CPSC 540: Machine Learning

Convex Optimization

Mark Schmidt

University of British Columbia

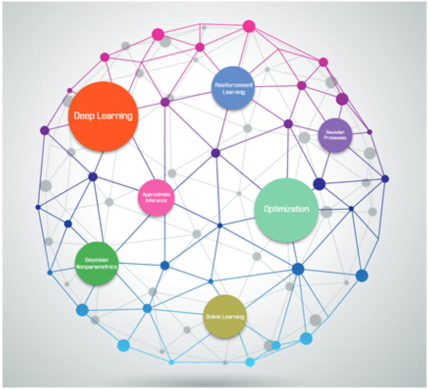
Winter 2018

Admin

- Auditting/registration forms:
  - Submit them at end of class, pick them up end of next class.
  - I need your prereq form before I'll sign registration forms.
  - I wrote comments on the back of some forms.
- Website/Piazza:
  - `https://www.cs.ubc.ca/~schmidtm/Courses/540-W18`.
  - `https://piazza.com/ubc.ca/winterterm22017/cpsc540`.
- Tutorials: start today after class.
- Office hours:
  - With me tomorrow from 3-4 in ICICS 146.
  - With TA Wednesday from 2-3 in DLC Table 4.
- Assignment 1 due Friday.
  - All questions now posted, see Piazza update thread for changes.

Current Hot Topics in Machine Learning

- Graph of most common keywords among ICML papers in 2015:



- Why is there so much focus on deep learning and optimization?

Why Study Optimization in CPSC 540?

- In machine learning, training is typically written as optimization:
  - We numerically optimize parameters w of model, given data.
- There are some exceptions:
  1. Methods based on counting and distances (KNN, random forests).
     - See CPSC 340.
  2. Methods based on averaging and integration (Bayesian learning).
     - Later in course.

  But even these models have parameters to optimize.
- But why study optimization? Can't I just use optimization libraries?
  - "\", linprog, quadprog, CVX, MOSEK, and so.

The Effect of Big Data and Big Models

- Datasets are getting huge, we might want to train on:
  - Entire medical image databases.
  - Every webpage on the internet.
  - Every product on Amazon.
  - Every rating on Netflix.
  - All flight data in history.
- With bigger datasets, we can build bigger models:
  - Complicated models can address complicated problems.
  - Regularized linear models on huge datasets are standard industry tool.
  - Deep learning allows us to learn features from huge datasets.

- But optimization becomes a bottleneck because of time/memory.
  - We can't afford $O(d^2)$ memory, or an $O(d^2)$ operation.
  - Going through huge datasets hundreds of times is too slow.
  - Evaluating huge models many times may be too slow.
- Next class we'll start large-scale machine learning.
  - But first we'll show how to use some "off the shelf" optimization methods.

Robust Regression in Matrix Notation

- Regression with the absolute error as the loss,

$$\mathop{\mathrm{argmin}}_{w \in \mathbb{R}^d} \sum_{i=1}^{n} |w^T x^i - y^i|.$$

- In CPSC 340 we argued that this is more robust to outliers than least squares.

- This objective is not quadratic, but can be minimized as a linear program.
  - Linear program: "minimizing a linear function with linear constraints".

$$\mathop{\mathrm{argmin}}_{w} w^T c, \quad \text{where } w \text{ satifiies constratins like } \quad w^T a_i \leq b_i.$$

- Our first step is re-writing absolute value using $|\alpha| = \max\{\alpha, -\alpha\}$,

$$\mathop{\mathrm{argmin}}_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \max\{w^T x^i - y^i, y^i - w^T x^i\}.$$

Robust Regression as a Linear Program

- So we've show that L1-regression is equivalent to

$$\underset{w\in\mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} \max\{w^Tx^i - y^i, y^i - w^Tx^i\}.$$

