COMS 4995-004: Optimization for Machine Learning
Homework 1.

Question 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable convex function. In this question, we will prove that $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$. We will prove this by showing that for all vectors $u \in \mathbb{R}^d$, we have $u^\top \nabla^2 f(x) u \geq 0$.

(a) **(2 points)** Let $x, y \in \mathbb{R}^d$. Prove that

$$\int_{t=0}^{1} (1-t) \frac{\partial^2 f(x+ty)}{\partial^2 t} dt = f(x+y) - f(x) - \nabla f(x)^\top y.$$

(Hint: think about integration by parts.)

(b) **(2 points)** Prove that

$$\frac{\partial^2 f(x+ty)}{\partial^2 t} = y^\top \nabla^2 f(x+ty) y.$$

(c) **(3 points)** Set $y = \alpha u$, for some $\alpha \in \mathbb{R}$. Using the convexity of f and parts (a) and (b), show that there exists a $t' \in [0,1]$ such that $u^\top \nabla^2 f(x + t'\alpha u) u \geq 0$. *(Hint: use the mean-value theorem on the integral in part (a).)*

(d) **(2 points)** Show that part (c) implies that $u^\top \nabla^2 f(x) u \geq 0$.

Solution.
1(a). Using integration by parts,

$$\int_{t=0}^{1} (1-t) \frac{\partial^2 f(x+ty)}{\partial^2 t} dt = \left[(1-t) \frac{\partial f(x+ty)}{\partial t} \right]_{t=0}^{1} - \int_{t=0}^{1} -1 \cdot \frac{\partial f(x+ty)}{\partial t} dt.$$

We have $\frac{\partial f(x+ty)}{\partial t} = \nabla f(x+ty)^\top y$ so the first term on the RHS above equals $-\nabla f(x)^\top y$. By the fundamental theorem of calculus, the second term equals $f(x+y) - f(y)$.

1(b). We have $\frac{\partial f(x+ty)}{\partial t} = \nabla f(x+ty)^\top y$. By the chain rule, $\frac{\partial \nabla f(x+ty)}{\partial t} = \nabla^2 f(x+ty) y$. Putting these together, we get $\frac{\partial^2 f(x+ty)}{\partial^2 t} = y^\top \nabla^2 f(x+ty) y$.

1(c). Assume that $\alpha \neq 0$. The case $\alpha = 0$ is handled in 1(d).

By the convexity of f, we have $f(x+y) - f(x) - \nabla f(x)^\top y \geq 0$. Consider the case when $f(x+y) - f(x) - \nabla f(x)^\top y > 0$. Thus $\int_{t=0}^{1} (1-t) \frac{\partial^2 f(x+ty)}{\partial^2 t} dt > 0$. By the intermediate value theorem, there exists a $t' \in [0,1]$ such that $(1-t') \left. \frac{\partial^2 f(x+ty)}{\partial^2 t} \right|_{t=t'} = \int_{t=0}^{1} (1-t) \frac{\partial^2 f(x+ty)}{\partial^2 t} dt > 0$. Thus, using 1(b), we have $(1-t')y^\top \nabla f(x+t'y)y > 0$. Setting $y = \alpha u$, we have $(1-t')\alpha^2 u^\top f(x+t'\alpha u)u > 0$, which implies that $u^\top f(x + t'\alpha u)u > 0$ for $\alpha \neq 0$.

Now we consider the case when $f(x + y) - f(x) - \nabla f(x)^\top y = 0$. Then $\int_{t=0}^{1}(1 - t)\frac{\partial^2 f(x+ty)}{\partial^2 t}dt = 0$. Since $\frac{\partial^2 f(x+ty)}{\partial^2 t} \geq 0$, we conclude that $\frac{\partial^2 f(x+ty)}{\partial^2 t} = 0$ for all $t \in [0, 1]$, which implies that $u^\top f(x + t'\alpha u)u = 0$ for all $t' \in [0, 1]$ when $\alpha \neq 0$.

1(d). Part 1(c) implies that for every $\alpha \in \mathbb{R}$, there exists a $t' \in [0, 1]$ such that $u^\top f(x+t'\alpha u)u \geq 0$. Now let $\alpha \to 0$. Note that $t'\alpha \to 0$ since $t' \in [0, 1]$. Assuming $\nabla^2 f(\cdot)$ is continuous[1], we conclude that $u^\top f(x)u \geq 0$.

Question 2. Consider the following training set: $S = \{(x_i, y_i) \in \mathbb{R}^3 \times \mathbb{R} \mid i = 1, 2, 3\}$, where

$$\begin{aligned}(x_1, y_1) &= ((2, 0, 0), 1)\\ (x_2, y_2) &= ((0, 1, 0), -1)\\ (x_3, y_3) &= ((0, 0, 0.5), 1).\end{aligned}$$

Suppose we want to train a linear predictor $f_w = \langle w, x\rangle$ for some weight vector $w \in \mathbb{R}^3$. Consider training the predictor using the following three loss functions and regularization functions:

(i) (Square loss with no regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, no regularization.

