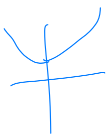
EE364b Spring 2020 Homework 1

Due Friday 4/17 at 11:59pm via Gradescope

1.1 (3 points) For each of the following convex functions, determine the subdifferential set at the specified point.



(a) $f(x_1, x_2, x_3) = \max\{|x_1|, |x_2|, |x_3|\}$ at $(x_1, x_2, x_3) = (0, 0, 0)$. $\{v \in \mathbb{R}^3 \mid v_i \in [-1,1]\}$

(b) $f(x) = e^{|x|}$ at $x = 0$ (x is a scalar). $[-1,1]$

(c) $f(x_1, x_2) = \max\{x_1 + x_2 - 1, x_1 - x_2 + 1\}$ at $(x_1, x_2) = (1, 1)$. $\begin{cases} f_1 = 1 \\ f_2 = 1 \end{cases}$ $\partial f = \{v \in \mathbb{R}^2 \mid v_1 = 1,\ v_2 \in [-1,1]\}$

1.2 (7 points) For each of the following convex functions, explain how to calculate a subgradient at a given x.

(e) $\partial f(x)_i = 1$ if x_i is in the top kth element

(a) $f(x) = \max_{i=1,\dots,m}(a_i^T x + b_i)$. $a_i \mid a_i^T x + b_i$

(b) $f(x) = \max_{i=1,\dots,m} |a_i^T x + b_i|$. $i \mid |a_i^T x + b| = f(x)$ $\partial f(x) = \begin{cases} a_i & \text{if } f(x) \ge 0 \\ -a_i & \text{if } f(x) < 0 \end{cases}$

(f)

(c) $f(x) = \max_{i=1,\dots,m} \left(-\log\left(a_i^T x + b_i\right)\right)$. You may assume x is in the domain of f.

if $Ax < b$ $0 \in \partial f(x)$

(d) $f(x) = \max_{0 \le t \le 1} p(t)$, where $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$. $\hat{t} \mid f(x) = p(t)$ $[t, \cdots t^{n-1}] \in \partial f$

if $a_i^T x > b_i$

$k = \arg\max_k (a_k^T x - b_k)$

$a_k \in \partial f(x)$

(e) $f(x) = x_{[1]} + \cdots + x_{[k]}$, where $x_{[i]}$ denotes the ith largest element of the vector x.

(f) $f(x) = \min_{Ay \preceq b} \|x - y\|^2$, *i.e.*, the square of the distance of x to the polyhedron defined by $Ay \preceq b$. You may assume that the inequalities $Ay \prec b$ are strictly feasible. (*Hint: You may use duality, and then use subgradient the rule for pointwise maximum*) Strong duality

(g) find $\hat{y} = f(x)$, $\hat{y} \in \partial f(x)$

(g) $f(x) = \max_{Ay \preceq b} y^T x$, *i.e.*, the optimal value of an LP as a function of the cost vector. (You can assume that the polyhedron defined by $Ay \preceq b$ is bounded.) (*Hint: You may use the subgradient rule for pointwise maximum*)

1.3 *(2 points) Convex functions that are not subdifferentiable.* Verify that the following functions, defined on the interval $[0, \infty)$, are convex, but not subdifferentiable at $x = 0$. (*Hint: You can prove by contradiction, i.e., assuming that the subgradient condition holds to reach a contradiction*)

(a) $f(0) = 1$, and $f(x) = 0$ for $x > 0$. infinite slop.

(b) $f(x) = -x^p$ for some $p \in (0, 1)$.

1.4 *(6 points) Conjugacy, subgradients and L_p-norms.* In the first part of this question, we show how conjugate functions are related to subgradients. Let $f : \mathbf{R}^n \to \mathbf{R}$ be convex and recall that its conjugate is $f^*(v) = \sup_x\{v^T x - f(x)\}$. Prove the following:

(a) For any v we have $v^T x \le f(x) + f^*(v)$ (this is sometimes called Young's inequality).

