes f

- f is smooth if $\frac{\partial f}{\partial x}$ is also continuous
- global minimum of f: value w s.t. $f(\vec{w}) \leq f(\vec{v})$ $\forall \vec{v}$
  - ↳ Sometimes need to settle for local minimum: $f(\vec{w}) \leq f(\vec{v})$, $\vec{v}_i \in$ "tiny ball centered @ $\vec{w}$"



Usually, finding local minima is easy, but finding global minima is hard or impossible



↳ exception: Convex functions: $\forall x, y \in \mathbb{R}^d$, line connecting $(x, f(x))$ & $(y, f(y))$ doesn't go below $f(\cdot)$

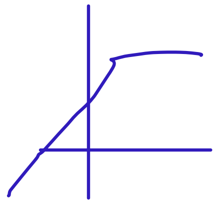
e.g. perceptron risk function is convex & nonsmooth

↳ Convex since it's a sum of convex parts

Consider a continuous, convex function over a closed domain:

3 possibilities:

1) no minima (goes to $-\infty$)



2) just one local minimum

perceptron risk function { 3) connected set of local minima

(that are all global minima w/ same objective values

} walk downhill

↳ risk = 0 in pie slice

Algos for smooth functions:

- gradient descent

$w = w - \varepsilon \nabla f(w)$

  - blind gradient descent: simply steps down, doesn't know how far to go down
  - SGD (also blind)
  - line search: looks ahead & tries to find minimum along the direction

- Newton's method (needs Hessian matrix of f)

  ↳ can be expensive, e.g. for NN's

- Nonlinear conjugate gradient method

  - has line search as subroutine

these algos find local minimum, not global

Algos for nonsmooth f:

- can still do gradient descent

Won't work: Newton's Method (needs 2nd order information)

- BFGS

line search:

repeat { 1) pick direction
2) find local minimum along that dimension/line by solving an optimization problem in 1D }

Some examples: 1) Secant method (smooth only)

2) 2nd derivatives (must be smooth)

3) direct line search

↳ golden section search (nonsmooth functions)

Constrained Optimization (smooth equality constraints)

isosurface
↳ find point on surface s.t. $g(\vec{w}) = 0$

goal: find $\vec{w}$ that minimizes $f(\vec{w})$ s.t. $g(\vec{w}) = 0$, g is smooth

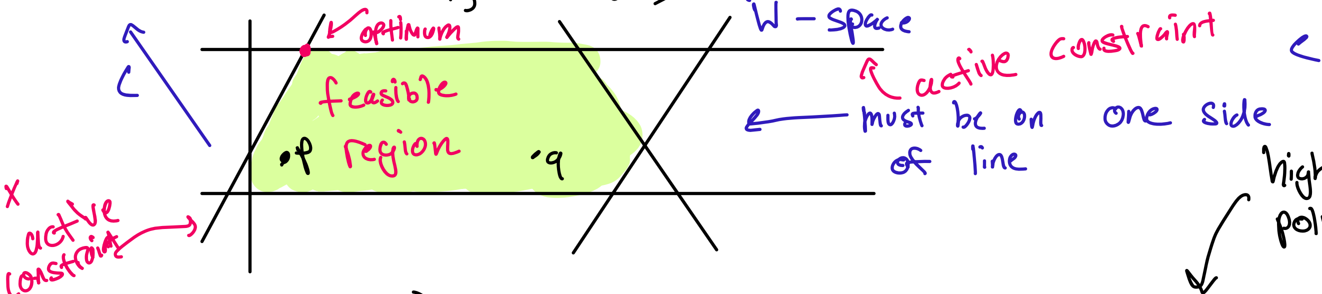
algorithm: lagrange multipliers (constrained smooth optimization problem $\longrightarrow$ smooth unconstrained optimization problem)

Linear Program: linear objective func. & linear inequality constraints

goal: find $\vec{w}$ that maximizes $c \cdot w$ s.t. $A\vec{w} \leq b$ ← component-wise constraints

$A \in \mathbb{R}^{n,d}$, $b \in \mathbb{R}^n \Longrightarrow n$ linear constraints

$\Longrightarrow A_i \cdot w \leq b_i, \; i \in [1, n]$



high-dimensional polyhedron

The set of points $\vec{w}$ that satisfy all constraints is a polytope called the feasible region F

The optimum is point in F that maximizes $c \cdot w$ (furthest in direction of c)

A point set P is convex if for every $p, q \in P$, line connecting p & q is entirely in the point set

The optimum generally achieves equality for some constraints (but not most of them)
↳ active constraints of the optimum

example: every feasible point (w, α) gives a linear classifier)

↳ not necessarily the best in test time

↳ 100% training accuracy

→ find w, α that maximizes 0 s.t. $y_i(x_i \cdot w + \alpha) \geq 1 \quad \forall i \in [1 \ldots n]$

IMPORTANT: linearly separable data iff feasible region $\neq$ empty set

⟶ also true for maximum margin classifier (quadratic program)

Algos for solving linear programs:

- Simplex (George Dantzig, 1947)
  - Walks from vertex to vertex in polytope
- Interior Point methods

Quadratic Program: quadratic, convex objective function

$\Longrightarrow$ hessian is PSD

quadratic term

goal: find $\vec{w}$ that minimizes $f(w) = w^T Q w + c^T w$

Symmetric Q, positive definite

s.t. same linear constraints as linear program

↳ only one local minimum (global minimum)

example: find maximum margin classifier