- Second step: introduce n variables r_i that upper bound the max functions,

$$\underset{w\in\mathbb{R}^d, r\in\mathbb{R}^n}{\text{argmin}} \sum_{i=1}^{n} r_i, \quad \text{with} \quad r_i \geq \max\{w^Tx^i - y^i, y^i - w^Tx^i\}, \forall i.$$

- This is a linear objective (in w and r) with non-linear constraints.
  - Note that we have $r_i = |w^Tx^i - y^i|$ at the solution.
  - Otherwise, either the constraints are violated or we could decrese r_i.
- To convert to a linear program, we need to convert to linear constraints.
  - Third step: split max constraints into individual linear constraints,

$$\underset{w\in\mathbb{R}^d,\, r\in\mathbb{R}^n}{\text{argmin}} \sum_{i=1}^{n} r_i, \quad \text{with} \quad r_i \geq w^Tx^i - y^i,\ r_i \geq y^i - w^Tx^i, \forall i.$$

Minimizing Absolute Values and Maxes

- We've shown that L1-norm regression can be written as a linear program,

$$\underset{w\in\mathbb{R}^d,\, r\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^n r_i, \quad \text{with} \quad r_i \geq w^T x^i - y^i,\ r_i \geq y^i - w^T x^i, \forall i,$$

- For medium-sized problems, we can solve this with Julia's *linprog*.
  - Linear programs are solvable in polynomial time.

- A general approach for minimizing absolute values and/or maximums:
  1. Replace absolute values with maximums.
  2. Replace maximums with new variables, constrain these to bound maixmums.
  3. Transform to linear constraints by splitting the maximum constraints.

Example: Support Vector Machine as a Quadratic Program

- The SVM optimization problem is

$$\operatorname*{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^n \max\{0, 1 - y^i w^T x^i\} + \frac{\lambda}{2}\|w\|^2,$$

- Introduce new variables to upper-bound the maxes,

$$\operatorname*{argmin}_{w \in \mathbb{R}^d, r \in \mathbb{R}^n} \sum_{i=1}^n r_i + \frac{\lambda}{2}\|w\|^2, \quad \text{with} \quad r_i \geq \max\{0, 1 - y^i w^T x^i\}, \forall i.$$

- Split the maxes into separate constraints,

$$\operatorname*{argmin}_{w \in \mathbb{R}^d, r \in \mathbb{R}^n} \sum_{i=1}^n r_i + \frac{\lambda}{2}\|w\|^2, \quad \text{with} \quad r_i \geq 0, \; r_i \geq 1 - y^i w^T x^i,$$

which is a quadratic program (quadratic objective with linear constraints).

Outline

General Lp-norm Losses

- Consider minimizing the regression loss

$$f(w) = \|Xw - y\|_p,$$

with a general Lp-norm, $\|r\|_p = \left(\sum_{i=1}^n |r_i|^p\right)^{\frac{1}{p}}$.

- With $p = 2$, we can minimize the function using linear algebra.
    - Squaring it gives least squares.
- With $p = 1$, we can minimize the function using linear programming.
- With $p = \infty$, we can also use linear programming.
- For $2 < p < \infty$, we can use gradient descent (next lecture).
    - Raise it to the power p to get a smooth problem.
- For $1 < p < 2$, there off-the-shelf methods to solve the problem.
- If we use $p < 1$ (which is not a norm), minimizing f is NP-hard.

Convex Optimization

- With $p \geq 1$ the problem is convex, while with $p < 1$ the problem is non-convex.
- A convex optimization problem can be written in the form

$$\min_{w \in \mathcal{C}} f(w),$$

where $\mathcal{C}$ is a convex set and f is a convex function.

- Convexity is usually a good indicator of tractability:
    - Minimizing convex functions is usually easy.
    - Minimizing non-convex functions is usually hard.
- Off-the-shelf software minimizes solves many convex problems (*MathProgBase*).

