DSAI5140: Accelerated Proximal Gradient Tutorial

Problems

Problem 1: (Supplementary information about convex functions and convex sets.)

(a) Prove that the following sets are convex sets.

i. Polyhedron
A (closed) polyhedron is a set of the form

$$\mathcal{P} \;=\; \{x \in \mathbb{R}^n \mid Ax \le b\},$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and the inequality is interpreted componentwise.

ii. ℓ_1-ball
For a radius $r > 0$, the ℓ_1-ball in $\mathbb{R}^n$ is

$$\mathcal{B}_1(r) \;=\; \{x \in \mathbb{R}^n \mid \|x\|_1 \le r\}, \qquad \|x\|_1 := \sum_{i=1}^{n} |x_i|.$$

iii. The positive semidefinite (PSD) cone
Let $\mathbb{S}^n$ be the space of $n \times n$ real symmetric matrices. The positive semidefinite (PSD) cone is

$$\mathbb{S}^n_+ \;=\; \{X \in \mathbb{S}^n \mid X \succeq 0\},$$

where $X \succeq 0$ means $v^\top X v \ge 0$ for all $v \in \mathbb{R}^n$.

(b) Convex function.

i. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by

$$f(x) := \|x\|_1.$$

Prove that f is convex, i.e., show that for all $t_1, t_2 \in \mathbb{R}$ and all $\theta \in [0, 1]$,

$$f\big(\theta x_1 + (1-\theta)x_2\big) \le \theta f(x_1) + (1-\theta) f(x_2).$$

ii. For convex functions f and g, prove that $f + g$ is also a convex function.

iii. The composition of convex functions $(f \circ g)(x) = f(g(x))$ is not always convex. Please try to construct a counterexample. (A sufficient condition for convexity is that the outer function is convex and nondecreasing and the inner function is convex.)

Problem 2: (Strong Convexity) A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-strongly convex ($\mu > 0$) if for all $x, y \in \mathbb{R}^n$, we have

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

(a) Show that if f is μ-strongly convex and has a minimizer x^*, then for all $x \in \mathbb{R}^n$,

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

(b) Consider the quadratic optimization problem $f(x) = \frac{1}{2}x^T Ax + b^T x$ in $\mathbb{R}^2$, where the Hessian matrix A and vector b are given by:

$$A = \begin{bmatrix} 6 & 2 \\ 2 & 9 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Determine the largest valid strong convexity parameter μ for this function. (Hint: Analyze the eigenvalues of A).

Problem 3: (Prox operators for ℓ_0 and ℓ_2) $\|x\|_0$ counts the number of nonzero components of x.

(a) Let $\lambda > 0$ be given. Derive a closed form for

$$\text{prox}_{\lambda\|\cdot\|_0}(v) = \arg\min_{x \in \mathbb{R}^n} \left\{ \lambda\|x\|_0 + \frac{1}{2}\|x - v\|_2^2 \right\}.$$

Hint: Reduce the problem to a scalar minimization for each coordinate. You should obtain a coordinate-wise thresholding rule.

(b) Let $\lambda > 0$ be given. Derive a closed form for

$$\text{prox}_{\lambda\|\cdot\|_2}(v) = \arg\min_{x \in \mathbb{R}^n} \left\{ \lambda\|x\|_2 + \frac{1}{2}\|x - v\|_2^2 \right\}.$$

Hint: Note that the l_2-norm is not differentiable at 0. If you have already learned about subdifferential, you can consider the optimality conditions for the minimization problem in the same way as for differentiable functions. Even if you don't know subdifferential, it's okay. You can rewrite $x = ru$ into minimization problem where $r = \|x\|_2 \geq 0$ and u is a unit vector ($\|u\|_2 = 1$).

(c) Calculate $\text{prox}_{\lambda\|\cdot\|_0}(\cdot)$, $\text{prox}_{\lambda\|\cdot\|_1}(\cdot)$ and $\text{prox}_{\lambda\|\cdot\|_2}(\cdot)$ for $v^{(1)} = (1,1)$ and $v^{(2)} = (2,0)$ when $\lambda = 1$.

Problem 4: Convergence of Proximal Gradient Descent (PGD) Consider the composite convex optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + g(x),$$

where f is convex, continuously differentiable, and has L-Lipschitz gradient, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y,$$

and g is proper, closed, and convex (possibly nonsmooth). Assume $X^\star := \arg\min F \neq \emptyset$.

Define the proximal gradient descent (PGD) iteration with step size $1/L$:

$$x_{k+1} = \text{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right).$$

Equivalently, define the quadratic upper model

$$Q_L(x; x_k) := f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|^2 + g(x),$$

and note that $x_{k+1} = \arg\min_x Q_L(x; x_k)$.

