CS 726: Homework #2

Elaine Chiu, Department of Statistics

Please typeset your solutions.

Note: You can use the results we have proved in class – no need to prove them again.

Q 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function that is μ-strongly convex for some $\mu > 0$. Let $L > 0$ and let $\mathcal{X}$ be a closed convex set.

1. Under what conditions (on μ, L, $\mathcal{X}$) can f be L-Lipschitz continuous on $\mathcal{X}$? [10pts]

2. Under what conditions (on μ, L, $\mathcal{X}$) can f be L-smooth on $\mathcal{X}$? [10pts]

Solution:

My attack to this problem is to look through all the upper and lower bounds for $f(\mathbf{x})$ and $f(\mathbf{y})$ and ensure the lower bound is less than the upper bound.

i Recall that the definition of μ-strongly convex:

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Recall that for $f : \mathbb{R}^d \to \mathbb{R}$ to be L-Lipschitz continuous, we need to have for all $\mathbf{x}$, $\mathbf{y} \in \mathcal{X}$:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$$

The definition of strong convexity gives that:

$$\begin{aligned}
&f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq f(\mathbf{x}) + \alpha(f(\mathbf{y}) - f(\mathbf{x})) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
&\to f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) + \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq \alpha(f(\mathbf{y}) - f(\mathbf{x})) + f(\mathbf{x}) \\
&\to f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) + \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 - f(\mathbf{x}) \leq \alpha(f(\mathbf{y}) - f(\mathbf{x})) \\
&\to \frac{f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) + \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 - f(\mathbf{x})}{\alpha} \leq f(\mathbf{y}) - f(\mathbf{x})
\end{aligned}$$

That is, the upper bound for $|f(\mathbf{y}) - f(\mathbf{x})|$ is $\frac{f((1-\alpha)\mathbf{x}+\alpha\mathbf{y})-f(\mathbf{x})}{\alpha} + (1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$ for all $\alpha \in (0, 1)$. However, we require f to be L-Lipschitz continuous. Thus we require:

$$\frac{f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) - f(\mathbf{x})}{\alpha} + (1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq L\|\mathbf{x} - \mathbf{y}\|$$

$\forall$x,y$\in \mathcal{X}$.

ii Recall that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to a norm $\|\cdot\|$ on a set $\mathcal{X} \in \mathbb{R}^d$ if there exists a constant $L \leq \infty$ such that $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

$\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Also, recall that there are a lot of equivalent conditions of strong convexity. The following four conditions are equivalent. (They all mean μ strong convexity.)

i $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$

ii $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex.

iii $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$

iv $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}) - \frac{\alpha(1-\alpha)\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$

Consider (iii), we have:

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$$

Thus we obtain the upper bound for the gradient. Recall from the course note proposition 1.1, we have the Holder's inequality: For any two vectors, $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$|\langle \mathbf{z}, \mathbf{x} \rangle| \leq \|\mathbf{z}\|_*\|\mathbf{x}\|$$

As a result,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*\|\mathbf{x} - \mathbf{y}\| \geq |\langle(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), (\mathbf{x} - \mathbf{y}))\rangle| \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$$

Thus,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \geq \mu\|\mathbf{x} - \mathbf{y}\|$$

However, we want that:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

Which implies $\mu \leq L$.

Q 2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function, let $\{L_1, \ldots, L_d\}$ be positive constants, and suppose that for all $i \in \{1, \ldots, d\}$, all $\delta \in \mathbb{R}$, and all $\mathbf{x} \in \mathbb{R}^d$, you have

$$|\nabla_i f(\mathbf{x} + \delta\mathbf{e}_i) - \nabla_i f(\mathbf{x})| \leq L_i|\delta|,$$

where $\mathbf{e}_i$ is the i^{th} standard basis vector (i.e., the vector with all zeros except for the i^{th} entry, which equals one) and ∇_i denotes the i^{th} entry of the gradient.