Convex Combinations and Differentiability Classes

- To define convex sets and functions, we use notion of convex combination:
  - A convex combination of two variables w and v is given by

$$\theta w + (1-\theta)v \quad \text{for any} \quad 0 \leq \theta \leq 1.$$

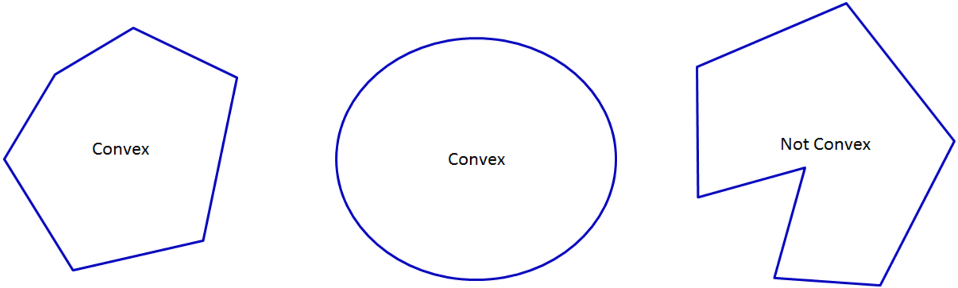
  - A convex combination of k variables $\{w_1, w_2, \dots, w_k\}$ is given by

$$\sum_{c=1}^{k} \theta_c w_c \quad \text{where} \quad \sum_{c=1}^{k} \theta_c = 1,\ \theta_c \geq 0.$$

- We're also going to use the notion of differentiability classes:
  - C^0 is the set of continuous functions.
  - C^1 is the set of continuous functions with continuous first-derivatives.
  - C^2 is the set of continuous functions with continuous first- and second-derivatives.

Convex Sets

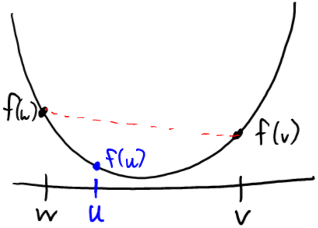
- A set $\mathcal{C}$ is convex if convex combinations of points in the set are also in the set.



- For all $w \in \mathcal{C}$ and $v \in \mathcal{C}$ we have $\underbrace{\theta w + (1-\theta)v}_{\text{convex comb}} \in \mathcal{C}$ for $0 \leq \theta \leq 1$.
- A trivial example is that $\mathbb{R}^d$ is convex.

Convex Functions

- A function f is convex if the area above the function is a convex set.
  - And its domain is convex.



- Equivalently, the function is always below the "chord" between two points.

$$f(\underbrace{\theta w + (1-\theta)v}_{\text{convex comb}}) \leq \underbrace{\theta f(w) + (1-\theta)f(v)}_{\text{"chord"}}, \quad \text{for all } w \in \mathcal{C}, v \in \mathcal{C}, 0 \leq \theta \leq 1.$$

- Extremely-useful property: all local minima of convex functions are global minima.
  - Indeed, $\nabla f(w) = 0$ means w is a global minima.

One-Dimensional Convex Functions

- A 1-variable twice-differentiable (C^2) function is convex iff $f''(w) \geq 0$ for all w.
- Examples:
  - Quadratic $w^2 + bw + c$ with $a \geq 0$.
  - Linear: $aw + b$.
  - Constant: b.
  - Exponential: $\exp(aw)$.
  - Negative logarithm: $-\log(w)$.
  - Negative entropy: $w \log w$, for $w > 0$.
  - Logistic loss: $\log(1 + \exp(-w))$.

