CPSC 440: Advanced Machine Learning

Convex Optimization

Mark Schmidt

University of British Columbia

Winter 2021

Last Time: Convex Optimization

- In machine learning we often need to solve convex optimization problems,

$$\operatorname*{argmin}_{w \in \mathcal{C}} f(w),$$

where f is a convex function and $\mathcal{C}$ is a convex set.
  - Key property: all local optima are global optima.

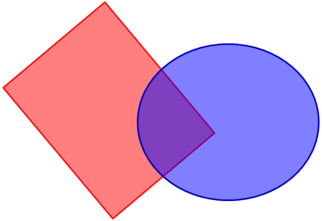
- We say set $\mathcal{C}$ is convex if convex combinations stay inside the set,

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

- Important examples in ML of simple convex sets:
  - $\mathbb{R}^d$, non-negative orthant, hyper-planes, half-spaces, and norm-balls.

Showing a Set is Convex from Intersections

- Useful property: the intersection of convex sets is convex.



- We can prove convexity of a set by showing it's an intersection of convex sets.
- Example: "linear programs" have constraints of the form $Aw \leq b$.
  - Each constraints $a_i^\top b_i$ defines a half-space.
  - Half-spaces are convex sets.
  - So the set of w satisfying $Aw \leq b$ is the intersection of convex sets.

Showing a Set is Convex from a Convex Function

- The set $\mathcal{C}$ is often the intersection of a set of inequalities of the form

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

  for some function g and some number τ.

- Sets defined like this are convex if g is a convex function (see bonus).
  - This follows from the definition of a convex function (next topic).

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

Digression: k-way Convex Combinations and Differentiability Classes

- A convex combination of 2 vectors w_1 and w_2 is given by

$$\theta w_1 + (1-\theta) w_2, \quad \text{where} \quad 0 \leq \theta \leq 1.$$

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

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

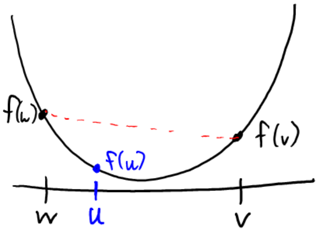
- We'll define convex functions for different differentiability classes:
  - C^0 is the set of continuous functions.
  - C^1 is the set of continuous functions with continuous first-derivatives.
  - C^2 is the set of continuous functions with continuous first- and second-derivatives.

Definitions of Convex Functions

- Four equivalent definitions of convex functions (depending on differentiability):
  1. A C^0 function is convex if the area above the function is a convex set.
  2. A C^0 function is convex if the function is always below its "chords" between points.
  3. A C^1 function is convex if the function is always above its tangent planes.
  4. A C^2 function is convex if it is curved upwards everwhere.
     - If the function is univariate this means $f''(w) \geq 0$ for all w.
- Univariate examples where you can show $f''(w) \geq 0$ for all w:
  - 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: $\log(1 + \exp(-w))$.

C^0 Definitions of Convex Functions

- A function f is convex iff the area above the function is a convex set.



- Equivalently, the function is always below its “chords” between points.

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

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

Convexity of Norms

- The C^0 definition can be used to show that all norms are convex:
  - If $f(w) = \|w\|_p$ for a generic norm, then we have

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

  so f is always below the "chord".
- See course webpage notes on norms if the above steps aren't familiar.

- Also note that all squared norms are convex.
  - These are all convex: $|w|, \|w\|, \|w\|_1, \|w\|^2, \|w_1\|^2, \|w\|_\infty, \ldots$

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: $$h(w) = \alpha f(w), \quad (\text{for } \alpha \geq 0)$$
  2. Sum: $$h(w) = f(w) + g(w).$$
  3. Maximum: $$h(w) = \max\{f(w), g(w)\}.$$
  4. Composition with linear: $$h(w) = f(Aw),$$

     where A is a matrix (or another "linear operator").

- Note that multiplication and composition do not preserve convexity in general.
  - $f(w)g(w)$ is not a convex function in general, even if f and g are convex.
  - $f(g(w))$ is not a convex function in general, even if f and g are convex.

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 linear.
- We can use these to quickly show that SVMs are convex,

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

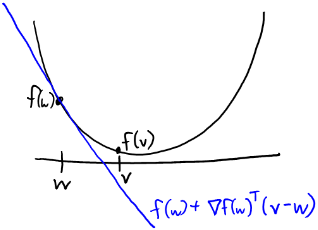
- Second term is squared norm multiplied by non-negative $\frac{\lambda}{2}$.
  - Squared norms are convex, and non-negative scaling perserves convexity.
- First term is sum(max(linear)). Linear is convex and sum/max preserve convexity.
- Since both terms are convex, and sums preserve convexity, SVMs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

C^1 Definition of Convex Functions

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

- A *differentiable* (C^1) function f is convex iff f is always above tangent planes.

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



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

C^2 Definition of Convex Functions

- The multivariate C^2 definition is based on the Hessian matrix, $\nabla^2 f(w)$.
  - The matrix of second partial derivatives,

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

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

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

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

Convexity of Twice-Differentiable Functions

- A C^2 function is convex iff:
$$\nabla^2 f(w) \succeq 0,$$
for all w in the domain ("curved upwards" in every direction).

