NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2019)

Brett Bernstein

February 18, 2019

Instructions: Following most lab and lecture sections, we will be providing concept checks for review. Each concept check will:

- List the lab/lecture learning objectives. You will be responsible for mastering these objectives, and demonstrating mastery through homework assignments, exams (midterm and final), and on the final course project.
- Include concept check questions. These questions are intended to reinforce the lab/lectures, and help you master the learning objectives.

You are strongly encourage to complete all concept check questions, and to discuss these (and related) problems on Piazza and at office hours. However, problems marked with a ($\star$) are considered optional.

Week 3 Lab: Concept Check Exercises

Convexity

Optional Learning Objectives

Convex optimization and Lagrangian duality will not covered on the midterm exam, so in some sense these objectives are optional.

- Define a convex set, a convex function, and a strictly convex function. (Don't forget that the domain of a convex function must be a convex set!)
- For an optimization problem, define the terms feasible set, feasible point, active constraint, optimal value, and optimal point.
- Give the form for a general inequality-constrained optimization problem (there are many ways to do this, but our convention is to have inequality constraints of the form $f_i(x) \leq 0$).

- Define the Lagrangian for this optimization problem, and explain how the Lagrangian encodes all the information in the original optimization problem.
- Write the primal and dual optimization problem in terms of the Lagrangian.

Convexity Concept Check Problems

1. If $A, B \subseteq \mathbb{R}^n$ are convex, then $A \cap B$ is convex.

   Solution. Let $x, y \in A \cap B$ and $t \in (0,1)$. Since A, B are convex, we have

   $$(1-t)x + ty \in A \quad \text{and} \quad (1-t)x + ty \in B.$$

   Thus $(1-t)x + ty \in A \cap B$.

2. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that $af + bg$ is convex if $a, b \geq 0$.

   Solution. Let $x, y \in \mathbb{R}^n$ and $\theta \in (0,1)$. Then

   $$\begin{aligned}(af+bg)((1-\theta)x+\theta y) &= af((1-\theta)x+\theta y) + bg((1-\theta)x+\theta y)\\ &\leq a[(1-\theta)f(x)+\theta f(y)] + b[(1-\theta)g(x)+\theta g(y)]\\ &= (1-\theta)(af+bg)(x) + \theta(af+bg)(y).\end{aligned}$$

3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Prove that if $\nabla f(x) = 0$ then x is a global minimizer.

   Solution. Suppose $\nabla f(x) = 0$. The gradient (or first-order) characterization of convexity says

   $$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

   for all y. If $\nabla f(x) = 0$ then this says $f(y) \geq f(x)$ for all x.

4. Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and x is a global minimizer, then it is the unique global minimizer.

   Solution. Suppose y is also a global minimizer with $y \neq x$. Then

   $$f((y+x)/2) < f(y)/2 + f(x)/2 = f(x)$$

   contradicting the fact that $f(x)$ was a global minimizer.

5. Prove that any affine function $f : \mathbb{R}^n \to \mathbb{R}$ is both convex and concave.

   Solution. Recall that f has the form $f(x) = w^T x + b$ where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then, for $x, y \in \mathbb{R}^n$ and $\theta \in (0,1)$,

   $$f((1-\theta)x+\theta y) = w^T((1-\theta)x+\theta y)+b = (1-\theta)(w^T x+b)+\theta(w^T y+b) = (1-\theta)f(x)+\theta f(y).$$

   This shows f is convex. But the same holds if we replace w with $-w$ and b with $-b$. Hence f is also concave.

6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and let $g : \mathbb{R}^m \to \mathbb{R}^n$ be affine. Then $f \circ g$ is convex.

   Solution. Write $g(x) = Ax + b$ where $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. For $x, y \in \mathbb{R}^m$ and $t \in (0, 1)$ we have

$$\begin{aligned} f(g((1-t)x + ty)) &= f((1-t)(Ax+b) + t(Ay+b)) \\ &\leq (1-t)f(Ax+b) + tf(Ay+b) \\ &= (1-t)f(g(x)) + tf(g(y)). \end{aligned}$$

7. ($\star\star$)

   (a) Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Show that f has one-sided left and right derivatives at every point.

   (b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that f has one-sided directional derivatives at every point.

   (c) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that if x is not a minimizer of f then f has a descent direction at x (i.e., a direction whose corresponding one-sided directional derivative is negative).

   Solution. We first prove the following lemma.

