Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 1

Student name: Dania Sana

Exercise 1: Separating Points on the Unit Interval

Consider a learning problem where the data source $\mathcal{X} = [0, 1]$ is the unit interval and each sample point $X \in \mathcal{X}$ is drawn uniformly from $\mathcal{X}$ and is labeled as zero is $X < p^*$ and labeled as 1 otherwise, where p^* is an unknown parameter. Suppose we want to model finding p^* with $0-1$-loss and the class of hypothesis is $\mathcal{H} = [0, 1]$. Provide a function $f(\cdot, \cdot)$ such that for any $0 < \epsilon, \delta < 1$ and given $n \geq f(\epsilon, \delta)$ many samples, any hypothesis $H \in \mathcal{H}$ with zero empirical risk has low expected risk with probability at least $1 - \delta$. That is:

$$\ell(H) \leq \epsilon.$$

In other words, if $n \geq f(\epsilon, \delta)$, then the probability of existence of a hypothesis with zero empirical risk but with expected more than ϵ is at most δ.

Solution:

The lecture handout 1, page 17 states that the $0-1$-loss tells us whether $X = \mathbf{x} \in \mathcal{X}$ is missclassified by $H \in \mathcal{H}$. According to the condition of the problem, $\mathbf{x}$ is classified by the optimal hypothesis $H^* \in \mathcal{H}$ as 0 if $\mathbf{x} < p^*$ and as 1 if $\mathbf{x} \geq p^*$. Therefore, every $H \in \mathcal{H}$ is accompanied by a particular $p \in [0, 1]$ that classifies a point $\mathbf{x} \in \mathcal{X}$ based on this particular p. For example let us choose $H_1 \in \mathcal{H}$ with $p_1 \in [0, 1]$. Then H_1 classifies $\mathbf{x} \in \mathcal{X}$ as 0 if $\mathbf{x} < p_1$ and as 1 if $\mathbf{x} \geq p_1$. Furthermore, we have that the $0-1$-loss function $\ell : \mathcal{H} \times \mathcal{X} \to \{0, 1\}$ takes the following values in our case:

- $\ell(H_1, \mathbf{x}) = 0$ if H_1 properly classifies $\mathbf{x}$, i.e., if its true label (determined by p^*) coincides with the label determined by H_1 (i.e., by p_1).
- $\ell(H_1, \mathbf{x}) = 1$, if the label determined by H_1 (p_1) is different from the true label of $\mathbf{x}$ (again, determined by p^*).

Since $X \in \mathcal{X} = [0, 1]$ is drawn uniformly (i.e., $X \sim U[0, 1]$) we have that: $P(X < p^*) = p^*$ and $P(X \geq p^*) = 1 - p^*$. Based on the previous argument, $\{X < p^*\} = \{label_X = 0\}$ and $\{X \geq p^*\} = \{label_X = 1\}$ so it holds that $P[label_X = 0] = p^*$ and $P[label_X = 1] = 1 - p^*$. Thus, we can equivalently define the loss function as $\ell(label_H, label_X)$ which is 0 if $label_H = label_X$ and 1 if $label_H \neq label_X$ where $label_X$ is the true label of $X \in \mathcal{X}$ and $label_H$ is the label the hypothesis $H \in \mathcal{H}$ assigns to X.

Given n-independent samples $X_1, X_2, X_3, ..., X_n \sim \mathcal{X}$, hypothesis $H \in \mathcal{H}$ and $0-1$-loss ℓ, the empirical risk is defined as:

$$\ell_n(H) = \frac{1}{n} \sum_{i=1}^{n} \ell(H, X_i)$$

Clearly for $H = H^*$, the empirical risk $\ell_n(H^*) = 0$ since $\ell(H^*, X_i) = 0$ for $\forall i = 1, 2, ..., n$ because H^* (accompanied by p^*) labels every X_i correctly. Thus there is at least one

hypothesis H $(= H^*)$ such that its empirical risk is 0.