Convexity of Norms

- All norms are convex:
  - If $f(w) = \|w\|_p$ for a generic norm, then we have

$$\begin{aligned} f(\theta w + (1-\theta)v) &= \|\theta w + (1-\theta)v\|_p \\ &\leq \|\theta w\|_p + \|(1-\theta)v\|_p && \text{(triangle inequality)} \\ &= |\theta| \cdot \|w\|_p + |1-\theta| \cdot \|v\|_p && \text{(absolute homogeneity)} \\ &= \theta\|w\|_p + (1-\theta)\|v\|_p && (0 \leq \theta \leq 1) \\ &= \theta f(w) + (1-\theta)f(v), && \text{(definition of } f\text{)} \end{aligned}$$

    so f is always below the "chord".
  - See course webpage notes on norms if the above steps aren't familiar.
- In addition, all squared norms are convex.
  - These are all convex: $|w|, \|w\|, \|w\|_1, \|w\|^2, \|w_1\|^2, \|w\|_\infty$,...

Operations that Preserve Convexity

- There are a few operations that preserve convexity.
  - Can show convexity by writing as sequence of convexity-preserving operations.

- If f and g are convex functions, the following preserve convexity:
  1. Non-negative scaling: $$h(w) = \alpha f(w).$$
  2. Sum: $$h(w) = f(w) + g(w).$$
  3. Maximum: $$h(w) = \max\{f(w), g(w)\}.$$
  4. Composition with affine map: $$h(w) = f(Aw + b),$$

     where an affine map $w \mapsto Aw + b$ is a multi-input multi-output linear function.
     - Like $g(w) = Aw + b$ which takes in a vector and outputs a vector.

- But note that composition $f(g(w))$ of convex f and g is not convex in general.

Convexity of SVMs

- If f and g are convex functions, the following preserve convexity:
  1. Non-negative scaling.
  2. Sum.
  3. Maximum.
  4. Composition with affine map.
- We can use these to quickly show that SVMs are convex,

$$f(w) = \sum_{i=1}^{n} \max\{0, 1 - y^i w^T x^i\} + \frac{\lambda}{2}\|w\|^2.$$

- Second term is squared norm multiplied by non-negative $\frac{\lambda}{2}$.
  - Squared norms are convex, and non-negative scaling perserves convexity.
- First term is sum(max(linear)). Linear is convex and sum/max preserve convexity.
- Since both terms are convex, and sums preserve convexity, SVMs are convex.

Convex Sets from Functions

- We often have constraints on our variables w.
  - How do we know if these constraints define a convex set?
- Consider the "sublevel set" of a convex function g,

$$\mathcal{C} = \{w \mid g(w) \leq \tau\},$$

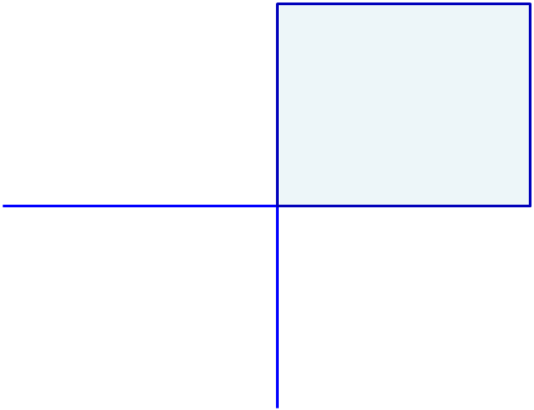
for some number τ.
- If g is a convex function, then $\mathcal{C}$ is a convex set.
  - This follows from the definitions:

$$g(\underbrace{\theta w + (1-\theta)v}_{\text{convex comb}}) \leq \underbrace{\theta g(w) + (1-\theta)g(v)}_{\text{by convexity}} \leq \underbrace{\theta\tau + (1-\theta)\tau}_{\text{definition of } g} = \tau.$$

- Example:
  - The set of w where $w^2 \leq 10$ forms a convex set by convexity of w^2, $[-\sqrt{10}, \sqrt{10}]$.