- This notation $A \succeq 0$ means that A is positive semidefinite.

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

Example: 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 for any w is given by

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

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

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

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

Showing that Function is Convex

- Most common approaches for showing that a function is convex:
  1. Show that f is constructed from operations that preserve convexity.
     - Non-negative scaling, sum, max, composition with linear.
  2. Show that $\nabla^2 f(w)$ is positive semi-definite for all w (for C^2 functions),
  $$\nabla^2 f(w) \succeq 0 \;\; \text{(zero matrix)}.$$
  3. Show that f is below chord for any convex combination of points.
  $$f(\theta w + (1-\theta)v \leq \theta f(w) + (1-\theta) f(v).$$
- Post-lecture slides: convexity of logistic regression from C^2 definition.
  - And how to write logistic regression gradient and Hessian in matrix notation.

Outline

Positive Semi-Definite, Positive Definite, Generalized Inequality

- The notation $A \succeq 0$ indicates that A is positive semi-definite.
  - The eigenvalues of A are all non-negative.
  - $v^\top A v \geq 0$ for all vectors v.

- The notation $A \succ 0$ indicates that A is positive definite.
  - The eigenvalues of A are all positive.
  - $v^\top A v > 0$ for all vectors $v \neq 0$.
  - This implies that A is invertible (bonus).

- The notation $A \succeq B$ indicates that $A - B$ is positive semi-definite.
  - The eigenvalues of $A - B$ are all non-negative.
  - $v^\top A v \geq v^\top B v$ for all vectors v.

MEMORIZE!

More Examples of Convex Functions

- Some convex sets based on these definitions that we'll use (for covariances):
  - 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 we'll use in this course:
  - $f(W) = -\log \det W$ for $W \succ 0$ (negative log-determinant).
  - $f(W, v) = v^\top W^{-1} v$ for $W \succ 0$.
  - $f(w) = \log(\sum_{j=1}^{d} \exp(w_j))$ (log-sum-exp function).

Positive Semi-Definite, Positive Definite, Generalized Inequality

- Note that not every matrix can be compared.
- With these matrices:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

  neither $A \succeq B$ nor $B \succeq A$ (the "generalized inequality" defines a "partial order").

- It's often useful to compare to the identity matrix I, which has eigenvalues 1.
  - So a matrix of the form μI for a scalar μ has all eigenvalues equal to μ.

- Writing $LI \succeq A \succeq \mu I$ means "eigenvalues of A are between μ and L".

Convexity, Strict Convexity, and Strong Convexity

- We say that a C^2 function is convex if for all w,
$$\nabla^2 f(w) \succeq 0,$$
and this implies any stationary point ($\nabla f(w) = 0$) is a global minimum.
- We say that a C^2 function is strictly convex if for all w,
$$\nabla^2 f(w) \succ 0,$$
and this implies there is at most one stationary point (and $\nabla^2 f(w)$ is invertible).
- We say that a C^2 function is strongly convex if for all w.
$$\nabla^2 f(w) \succeq \mu I, \quad \text{for some } \mu > 0,$$
and this implies there exists a minimum (if domain $\mathcal{C}$ is closed).
  - Strong convxity affects speed of gradient descent, and how much data you need.

Convexity, Strict Convexity, and Strong Convexity

- These definitions simplify for univariate functions:
  - Convex: $f''(w) \geq 0$.
  - Strictly convex: $f''(w) > 0$.
  - Strongly convex: $f''(w) \geq \mu$ for $\mu > 0$.

- Examples:
  - Convex: $f(w) = w$.
    - Since $f''(w) = 0$.
  - Strictly convex: $f(w) = \exp(w)$.
    - Since $f''(w) = \exp(w) > 0$.
  - Strongly convex: $f(w) = \frac{1}{2}w^2$.
    - Since $f''(w) = 1$ so it is strongly convex with $\mu = 1$.

Strict Convexity of L2-Regularized Least Squares

- In L2-regularized least squares, the Hessian matrix is
$$\nabla^2 f(w) = (X^\top X + \lambda I).$$
- We can show that this is positive-definite, so the problem is strictly convex,
$$v^\top \nabla^2 f(w) v = v^\top (X^\top X + \lambda I) v = \underbrace{\|Xv\|^2}_{\geq 0} + \underbrace{\lambda \|v\|^2}_{>0} > 0,$$
where we used that $\lambda > 0$ and $\|v\| > 0$ for $v \neq 0$.
- This implies that the matrix $(X^\top X + \lambda I)$ is invertible, and solution is unique.
  - Similar argument shows it's strongly-convex with $\mu = \lambda$.
  - Value μ can be larger if columns of X are independent (no collinearity).
    - In this case, $\|Xv\| \neq 0$ for $v \neq 0$ so even least squares is strongly-convex.

Strong-Convexity Discussion

- We can also define strict and strong convexity for C^1 and C^0 functions (bonus).
  - And note that (strong convexity) implies (strict convexity) implies (convexity).