Way 1:

Now we need to find a function f such that for any $0 < \epsilon$, $\delta < 1$ and $n \geq f(\epsilon, \delta)$ many samples, any hypothesis H with 0 empirical risk, the following holds:

$$\ell(H) > \epsilon$$

with probability at most δ. We can use the Chebyshev's inequality which states:
If X is a random variable with finite expected value and variance σ^2 then $\forall k > 0$, the following holds:

$$P(|X - \mathbb{E}[X]| \geq k) \leq \frac{\sigma^2}{k^2} \tag{1}$$

If instead of X, in (1), we put $\frac{1}{n}\sum_{i=1}^{n} \ell(H, X_i)$, then $E_{\mathcal{X}}[\frac{1}{n}\sum_{i=1}^{n} \ell(H, X_i)] = \ell(H)$. Since $\frac{1}{n}\sum_{i=1}^{n} \ell(H, X_i) = 0$, we let $k = \epsilon$ and the Chebyshev's inequality becomes:

$$P(|\ell(H)| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \tag{2}$$

where $\sigma^2 = Var(\frac{1}{n}\sum_{i=1}^{n} \ell(H, X_i))$. We are now left to find/bound σ^2. Since $X_1, X_2, ..., X_n$ are independent, $\ell(H, X_1), \ell(H, X_2), ..., \ell(H, X_n)$ are also independent thus $\sigma^2 = Var(\frac{1}{n}\sum_{i=1}^{n} \ell(H, X_i)) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\ell(H, X_i))$. Moreover, we have the following:

$$\begin{aligned}
Var(\ell(H, X_i)) &= \mathbb{E}[\ell(H, X_i)^2] - \mathbb{E}[\ell(H, X_i)]^2 && (3)\\
&= \ell(H, label_{X_i} = 0)p^* + \ell(H, label_{X_i} = 1)(1 - p^*) - && (4)\\
&- [\ell(H, label_{X_i} = 0)p^* + \ell(H, label_{X_i} = 1)(1 - p^*)]^2 && (5)\\
&= \ell(H, label_{X_i} = 0)p^*(1 - p^*) + \ell(H, label_{X_i} = 1)p^*(1 - p^*) - && (6)\\
&- 2p^*(1 - p^*)\ell(H, label_{X_i} = 0)\ell(H, label_{X_i} = 1) && (7)\\
&= p^*(1 - p^*) && (8)
\end{aligned}$$

where the last step comes from the fact that when H is fixed, it classifies X_i either with label 1 or label 0 thus if $\ell(H, label_{X_i} = 0) = 0$ then $\ell(H, label_{X_i} = 1) = 1$ and vice versa. $f(p^*) = p^* - (p^*)^2$ achieves its maximum at $p^* = \frac{1}{2}$ (since $f'(p^*) = 1 - 2p^*$ and $f''(p^*) = -2 < 0$ thus for that p^*, such that $f'(p^*) = 0$, the function f, achieves its maximum. This value is $p^* = \frac{1}{2}$.), therefore $Var(\ell(H, X_i)) \leq (\frac{1}{2})^2 = \frac{1}{4}$. Finally we can further bound (2) by $\frac{1}{4n\epsilon^2}$. Thus, in order for this probability in (2) to be smaller than or equal to δ, it is sufficient for the following to hold:

$$\frac{1}{4n\epsilon^2} \leq \delta, \quad i.e., \quad \frac{1}{4\delta\epsilon^2} \leq n$$

We finally let $f(\epsilon, \delta) = \frac{1}{4\delta\epsilon^2}$. For this function it holds that:

$$P(|\ell(H)| > \epsilon) \leq P(|\ell(H)| \geq \epsilon) \leq \delta$$

where the first inequality holds because $\{|\ell(H)| > \epsilon\} \subseteq \{|\ell(H)| \geq \epsilon\}$

Way 2:
We could use the Theorem 1.6 from the section on the Vapnik-Chervonenkis theory in the lecture notes, according to which $\ell(H) > \epsilon$ with probability at most $4\mathcal{H}(2n) \cdot exp(-\epsilon^2 n/8)$ (since the empirical risk is 0). The idea is to bound this probability by δ and find the function $f(\epsilon, \delta)$ which satisfies the condition of the problem. We need to estimate the growth function $\mathcal{H}(n)$ of our learning problem, according to the definition of the lecture notes, page 22, which states that given n-training samples $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and the class of hypothesis $\mathcal{H}$ we have:

$$\mathcal{H}(n) := max\{|\mathcal{H} \cap \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}| : \mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}\}$$