(a) $f(x) + f^*(v)$
$= f(x) + \sup_z \{v^T z - f(z)\}$
$\ge f(x) + v^T x - f(x)$
$= v^T x$

b) $f(x) + f^*(g)$
$= f(x) + \sup_z \{z^T g - f(z)\}$

① if $g^T x = f(x) + f^*(g)$
$g^T x = f(x) + \sup_z \{z^T g - f(z)\}$
$g^T x - f(x) = \sup_z \{z^T g - f(z)\}$
$\therefore g^T y - f(y) \le g^T x - f(x)$ for any y
$\to f(y) \ge f(x) + g^T(y-x)$
$\therefore g \in \partial f(x)$

② if $g \in \partial f(x)$
$f(y) \ge f(x) + g^T(y-x)$ for any y
$g^T x - f(x) \ge g^T y - f(y)$ for any y
$\therefore g^T x - f(x) = \sup_z \{g^T z - f(z)\}$
$\therefore g^T x = f(x) + f^*(g)$

(b) We have $g^T x = f(x) + f^*(g)$ if and only if $g \in \partial f(x)$.

Note that (you do not need to prove this) if f is closed, so that $f(x) = f^{**}(x)$, result (b) implies the duality relationship that $g \in \partial f(x)$ if and only if $x \in \partial f^*(g)$ if and only if $g^T x = f(x) + f^*(g)$.

In the second part of this question, we apply the result (b) to characterize the subdifferentials of the function $f(x) = \|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$, where $p \ge 1$. We denote $q = \frac{p}{p-1}$ if $p > 1$ and $q = +\infty$ if $p = 1$. Note that $\frac{1}{p} + \frac{1}{q} = 1$.

(c) Show that for any v we have $f^*(v) = \mathcal{I}_q(v)$ where $\mathcal{I}_q(v) = 0$ if $\|v\|_p \le 1$ and $\mathcal{I}_q(v) = +\infty$ if $\|v\|_p > 1$.

(d) Deduce from (b) and (c) that for any x and any g, we have $g \in \partial f(x)$ if and only if $g^T x = \|x\|_p$ and $\|g\|_q \le 1$.

(e) Determine $\partial f(0)$ for $p = 1, 2, +\infty$.

In the final part of this question, we extend the case $p = 1$ in the context of symmetric matrices. Denote $\mathbf{S}$ the set of $n \times n$ real symmetric matrices. For $X \in \mathbf{S}$, recall the definition of its nuclear norm $\|X\|_* = \sum_{i=1}^n |\lambda_i(X)|$ where $\lambda_1(X), \dots, \lambda_n(X)$ are the eigenvalues of X and its operator norm $\|X\| = \sup_{i=1,\dots,n} |\lambda_i(X)|$.

(f) Consider $f(X) = \|X\|_*$. Show that $\partial f(0) = \{Z \in \mathbf{S} \mid \|Z\| \le 1\}$. Determine $\partial f(X)$ for an arbitrary $X \in \mathbf{S}$ in terms of the eigenvalues and eigenvectors of X.

1.5 *Optional (extra credit, 6 points). Non-convex non-differentiable functions, Clarke subdifferentials and Neural Networks.* Let $f : \mathbf{R}^n \to \mathbf{R}$ be a given function that we do not assume to be convex nor to be differentiable (e.g., a deep neural network with ReLU activation functions), so that the subdifferential $\partial f(x) = \{g \in \mathbf{R}^n \mid f(y) \ge f(x) + g^\top (y - x)\ \forall y\}$ is possibly an empty set. In this question, we explore a more general notion of subdifferentials, namely, Clarke subdifferentials, originally referred to as generalized gradients [Cla75].

We make the following technical assumption: we assume that f is locally Lipschitz, i.e., for any $x \in \mathbf{R}^n$, there exists $\eta > 0$ and $L_x > 0$ such that $|f(y) - f(z)| \le L_x \|y - z\|_2$ for any y, z such that $\|x - y\|_2, \|x - z\|_2 \le \eta$. Then, it follows that the function f is differentiable almost everywhere with respect to the Lebesgue measure (this result is sometimes referred to as Rademacher's theorem [BL10]). We denote by D the subset of $\mathbf{R}^n$ where f is differentiable. In other words, if we consider a bounded open set B in $\mathbf{R}^n$ and we pick x uniformly at random in B, then f is differentiable at x with probability equal to 1.

