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 \((x_1, x_2, x_3) = (0, 0, 0)\).
(b) \( f(x) = e^{|x|} \) at \(x = 0\) (\(x\) is a scalar).
(c) \( f(x_1, x_2) = \max\{x_1 + x_2 - 1, x_1 - x_2 + 1\} \) at \((x_1, x_2) = (1, 1)\).

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

(a) \( f(x) = \max_{i=1, \ldots, m} (a_i^T x + b_i) \).
(b) \( f(x) = \max_{i=1, \ldots, m} |a_i^T x + b_i| \).
(c) \( f(x) = \max_{i=1, \ldots, m} (-\log (a_i^T x + b_i)) \). You may assume \(x\) is in the domain of \(f\).
(d) \( f(x) = \max_{0 \leq t \leq 1} p(t) \), where \(p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}\).
(e) \( f(x) = x_{[1]} + \cdots + x_{[k]} \), where \(x_{[i]}\) denotes the \(i\)th largest element of the vector \(x\).
(f) \( f(x) = \min_{Ay \leq b} \|x - y\|^2 \), i.e., the square of the distance of \(x\) to the polyhedron defined by \(Ay \leq b\). You may assume that the inequalities \(Ay \leq b\) are strictly feasible. (\(\text{Hint: You may use duality, and then use subgradient the rule for pointwise maximum}\))
(g) \( f(x) = \max_{Ay \leq b} y^T x \), i.e., the optimal value of an LP as a function of the cost vector. (You can assume that the polyhedron defined by \(Ay \leq b\) is bounded.) (\(\text{Hint: You may use the subgradient rule for pointwise maximum}\))

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 \(x = 0\). (\(\text{Hint: You can prove by contradiction, i.e., assuming that the subgradient condition holds to reach a contradiction}\))

(a) \( f(0) = 1, \text{ and } f(x) = 0 \) for \(x > 0\).
(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 \( f: \mathbb{R}^n \to \mathbb{R} \) 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 \(v^T x \leq f(x) + f^*(v)\) (this is sometimes called Young’s inequality).
1.5 Optional (extra credit, 6 points). Non-convex non-differentiable functions, Clarke subdifferentials and Neural Networks. Let $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\partial f(x) = \{ g \in \mathbb{R}^n \mid f(y) \geq f(x) + g^T(y - x) \ \forall y \}$ 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 $x \in \mathbb{R}^n$, there exists $\eta > 0$ and $L_x > 0$ such that $|f(y) - f(z)| \leq L_x \|y - z\|_2$ for any $y, z$ such that $\|x - y\|_2, \|x - z\|_2 \leq \eta$. 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 $\mathbb{R}^n$ where $f$ is differentiable. In other words, if we consider a bounded open set $B$ in $\mathbb{R}^n$ 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

$$\partial_C f(x) = \text{Co} \left\{ \lim_{k \to \infty} \nabla f(x_k) \mid x_k \to x, x_k \in D, \lim_{k \to \infty} \nabla f(x_k) \text{ exists} \right\} .$$
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 \( X \in \mathbb{R}^{n \times d} \) be a given data matrix and \( y \in \mathbb{R}^n \) be a vector of real-valued observations. For the neural network parameters \( u_1, \ldots, u_m \in \mathbb{R}^d \) and \( \alpha_1, \ldots, \alpha_m \in \mathbb{R} \), consider the loss function

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

where we have introduced the component-wise ReLU activation function \( \sigma \) defined as \( \sigma(z) = (\max\{z_1, 0\}, \ldots, \max\{z_n, 0\}) \in \mathbb{R}^n \) for \( z = (z_1, \ldots, z_n) \in \mathbb{R}^n \). Show that \( 0 \in \partial_C f(0, \ldots, 0, 0, \ldots, 0) \).
References