Prove that for all $i \in \{1, \ldots, d\}$, all $\delta \in \mathbb{R}$, and all $\mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x} + \delta\mathbf{e}_i) - f(\mathbf{x}) \leq \delta\nabla_i f(\mathbf{x}) + \frac{L_i}{2}|\delta|^2. \qquad \text{[10pts]}$$

Now consider the following randomized coordinate descent update rule:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_{i_k}\nabla_{i_k} f(\mathbf{x}_k)\mathbf{e}_{i_k},$$

where i_k is chosen uniformly at random from the set $\{1, 2, ..., d\}$ (and independently from any prior random choices) and α_{i_k} is the step size you are asked to determine. Prove that there exists the choice of the step sizes $\alpha_i > 0$, $i \in \{1, \ldots, d\}$, and a constant $\beta > 0$ such that:

$$\mathbb{E}_{i_k \sim \mathtt{Unif}(\{1,\ldots,d\})}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)] \leq -\frac{\beta}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

How would you choose α_{i_k}'s? What is the largest β you can get this way? [20pts]

Prove that if f is bounded below by some $f^* > -\infty$, then

$$\min_{0 \leq k \leq K} \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\beta(K+1)},$$

where the expectation is taken w.r.t. all the random choices the algorithm takes (i.e., over all $i_1, i_2, \ldots, i_K$). [10pts]

Solution:

I am going to answer this question by proving several claims. I think this function is just an extension of the L-smooth function on a single coordinate, so I want to design a function to extend the result.

Claim 1 : Define $g_j(\delta) = f(\mathbf{x} + \delta e_j)$. For $j \in \{1, 2, ...d\}$. we have $g_j(\delta)$ is L_j smooth.

Recall that to have L_j smooth, we require to have for all $\forall$ $\mathbf{x}$ $\mathbf{y} \in \mathcal{X}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L_j \|\mathbf{x} - \mathbf{y}\|$, and in our particular case, we only consider the points in the form of $\mathbf{x} + \delta e_j$, $\delta > 0$.

Thus we require, $\forall \delta > 0$, $|\nabla g_j(\mathbf{x} + \delta) - \nabla g_j(\mathbf{x})| \leq L_j \|\delta\|$.

However, $|\nabla g_j(\mathbf{x} + \delta) - \nabla g_j(\mathbf{x})| = |\nabla f(x + \delta e_j) - \nabla f(\mathbf{x})| \leq L_j \|\delta\|$ as given.

Claim 2 $f(\mathbf{x} + \delta e_j) \leq f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \frac{L_j}{2}\|\delta\|^2$

Based on Equation 3.3, 3.4, 3.5 from the textbook from Wright and Recht, for a L-smooth function f, assuming that the updating rule is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$$

By applying Lemma 2.2 from the textbook, which states that: Given a L-smooth function for $\mathbf{x}, \mathbf{y} \in \text{domain}(f)$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

as a result, for any $j \in \{1, 2, ...d\}$, we know for any $\delta \in \mathbb{R}$, $g_j(\delta)$ is L_j smooth, so we have

$$g_j(\delta) \leq g_j(0) + \nabla g(0)^T(\delta - 0) + \frac{L_j}{2}\|\delta - 0\|^2$$

Implies that

$$f(\mathbf{x} + \delta e_j) \leq f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \frac{L_j}{2}\|\delta\|^2$$

Claim 3 $\alpha'_{i_k}s$ could be $\frac{1}{L_j}$ as the best option. As, if we have L-smooth function, if we have updating rule $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$, if we choose $\alpha = \frac{1}{L}$, then we have

$$f(\mathbf{x}_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)^2\|$$

Proof: By Claim 2, we have got the

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Take $\mathbf{y} = \mathbf{x}_{k+1} = x_k - \alpha \nabla f(x_k)$, $\mathbf{x} = \mathbf{x}_k$, we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &= f(\mathbf{x}_k - \alpha \nabla f(x_k)) \\
&\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(-\alpha \nabla f(x_k)) + \frac{L}{2}\|\alpha \nabla f(\mathbf{x}_k)\|^2 \\
&= f(\mathbf{x}_k) - \alpha \|\nabla f(\mathbf{x}_k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f(x_k)\|^2 \\
&= f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2
\end{aligned}$$

if $\alpha = \frac{1}{L}$.

