CS 726: Homework #2

Posted: 09/30/2022, due: 10/11/2022 by 8pm CT on Canvas

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:

(i) The necessary condition is that $\mathcal{X}$ should be bounded. It is not possible for f to simultaneously be μ-strongly convex and L-Lipschitz continuous on the entire $\mathbb{R}^d$ for any $\mu, L > 0$ (μ and L are finite). To see this, let $\mathbf{x}$ be such that $\nabla f(\mathbf{x}) \neq \mathbf{0}$ (such a point must exists, as the minimum of a strongly convex function is unique and the function is defined on the entire $\mathbb{R}^d$) and let $\mathbf{y} = \mathbf{x} + \alpha \nabla f(\mathbf{x})$, for some $\alpha > 0$. Strong convexity then gives us:

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \alpha\left(1 + \alpha\frac{\mu}{2}\right)\|\nabla f(\mathbf{x})\|_2^2, \tag{1}$$

while Lipschitz continuity implies:

$$f(\mathbf{y}) - f(\mathbf{x}) \leq L\alpha\|\nabla f(\mathbf{x})\|_2. \tag{2}$$

But, for, e.g., $\alpha \geq \frac{2}{\mu}\frac{L}{\|\nabla f(\mathbf{x})\|_2}$, the lower bound on $f(\mathbf{y}) - f(\mathbf{x})$ from Eq. (1) is larger than the upper bound on $f(\mathbf{y}) - f(\mathbf{x})$ from Eq. (2), which is a contradiction.

(ii) L should be greater than μ. If f is L-smooth and μ-strongly convex at the same time, we have the following inequality for any $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{X}$:

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

If $L < \mu$ the inequality cannot hold so f cannot be L-smooth and μ-strongly convex.

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 $\mathbf{x} \in \mathbb{R}^d$,

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

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 $\alpha_i > 0$, $i \in \{1, \dots, d\}$, and a constant $\beta > 0$ such that:

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

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 $i_1, i_2, \dots, i_K$). [10pts]

Solution:

(i) Note first that $\langle \nabla f(\mathbf{y}), \mathbf{e}_i \rangle = \nabla_i f(\mathbf{y})$ since $\mathbf{e}_i$ is an all-zero vector except for its i-th entry. Then the claim can be proved using Taylor's theorem:

$$\begin{aligned}
f(\mathbf{x} + \delta \mathbf{e}_i) &= f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t\delta \mathbf{e}_i), \delta \mathbf{e}_i \rangle \, \mathrm{d}t \\
&= f(\mathbf{x}) + \int_0^1 \delta \nabla_i f(\mathbf{x} + t\delta \mathbf{e}_i) \mathrm{d}t \\
&= f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \int_0^1 \delta(\nabla_i f(\mathbf{x} + t\delta \mathbf{e}_i) - \nabla_i f(\mathbf{x})) \mathrm{d}t \\
&\leq f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \int_0^1 |\delta| \cdot |\nabla_i f(\mathbf{x} + t\delta \mathbf{e}_i) - \nabla_i f(\mathbf{x})| \mathrm{d}t \\
&\leq f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \int_0^1 |\delta|^2 L_i t \mathrm{d}t \\
&= f(\mathbf{x}) + \delta \nabla_i f(\mathbf{x}) + \frac{L_i}{2}|\delta|^2
\end{aligned}$$

(ii) Plugging the update rule in the inequality we get in (i), we get:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\alpha_{i_k}(\nabla_{i_k} f(\mathbf{x}_k))^2 + \frac{L_{i_k}}{2}(\alpha_{i_k} \nabla_{i_k} f(\mathbf{x}_k))^2,$$

where i_k is chosen uniformly from $\{1, 2, \cdots, d\}$. Now we minimize the RHS by choosing $\alpha_{i_k} = \frac{1}{L_{i_k}}$, then:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{1}{2L_{i_k}}(\nabla_{i_k} f(\mathbf{x}_k))^2$$