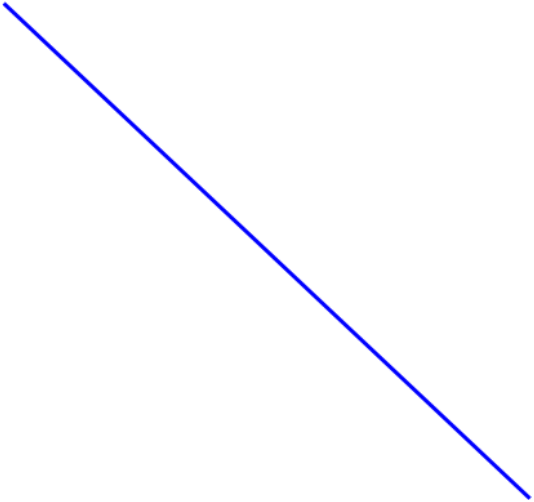
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



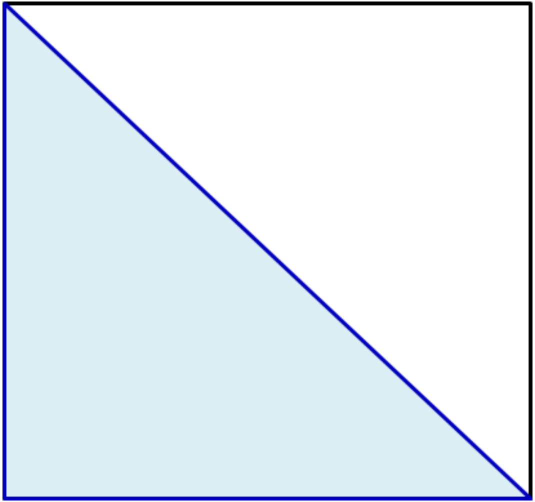
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



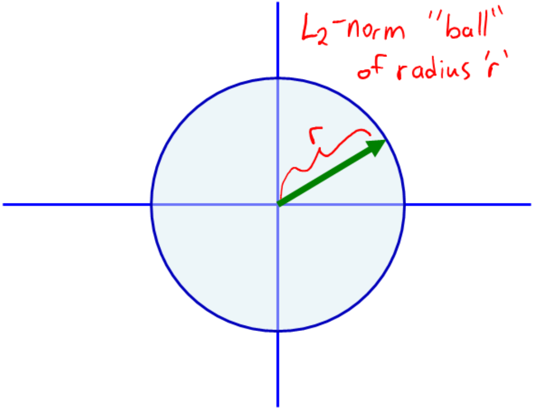
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



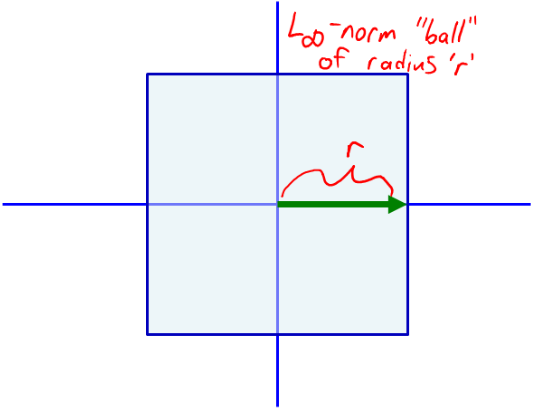
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



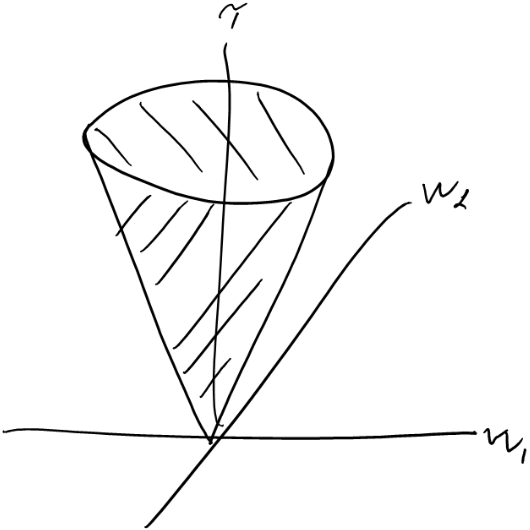
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