And applying this to $g_j(\delta) = f(\mathbf{x} + \delta e_j)$, which is L_j smooth, we get:

$$g_j(\mathbf{x}_{k+1}) \leq g_j(\mathbf{x}_k) - \frac{1}{2L_j}\|\nabla g_j(\mathbf{x}_k)\|^2$$

implies that

$$f(\mathbf{x}_k - \alpha\nabla_j f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - \frac{1}{2L_j}\|\nabla_j^2 f(\mathbf{x}_k)\|$$

if for all $j \in \{1, 2, ...d\}$, $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_j \nabla_j f(\mathbf{x}_k)$.

Claim 4

$$\mathbb{E}_{i_k \sim \mathrm{Unif}(\{1,\ldots,d\})}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)] \leq -\frac{\beta}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Proof: With the above derivation, we have

$$\begin{aligned}
\mathbb{E}_{i_k}[f(\mathbf{x}_{k+1}) - (\mathbf{x}_k)] &= \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}(f(\mathbf{x}_{k+1}) - f(x_k)) \\
&= \frac{1}{d}\sum_{i=1}^{d}\frac{-2}{L_i}\|\nabla_i^2 f(\mathbf{x}_k)\| \\
&\leq \frac{-1}{2d}\|\nabla f(\mathbf{x}_k)\|_2^2
\end{aligned}$$

β in this case is $\frac{1}{d}$.

Claim 6

$$\min_{0\leq k\leq K}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\beta(K+1)},$$

Proof:

$$\mathbb{E}(f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)) \leq \frac{-\beta}{2}\|\nabla f(\mathbf{x}_k)\|_2^2$$

implies:

$$\mathbb{E}(f(\mathbf{x}_1) - f(\mathbf{x}_0)) \leq \frac{-\beta}{2}\|\nabla f(\mathbf{x}_0)\|_2^2$$

$$\mathbb{E}(f(\mathbf{x}_2) - f(\mathbf{x}_1)) \leq \frac{-\beta}{2}\|\nabla f(\mathbf{x}_1)\|_2^2$$

...

$$\mathbb{E}(f(\mathbf{x}_K) - f(\mathbf{x}_{K-1})) \leq \frac{-\beta}{2}\|\nabla f(\mathbf{x}_{K-1})\|_2^2$$

Summing over the inequalities, to get,

$$f^* \leq f(\mathbf{x}_K) \leq f(\mathbf{x}_0) - \sum_{i=0}^{K} \frac{\beta}{2} \|f(\mathbf{x}_i)\|_2^2$$

The left inequality comes from f^* is the global minimizer. Thus:

$$f(\mathbf{x}_0) - f^* \geq \sum_{i=0}^{K} \frac{\beta}{2} \|\nabla f(\mathbf{x}_i)\|_2^2 \geq \frac{\beta}{2} \min_{0 \leq I \leq K} (K+1) \|\nabla f(\mathbf{x}_i)\|_2^2$$

$$\frac{2(f(\mathbf{x}_0) - f^*)}{\beta(K+1)} \geq \|\nabla f(\mathbf{x}_k)\|_2^2$$

Q 3 (Bregman Divergence)**.** Bregman divergence of a continuously differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$ is a function of two points defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle .$$

Equivalently, you can view Bregman divergence as the error in the first-order approximation of a function:

$$\psi(\mathbf{x}) = \psi(\mathbf{y}) + \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + D_\psi(\mathbf{x}, \mathbf{y}).$$

(i) What is the Bregman divergence of a simple quadratic function $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$, where $\mathbf{x}_0 \in \mathbb{R}^d$ is a given point? [5pts]

(ii) Given $\mathbf{z} \in \mathbb{R}^d$ and some continuously differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$, what is the Bregman divergence of function $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{z}, \mathbf{x} \rangle$? [5pts]

(iii) Use Part (ii) and the definition of Bregman divergence to prove the following 3-point identity:

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d) : \quad D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla \psi(\mathbf{z}) - \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}). \quad [5pts]$$

Hint: Consider fixing $\mathbf{y}$ and viewing $D_\psi(\mathbf{x}, \mathbf{y})$ as a function of the first argument only.

