DSAI5104: Gradient Descent Tutorial Solutions

Problem 1: LASSO and Regularization

(a) **Convexity and Non-differentiability**

Let the objective function be $F(\beta) = f(\beta) + g(\beta)$, where:

$$f(\beta) = \frac{1}{2}\|X\beta - b\|_2^2 \quad \text{and} \quad g(\beta) = \lambda\|\beta\|_1 = \lambda\sum_{i=1}^{n}|\beta_i|.$$

Convexity:

- The term $f(\beta)$ is a quadratic form. Its Hessian is $\nabla^2 f(\beta) = X^\top X$. Since $X^\top X$ is always positive semi-definite (PSD), $f(\beta)$ is convex.
- The term $g(\beta)$ is a norm. By the triangle inequality $\|\alpha x + (1-\alpha)y\| \leq \alpha\|x\| + (1-\alpha)\|y\|$, all norms are convex functions.
- Since the sum of convex functions is convex, $F(\beta)$ is convex.

Non-differentiability: The term $g(\beta)$ involves the absolute value function $|x|$. The function $h(x) = |x|$ is not differentiable at $x = 0$ because the left and right limits of the derivative differ:

$$\lim_{h\to 0^-}\frac{|h| - 0}{h} = -1, \quad \lim_{h\to 0^+}\frac{|h| - 0}{h} = 1.$$

Therefore, the objective function $F(\beta)$ is non-differentiable at any point β where at least one component $\beta_j = 0$.

(b) **Gradient of the Loss Term**

Expanding the term $f(\beta)$:

$$f(\beta) = \frac{1}{2}(X\beta - b)^\top(X\beta - b) = \frac{1}{2}\left(\beta^\top X^\top X\beta - 2\beta^\top X^\top b + b^\top b\right).$$

Taking the derivative with respect to β:

$$\nabla f(\beta) = \frac{1}{2}\left(2X^\top X\beta - 2X^\top b\right) = X^\top X\beta - X^\top b = X^\top(X\beta - b).$$

(c) **Optimization Challenges and Sparsity**

Challenge: Standard Gradient Descent requires calculating $\nabla F(\beta)$. Since the gradient is undefined at $\beta_j = 0$, the algorithm may fail or oscillate near zero. We typically require Proximal Gradient methods (like ISTA) or Subgradient methods to handle the singularity.

Sparsity: The ℓ_1-norm ball $\{\beta \mid \|\beta\|_1 \leq C\}$ is a "diamond" shape (polytope) with sharp corners on the coordinate axes. The ℓ_2-norm ball is a sphere. When the elliptical contours of the loss function $f(\beta)$ expand to touch the regularization constraint:

- They are highly likely to touch a "corner" of the ℓ_1 diamond, setting some coordinates exactly to zero (sparsity).
- They will almost always touch the smooth surface of the ℓ_2 sphere at a non-axis point, resulting in small but non-zero values for all coefficients.

Problem 2: Convexity and Gradient Descent Directions

(a) **Descent Direction Proof**

Using the first-order Taylor expansion (or Mean Value Theorem) for f near x^k:

$$f(x^k + \alpha d^k) = f(x^k) + \alpha \nabla f(x^k)^\top d^k + o(\alpha).$$

Substitute the gradient descent direction $d^k = -\nabla f(x^k)$:

$$f(x^k + \alpha d^k) = f(x^k) - \alpha \|\nabla f(x^k)\|^2 + o(\alpha).$$

Since $\nabla f(x^k) \neq 0$, we have $\|\nabla f(x^k)\|^2 > 0$. For sufficiently small $\alpha > 0$, the linear term $-\alpha \|\nabla f(x^k)\|^2$ dominates the remainder $o(\alpha)$, implying:

$$f(x^k + \alpha d^k) < f(x^k).$$

(b) **Strong Convexity Convergence Rate**

Let f be μ-strongly convex. By definition:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Let $x = x^k$ and minimize both sides with respect to y. The unconstrained minimum of the quadratic RHS occurs at $y^* = x^k - \frac{1}{\mu}\nabla f(x^k)$. Substituting this back yields the Polyak-Lojasiewicz (PL) inequality:

$$f(x^*) \geq f(x^k) - \frac{1}{2\mu}\|\nabla f(x^k)\|^2 \implies \|\nabla f(x^k)\|^2 \geq 2\mu(f(x^k) - f(x^*)).$$

Now, using the standard descent lemma (which holds for suitable step size $\alpha \leq 1/L$):

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2.$$

Subtract $f(x^*)$ from both sides and substitute the PL lower bound for $\|\nabla f(x^k)\|^2$:

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{\alpha}{2}\left(2\mu(f(x^k) - f(x^*))\right) \\ f(x^{k+1}) - f(x^*) &\leq (f(x^k) - f(x^*))(1 - \mu\alpha). \end{aligned}$$

