NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2018)

Brett Bernstein

February 12, 2018

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 4 Lab: Concept Check Exercises

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