Mô hình hoá và tối ưu hoá trong học máy

Homework 2

▼ 1. Gradient descent convergence analysis

▼ 1.1. Nonconvex case

Here we will assume nothing about convexity of f. We will show that gradient descent reaches an ϵ-substationary point x, such that $\|\nabla f(x)\|_2 \leq \epsilon$, in $O(1/\epsilon^2)$ iterations. Important note: you may use here that

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2, \quad \text{for all } x, y. \tag{1}$$

Recall that you assumed convexity and twice differentiability of f on Homework 1 to show that the above is equivalent to the L-Lipschitz condition on ∇f. But (1) is in fact a consequence of ∇f being L-Lipschitz, and does not actually require convexity or twice differentiability of f.

▼ a.

(a, 2 pts) Plug in $y = x^+ = x - t\nabla f(x)$ to (1) to show that

$$f(x^+) \leq f(x) - \left(1 - \frac{Lt}{2}\right)t\|\nabla f(x)\|_2^2.$$

$$\begin{cases} f(y) \leq f(x) + \nabla f(x)^T(y-x) + \dfrac{L}{2}\|y-x\|_2^2 \\ y = x^+ = x - t\nabla f(x) \end{cases}$$

$$\Rightarrow f(x^+) \leq f(x) + \nabla f(x)^T(-t\nabla f(x)) + \frac{L}{2}\|-t\nabla f(x)\|_2^2$$

$$\Leftrightarrow f(x^+) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{t^2 L}{2}\|\nabla f(x)\|_2^2$$

$$\Leftrightarrow f(x^+) \leq f(x) - \left(1 - \frac{Lt}{2}\right)t\|\nabla f(x)\|_2^2$$

▼ b.

(b, 2 pts) Use $t \leq 1/L$, and rearrange the previous result, to get

$$\|\nabla f(x)\|_2^2 \leq \frac{2}{t}(f(x) - f(x^+)).$$

$$\begin{cases} f(x^+) \leq f(x) - \left(1 - \dfrac{Lt}{2}\right)t\|\nabla f(x)\|_2^2 \\ t \leq \dfrac{1}{L} \end{cases}$$

$$\Rightarrow f(x^+) \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2$$

$$\Leftrightarrow \|\nabla f(x)\|_2^2 \leq \frac{2}{t}(f(x) - f(x^+))$$

▼ c.

(c, 2 pts) Sum the previous result over all iterations from $1, \ldots, k+1$ to establish

$$\sum_{i=0}^{k}\|\nabla f(x^{(i)})\|_2^2 \leq \frac{2}{t}(f(x^{(0)}) - f^\star).$$

$$
\begin{aligned}
&f(x^+) \le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\
&\Rightarrow f(x) \ge f(x^+) + \frac{t}{2}\|\nabla f(x)\|_2^2 \\
&\Rightarrow f(x^{(0)}) \ge f(x^{(1)}) + \frac{t}{2}\|\nabla f(x^{(0)})\|_2^2 \\
&\ge f(x^{(2)}) + \frac{t}{2}\|\nabla f(x^{(1)})\|_2^2 + \frac{t}{2}\|\nabla f(x^{(0)})\|_2^2 \ge \dots \\
&\ge f^* + \frac{t}{2}\sum_{i=0}^{k}\|\nabla f(x^{(i)}\|_2^2 \\
&\Rightarrow \sum_{i=0}^{k}\|\nabla f(x)^{(i)}\|_2^2 \le \frac{2}{t}(f(x^{(0)} - f^*)
\end{aligned}
$$

▼ d.