Taking the expectation w.r.t. i_k on both sides, we have:

$$\mathbb{E}_{i_k}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)] \leq \mathbb{E}_{i_k}[-\frac{1}{2L_{i_k}}(\nabla_{i_k} f(\mathbf{x}_k))^2] = \frac{1}{d}\sum_{i_k=1}^{d} -\frac{1}{2L_{i_k}}(\nabla_{i_k} f(\mathbf{x}_k))^2 \leq -\frac{1}{2Ld}\|\nabla f(\mathbf{x}_k)\|_2^2,$$

where $L = \max_i L_i$. Therefore, the second claim is proved, and the largest β is $\frac{1}{2Ld}$.

(iii) Note first that

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

Then by summing the claim from part (ii) from $k = 0$ to K, we have:

$$\sum_{k=0}^{K} \mathbb{E}\big[-\frac{\beta}{2} \|\nabla f(\mathbf{x}_k)\|_2^2 \big] \geq \mathbb{E}\Big[\sum_{k=0}^{K} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \Big] = \mathbb{E}[f(\mathbf{x}_K) - f(\mathbf{x}_0)] \geq \mathbb{E}[f^* - f(\mathbf{x}_0)].$$

Combining these results, we arrive at:

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

which finally gives:

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

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:

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

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, $\forall \mathbf{x} \in \mathcal{X}$,

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

Solution:

Observe that the Bregman divergence of either a linear or a constant function is zero.

(i) Note that $\|\mathbf{x} - \mathbf{x}_0\|_2^2 = \langle \mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0 \rangle$, therefore, expanding square norms into inner products, we have

$$\|\mathbf{x} - \mathbf{x}_0\|_2^2 - \|\mathbf{y} - \mathbf{x}_0\|^2 = \langle \mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0 \rangle - \langle \mathbf{y} - \mathbf{x}_0, \mathbf{y} - \mathbf{x}_0 \rangle = \langle \mathbf{x} - 2\mathbf{x}_0 + \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle .$$

Moreover, recall that $\nabla \|\mathbf{x} - \mathbf{x}_0\|_2^2 = 2(\mathbf{x} - \mathbf{x}_0)$. Thus, plugging ψ into the definition of Bregman divergence, we have directly:

$$\begin{aligned} D_\psi(\mathbf{x}, \mathbf{y}) &= \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{x}_0\|_2^2 - \langle \mathbf{y} - \mathbf{x}_0, \mathbf{x} - \mathbf{y} \rangle \\ &= \left\langle \frac{1}{2}\mathbf{x} - \mathbf{x}_0 + \frac{1}{2}\mathbf{y}, \mathbf{x} - \mathbf{y} \right\rangle - \langle \mathbf{y} - \mathbf{x}_0, \mathbf{x} - \mathbf{y} \rangle \\ &= \frac{1}{2}\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \\ &= \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 . \end{aligned}$$

(ii) We have already discussed that the Bregman divergence of a linear function is zero. Moreover, note that from the definition of Bregman Divergence, for any two functions f, g, we have $D_{f+g}(\mathbf{x}, \mathbf{y}) = D_f(\mathbf{x}, \mathbf{y}) + D_g(\mathbf{x}, \mathbf{y})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Thus $D_\psi(\mathbf{x}, \mathbf{y}) = D_\phi(\mathbf{x}, \mathbf{y})$.

(iii) Let $\phi(\mathbf{x}) = D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$. As a function of $\mathbf{x}$, all the terms in the definition of $\phi(\mathbf{x})$ apart from $\psi(\mathbf{x})$ are either linear or constant. Thus, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$: $D_\phi(\mathbf{x}, \mathbf{z}) = D_\psi(\mathbf{x}, \mathbf{z})$. By the definition of Bregman divergence (w.r.t. ϕ):

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

It remains to rearrange the terms in the last equality.