(iv) Suppose I give you the following function: $h(\mathbf{x}) = \langle \mathbf{z}, \mathbf{x} \rangle + D_\psi(\mathbf{x}, \bar{\mathbf{x}})$, where $\mathbf{z} \in \mathbb{R}^d$ and $\bar{\mathbf{x}} \in \mathbb{R}^d$ are given, fixed vectors. Let $\mathcal{X}$ be a closed convex set. Define $\mathbf{y} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})$. Prove that, $\forall \mathbf{x} \in \mathcal{X}$,

$$h(\mathbf{x}) \geq \langle \mathbf{z}, \mathbf{y} \rangle + D_\psi(\mathbf{y}, \bar{\mathbf{x}}) + D_\psi(\mathbf{x}, \mathbf{y}). \quad [5pts]$$

Solution:

i The Bregman Divergence of $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$.

$$\begin{aligned} D_\psi(\mathbf{x}, \mathbf{y}) &= \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \\ &= \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{y} - \mathbf{x}_0)^T(\mathbf{y} - \mathbf{x}_0) - \left\langle \nabla \frac{1}{2}(\mathbf{y} - \mathbf{x}_0)^T(\mathbf{y} - \mathbf{x}_0), (\mathbf{x} - \mathbf{y}) \right\rangle \end{aligned}$$

And recall that vector calculus gives:

$$\frac{\partial(u \cdot v)}{\partial x} = \frac{\partial u^T v}{\partial x} = \frac{\partial u}{\partial x} v + \frac{\partial v}{\partial x} u$$

Thus, the above becomes:

$$D_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) - \frac{1}{2}(\mathbf{y} - \mathbf{x}_0)^T(\mathbf{y} - \mathbf{x}_0) - \langle (\mathbf{y} - \mathbf{x}_0), (\mathbf{x} - \mathbf{y}) \rangle$$

ii The Bregman Divergence of $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{z}, \mathbf{x} \rangle$

$$\begin{aligned} D_\psi(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \\ &= \psi(\mathbf{x}) + \langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{y}) - \langle \mathbf{z}, \mathbf{y} \rangle - \langle \nabla(\psi(\mathbf{y}) + \langle \mathbf{z}, \mathbf{y} \rangle), \mathbf{x} - \mathbf{y}) \rangle \\ &= \psi(\mathbf{x}) + \langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{y}) - \langle \mathbf{z}, \mathbf{y} \rangle - \langle \nabla\psi(\mathbf{y}), (\mathbf{x} - \mathbf{y}) \rangle - \langle \mathbf{z}, \mathbf{x} - \mathbf{y} \rangle \\ &= \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \end{aligned}$$

Which is the same as the Bregman divergence of the ψ function.

iii

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d): \quad D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}).$$

I do this simply by expanding both sides.

$$D_\psi(\mathbf{x}, \mathbf{y}) = \ \psi(\mathbf{x}) - \psi(\mathbf{y}) + \langle (\nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}) \rangle$$

For the right-hand side:

$$\begin{aligned} &D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}) \\ &= \psi(\mathbf{z}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{z} - \mathbf{y} \rangle + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{z}) - \langle \nabla\psi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ &= \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \end{aligned}$$

iv Suppose we have the following function: $h(\mathbf{x}) = \langle \mathbf{z}, \mathbf{x} \rangle + D_\psi(\mathbf{x}, \bar{\mathbf{x}})$, where $\mathbf{z} \in \mathbb{R}^d$ and $\bar{\mathbf{x}} \in \mathbb{R}^d$ are given, fixed vectors. Let $\mathcal{X}$ be a closed convex set. Define $\mathbf{y} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})$. Prove that, $\forall \mathbf{x} \in \mathcal{X}$,

$$h(\mathbf{x}) \geq \langle \mathbf{z}, \mathbf{y} \rangle + D_\psi(\mathbf{y}, \bar{\mathbf{x}}) + D_\psi(\mathbf{x}, \mathbf{y}).$$

Since $\mathbf{y}$ is the point that minimizes the function, the slope around $\mathbf{y}$ can only be positive or to be clear, for all $\mathbf{x} \neq \mathbf{y}$ we need to have

$$\langle \partial h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$$

$\forall \mathbf{x}$ and since

$$\partial h(\mathbf{y}) = \partial \langle \mathbf{z}, \mathbf{y} \rangle + \frac{\partial D_\psi(\mathbf{y}, \overline{\mathbf{x}})}{\partial \mathbf{y}}$$

notice that we have:

$$\begin{aligned} \frac{\partial D_\psi(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}} (\psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle) \\ &= \frac{\partial}{\partial \mathbf{x}} \psi(\mathbf{x}) - \nabla\psi(\mathbf{y}) \\ &= \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{y}) \end{aligned}$$