(d, 2 pts) Lower bound the sum in the previous result to get

$$\min_{i=0,\ldots,k}\|\nabla f(x^{(i)})\|_2 \le \sqrt{\frac{2}{t(k+1)}(f(x^{(0)}) - f^\star)},$$

which establishes the desired $O(1/\epsilon^2)$ rate for achieving ϵ-substationarity.

$$
\begin{aligned}
&\sum_{i=0}^{k}\|\nabla f(x)^{(i)}\|_2^2 \ge (k+1)\min_{i=0,\ldots,k}\|\nabla f(x^{(i)}\|_2^2 \\
&\Rightarrow (k+1)\min_{i=0,\ldots,k}\|\nabla f(x^{(i)}\|_2^2 \le \frac{2}{t}(f(x^{(0)} - f^*) \\
&\Rightarrow \min_{i=0,\ldots,k}\|\nabla f(x^{(i)}\|_2 \le \sqrt{\frac{2}{t(k+1)}(f(x^{(0)} - f^*)}
\end{aligned}
$$

▼ 1.2. Convex case

Now we will assume that f is convex. We will show that gradient descent reaches an ϵ-suboptimal point x, such that $f(x) - f^\star \le \epsilon$, in $O(1/\epsilon)$ iterations. Going back to part (b) from the nonconvex case, we can rearrange this to get

$$f(x^+) \le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2. \tag{2}$$

Note that, by this property, we see that gradient descent is indeed a descent method for $t \le 1/L$ (it decreases the criterion at each iteration).

▼ a.

(a, 3 pts) Starting with (2), apply the first-order condition for convexity of f, to show

$$f(x^+) \le f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2.$$

$$
\begin{aligned}
&\begin{cases} f(x^+) \le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ f^* \ge f(x) + \nabla f(x)^T(x^* - x) \end{cases} \\
&\Rightarrow f(x^+) \le f^* + \nabla f(x)^T(x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2
\end{aligned}
$$

▼ b.

(b, 3 pts) From the previous result, show that

$$f(x^+) \leq f^\star + \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right).$$

- $x^* = x^+ - t\nabla f(x^+) = x - t\nabla f(x) - t\nabla f(x^+)$
  $\Rightarrow t(\nabla f(x) + \nabla f(x^+)) = x - x^*$
- $f^* + \nabla f(x)^T(x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2$
  $= f^* + t\nabla f(x)^T(\nabla f(x) + \nabla f(x^+)) - \frac{t}{2}\|\nabla f(x)\|_2^2$
  $= f^* + t\|\nabla f(x)\|_2^2 + t\nabla f(x)^T\nabla f(x^+) - \frac{t}{2}\|\nabla f(x)\|_2^2$
  $= f^* + \frac{1}{2t}(t^2\|\nabla f(x)\|_2^2 + 2t^2\nabla f(x)^T\nabla f(x^+) + t^2\|\nabla f(x^+)\|_2^2 - t^2\|\nabla f(x^+)\|_2^2)$
  $= f^* + \frac{1}{2t}(\|t\nabla f(x) + t\nabla f(x^+)\|_2^2 - \|t\nabla f(x^+)\|_2^2)$
  $= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$
- $f(x^+) \leq f^* + \nabla f(x)^T(x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2$
  $\Rightarrow f(x^+) \leq f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$

▼ c.

(c, 2 pts) Sum the previous result over all iterations $1, \ldots, k$ to get

$$\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2.$$

$f(x^{(k+1)}) \leq f^* + \frac{1}{2t}(\|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2)$

$\Rightarrow \frac{1}{2t}\|x^{(k)} - x^*\|_2^2 \geq f(x^{(k+1)}) - f^* + \frac{1}{2t}\|x^{(k+1)} - x^*\|_2^2$

$\Rightarrow \frac{1}{2t}\|x^{(0)} - x^*\|_2^2 \geq f(x^{(1)}) - f^* + \frac{1}{2t}\|x^{(1)} - x^*\|_2^2$

$\geq f(x^{(1)}) - f^* + f(x^{(2)}) - f^* + \frac{1}{2t}\|x^{(2)} - x^*\|_2^2 \geq \ldots$

$\geq \sum_{i=1}^{k}(f(x^{(i)} - f^*)$

$\Rightarrow \sum_{i=1}^{k}(f(x^{(i)} - f^*) \leq \frac{1}{2t}\|x^{(0)} - x^*\|_2^2$

▼ d.