   Lemma 1. *If $f : \mathbb{R} \to \mathbb{R}$ is convex and $x < y < z$ then*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x}.$$

   Proof. Let $t \in (0, 1)$ satisfy $(1 - t)x + tz = y$. By convexity we have

$$f(y) = f((1-t)x + tz) \leq (1-t)f(x) + tf(z)$$

   giving

$$\frac{f(y) - f(x)}{y - x} \leq \frac{(1-t)f(x) + tf(z) - f(x)}{(1-t)x + tz - x} = \frac{t(f(z) - f(x))}{t(z - x)} = \frac{f(z) - f(x)}{z - x}.$$

   □

   (a) For the right derivative, we will show

$$\lim_{y \downarrow x} \frac{f(y) - f(x)}{y - x} = \inf_{y > x} \frac{f(y) - f(x)}{y - x} =: L.$$

   Fix $\epsilon > 0$ and choose $y' > x$ so that

$$\frac{f(y') - f(x)}{y' - x} < L + \epsilon.$$

Letting $\delta = y' - x$, the lemma shows that

$$\frac{f(y) - f(x)}{y - x} < L + \epsilon$$

for any $y < x + \delta$ proving the limit exists.

For the left derivative, we could repeat the above, or note that $g(t) = 2x - t$ is affine, so $f \circ g$ is convex. By the above

$$\lim_{y \downarrow x} \frac{f(g(y)) - f(g(x))}{y - x} = \lim_{y \downarrow x} \frac{f(2x - y) - f(x)}{y - x} = \lim_{h \downarrow 0} \frac{f(x - h) - f(x)}{h}$$

exists, where $h = y - x$. This proves the left derivative exists as well.

(b) Fix $x, v \in \mathbb{R}^n$ and let $g : \mathbb{R} \to \mathbb{R}^n$ be defined by $g(t) = x + tv$. Then $f \circ g$ is convex, and thus the previous part applies. But the right derivative of g at 0 is the one-sided directional derivative of f at x in the direction v:

$$\lim_{h \downarrow 0} \frac{f(g(h)) - f(g(0))}{h} = \lim_{h \downarrow 0} \frac{f(x + hv) - f(x)}{h}.$$

(c) Let y be a minimizer of f and let $g(t) = x + t(y - x)$. By the arguments in the first part above, the value

$$\frac{f(g(1)) - f(g(0))}{1 - 0} = f(y) - f(x) < 0$$

is an upper bound on the right derivative of g at 0. But this is a directional derivative, by the argument in the second part above.

Convex Optimization Problems

1. Suppose there are mn people forming m rows with n columns. Let a denote the height of the tallest person taken from the shortest people in each column. Let b denote the height of the shortest person taken from the tallest people in each row. What is the relationship between a and b?

   Solution. Let H_{ij} denote the height of the person in row i and column j. Then

   $$a = \max_j \min_i H_{ij} \leq \min_i \max_j H_{ij} = b,$$

   by the max-min inequality.

2. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be given data. You want to find the center and radius of the smallest sphere that encloses all of the points. Express this problem as a convex optimization problem.

Solution.

$$\begin{array}{ll} \text{minimize}_{r,c} & r \\ \text{subject to} & \|x_i - c\|_2 \leq r \quad \text{for } i = 1, \ldots, n. \end{array}$$

This problem is convex since norms are convex, so $f_i(c) = \|x_i - c\|_2$ is convex (composition of convex with affine).

3. Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ and $y_1, \ldots, y_n \in \{-1, 1\}$. Here we look at y_i as the label of x_i. We say the data points are linearly separable if there is a vector $v \in \mathbb{R}^d$ and $a \in \mathbb{R}$ such that $v^T x_i > a$ when $y_i = 1$ and $v^T x_i < a$ for $y_i = -1$. Give a method for determining if the given data points are linearly separable.

   Solution. Solve the hard-margin SVM problem

   $$\begin{array}{ll} \text{minimize}_{w,b} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i = 1, \ldots, n. \end{array}$$

   If the resulting problem is feasible, then the data is linearly separable.

4. Consider the Ivanov form of ridge regression:

   $$\begin{array}{ll} \text{minimize} & \|Ax - y\|_2^2 \\ \text{subject to} & \|x\|_2^2 \leq r^2, \end{array}$$

   where $r > 0$, $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ are fixed.

   (a) What is the Lagrangian?

   (b) What do you get when you take the supremum of the Lagrangian over the feasible values for the dual variables?

   Solution.

   (a) $L(x, \lambda) = \|Ax - y\|_2^2 + \lambda(\|x\|_2^2 - r^2)$. Note that this is a shifted version of the Tikhonov objective.

   (b)

   $$\sup_{\lambda \geq 0} L(x, \lambda) = \begin{cases} +\infty & \text{if } \|x\|_2^2 > r^2, \\ \|Ax - y\|_2^2 & \text{otherwise.} \end{cases}$$

   Note that the original Ivanov minimization is then just

   $$\inf_x \sup_{\lambda \geq 0} L(x, \lambda).$$

Subgradients

1. ($\star$) If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable at x, the $\partial f(x) = \{\nabla f(x)\}$.