(ii) (Square loss with ℓ_1 regularization) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, regularization $R(w) = \|w\|_1$, regularization constant $\lambda = 1$.

(iii) (Logistic loss with no regularization) loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, no regularization.

(iii) (Logistic loss with ℓ_2 regularization) loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, regularization $R(w) = \frac{1}{2}\|w\|_2^2$, regularization constant $\lambda = 1$.

For training loss function in each of the above cases, answer the following questions:

1. **(2 points per function)** Give formulas for the gradient (or subgradient, if the function is not differentiable) and Hessian (if it exists) as a function of w.
2. **(1 point per function)** Is the function strongly convex? If yes, compute a lower bound on the strong convexity constant μ. Try to make it as tight as possible.
3. **(1 point per function)** Is the function smooth? If yes, compute an upper bound on the smoothness constant β. Try to make it as tight as possible.

Solution.
The training loss function is $L(w) = \frac{1}{3}\sum_{i=1}^{3} \ell(\langle w, x_i\rangle, y_i) + \lambda R(w)$. Using the chain rule, the gradient is

$$\nabla L(w) = \frac{1}{3}\sum_{i=1}^{3} \ell'(\langle w, x_i\rangle, y_i)x_i + \lambda \nabla R(w),$$

[1]This was inadvertently not specified in the problem description. As mentioned on Piazza it's fine to make this assumption.

where $\ell'(\hat{y}, y) := \frac{d\ell(\hat{y},y)}{d\hat{y}}$ if $\ell(\cdot, y)$ is differentiable at $\hat{y}$ or a subderivative otherwise, and $\nabla R(w)$ is the gradient of R at w if it is differentiable at w or the subgradient otherwise, with some abuse of notation. Similarly, again using the chain rule, the Hessian is

$$\nabla L(w) = \frac{1}{3}\sum_{i=1}^{3} \ell''(\langle w, x_i\rangle, y_i)x_i x_i^\top + \lambda \nabla^2 R(w),$$

where $\ell''(\hat{y}, y) := \frac{d^2\ell(\hat{y},y)}{d\hat{y}^2}$, assuming the second derivatives exist. We now apply these formulas to the specific loss and regularization functions in the problem.

(i) Loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, no regularization.

1. Here, $\ell'(\hat{y}, y) = 2(\hat{y} - y)$ and $\ell''(\hat{y}, y) = 2$. Thus we have

   $$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3}(\langle w, x_i\rangle - y_i)x_i$$

   and

   $$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3} x_i x_i^\top = \frac{2}{3}\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}.$$

2. $L(w)$ is strongly convex since its Hessian given above is positive definite. The smallest eigenvalue of the Hessian is $\frac{2}{3} \cdot 0.25 = \frac{1}{6}$, so the tightest strong convexity constant equals $\frac{1}{6}$.

3. $L(w)$ is smooth since all eigenvalues of its Hessian are bounded by $\frac{8}{3}$. The tightest smoothness constant equals $\frac{8}{3}$.

(ii) loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$, regularization $R(w) = \|w\|_1$, regularization constant $\lambda = 1$.

1. Here, $\ell'(\hat{y}, y) = 2(\hat{y} - y)$ and $\ell''(\hat{y}, y) = 2$. $\|w\|_1$ is not differentiable whenever there is a coordinate that equals 0. For any coordinate $w_i \neq 0$, we have $\frac{\partial \|w\|_1}{\partial w_i} = \frac{w_i}{|w_i|}$, and for any coordinate $w_i = 0$, the subdifferential set w.r.t. w_i is $[-1, 1]$. Thus one possible subgradient of $\|w\|_1$ is $\langle \mathrm{sgn}(w_1), \mathrm{sgn}(w_2), \ldots, \mathrm{sgn}(w_d)\rangle$, where $\mathrm{sgn} : \mathbb{R} \to [-1, 1]$ is defined as

   $$\mathrm{sgn}(u) = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0. \end{cases}$$

   The setting $\mathrm{sgn}(0) = 0$ is an arbitrary choice, it can be set to any number in $[-1, 1]$.

   A subgradient of $L(w)$ can thus be given as

   $$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3}(\langle w, x_i\rangle - y_i)x_i + \langle \mathrm{sgn}(w_1), \mathrm{sgn}(w_2), \ldots, \mathrm{sgn}(w_d)\rangle.$$

   The Hessian only exists at points w which have no zero coordinates. At such points, $\nabla^2\|w\|_1 = 0$, and this at such points,

   $$\nabla L(w) = \frac{2}{3}\sum_{i=1}^{3} x_i x_i^\top = \frac{2}{3}\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}.$$

2. *(Note: in class, we have only defined strong convexity for differentiable functions. Hence the following answer will be considered valid.)* $L(w)$ is not strongly convex since it is not differentiable everywhere.