(d, 2 pts) Use the fact that gradient descent is a descent method to lower bound the sum above, and conclude

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk},$$

which establishes the desired $O(1/\epsilon)$ rate for achieving ϵ-suboptimality.

$f(x^{(k)} \leq f(x^{(i)} \; \forall i = 1, \ldots, k$

$\Rightarrow \sum_{i=1}^{k}(f(x^{(i)} - f^*) \geq k(f(x^{(k)} - f^*)$

$$\Rightarrow f(x^{(k)}) - f^* \le \frac{\|x^{(0)} - x^*\|_2^2}{2tk}$$

▼ 2. Properties and examples of subgradients

▼ a.

(a, 2 pts) Show that $\partial f(x)$ is a closed and convex set for any function f (not necessarily convex) and any point x in its domain.

- $\partial f(x) = \{g \in \mathbb{R}^n : f(y) \ge f(x) + g^T(y-x)\ \forall y\}$
- Chứng minh $\partial f(x)$ là tập lồi.

  Chọn $g_1, g_2 \in \partial f(x) \Rightarrow \begin{cases} f(y) \ge f(x) + g_1^T(y-x) \\ f(y) \ge f(x) + g_2^T(y-x) \end{cases}$

  Xét $tg_1 + (1-t)g_2$:

  $$\begin{aligned} & f(x) + (tg_1 + (1-t)g_2)^T(y-x) \\ & = t(f(x) + g_1^T(y-x)) + (1-t)(f(x) + g_2^T(y-x)) \\ & \le tf(y) + (1-t)f(y) = f(y) \end{aligned}$$

  $\Rightarrow tg_1 + (1-t)g_2 \in \partial f(x)$

  $\Rightarrow \partial f(x)$ là tập lồi.

- Chứng minh $\partial f(x)$ là tập đóng.

  Đặt $S_y = \{g \in \mathbb{R}^n : f(y) \ge f(x) + g^T(y-x)\}$

  $\Rightarrow S_y$ là halfspace $\Rightarrow S_y$ là tập đóng

  mà $\partial f(x) = \bigcap_y S_y \Rightarrow \partial f(x)$ là tập đóng.

▼ b.

(b, 2 pts) Show that $g \in \partial f(x)$ if and only if $(g, -1)$ defines supporting hyperplane to epigraph of f at $(x, f(x))$ (i.e., $(g, -1)$ is the normal vector to this hyperplane).

- Phần trên đồ thị của f (epigraph of f):

  $epi(f) = \{(x,t) : x \in dom(f), f(x) \le t\}$

- Chứng minh $g \in \partial f(x) \Rightarrow (g, -1)$ định nghĩa một hyperplane tiếp xúc với phần trên đồ thị của f tại điểm $(x, f(x))$.

  $g \in \partial f(x) \Rightarrow f(y) \ge f(x) + g^T(y-x)$

  Nếu $(y,t) \in epi(f) \Rightarrow t \ge f(y) \ge f(x) + g^T(y-x)$

  $\Rightarrow g^T(y-x) - f(y) + f(x) \le 0$

  $$\Rightarrow \begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0 \qquad (1)$$

  $\Rightarrow A = \{(y,t) : (y,t)$ thoả mãn (1) $\}$ là một halfspace có bờ là một hyperplane tiếp xúc với $epi(f)$ tại điểm $(x, f(x))$.

  $\Rightarrow (g, -1)$ định nghĩa một hyperplane tiếp xúc với phần trên đồ thị của f tại điểm $(x, f(x))$.

- Chứng minh $(g, -1)$ định nghĩa một hyperplane tiếp xúc với phần trên đồ thị của f tại điểm $(x, f(x)) \Rightarrow g \in \partial f(x)$.

  $(g, -1)$ định nghĩa một hyperplane tiếp xúc với phần trên đồ thị của f tại điểm $(x, f(x)) \Rightarrow g \in \partial f(x)$

  $$\Rightarrow \begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ f(y) \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) = 0 \qquad \text{(P)}$$

  Chọn phần không gian A bờ là (P) sao cho $epi(f) \notin A$

  $$\Rightarrow \begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ f(y) \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0$$

  $\Rightarrow g^T(y-x) - f(y) + f(x) \le 0$

  $\Rightarrow f(y) \ge f(x) + g^T(y-x)$

$\Rightarrow g \in \partial f(x).$

▼ c.