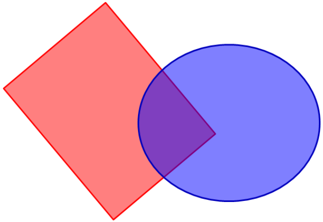
Examples of Simple Convex Sets

- Real space $\mathbb{R}^d$.
- Positive orthant $\mathbb{R}^d_+ : \{w \mid w \geq 0\}$.
- Hyper-plane: $\{w \mid a^T w = b\}$.
- Half-space: $\{w \mid a^T w \leq b\}$.
- Norm-ball: $\{w \mid \|w\|_p \leq \tau\}$.
- Norm-cone: $\{(w, \tau) \mid \|w\|_p \leq \tau\}$.



Showing a Set is Convex from Intersections

- The intersection of convex sets is convex.
  - Proof is trivial: convex combinations in the intersection are in the intersection.



- We can prove convexity of a set by showing it's an intersection of convex sets.

- Example: the w satisfying any number of linear constraints forms a convex set:

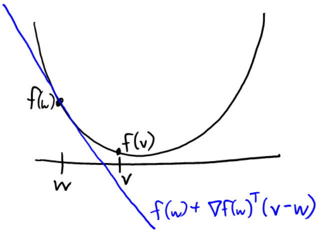
$$d \leq Aw \leq b$$
$$LB \leq w \leq UB.$$

Differentiable Convex Functions

- Convex functions must be continuous, and have a domain that is a convex set.
  - But they may be non-differentiable.

- For differentiable convex functions, there is a third equivalent definiton:
  - A *differentiable* f is convex iff f is always above tangent.

$$f(v) \geq f(w) + \nabla f(w)^T(v - w), \quad \forall w \in \mathcal{C}, v \in \mathcal{C}.$$



- Notice that $\nabla f(w) = 0$ implies $f(v) \geq f(w)$ for all v, so w is a global minimizer.

Convexity of Twice-Differentiable Functions

- For C^2 functions, there is an equivalent definition of convexity.
- It requires defining the Hessian matrix, $\nabla^2 f(w)$.
  - The matrix of second partial derivaitves,

$$\nabla^2 f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1 \partial w_1} f(w) & \frac{\partial}{\partial w_1 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_1 \partial w_d} f(w) \\ \frac{\partial}{\partial w_2 \partial w_1} f(w) & \frac{\partial}{\partial w_2 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_2 \partial w_d} f(w) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_d \partial w_1} f(w) & \frac{\partial}{\partial w_d \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_d \partial w_d} f(w) \end{bmatrix}$$

- In the case of least squares, we can write the Hessian as

$$\nabla^2 f(w) = X^T X,$$

see course webpage notes on the gradients/Hessians of linear/quadratic functions.

Convexity of Twice-Differentiable Functions

- A multivariate C^2 function is convex iff:

$$\nabla^2 f(w) \succeq 0,$$

  for all w.
- This notation $A \succeq 0$ means that A is positive semidefinite.
- This condition means the function is flat or "curved upwards" in *every* direction.

- Two equivalent definitions of a positive semidefinite matrix A:
  1. All eigenvalues of A are non-negative.
  2. The quadratic $v^T A v$ is non-negative for all vectors v.

Convexity and Least Squares

- We can use twice-differentiable condition to show convexity of least squares,

$$f(w) = \frac{1}{2}\|Xw - y\|^2.$$

- The Hessian of this objective is given by

$$\nabla^2 f(w) = X^T X.$$

- So we want to show that $X^T X \succeq 0$ or equivalently that $v^T X^T X v \geq 0$ for all v.
- We can show this by non-negativity of norms,

$$v^T X^T X v = \underbrace{(Xv)^T (Xv)}_{u^T u} = \underbrace{\|Xv\|^2}_{\|u\|^2} \geq 0,$$

so least squares is convex and solving $\nabla f(w) = 0$ gives *global minimum*.