Without loss of generality, suppose $\mathbf{x}_1 \leq \mathbf{x}_2 \leq \mathbf{x}_3 \leq ... \leq \mathbf{x}_n$. Suppose that the true labels of these training samples are $\{\mathbf{x}_1 \to 0, \mathbf{x}_2 \to 0, ..., \mathbf{x}_m \to 0, \mathbf{x}_{m+1} \to 1, ..., \mathbf{x}_n \to 1\}$. A hypothesis $H \in \mathcal{H}$ can classify these training samples as: $\{\mathbf{x}_1 \to 0, ..., \mathbf{x}_t \to 0, \mathbf{x}_{t+1} \to 1, ..., \mathbf{x}_n \to 1\}$ for some t. (In the intersection $H \cap \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are these samples that the hypothesis H assigns correctly). We, therefore, have the following:

$$\mathcal{H} \cap \{\mathbf{x}_1, ..., \mathbf{x}_n\} = {}_{t=0,1,2,...,n}\{{}_{s=min(t,m)}\{\mathbf{x}_1, ..., \mathbf{x}_s\} \cup_{k=max(t+1,m+1)} \{\mathbf{x}_k, ..., \mathbf{x}_n\}\}$$

If $t = 0$ (i.e., there is no element labelled as 0 by a particular hypothesis), then ${}_{s=min(t,m)}\{\mathbf{x}_1, ..., \mathbf{x}_s\} = \phi$ and if $t = n$, (there is no element labelled as 1 by a particular hypothesis) then ${}_{max(t+1,m+1)}\{\mathbf{x}_k, ..., \mathbf{x}_n\} = \phi$. Thus $\mathcal{H}(n) = n + 1$ (since t can take values 0,1,2,..,n). Using the fact from the homework 0 that there is a Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$, then for $x \geq 0$, $e^x \geq 1 + x + \frac{x^2}{2} \geq \frac{x^2}{2}$. We substitute $x = \frac{\epsilon^2 n}{8}$ and obtain the inequality $e^{\frac{\epsilon^2 n}{8}} \geq \frac{\epsilon^4 n^2}{128}$ and $\frac{1}{e^{\frac{\epsilon^2 n}{8}}} \leq \frac{128}{\epsilon^4 n^2}$ which we substitute below together with the inequality $8n + 4 \leq 12n$, since $n \geq 1$:

$$4\mathcal{H}(2n) \cdot exp(-\epsilon^2 n/8) = 4(2n+1)exp(-\epsilon^2 n/8) \leq \frac{12n \cdot 128}{\epsilon^4 n^2} = \frac{1536}{\epsilon^4 n}$$

We just need to let $\frac{1536}{\epsilon^4 n} \leq \delta$ and thus $n \geq \frac{1536}{\epsilon^4 \delta}$. Finally, take $f(\epsilon, \delta) = \frac{1536}{\epsilon^4 \delta}$.
Note: I would really much appreciate if you take into grading only the way which is most suitable. Thank you!

Exercise 2: Continuous Convex Functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. Show that the following are equivalent:

a) f is a convex function

b) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following holds:

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leq \frac{f(\mathbf{x}) + f(\mathbf{y})}{2}$$

Solution 2): I need to show that (a) and (b) are equivalent. Thus I must prove that $(a) \to (b)$ and $(b) \to (a)$.

- Start with $(a) \rightarrow (b)$:

  Suppose that the function f is convex. Then for $0 \leq \lambda \leq 1$ we have:

  $$f(\mathbf{y}+\lambda(\mathbf{x}-\mathbf{y})) = f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}) = f(\mathbf{y})+\lambda(f(\mathbf{x})-f(\mathbf{y}))$$

  In other words for $0 \leq \lambda \leq 1$ the following holds:

  $$f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})) - f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) \geq 0$$