(c, 2 pts) For a convex function f, show that if $x \in U$ where U is a open neighborhood in its domain, then

$$f(y) \geq f(x) + g^T(y-x), \quad \text{for all } y \in U \quad \Rightarrow \quad g \in \partial f(x).$$

In other words, if the tangent line inequality holds in a local open neighborhood of x, then it holds globally.

- Để chứng minh $g \in \partial f(x)$, ta cần chứng minh $f(y) \geq f(x) + g^T(y-x) \;\forall\; y \notin U$
- Chọn điểm $z \notin U$

  Do U là tập mở $\Rightarrow\; \exists\; 0 < t < 1 : y = tx + (1-t)z \in U$

  $\Rightarrow f(y) \geq f(x) + g^T(y-x) \;\;\forall\;\; y \in U$

  $\Rightarrow f(tx + (1-t)z) \geq f(x) + g^T[tx + (1-t)z - x]$

  mà f là hàm lồi $\Rightarrow f(tx + (1-t)z) \leq tf(x) + (1-t)f(z)$

  $\Rightarrow tf(x) + (1-t)f(z) \geq f(x) + g^T[tx + (1-t)z - x]$

  $\Leftrightarrow (t-1)f(x) + (1-t)f(z) \geq g^T[(t-1)x + (1-t)z]$

  $\Leftrightarrow (1-t)[f(z) - f(x)] \geq (1-t)g^T(z-x)$

  $\Leftrightarrow f(z) \geq f(x) + g^T(z-x) \;\forall\; z \notin U$

$\Rightarrow g \in \partial f(x)$

▼ d.

(d, 1 pt) For a convex function f and subgradients $g_x \in \partial f(x)$, $g_y \in \partial f(y)$, prove that

$$(g_x - g_y)^T(x-y) \geq 0.$$

This property is called *monotonicity* of the subdifferential ∂f.

$g_x \in \partial f(x) \Rightarrow f(y) \geq f(x) + g_x^T(y-x)$

$g_y \in \partial f(y) \Rightarrow f(x) \geq f(y) + g_y^T(x-y)$

$\Rightarrow (g_x - g_y)^T(x-y) \geq 0$

▼ e.

(e, 2 pts) For $f(x) = \|x\|_2$, show that all subgradients $g \in \mathbb{R}^n$ at a point $x \in \mathbb{R}^n$ are of the form

$$g \in \begin{cases} \{x/\|x\|_2\} & x \neq 0 \\ \{v : \|v\|_2 \leq 1\} & x = 0. \end{cases}$$

- Với $x \neq 0 \Rightarrow \nabla f(x) = \dfrac{x}{\|x\|_2} \Rightarrow g \in \left\{\dfrac{x}{\|x\|_2}\right\}$
- Với $x = 0$

  $\Rightarrow g \in \partial f(0) = \{v : f(y) \geq f(0) + v^T(y-0) \quad \forall y\}$

  $\Rightarrow \|y\|_2 \geq v^T y$

  $\Rightarrow \|v\|_2 \leq 1$

  $\Rightarrow g \in \{v : \|v\|_2 \leq 1\}$

▼ f.

(f, 3 pts) For $f(x) = \max_{s \in S} f_s(x)$, where each f_s is convex, show that

$$\partial f(x) \supseteq \operatorname{conv}\left(\bigcup_{s: f_s(x) = f(x)} \partial f_s(x) \right).$$

.Xét $f(x) = \max_{s \in S} f_s(x)$:

$\Rightarrow f(y) \geq f_s(y) \geq f_s(x) + g^T(y - x) = f(x) + g^T(y - x)$

$\Rightarrow \forall\, g \in \partial f_s(x) : g \in \partial f(x)$

$\Rightarrow \bigcup\limits_{s: f_s(x) = f(x)} \partial f_s(x) \subseteq \partial f(x)$

mà $\partial f(x)$ là tập lồi

$\Rightarrow conv\left(\bigcup\limits_{s: f_s(x) = f(x)} \partial f_s(x) \right) \subseteq \partial f(x)$

▼ 3. Properties and examples of proximal operators

We will inspect various properties and examples of proximal operators. Unless otherwise specified, take h to be a convex function with domain $\operatorname{dom}(h) = \mathbb{R}^n$, and $t > 0$ be arbitrary, and consider its associated proximal operator

$$\operatorname{prox}_{h,t}(x) = \underset{z}{\operatorname{argmin}}\ \frac{1}{2t}\|x - z\|_2^2 + h(z).$$

▼ a.

(a, 3 pts) Prove that $\operatorname{prox}_{h,t}$ is a well-defined function on $\mathbb{R}^n$, that is, each point $x \in \mathbb{R}^n$ gets mapped to a unique value $\operatorname{prox}_{h,t}(x)$.

- Để chứng minh $prox_{h,t}(x)$ là hàm "well-defined", ta cần chứng minh $f(z) = \dfrac{1}{2t}\|x - z\|_2^2 + h(z)$ là hàm lồi chặt.
- Thật vậy, với $0 < \alpha < 1$, ta có:

  $f(\alpha z_1 + (1 - \alpha) z_2)$

  $= \dfrac{1}{2t}\|x - (\alpha z_1 + (1 - \alpha) z_2)\|_2^2 + h(\alpha z_1 + (1 - \alpha) z_2)$

  $= \dfrac{1}{2t}\|\alpha(x - z_1) + (1 - \alpha)(x - z_2)\|_2^2 + h(\alpha z_1 + (1 - \alpha) z_2)$

  $\leq \alpha \left(\dfrac{1}{2t}\|x - z_1\|_2^2 + h(z_1) \right) + (1 - \alpha) \left(\dfrac{1}{2t}\|x - z_2\|_2^2 + h(z_2) \right) - \dfrac{1}{2t}\alpha(1 - \alpha)\|z_1 - z_2\|_2^2$

  $< \alpha f(z_1) + (1 - \alpha) f(z_2)$ (Do z_1, z_2 là hai điểm phân biệt)

  $\Rightarrow f(z)$ là hàm lồi chặt (strictly convex function)

  $\Rightarrow \arg\min\limits_z f(z)$ chỉ có duy nhất một nghiệm

  $\Rightarrow prox_{h,t}$ là hàm "well-defined".

▼ b.

(b, 2 pts) Prove that $\operatorname{prox}_{h,t}(x) = u$ if and only if

$$h(y) \geq h(u) + \frac{1}{t}(x - u)^T(y - u), \quad \text{for all } y.$$

Hint: use subgradient optimality.

$prox_{h,t}(x) = \arg\min\limits_z \dfrac{1}{2t}\|x - z\|_2^2 + h(z) = u$

$$\Leftrightarrow \frac{1}{2t}\|x-u\|_2^2 + h(u) = \min_y \frac{1}{2t}\|x-y\|_2^2 + h(y)\ \forall\ y$$

$$\Leftrightarrow 0 \in \left\{-\frac{1}{t}(x-u)\right\} + \partial h(u)$$

$$\Leftrightarrow \frac{1}{t}(x-u) \in \partial h(u)$$

$$\Leftrightarrow h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u)\ \forall\ y$$

▼ c.

(c, 6 pts) Prove that $\text{prox}_{h,t}$ is nonexpansive, meaning

$$\|\text{prox}_{h,t}(x) - \text{prox}_{h,t}(y)\|_2 \leq \|x-y\|_2, \quad \text{for all } x, y.$$

Hint: use the previous question, and the monotonicity of subgradients from Q2(d).

