The University of Texas at Austin
Optimization

Homework 8

Constantine Caramanis, Sujay Sanghavi

1. Subgradients

Show the following for sub-gradients

(a) If $g(x) = f(Ax + b)$, then $\partial g(x) \supseteq A^T \partial f(Ax + b)$.

(b) If $f(x) = \max_{1 \leq i \leq m} f_i(x)$, then $\partial f(x) \supseteq \text{conv}\left(\bigcup_{i:f_i(x)=f(x)} \partial f_i(x)\right)$

In fact, the containments above are equalities, but the reverse inclusions are more delicate, so we are only asking you to show one inclusion. For your interest, we will detail the argument for both directions on the solution.

Solution

(a) Proof that $\partial g(x) \supseteq A^T \partial f(Ax + b)$

Let $g(x) = f(Ax + b)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, A is a matrix, and b is a vector.

Proof: Let $v \in \partial f(Ax + b)$. By the definition of the subgradient, for all $y \in \mathbb{R}^n$:

$$f(y) \geq f(Ax + b) + v^T(y - (Ax + b)).$$

Let $z \in \mathbb{R}^m$ (since $x \in \mathbb{R}^m$), and consider $y = Az + b$. Substituting into the inequality:

$$f(Az + b) \geq f(Ax + b) + v^T(A(z - x)).$$

Since $g(z) = f(Az + b)$, we have:

$$g(z) \geq g(x) + (A^T v)^T(z - x).$$

This shows that $A^T v$ is a subgradient of g at x, i.e., $A^T v \in \partial g(x)$. Therefore,

$$\partial g(x) \supseteq A^T \partial f(Ax + b).$$

(b) Proof that $\partial f(x) \supseteq$ conv $\left(\bigcup_{i:f_i(x)=f(x)} \partial f_i(x)\right)$

Let $S = \{i : f_i(x) = f(x)\}$ and let $v = \sum_{i \in S} \theta_i v_i$ where $v_i \in \partial f_i(x)$ and $\sum_{i \in S} \theta_i = 1$, $\theta_i \geq 0$.

Proof:

1. For any $z \in \mathbb{R}^n$:
$$f(z) = \max_i f_i(z) \geq f_i(z) \geq f_i(x) + v_i^T(z - x).$$
2. For $i \in S$:
$$f(z) \geq f_i(x) + v_i^T(z - x) = f(x) + v_i^T(z - x).$$
3. Multiplying by θ_i and summing:
$$f(z) \geq f(x) + [\sum_{i \in S} \theta_i v_i]^T(z - x).$$
4. Substituting v, we get:
$$f(z) \geq f(x) + v^T(z - x).$$

This inequality holds for all $z \in \mathbb{R}^n$, which means $v \in \partial f(x)$.
Since v is an arbitrary convex combination of elements from $\bigcup_{i \in S} \partial f_i(x)$, we conclude:

$$\text{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x)\right) \subseteq \partial f(x).$$

Therefore,

$$\partial f(x) \supseteq \text{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x)\right).$$

Conclusion: Both statements have been proven, demonstrating the relationships between the subgradients in each case.

2. More Subgradients

Show that for any convex function f, and for any points x, y, and $u \in \partial f(x)$, $v \in \partial f(y)$,

$$\langle u - v, x - y \rangle \geq 0.$$

Solution: To prove this inequality, we'll use the definition of subgradients and properties of convex functions.

Definitions and Properties:

1. **Subgradient at a point x:** A vector $u \in \mathbb{R}^n$ is a subgradient of f at x if
$$f(z) \geq f(x) + \langle u, z - x \rangle \quad \text{for all } z \in \mathbb{R}^n.$$
2. **Subdifferential $\partial f(x)$:** The set of all subgradients at x.
3. **Convex Function:** A function f is convex if its epigraph is a convex set, which implies the subgradient inequality holds.

Proof:

1. Since $u \in \partial f(x)$, we have:
$$f(y) \geq f(x) + \langle u, y - x \rangle. \tag{1}$$
Similarly, since $v \in \partial f(y)$, we have:
$$f(x) \geq f(y) + \langle v, x - y \rangle. \tag{2}$$

2. Adding inequalities (1) and (2):
$$(f(y) - f(x)) + (f(x) - f(y)) \geq \langle u, y - x \rangle + \langle v, x - y \rangle.$$
Simplifying the left side:
$$0 \geq \langle u, y - x \rangle + \langle v, x - y \rangle. \tag{3}$$

3. Note that $\langle v, x - y \rangle = -\langle v, y - x \rangle$, so:
$$\langle u, y - x \rangle + \langle v, x - y \rangle = \langle u, y - x \rangle - \langle v, y - x \rangle = \langle u - v, y - x \rangle.$$

4. Therefore:
$$0 \geq -\langle u - v, x - y \rangle.$$

5. Multiplying both sides by -1 (inequality direction reverses):
$$\langle u - v, x - y \rangle \geq 0.$$

Conclusion: We have shown that for any convex function f and any points $x, y \in \mathbb{R}^n$, with $u \in \partial f(x)$ and $v \in \partial f(y)$, the inequality $\langle u - v, x - y \rangle \geq 0$ holds. This confirms that the subdifferential of a convex function is a monotone operator.

This shows that the subdifferential is a monotone operator for any convex function. We will not use this specific property in this class, but it is heavily used for more advanced aspects of convex optimization and convex analysis.