- For example, we say that a C^0 function f is strongly convex if the function

$$f(w) - \frac{\mu}{2}\|w\|^2,$$

  is a convex function for some $\mu > 0$.
  - "If you 'un-regularize' by μ then it's still convex."

- If we have a convex loss f, adding L2-regularization makes it strongly-convex,

$$f(w) + \frac{\lambda}{2}\|w\|^2,$$

  with μ being at least λ.
  - So L2-regularization guarantees a solution exists, and that it is unique.

Summary

- 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.

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

- Strict and strong convexity guarantee uniqueness and existense of solutions.
  - Adding L2-regularization to a convex function gives you these.
- Post-lecture slides: matrix notation and convexity of logistic regerssion.
  - This will help with your assignments.
- How much data do we need?

Example: Convexity of Logistic Regression

- Consider the binary logistic regression model,

$$f(w) = \sum_{i=1}^{n} \log(1 + \exp(-y^i w^T x^i)).$$

- With some tedious manipulations, gradient in matrix notation is

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

  where the vector r has elements $r_i = -y^i h(-y^i w^T x^i)$.
  - And h is the sigmoid function, $h(\alpha) = 1/1 + \exp(-\alpha)$.

- We know the gradient has this form from the multivariate chain rule.
  - Functions for the form $f(Xw)$ always have $\nabla f(w) = X^T r$ (see bonus slide).

Example: Convexity of Logistic Regression

- With some more tedious manipulations we get the Hessian in matrix notation as

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

where D is a diagonal matrix with $d_{ii} = h(y_i w^T x^i) h(-y^i w^T x^i)$.

  - The $f(Xw)$ structure leads to a $X^T D X$ Hessian structure.
  - For other problems D may not be diagonal.

- Since the sigmoid function h is non-negative, we can compute $D^{\frac{1}{2}}$, and

$$v^T X^T D X v = v^T X^T D^{\frac{1}{2}} D^{\frac{1}{2}} X v = (D^{\frac{1}{2}} X v)^T (D^{\frac{1}{2}} X v) = \|X D^{\frac{1}{2}} v\|^2 \geq 0,$$

so $X^T D X$ is positive semidefinite and logistic regression is convex.

Showing that Hyper-Planes are Convex

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

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

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

Convex Sets from Functions

- For sets of the form

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

If g is a convex function, then $\mathcal{C}$ is a convex set:

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

which means convex combinations are in the set.

Multivariate Chain Rule

- If $g : \mathbb{R}^d \mapsto \mathbb{R}^n$ and $f : \mathbb{R}^n \mapsto \mathbb{R}$, then $h(x) = f(g(x))$ has gradient

$$\nabla h(x) = \nabla g(x)^T \nabla f(g(x)),$$

where $\nabla g(x)$ is the Jacobian (since g is multi-output).

- If g is an affine map $x \mapsto Ax + b$ so that $h(x) = f(Ax + b)$ then we obtain

$$\nabla h(x) = A^T \nabla f(Ax + b).$$

- Further, for the Hessian we have

$$\nabla^2 h(x) = A^T \nabla^2 f(Ax + b) A.$$

Positive-Definite implies Invertibility

- If $A \succ 0$, then all the eigenvalues of A are positive.
- If each eigenvalue is positive, the product of the eigenvalues is positive.
- The product of the eigenvalues is equal to the determinant.
- Thus, the determinant is positive.
- The determinant not being 0 implies the matrix is invertible.

Strong Convexity of L2-Regularized Least Squares

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

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

$$v^\top \nabla^2 f(w) v = v^\top (X \top X + \lambda I) v = \underbrace{\|Xv\|^2}_{} + v^\top (\lambda I) v \geq v^\top (\lambda I) v,$$

  so we've shown that $\nabla^2 f(w) \succeq \lambda I$, which implies strong-convexity with $\mu = \lambda$.

- This implies that a solution exists, and that the solution is unique.

- Note that we have strong convexity with $\mu > \lambda$ if $X^\top X$ is positive definite.
  - Which happens iff the features are independent (not collinear).

Strictly-Convex Functions

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

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

$$f(v) > f(w) + \nabla f(w)^\top (v - w) \qquad (C^1)$$

$$\nabla^2 f(w) \succ 0 \qquad (C^2)$$

- 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.

A C^0 Definition of Strict and Strong Convexity

- There are many equivalent definitions of the convexities, here is one set for C^0 functions:
  - Convex (usual definition):
$$f(\theta w + (1-\theta)v) \leq \theta f(w) + (1-\theta)f(v).$$
  - Strictly convex (strict version, exclusindg $\theta = 0$ or $\theta = 1$):
$$f(\theta w + (1-\theta)v) < \theta f(w) + (1-\theta)f(v).$$
  - Strong convexity (need an "extra" bit of decrease as you move away from endpoints):
$$f(\theta w + (1-\theta)v) \leq \theta f(w) + (1-\theta)f(v) - \frac{\theta(1-\theta)\mu}{2}\|w-v\|^2.$$