Also acceptable solution: By definition,

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

(iv) By the definition of Bregman divergence,

$$h(\mathbf{x}) = h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + D_h(\mathbf{x}, \mathbf{y}). \tag{3}$$

Since $\mathbf{y} = \operatorname{argmin}_{\mathbf{x}' \in \mathcal{X}} h(\mathbf{x}')$, by the first-order necessary condition, it follows that $\langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$. Further, by Part (ii), $D_h(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{x}, \mathbf{y})$. Plugging these two expressions into Eq. (3),

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

where the last equality is by the definition of h.

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:

First we want to argue that $h_{\mathbf{z}}(\mathbf{x})$ is convex and continuously differentiable. $\|\cdot\|_p$ is convex because by triangle inequality we have for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$, and, thus, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\forall \alpha \in (0,1)$:

$$\|(1-\alpha)\mathbf{x} + \alpha\mathbf{y}\|_p \leq (1-\alpha)\|\mathbf{x}\|_p + \alpha\|\mathbf{y}\|_p.$$

Now, to prove that $\|\cdot\|_p^2$ is convex, we first apply convexity of $\|\cdot\|_p$ to show that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\forall \alpha \in (0,1)$:

$$\|(1-\alpha)\mathbf{x} + \alpha\mathbf{y}\|_p^2 \leq \Big((1-\alpha)\|\mathbf{x}\|_p + \alpha\|\mathbf{y}\|_p\Big)^2,$$

which is true because the quadratic function of a non-negative argument is monotonically increasing. Now, we can view the R.H.S. of the last inequality as the following function of one-dimensional variables a and b: $((1-\alpha)+\alpha b)^2$, where $a = \|\mathbf{x}\|_p$, $b = \|\mathbf{y}\|_p$. Using Jensen's inequality:

$$((1-\alpha)a + \alpha b)^2 \leq (1-\alpha)a^2 + \alpha b^2,$$

as the quadratic function is convex, so we finally arrive at:

$$\|(1-\alpha)\mathbf{x} + \alpha\mathbf{y}\|_p^2 \leq (1-\alpha)\|\mathbf{x}\|_p^2 + \alpha\|\mathbf{y}\|_p^2.$$

$h_{\mathbf{z}}(\mathbf{x})$ is convex since the sum of a linear function and a convex function is convex. Since we have $p > 1$, we also have that $h_{\mathbf{z}}(\mathbf{x})$ is continuously differentiable on $\mathbf{x}$ over $\mathbb{R}^d$.

Therefore, we know that any minimizer must satisfy $\nabla h_{\mathbf{z}}(\mathbf{x}) = \mathbf{0}$ and that it must be a global minimum too. We are going to verify that $\hat{\mathbf{x}} = -\nabla(\frac{1}{2}\|\mathbf{z}\|_q^2)$ is indeed a minimizer. In particular, we have

$$(\nabla h_{\mathbf{z}}(\mathbf{x}))_i = z_i + \operatorname{sign}(x_i)\,\|\mathbf{x}\|_p^{2-p}\,|x_i|^{p-1} \quad \text{and} \quad (\hat{\mathbf{x}})_i = -\operatorname{sign}(z_i)\,\|\mathbf{z}\|_q^{2-q}\,|z_i|^{q-1}.$$

Substituting $\hat{\mathbf{x}}$ into $\nabla h_{\mathbf{z}}(\mathbf{x})$ gives us

$$(\nabla h_{\mathbf{z}}(\hat{\mathbf{x}}))_i = z_i - \operatorname{sign}(z_i)\,\|\mathbf{z}\|_q^{2-p}\,\|\mathbf{z}\|_q^{(2-q)(p-1)}\,|z_i|^{(q-1)(p-1)}.$$