Đặt $\begin{cases} prox_{h,t}(x) = u \\ prox_{h,t}(y) = v \end{cases}$

$$\Rightarrow \begin{cases} \frac{1}{t}(x-u) \in \partial h(u) \\ \frac{1}{t}(y-v) \in \partial h(v) \end{cases}$$

$$\Rightarrow (x-u-y+v)^T(u-v) \geq 0$$

$$\Rightarrow \|u-v\|_2^2 \leq (x-y)^T(u-v) \leq \|x-y\|_2 \cdot \|u-v\|_2$$

$$\Rightarrow \|u-v\|_2 \leq \|x-y\|_2$$

$$\Rightarrow \|prox_{h,t}(x) - prox_{h,t}(y)\|_2 \leq \|x-y\|_2$$

▼ d.

(d, 3 pts) The proximal minimization algorithm (a special case of proximal gradient descent) repeats the updates:

$$x^{(k+1)} = \text{prox}_{h,t}(x^{(k)}), \quad k = 1, 2, 3, \ldots.$$

Write out these updates when applied to $h(x) = \frac{1}{2}x^TAx - b^Tx$, where $A \in \mathbb{S}^n$. Show that this is equivalent to the *iterative refinement* algorithm for solving the linear system $Ax = b$:

$$x^{(k+1)} = x^{(k)} + (A + \epsilon I)^{-1}(b - Ax^{(k)}), \quad k = 1, 2, 3, \ldots,$$

where $\epsilon > 0$ is some constant. **Bonus (1 pt):** assuming that proximal minimization converges to the minimizer of $h(x) = \frac{1}{2}x^TAx - b^Tx$ (which is does, under suitable step sizes), what would the iterations of iterative refinement converge to in the case when A is singular, $Ax = b$, and $x^{(0)} = 0$?

- Tính x^+:

$$x^+ = prox_{h,t}(x) = \arg\min_z \left(\frac{1}{2t}\|x-z\|_2^2 + h(z)\right)$$

$$= \arg\min_z \left(\frac{1}{2t}\|x-z\|_2^2 + \frac{1}{2}z^TAz - b^Tz\right)$$

$$= \left(\frac{1}{t}I + A\right)^{-1}\left(\frac{1}{t}x + b\right)$$

- Chứng minh $x^+ = x + (A + \epsilon I)^{-1}(b - Ax)$

$$x^+ = \left(\frac{1}{t}I + \frac{1}{2}\left(A + A^T\right)\right)^{-1}\left(\frac{1}{t}x + b\right)$$

$$= \left(\frac{1}{t}I + A\right)^{-1}\left(\frac{1}{t}x + Ax + b - Ax\right)$$

$$= \left(\frac{1}{t}I + A\right)^{-1} \cdot \left(\frac{1}{t}I + A\right) \cdot x + \left(\frac{1}{t}I + A\right)^{-1} \cdot (b - Ax)$$

$$= x + \left(\frac{1}{t}I + A\right)^{-1} \cdot (b - Ax)$$

Đặt $\epsilon = \frac{1}{t}(\epsilon > 0) \Rightarrow x^+ = x + (A + \epsilon I)^{-1}(b - Ax)$

▼ e.

(e, 8 pts) For a matrix-variate function h, we define its proximal operator as

$$\operatorname{prox}_{h,t}(X) = \underset{Z}{\operatorname{argmin}} \ \frac{1}{2t}\|X - Z\|_F^2 + h(Z),$$

For $h(X) = \|X\|_{\text{tr}}$, show that the proximal operator evaluated at $X = U\Sigma V^T$ (this is an SVD of X) is so-called matrix soft-thresholding,

$$\operatorname{prox}_{h,t}(X) = U\Sigma_t V^T, \quad \text{where } \Sigma_t = \operatorname{diag}\Big((\Sigma_{11} - t)_+, \dots, (\Sigma_{nn} - t)_+\Big),$$

and $x_+ = \max\{x, 0\}$ denotes the positive part of x. Hint: start with subgradient optimality as you developed in Q3(b), and use the subgradients of the trace norm from Q2(g).