   (Note: strong convexity of a function f can be more generally defined as the following condition for any two points x, y: $f(y) \geq f(x) + g^\top(y-x) + \frac{\alpha}{2}\|y-x\|^2$, where g is a subgradient of f at x. We will adopt this definition moving forward in the class. Several students have given the following answer assuming this definition. This is also a valid answer.) We have $L(w) = \frac{1}{3}\sum_{i=1}^{3}(\langle w, x_i\rangle - y_i)^2 + \|w\|_1$. While the $\|w\|_1$ is just convex but not strongly convex, as in part (i) of this question, $\frac{1}{3}\sum_{i=1}^{3}(\langle w, x_i\rangle - y_i)^2$ is $\frac{1}{6}$-strongly convex. Thus, $L(w)$ is also $\frac{1}{6}$-strongly convex.

3. $L(w)$ is not smooth since it is not differentiable everywhere. *(Note: unlike strong convexity, a smooth function is automatically differentiable everywhere, hence it doesn't make sense to define it in terms of subgradients.)*

(iii) Loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, no regularization.

1. Here, $\ell'(\hat{y}, y) = \frac{-\exp(-\hat{y}y)y}{1+\exp(-\hat{y}y)}$ and $\ell''(\hat{y}, y) = \frac{\exp(-\hat{y}y)y^2}{(1+\exp(-\hat{y}y))^2}$. Thus we have

$$\nabla L(w) = \frac{1}{3}\sum_{i=1}^{3}\frac{-\exp(-\langle w, x_i\rangle y_i)y_i}{1+\exp(-\langle w, x_i\rangle y_i)}x_i$$

   and

$$\nabla^2 L(w) = \frac{1}{3}\sum_{i=1}^{3}\frac{\exp(-\langle w, x_i\rangle y_i)y_i^2}{(1+\exp(-\langle w, x_i\rangle y_i))^2}x_i x_i^\top.$$

2. Using the specific values of (x_i, y_i), we can also write the Hessian as

$$\nabla^2 L(w) = \operatorname{diag}\left(\frac{4\exp(-2w_1)}{3(1+\exp(-2w_1))^2}, \frac{\exp(w_2)}{3(1+\exp(w_2))^2}, \frac{0.25\exp(-0.5w_3)}{3(1+\exp(-0.5w_3))^2)}\right),$$

   where $\operatorname{diag}(a, b, c)$ is the diagonal matrix with a, b, c on the diagonal. Since the Hessian is a diagonal matrix, its eigenvalues are exactly the diagonal entries. Now if we let $w_1 \to -\infty$, then the first diagonal entry goes to 0, which means that there is no $\alpha > 0$ such that $\nabla^2 L(w) \succeq \alpha\mathbf{I}$ for all w. Hence, $L(w)$ is not strongly convex.

3. To analyze the smoothness of $L(w)$, we note that all the eigenvalues of $\nabla^2 L(w)$ are of the form $\frac{cu}{(1+u)^2}$, where c is a constant and $u \geq 0$. Note that[2] $\frac{u}{(1+u)^2} \leq \frac{1}{4}$ for all u, with equality when $u = 1$. Thus $\frac{cu}{(1+u)^2} \leq \frac{c}{4}$, and hence the diagonal entries of $\nabla^2 L(w)$ are bounded by $\frac{1}{3}, \frac{1}{12}, \frac{1}{48}$ respectively, with all these bounds simultaneously attained when $w_1 = w_2 = w_3 = 0$. Thus, the tightest smoothness constant is $\frac{1}{3}$.

(iii) Loss function $\ell(\hat{y}, y) = \log(1 + \exp(-\hat{y}y))$, regularization $R(w) = \frac{1}{2}\|w\|_2^2$, regularization constant $\lambda = 1$.

[2] $\frac{u}{(1+u)^2} \leq \frac{1}{4} \Leftrightarrow 4u \leq (1+u)^2 \Leftrightarrow 0 \leq (1-u)^2$

1. We can reuse the calculations from (iii) along with the facts that $\nabla R(w) = w$ and $\nabla^2 R(w) = \mathbf{I}$ to get

$$\nabla L(w) = \frac{1}{3}\sum_{i=1}^{3} \frac{-\exp(-\langle w, x_i\rangle y_i) y_i}{1+\exp(-\langle w, x_i\rangle y_i)} x_i + w$$

and

$$\nabla^2 L(w) = \frac{1}{3}\sum_{i=1}^{3} \frac{\exp(-\langle w, x_i\rangle y_i) y_i^2}{(1+\exp(-\langle w, x_i\rangle y_i))^2} x_i x_i^\top + \mathbf{I}.$$

2. We can rewrite the Hessian as

$$\nabla^2 L(w) = \operatorname{diag}\left(\frac{4\exp(-2w_1)}{3(1+\exp(-2w_1))^2} + 1, \frac{\exp(w_2)}{3(1+\exp(w_2))^2} + 1, \frac{0.25\exp(-0.5w_3)}{3(1+\exp(-0.5w_3))^2)} + 1\right).$$

All eigenvalues of the Hessian are at least 1, attained when $w_1 \to -\infty$. Thus, the tightest strong convexity constant is 1.

3. Reasoning as in (iii), the eigenvalues of the Hessian are bounded by $\frac{4}{3}, \frac{13}{12}, \frac{49}{48}$ respectively. Thus the tightest smoothness constant is $\frac{4}{3}$.