The Clarke subdifferential of f at x is defined as

$$\partial_C f(x) = \mathbf{Co} \left\{ \lim_{k \to \infty} \nabla f(x_k) \mid x_k \to x,\, x_k \in D,\, \lim_{k \to \infty} \nabla f(x_k) \text{ exists} \right\}.$$

The goal of this exercise is to characterize some basic properties of Clarke subdifferentials, relate $\partial_C f(x)$ to $\partial f(x)$ and study some implications of the condition $0 \in \partial_C f(x)$, which is necessary and sufficient for global optimality in the convex case. Prove the following:

(a) If f is a continuously differentiable function then $\partial_C f(x) = \{\nabla f(x)\}$.

(b) If f is convex then $\partial_C f(x) \subseteq \partial f(x)$. *(Optional, no credit)* Show that equality actually holds, i.e., $\partial_C f(x) = \partial f(x)$. *Hint: Suppose by contradiction that there exists $g \in \partial f(x)$ such that $g \notin \partial_C f(x)$. Set $h(x) = f(x) - g^T x$. Show that $0 \in \partial h(x)$ and $0 \notin \partial_C h(x)$. Use the hyperplane separation theorem to conclude.*

We say that x is *Clarke stationary* if $0 \in \partial_C f(x)$. If f is convex, then, from (b), we know that x is a global minimizer of f. For a non-convex function f, this property does not extend in general as we explore next.

(c) Suppose that x is a local minimum (resp. maximum) of f, i.e., there exists a radius $\eta > 0$ such that $f(y) \geq f(x)$ (resp. $f(y) \leq f(x)$) for any y such that $\|y - x\|_2 \leq \eta$. Show that x is Clarke stationary. *Hint: suppose by contradiction that $0 \notin \partial_C f(x)$ and conclude by using the hyperplane separating theorem with the convex sets $\partial_C f(x)$ and $\{0\}$.*

(d) Suppose that $\inf_x f(x) > -\infty$ and that $\inf_x f(x)$ is attained. Show that if x is the *unique* Clarke stationary point of f, then x is the unique global minimizer of f.

Finally, we study two examples of non-convex non-differentiable functions: a two-dimensional input function which has a unique Clarke stationary point that is the global minimizer, and, a neural network training loss which has a spurious Clarke stationary point at $(0, \ldots, 0)$.

(e) Consider the function with two-dimensional inputs $f(x_1, x_2) = 10\,|x_2 - x_1^2| + (1 - x_1)^2$. Show that the unique Clarke stationary point of f is $(x_1, x_2) = (1, 1)$ and that it is the unique global minimizer of f.

(f) Consider a supervised learning setting with a neural network parameterization: let $X \in \mathbf{R}^{n \times d}$ be a given data matrix and $y \in \mathbf{R}^n$ be a vector of real-valued observations. For the neural network parameters $u_1, \ldots, u_m \in \mathbf{R}^d$ and $\alpha_1, \ldots, \alpha_m \in \mathbf{R}$, consider the loss function

$$f(u_1, \ldots, u_m, \alpha_1, \ldots, \alpha_m) = \|y - \sum_{i=1}^{m} \sigma(X u_i) \alpha_i\|_2^2 \,,$$

where we have introduced the component-wise ReLU activation function σ defined as $\sigma(z) = (\max\{z_1, 0\}, \ldots, \max\{z_n, 0\}) \in \mathbf{R}^n$ for $z = (z_1, \ldots, z_n) \in \mathbf{R}^n$. Show that $0 \in \partial f_C(0, \ldots, 0, 0, \ldots, 0)$.

References

[BL10] Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples.* Springer Science & Business Media, 2010.

[Cla75] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.