Đặt $prox_{h,t}(X) = M$

$\Rightarrow M \in X - t\partial\|M\|_{tr}$

$\Rightarrow M \in U\Sigma V^T - t\{UV^T + W : \|W\|_{op} \leq 1, U^T W = 0, WV = 0\}$

$\Rightarrow M = U\Sigma_t V^T$, với $(\Sigma_t)_{ii} = \max\{\Sigma_{ii} - t, 0\}$

$\Rightarrow X_{ij} = \max\{X_{ij}, 0\}$

▼ 4. Group lasso logistic regression

Suppose we have features $X \in \mathbb{R}^{n\times(p+1)}$ that we divide into J groups:

$$X = \Big[\mathbb{1} \ X_{(1)} \ X_{(2)} \ \cdots \ X_{(J)}\Big],$$

where $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^n$ and each $X_{(j)} \in \mathbb{R}^{n\times p_j}$. To achieve sparsity over groups of features, rather than individual features, we can use a *group lasso* penalty. Write $\beta = (\beta_0, \beta_{(1)}, \dots, \beta_{(J)}) \in \mathbb{R}^{p+1}$, where β_0 is an intercept term and each $\beta_{(j)} \in \mathbb{R}^{p_j}$. Consider the problem

$$\min_{\beta} \ g(\beta) + \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2, \tag{3}$$

where g is a loss function and $\lambda \geq 0$ is a tuning parameter. The penalty $h(\beta) = \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2$ is called the group lasso penalty. A common choice for w_j is $\sqrt{p_j}$ to adjust for the group size.

▼ a.

(a, 3 pts) Derive the proximal operator $\operatorname{prox}_{h,t}(\beta)$ for the group lasso penalty defined above.

$$prox_{h,t}(\beta) = \arg\min_{z} \frac{1}{2t}\|\beta - z\|_2^2 + \lambda \sum_{j=1}^{J} w_j \|z_{(j)}\|_2$$

$$prox_{h,t}(\beta)_{(j)} = prox_{tw_j\lambda\|\cdot\|_2}(\beta)_{(j)} = \beta_{(j)} - tw_j\lambda \cdot proj_{\mathrm{B}_{\|\cdot\|_2}}\left(\frac{\beta}{tw_j\lambda}\right)_{(j)}$$

$$= \begin{cases} \beta_{(j)} - tw_j\lambda \cdot \dfrac{\frac{\beta_{(j)}}{tw_j\lambda}}{\left\|\frac{\beta_{(j)}}{tw_j\lambda}\right\|_2} & if \ \left\|\frac{\beta_{(j)}}{tw_j\lambda}\right\|_2 > 1 \\ \beta_{(j)} - tw_j\lambda \cdot \dfrac{\beta_{(j)}}{tw_j\lambda} & if \ \left\|\frac{\beta_{(j)}}{tw_j\lambda}\right\|_2 \leq 1 \end{cases}$$

$$= \max\left(0, 1 - \frac{tw_j\lambda}{\|\beta_{(j)}\|_2}\right)\beta_{(j)}$$

▼ b.

(b, 2 pts) Let $y \in \{0,1\}^n$ be a binary label, and let g be the logistic loss

$$g(\beta) = -\sum_{i=1}^{n} y_i (X\beta)_i + \sum_{i=1}^{n} \log(1 + \exp\{(X\beta)_i\}),$$

Write out the steps for proximal gradient descent applied to the logistic group lasso problem (3) in explicit detail.

- $(\nabla g(\beta))_k = \left[\sum_{i=1}^{n} X_{ik} \left(\frac{e^{(X\beta)_i}}{1 + e^{(X\beta)_i}} - y_i \right)\right]$ với $k = 1, 2, ..., p+1$

  $\Rightarrow \nabla g(\beta) = \left(\frac{e^{(X\beta)}}{1 + e^{(X\beta)}} - y \right)^T X$
- $x^+_{(j)} = prox_{h,t}(x - t\nabla g(x))_{(j)}$

  $= \max\left(0, 1 - \frac{tw_j \lambda}{\|(x - t\nabla g(x))_{(j)}\|_2}\right)(x - t\nabla g(x))_{(j)}$