3. Coordinate Descent

(a) Give an example that shows that coordinate descent may not find the optimum of a convex function. That is, provide a simple function f and a point x such that coordinate descent starting from x will not get to the global minimum of f.

(b) Let $f(x, y) = x^2 + 3xy$, where x, y are scalars. Note that f is not convex. Would coordinate descent with exact line search always converge to a stationary point?

Solution

(a) Example Showing Coordinate Descent May Not Find the Optimum of a Convex Function

Consider the function $f(x, y) = (x + y - 1)^2$. Let's analyze why coordinate descent starting from $(x_0, y_0) = (0, 1)$ fails to find the global minimum.

Analysis:

1. The function is convex as it's a squared linear function.
2. The global minimum occurs along the line $x + y = 1$.
3. Starting from $(0, 1)$:
   - When minimizing over x with $y = 1$: $f(x, 1) = (x + 1 - 1)^2 = x^2$
   - This gives $x_1 = 0$ (minimum of x^2)
   - When minimizing over y with $x = 0$: $f(0, y) = (y - 1)^2$
   - This gives $y_1 = 1$ (minimum of $(y - 1)^2$)
4. The algorithm stays at $(0, 1)$ indefinitely.
5. However, this is not the global minimum, as $f(0.5, 0.5) = 0 < f(0, 1) = 0$.

(b) Analysis of $f(x, y) = x^2 + 3xy$

Analysis of Partial Derivatives: The partial derivatives are:

$$\frac{\partial f}{\partial x} = 2x + 3y$$
$$\frac{\partial f}{\partial y} = 3x$$

Coordinate Descent Steps:

1. **Minimize over x with fixed y:**
   - Set $\frac{\partial f}{\partial x} = 0$: $2x + 3y = 0$
   - Optimal $x = -\frac{3}{2}y$
2. **Minimize over y with fixed x:**
   - If $x^{(1)} \neq 0$, then $\frac{\partial f}{\partial y} = 3x^{(1)} \neq 0$
   - This implies no finite minimizer exists for y
   - The function decreases without bound as:
     - $y \to -\infty$ if $x^{(1)} > 0$
     - $y \to \infty$ if $x^{(1)} < 0$

Key Observations:

- The function has a saddle point at $(0, 0)$
- Starting from any point where $y^{(0)} \neq 0$:
  - First iteration gives $x^{(1)} = -\frac{3}{2}y^{(0)} \neq 0$
  - Second iteration leads to unbounded descent in y
- Only when starting exactly at $y^{(0)} = 0$ will the method stay at the stationary point $(0, 0)$

Conclusion: No, coordinate descent with exact line search would not always converge to a stationary point for this function. The method can fail due to the function's non-convexity and the presence of a saddle point, causing the iterations to diverge or fail to find a stationary point.

4. Frank Wolfe and PGD

In the lectures, we showed an example where Frank-Wolge and projected gradient descent (PGD) behave very differently. Replicate that here, and show your plots of the trajectory obtained.

$$\begin{aligned} \min \quad & 100x_1^2 + x_2^2 + (x_3 - 20)^2 \\ \text{s.t.} \quad & x_1 + x_2 + x_3/20 = 1 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Solution

To replicate the example where Frank-Wolfe (FW) and Projected Gradient Descent (PGD) behave differently, we will solve the following constrained optimization problem:

$$\begin{aligned} \min \quad & f(x) = 100x_1^2 + x_2^2 + (x_3 - 20)^2 \\ \text{s.t.} \quad & x_1 + x_2 + \frac{x_3}{20} = 1 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Step 1: Reformulate the Problem

First, we simplify the problem by expressing x_3 in terms of x_1 and x_2:

$$x_3 = 20(1 - x_1 - x_2).$$

Step 2: Compare Trajectories

Frank-Wolfe Trajectory:

- Moves directly towards the minimizer by solving a linear problem at each iteration
- Makes large steps when possible, reaching the optimal point quickly

Projected Gradient Descent Trajectory:

- Takes a gradient step followed by a projection, which can lead to smaller effective steps if the gradient points outside the feasible region
- May require careful selection of the step size α_k to ensure convergence

Example Iteration:

- **Starting Point:** $x^{(0)} = (0.5, 0.5)$
- **Gradient at $x^{(0)}$:** Same as FW, $\nabla f(x^{(0)}) = (900, 801)$
- **Gradient Step with $\alpha_0 = 0.001$:**

$$y^{(0)} = (0.5, 0.5) - 0.001 \times (900, 801) = (-0.4, -0.301).$$

- **Projection onto D:** Since $y^{(0)}$ has negative components, the projection is $x^{(1)} = (0, 0)$
- **Result:** Similar to FW, PGD reaches $x^* = (0, 0)$ in one iteration

Observation

- In this specific problem, both FW and PGD converge to the optimal solution in one iteration when starting from $(0.5, 0.5)$
- However, in general, FW can exploit the structure of the feasible region to make more efficient moves, while PGD might struggle with projections that limit the step size

Conclusion

By applying both FW and PGD to the given problem, we observe that:

- **Frank-Wolfe** effectively leverages the linear structure of the feasible set to make significant progress towards the optimum
- **Projected Gradient Descent** may face challenges due to the need to project after each gradient step, which can slow down convergence if not properly managed

This example demonstrates how FW and PGD can behave differently depending on the problem structure and highlights the importance of choosing an appropriate optimization method for a given problem.