Problem 3: Lipschitz Continuity

(a) **Gradient Bound**

By definition, $|f(y) - f(x)| \leq L\|y - x\|$. Consider $y = x + tu$ for a unit vector u.

$$\left|\frac{f(x+tu) - f(x)}{t}\right| \leq L\left\|\frac{tu}{t}\right\| = L\|u\| = L.$$

Taking the limit as $t \to 0$, we get the directional derivative $|\nabla f(x)^\top u| \leq L$. Choosing $u = \frac{\nabla f(x)}{\|\nabla f(x)\|}$ gives:

$$\|\nabla f(x)\| \leq L.$$

(b) **Huber Loss Lipschitz Constant**

The derivative of $h_\delta(x)$ is:

$$h'_\delta(x) = \begin{cases} x & |x| \leq \delta \\ \delta \cdot \operatorname{sgn}(x) & |x| > \delta \end{cases}$$

We check the magnitude:

- If $|x| \leq \delta$, then $|h'_\delta(x)| = |x| \leq \delta$.
- If $|x| > \delta$, then $|h'_\delta(x)| = |\delta \cdot (\pm 1)| = \delta$.

In all cases, $|h'_\delta(x)| \leq \delta$. Thus, the function is L-Lipschitz with $L = \delta$.

Problem 4: Smoothness and Iteration compute

Let

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \qquad f(x) = \frac{1}{2} x^\top A x + b^\top x, \quad x \in \mathbb{R}^2.$$

(a) **Lipschitz Constant of Gradient**

We have

$$\nabla f(x) = Ax + b.$$

For any $x, y \in \mathbb{R}^2$,

$$\|\nabla f(x) - \nabla f(y)\| = \|A(x-y)\| \le \|A\|_2 \, \|x - y\|.$$

Since A is symmetric positive definite, $\|A\|_2 = \lambda_{\max}(A)$. Compute the eigenvalues:

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda) - 1 = \lambda^2 - 5\lambda + 5,$$

hence

$$\lambda_{1,2} = \frac{5 \pm \sqrt{5}}{2}.$$

Therefore the smallest Lipschitz constant is

$$L = \lambda_{\max}(A) = \frac{5 + \sqrt{5}}{2}.$$

(b) **Iteration compute**

First compute the gradient at x^k:

$$g_k = \nabla f(x^k) = Ax^k + b = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

Hence the gradient descent update (with $\alpha = \frac{1}{L}$) gives

$$x^{k+1} = x^k - \frac{1}{L} g_k = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{1}{L} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 - \frac{3}{L} \\ \frac{1}{L} \end{pmatrix}.$$

Problem 5: Convergence Rate Comparison

Parameters: $\epsilon = 10^{-4}$, $L = 10$, $D = 1$.

Case 1: Convex, Lipschitz (Non-smooth)

$$T_1 = \frac{D^2 L}{\epsilon^2} = \frac{1^2 \cdot 10}{(10^{-4})^2} = \frac{10}{10^{-8}} = 10^9 \text{ iterations}.$$

Case 2: Convex, L-smooth

$$T_2 = \frac{LD^2}{2\epsilon} = \frac{10 \cdot 1^2}{2 \cdot 10^{-4}} = \frac{5}{10^{-4}} = 50,000 \text{ iterations.}$$

Discussion: $T_1 = 1,000,000,000$ vs $T_2 = 50,000$. The smooth assumption allows for a drastically faster convergence rate ($O(1/\epsilon)$ vs $O(1/\epsilon^2)$) because gradients vary continuously, allowing larger, consistent steps towards the minimum.

Problem 6: Finite Termination

(a) **Update Rule derivation**

$$x^{k+1} = x^k - \alpha(Ax^k + b) = (I - \alpha A)x^k - \alpha b.$$

(b) **Case $A = cI$**

Here $L = c$, so step size $\alpha = 1/c$. The minimizer is $x^* = -A^{-1}b = -\frac{1}{c}b$. Start at arbitrary x^0:

$$\begin{aligned} x^1 &= (I - \frac{1}{c}(cI))x^0 - \frac{1}{c}b \\ &= (I - I)x^0 - \frac{1}{c}b \\ &= 0 - \frac{1}{c}b = x^*. \end{aligned}$$

Convergence occurs in exactly 1 step.

(c) **Piecewise-Linear-Quadratic Function**

Finite termination in 2 steps occurred because:

(a) **Step 1:** The iterate moved from the linear region into the quadratic region.

(b) **Step 2:** Once inside the quadratic region (where curvature is constant), the step size $\alpha = 1/L$ matched the inverse Hessian exactly (as in part b), leading immediately to the minimizer.

The function is **not** globally quadratic because the Hessian is not constant everywhere (it is A in the center and 0 in the linear regions).