The goal of this problem is to prove its convergence. You can try to prove it directly, or refer to the following proof sequence.

(a) Using the smoothness of f, prove

$$F(x_{k+1}) \le Q_L(x_{k+1}; x_k).$$

(b) Show that $Q_L(\cdot; x_k)$ is L-strongly convex. Use this fact to prove the inequality: for all $x \in \mathbb{R}^n$,

$$Q_L(x; x_k) \ge Q_L(x_{k+1}; x_k) + \frac{L}{2}\|x - x_{k+1}\|^2.$$

(c) **(Monotonic decrease)** Combine the results of (1)–(2) to show that for all x,

$$F(x_{k+1}) \le Q_L(x; x_k) - \frac{L}{2}\|x - x_{k+1}\|^2.$$

Then use convexity of f to prove

$$F(x_{k+1}) \le F(x) + \frac{L}{2}\Big(\|x - x_k\|^2 - \|x - x_{k+1}\|^2\Big), \quad \forall x.$$

(d) **($O(1/k)$ convergence of objective values)** Let $x^\star \in X^\star$.

$$F(x_K) - F^\star \leq \frac{L}{2K}\|x_0 - x^\star\|^2,$$

where $F^\star = \min_x F(x)$.

Problem 5: (Background: Backtracking line search when L is unknown) In class we often set the stepsize as $\alpha = \frac{1}{L}$, but in practice L may be hard to estimate. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, and assume ∇f is L-Lipschitz for some (unknown) $L > 0$.

We consider the following *backtracking gradient descent* rule. Fix parameters $\eta > 1$ and an initial guess $\widehat{L}_0 > 0$. Given x^k and current guess $\widehat{L}$, repeat:

$$\text{(trial step)} \quad x^{k+1}(\widehat{L}) = x^k - \frac{1}{\widehat{L}}\nabla f(x^k),$$

and *accept* this $\widehat{L}$ if

$$f\Big(x^{k+1}(\widehat{L})\Big) \;\leq\; f(x^k) \;-\; \frac{1}{2\widehat{L}}\|\nabla f(x^k)\|_2^2. \tag{BT}$$

If (BT) fails, set $\widehat{L} \leftarrow \eta\widehat{L}$ and try again.

(a) Prove that for any $x \in \mathbb{R}^n$ and any $\widehat{L} \geq L$,

$$f\left(x - \frac{1}{\widehat{L}}\nabla f(x)\right) \;\leq\; f(x) \;-\; \frac{1}{2\widehat{L}}\|\nabla f(x)\|_2^2.$$

Remark: The backtracking loop must accept after finitely many trials at every iteration k.

(b) Suppose the initial guess at iteration k is $\widehat{L} = \widehat{L}_k$. Estimate the number of times you multiply by η before acceptance at most.

Problem 6: (Objective value of APGD is not necessarily monotone)

A key idea of APGD is to introduce an extrapolated point ("momentum"):

$$\begin{cases} y^k = x^k + \beta\,(x^k - x^{k-1}), \\ x^{k+1} = y^k - \alpha\nabla f(y^k). \end{cases}$$

Consider the simplest convex smooth function

$$f(x) = \frac{1}{2}x^2 \quad (x \in \mathbb{R}), \qquad \nabla f(x) = x,$$

and choose $\alpha = \frac{1}{2}$ and $\beta = 4$. Let $x^{-1} = x^0 = 1$.

(a) Compute x^1 and x^2 explicitly and show that the non-monotonicity of objective value.

(b) In one or two sentences, explain intuitively why introducing the extrapolated point y_k can cause overshooting (use only geometric/algorithmic intuition; no convergence proof required).

(Code implementation.)

(a) **(A simple restart rule)** A common practical fix is a *restart*: if $f(x_{k+1}) > f(x_k)$, then "reset the momentum" by setting

$$x_k = x_{k+1}, \qquad x_{k-1} = x_k,$$

(equivalently, force the next step to use $y_k = x_k$).

You can try writing code and observing how this restart mechanism behaves; you can also try different, more complex functions.

(b) **(Complexity sanity check)** The rates $O(1/k)$ (PGD) vs. $O(1/k^2)$ (APGD) can roughly estimate how many iterations are needed to reach accuracy $\varepsilon = 10^{-4}$:

$$\text{PGD: } \frac{1}{k} \lesssim \varepsilon, \qquad \text{APGD: } \frac{1}{k^2} \lesssim \varepsilon.$$

Try to write code to compare the convergence speed of the two methods, you can solve Lasso problem or change the l_1 norm to l_0 and l_2 norm or other more complex functions.