   Solution. By the gradient (first-order) conditions for convexity, we know that $\nabla f(x) \in \partial f(x)$. Next suppose $g \in \partial f(x)$. This means that for all $v \in \mathbb{R}^n$ and $h \in \mathbb{R}$ we have

   $$f(x + hv) \geq f(x) + hg^T v \implies \frac{f(x + hv) - f(x)}{h} \geq g^T v.$$

   Using $-h$ in place of h gives

   $$f(x - hv) \geq f(x) - hg^T v \implies g^T v \geq \frac{f(x - hv) - f(x)}{-h}.$$

   Taking limits as $h \to 0$ gives

   $$\nabla f(x)^T v \geq g^T v \geq \nabla f(x)^T v.$$

   Thus all terms are equal. Subtracting gives

   $$(\nabla f(x) - g)^T v = 0,$$

   which holds for all $v \in \mathbb{R}^n$. Letting $v = \nabla f(x) - g$ proves

   $$\|\nabla f(x) - g\|_2^2 = 0$$

   giving the result.

2. Fix $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$. Then the subdifferential $\partial f(x)$ is a convex set.

   Solution. Let $g_1, g_2 \in \partial f(x)$ and $t \in (0, 1)$. We must show $(1 - t)g_1 + tg_2$ is a subgradient. Note that, for any $y \in \mathbb{R}^n$, we have

   $$\begin{aligned} f(x) + ((1 - t)g_1 + tg_2)^T(y - x) &= (1 - t)(f(x) + g_1^T(y - x)) + t(f(x) + g_2^T(y - x)) \\ &\leq (1 - t)f(y) + tf(y) \\ &= f(y). \end{aligned}$$

3. (a) True or False: A subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at x is normal to a hyperplane that globally understimates the graph of f.

   (b) True or False: If $g \in \partial f(x)$ then $-g$ is a descent direction of f.

   (c) True or False: For $f : \mathbb{R} \to \mathbb{R}$, if $1, -1 \in \partial f(x)$ then x is a global minimizer of f.

   (d) True or False: Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $g \in \partial f(x)$. Then $\alpha g \in \partial f(x)$ for all $\alpha \in [0, 1]$.

   (e) True or False: If the sublevel sets of a function are convex, then the function is convex.

Solution.

(a) False. The underestimating hyperplane is a subset of $\mathbb{R}^{n+1}$ but a subgradient is an element of $\mathbb{R}^n$.

(b) False. In lab we considered $f(x_1, x_2) = |x_1| + 2|x_2|$ and noted that $(1, -2) \in \partial f(3, 0)$ but $(-1, 2)$ is not a descent direction.

(c) True. The subdifferential of f at x is convex, and thus contains 0. If 0 is a subgradient of f at x, then x is a global minimizer.

(d) False. Suppose $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$. Then $\partial f(1) = \{2\}$, and thus doesn't contain 2α for $\alpha \in [0, 1)$.

(e) False. A counterexample is $f(x) = -e^{-x^2}$. The converse is true though. Functions that have convex sublevel sets are called *quasiconvex*.

4. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x_1, x_2) = |x_1| + 2|x_2|$. Compute $\partial f(x_1, x_2)$ for each $x_1, x_2 \in \mathbb{R}^2$.

Solution. Write $f(x_1, x_2) = f_1(x_1, x_2) + f_2(x_1, x_2)$ where $f_1(x_1, x_2) = |x_1|$ and $f_2(x_1, x_2) = 2|x_2|$. When $x_1 \neq 0$ we have $\partial f_1(x_1, x_2) = \{(\operatorname{sgn}(x_1), 0)^T\}$ and when $x_1 = 0$ we have

$$\partial f_1(x_1, x_2) = \{(b, 0)^T \mid b \in [-1, 1]\}.$$

When $x_2 \neq 0$ we have $\partial f_2(x_1, x_2) = \{(0, 2\operatorname{sgn}(x_2))^T\}$ and when $x_2 = 0$ we have

$$\partial f_2(x_1, x_2) = \{(0, c)^T \mid c \in [-2, 2]\}.$$

Combining we have

$$\partial f(x_1, x_2) = \partial f_1(x_1, x_2) + \partial f_2(x_1, x_2),$$

where we are summing sets. Recall that if $A, B \subseteq \mathbb{R}^n$ then

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

This gives 4 cases:

(a) If $x_1, x_2 \neq 0$ this gives $\partial f(x_1, x_2) = \{(\operatorname{sgn}(x_1), 2\operatorname{sgn}(x_2))^T\}$.

(b) If $x_1 = 0$ and $x_2 \neq 0$ we have $\partial f(x_1, x_2) = \{(b, 2\operatorname{sgn}(x_2))^T \mid b \in [-1, 1]\}$.

(c) If $x_1 \neq 0$ and $x_2 = 0$ we have $\partial f(x_1, x_2) = \{(\operatorname{sgn}(x_1), c)^T \mid c \in [-2, 2]\}$.

(d) If $x_1 = 0$ and $x_2 = 0$ we have $\partial f(x_1, x_2) = \{(b, c)^T \mid b \in [-1, 1], c \in [-2, 2]\}$.