Thus, we have:

$$\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} = \frac{\partial \langle \mathbf{z}, \mathbf{y} \rangle}{\partial \mathbf{y}} + \nabla\psi(\mathbf{y}) - \nabla\psi(\overline{\mathbf{x}})$$

We then can claim that:

$$\left\langle \frac{\partial h(\mathbf{y})}{\partial \mathbf{y}}, \mathbf{x} - \mathbf{y} \right\rangle = \left\langle \frac{\partial \langle \mathbf{z}, \mathbf{y} \rangle}{\partial \mathbf{y}} + \nabla\psi(\mathbf{y}) - \nabla\psi(\overline{\mathbf{x}}), \mathbf{x} - \mathbf{y} \right\rangle \geq 0$$

Now, let's prove the inequality:

$$\begin{aligned}
\langle \mathbf{z}, \mathbf{x} \rangle &\geq \langle \mathbf{z}, \mathbf{y} \rangle + \left\langle \frac{\partial h(\mathbf{y})}{\partial \mathbf{y}}, \mathbf{y} - \mathbf{x} \right\rangle \\
&\geq \langle \mathbf{z}, \mathbf{y} \rangle + \langle \nabla\psi(\mathbf{y}) - \nabla\psi\overline{\mathbf{x}}, \mathbf{y} - \mathbf{x} \rangle \\
&= \langle \mathbf{z}, \mathbf{y} \rangle - \langle \nabla\psi(\overline{\mathbf{x}}), \mathbf{y} - \overline{\mathbf{x}} \rangle + \psi(\mathbf{y}) - \psi(\overline{\mathbf{x}}) + \langle \nabla\psi(\overline{\mathbf{x}}), \mathbf{x} - \overline{\mathbf{x}} \rangle - \psi(\mathbf{x}) + \psi(\overline{\mathbf{x}}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{y})
\end{aligned}$$

Implies:

$$\langle \mathbf{z}, \mathbf{x} \rangle \geq \langle \mathbf{z}, \mathbf{y} \rangle + D_\psi(\mathbf{y}, \overline{\mathbf{x}}) - D_\psi(\mathbf{x}, \overline{\mathbf{x}}) + D_\psi(\mathbf{x}, \mathbf{y})$$

Implies:

$$\langle \mathbf{z}, \mathbf{x} \rangle + D_\psi(\mathbf{x}, \overline{\mathbf{x}}) \geq \langle \mathbf{z}, \mathbf{y} \rangle + D_\psi(\mathbf{y}, \overline{\mathbf{x}}) + D_\psi(\mathbf{x}, \mathbf{y})$$

Implies:

$$h(\mathbf{x}) \geq \langle \mathbf{z}, \mathbf{y} \rangle + D_\psi(\mathbf{y}, \overline{\mathbf{x}}) + D_\psi(\mathbf{x}, \mathbf{y})$$

Q 4 (Gradient descent with ℓ_p norms)**.** Let $p > 1$ be a parameter and let $q = \frac{p}{p-1}$ (so that $\frac{1}{p} + \frac{1}{q} = 1$). Prove that the following function:

$$h_{\mathbf{z}}(\mathbf{x}) = \langle \mathbf{z}, \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{x}\|_p^2$$

is minimized for $\mathbf{x} = -\nabla(\frac{1}{2}\|\mathbf{z}\|_q^2)$ and that $\min_{\mathbf{x}\in\mathbb{R}^d} h_{\mathbf{z}}(\mathbf{x}) = -\frac{1}{2}\|\mathbf{z}\|_q^2$.