since $\|\hat{\mathbf{x}}\|_p = \|\mathbf{z}\|_q^{(2-q)(2-p)}\left(\sum_{i=1}^d |z_i|^{p/(p-1)}\right)^{\frac{1}{p}} = \|\mathbf{z}\|_q$. Notice that $(2-p)+(2-q)(p-1) = 0$ and $(q-1)(p-1) = 1$, hence the terms with $\|\mathbf{z}\|_q$ cancel out and we are left with $(\nabla h_{\mathbf{z}}(\hat{\mathbf{x}}))_i = z_i - \operatorname{sign}(z_i)\,|z_i| = 0$ for any $z_i \in \mathbb{R}$ and $i \in \{1, \ldots, d\}$. We have verified that $\hat{\mathbf{x}}$ is a minimizer, and for the minimum value, we have

$$h_{\mathbf{z}}(\hat{\mathbf{x}}) = -\sum_{i=1}^d \left(\,\|\mathbf{z}\|_q^{2-q}\,|z_i|^q\,\right) + \frac{1}{2}\|\mathbf{z}\|_q^2 = -\frac{1}{2}\|\mathbf{z}\|_q^2.$$

Now since f is L-smooth w.r.t. p-norm, we have for all $\mathbf{x}_k, \mathbf{x}_{k+1} \in \mathbb{R}^d$

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

We want to choose $\mathbf{x}_{k+1}$ such that it minimizes the R.H.S. of the above inequality which gives us the best progress guarantee for this iteration. Notice that this exactly how $\mathbf{x}_{k+1}$ is defined in the problem statement. Rearranging the above inequality, we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \min_{\mathbf{u}\in\mathbb{R}^d}\left\{\langle \nabla f(\mathbf{x}_k), \mathbf{u} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{x}_k\|_p^2\right\} \\
&= L\min_{\mathbf{u}\in\mathbb{R}^d}\left\{\langle \frac{1}{L}\nabla f(\mathbf{x}_k), \mathbf{u} - \mathbf{x}_k\rangle + \frac{1}{2}\|\mathbf{u} - \mathbf{x}_k\|_p^2\right\} \\
&= L\Big(-\frac{1}{2}\left\|\frac{1}{L}\nabla f(\mathbf{x}_k)\right\|_q^2\Big) \\
&= -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_q^2
\end{aligned}$$

as required. Now suppose that f is bounded below by $\bar{f}$, we have

$$\frac{1}{2L}\sum_{i=0}^{k}\|\nabla f(\mathbf{x}_i)\|_q^2 \leq \sum_{i=0}^{k}(f(\mathbf{x}_i) - f(\mathbf{x}_{i+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_k) \leq f(\mathbf{x}_0) - \bar{f}.$$

Using similar arguments as described in class, we can conclude that

$$\min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_i)\|_q \leq \sqrt{\frac{2L(f(\mathbf{x}_0) - \bar{f})}{k}}.$$

If we had chosen $\mathbf{x}_{k+1}$ using standard gradient descent w.r.t. to $\|\cdot\|_2$ instead we would have gotten

$$\min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_i)\|_q \leq \min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_i)\|_2 \leq \sqrt{\frac{2L_2(f(\mathbf{x}_0) - \bar{f})}{k}} \leq d^{\frac{1}{p}-\frac{1}{2}}\sqrt{\frac{2L(f(\mathbf{x}_0) - \bar{f})}{k}}$$

for $1 < p \leq 2$ and

$$\min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_i)\|_q \leq d^{\frac{1}{q}-\frac{1}{2}}\min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_i)\|_2 \leq d^{\frac{1}{q}-\frac{1}{2}}\sqrt{\frac{2L_2(f(\mathbf{x}_0) - \bar{f})}{k}} \leq d^{\frac{1}{q}-\frac{1}{2}}\sqrt{\frac{2L(f(\mathbf{x}_0) - \bar{f})}{k}}$$

for $p > 2$, where L_2 is the smoothness parameter with respect to 2-norm. This means in the general if we do gradient descent with respect to the correct norm, then we can save a polynomial d factor in the worst case, which is significant when d is large.