Strict Convexity and Positive-Definite Matrices

- We say that a C^2 function is strictly convex iff for all w we have

$$\nabla^2 f(w) \succ 0,$$

meaning that the Hessian is positive definite everywhere.

- Equivalent definitions of a positive definite matrix A:
  1. The eigevalues of A are all positive.
  2. $v^T A v > 0$ for all $v \neq 0$.

- Why do we care about strict convexity?
  - Positive-definite matrices are invertible, so $[\nabla^2 f(w)]^{-1}$ exists.
  - There can be at most one global optimum (so it's unique, if one exists).

Strict Convexity and L2-Regularized Least Squares

- In L2-regularized least squares, the Hessian matrix is

$$\nabla^2 f(w) = (X^T X + \lambda I).$$

- This matrix is positive-definite.

$$v^T(X^T X + \lambda I)v = \underbrace{\|Xv\|^2}_{\geq 0} + \underbrace{\lambda\|v\|^2}_{>0} > 0,$$

  which follows from properties of norms:
  - Both terms are non-negative because they're norms.
  - Second term $\|v\|$ is positive because $v \neq 0$ and $\lambda > 0$.

- This implies that:
  - The solution is unique.
  - The matrix $(X^T X + \lambda I)$ is invertible.

Summary

- Converting non-smooth problems involving max to constrained smooth problems.
- Convex optimization problems are a class that we can usually efficiently solve.
- Showing functions and sets are convex.
  - Either from definitions or convexity-preserving operations.
- C^2 definition of convex functions that the Hessian is positive semidefinite.

- How many iterations of gradient descent do we need?

Showing a Set is Convex from Defintion

- We can prove convexity of a set from the definition:
  - Choose a generic w and v in $\mathcal{C}$, show that generic u between them is in the set.

- Hyper-plane example: $\mathcal{C} = \{w \mid a^T w = b\}$.
  - If $w \in \mathcal{C}$ and $v \in \mathcal{C}$, then we have $a^T w = b$ and $a^T v = b$.
  - To show $\mathcal{C}$ is convex, we can show that $a^T u = b$ for u between w and v.

$$\begin{aligned} a^T u &= a^T(\theta w + (1-\theta)v) \\ &= \theta(a^T w) + (1-\theta)(a^T v) \\ &= \theta b + (1-\theta)b = b. \end{aligned}$$

- Alternately, you could use that linear functions $a^T w$ are convex, and $\mathcal{C}$ is the intersection of $\{w \mid a^T w \leq b\}$ and $\{w \mid a^T w \geq b\}$.

Strictly-Convex Functions

- A function is strictly-convex if the convexity definitions hold strictly:

$$f(\theta w + (1-\theta)v) < \theta f(w) + (1-\theta)f(v), \quad 0 < \theta < 1 \tag{C^0}$$

$$f(v) > f(w) + \nabla f(w)^T(v - w) \tag{C^1}$$

$$\nabla^2 f(w) \succ 0 \tag{C^2}$$

- Function is always strictly below any chord, strictly above any tangent, and curved upwards in every direction.

- Strictly-convex function have at most one global minimum:
  - w and v can't both be global minima if $w \neq v$:
    it would imply convex combinations u of w and v would have $f(u)$ below the global minimum.

More Examples of Convex Functions

- Examples of more exotic convex sets over matrix variables:
  - The set of positive semidefinite matrices $\{W \mid W \succeq 0\}$.
  - The set of positive definite matrices $\{W \mid W \succ 0\}$.

- Some more exotic examples of convex functions:
  - $f(w) = \log(\sum_{j=1}^{d} \exp(w_j))$ (log-sum-exp function).
  - $f(W) = \log \det W$ for $W \succ 0$ (log-determinant).
  - $f(W, v) = v^T W^{-1} v$ for $W \succ 0$.