Now let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that is L-smooth w.r.t. $\|\cdot\|_p$, for some L, i.e.,

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \quad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_q \leq L\|\mathbf{x} - \mathbf{y}\|_p.$$

Consider the following update rule:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{u} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{x}_k\|_p^2 \right\}.$$

Use the first part of the question to argue that:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_q^2.$$

Assuming that f is bounded below, derive the bound for $\min_{0\leq i\leq k} \|\nabla f(\mathbf{x}_i)\|_q$ similar to the one that was derived in class for $p = 2$. What is the best bound you could have gotten for $\min_{0\leq i\leq k} \|\nabla f(\mathbf{x}_i)\|_q$ if instead of the approach used in this question, you used standard gradient descent (w.r.t. $\|\cdot\|_2$) that we analyzed in class? [20pts]

Solution:

I think this question is related to duality and minimization. Recall that:

$$\|\mathbf{x}\|_* := \max_{\|\mathbf{z}\|\leq 1} \mathbf{z}^T\mathbf{x}$$

for L^p norm,

$$(\|\mathbf{x}\|_p)_* = \|\mathbf{x}\|_q$$

for $\frac{1}{p} + \frac{1}{q} = 1$

Recall that Holder's inequality gives:

$$|\mathbf{z}^T\mathbf{x}\| \leq \|\mathbf{z}\|\|\mathbf{x}\|_*$$

Thus, to solve the minimization problem:

$$\min_{\mathbf{x}} h_{\mathbf{z}}\mathbf{x} = \langle \mathbf{z}, \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{x}\|_p^2$$

we can see this as equivalent to solving:

$$\min_{\mathbf{y}} h_{\mathbf{z}}(\mathbf{y}) = -\langle \mathbf{z}, \mathbf{y} \rangle + \frac{1}{2}\|\mathbf{y}\|_p^2$$

for $\mathbf{y} = -\mathbf{x}$. And in this case, by Holder's inequality, we have:

$$\langle \mathbf{z}, \mathbf{y} \rangle \leq |\mathbf{z}^T\mathbf{y}| \leq \|\mathbf{z}\|_q\|\mathbf{y}\|_p$$

And thus:

$$-\langle \mathbf{z}, \mathbf{y} \rangle + \frac{1}{2}\|\mathbf{y}\|_p^2 \geq -\|\mathbf{z}\|_q\|\mathbf{y}\|_p + \frac{1}{2}\|\mathbf{y}\|_p^2$$

Notice that $\|\mathbf{z}\|_q$ is fixed in this problem and we need to choose $\mathbf{y}$ and thus $\mathbf{x}$. Let $J = \|\mathbf{y}\|_p$, the function is then:

$$-\|\mathbf{z}\|_q J + \frac{1}{2}J^2 = \frac{1}{2}(J - \|\mathbf{z}\|_q)^2 - \frac{1}{2}\|\mathbf{z}\|_q^2$$

Since the above involves a positive quadratic term, thus we show the minimum is $-\frac{1}{2}\|\mathbf{z}\|_q^2$.

Now recall that to have the bound

$$-\langle \mathbf{z}, \mathbf{y} \rangle \geq -\|\mathbf{z}\|_q\|\mathbf{y}\|_p$$

tight, that is, if we want to have:

$$-\langle \mathbf{z}, \mathbf{y} \rangle = -\|\mathbf{z}\|_q\|\mathbf{y}\|_p$$

then we requires $|\mathbf{z}_i|^q = \alpha|\mathbf{y}_i|^p$, for all $i \in \{1, 2, ...d\}$. We could notice this can be achieved by the proposed relationship that $\mathbf{y}_i = \nabla_i(\frac{1}{2}\|\mathbf{z}\|_q^2)$, thus $\mathbf{x}_i = -\mathbf{y}_i = -\nabla_i(\frac{1}{2}\|\mathbf{z}\|_q^2)$. Because:

$$\begin{aligned} -\nabla_i(\frac{1}{2}\|\mathbf{z}\|_q^2) &= -q\mathbf{z}_i^{q-1}(z_1^q + z_2^q + ...z_d^q)^{\frac{2}{q}-1} \\ &\propto -q\mathbf{z}_i^{q-1} \end{aligned}$$

Thus,

$$\mathbf{y}_i^p \propto q^p\mathbf{z}_i^{(q-1)p} = q^p\mathbf{z}_i^{(q-1)\frac{q}{q-1}} = q^p\mathbf{z}_i^q$$