  Since $F(\lambda) = f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})) - f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) \geq 0$ for $0 \leq \lambda \leq 1$ and it is continuous (because f is, thus taking the integral is elligible), also the integral of $F(\lambda)$ over $[0, 1]$ will be non-negative. We then have:

  $$\int_0^1 F(\lambda)\, d\lambda = \int_0^1 (f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})) - f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})))\, d\lambda \geq 0.$$

  We get the following:

  $$\frac{f(\mathbf{x}) + f(\mathbf{y})}{2} = f(\mathbf{y}) + \frac{(f(\mathbf{x}) - f(\mathbf{y}))}{2} = \int_0^1 (f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})))\, d\lambda \geq \int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda$$

  Thus I just proved direction $(a) \rightarrow (b)$

- Direction $(b) \rightarrow (a)$:

  Suppose f is not convex; then there exist $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda_0 \in (0, 1)$ such that:

  $$f(\lambda_0\mathbf{x} + (1 - \lambda_0)\mathbf{y}) > \lambda_0 f(\mathbf{x}) + (1 - \lambda_0)f(\mathbf{y})$$

  Consider the function $F(\lambda) = f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) - \lambda f(\mathbf{x}) - (1 - \lambda)f(\mathbf{y})$ which is continuous since f is continuous. According to my assumption $F(\lambda_0) > 0$. Moreover, we have that $F(0) = F(1) = 0$. Now, let $m_1 \in [0, \lambda_0)$ be the largest value less than λ_0 such that $F(m_1) = 0$ and let $m_2 \in (\lambda_0, 1]$ be the smallest value greater than λ_0 such that $F(m_2) = 0$. (We know that there is such m_1 and m_2 since $F(0) = F(1) = 0$, i.e., if there is no $m_1 \in (0, \lambda_0)$ and no $m_2 \in (\lambda_0, 1)$ such that $F(m_1) = F(m_2) = 0$, just choose $m_1 = 0$ and $m_2 = 1$). On the interval $\lambda \in (m_1, m_2)$ (without including m_1, m_2) we have:

  $$F(\lambda) = f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) - \lambda f(\mathbf{x}) - (1 - \lambda)f(\mathbf{y}) > 0.$$

  because of how we chose m_1 and m_2 and using the fact that F is continuous (F can't have jumps from the negative (y) to positive (y) without touching the x-axis (y=0) which implies that based on how m_1 and m_2 are chosen $F(\lambda) > 0$ on (m_1, m_2)). From the condition we pose, $F(m_1) = F(m_2) = 0$, thus:

  $$f(m_1\mathbf{x} + (1 - m_1)\mathbf{y}) = m_1 f(\mathbf{x}) + (1 - m_1)f(\mathbf{y}) \tag{9}$$

  $$f(m_2\mathbf{x} + (1 - m_2)\mathbf{y}) = m_2 f(\mathbf{x}) + (1 - m_2)f(\mathbf{y}) \tag{10}$$

Now, define $a = m_1\mathbf{x} + (1 - m_1)\mathbf{y}$ and $b = m_2\mathbf{x} + (1 - m_2)\mathbf{y}$. I claim that for $\beta \in (0, 1)$ we have:

$$f(\beta a + (1 - \beta)b) > \beta f(a) + (1 - \beta)f(b) \tag{11}$$

Indeed this is true since:

$$\begin{aligned}
f(\beta a + (1 - \beta)b) &= f(\beta(m_1\mathbf{x} + (1 - m_1)\mathbf{y}) + (1 - \beta)(m_2\mathbf{x} + (1 - m_2)\mathbf{y})) && (12)\\
&= f((\beta m_1 + (1 - \beta)m_2)\mathbf{x} + (\beta(1 - m_1) + (1 - \beta)(1 - m_2))\mathbf{y}) && (13)\\
&> (\beta m_1 + (1 - \beta)m_2)f(\mathbf{x}) + (\beta(1 - m_1) + (1 - \beta)(1 - m_2))f(\mathbf{y}) && (14)\\
&= \beta(m_1 f(\mathbf{x}) + (1 - m_1)f(\mathbf{y})) + (1 - \beta)(m_2 f(\mathbf{x}) + (1 - m_2)f(\mathbf{y})) && (15)\\
&= \beta f(m_1\mathbf{x} + (1 - m_1)\mathbf{y}) + (1 - \beta)f(m_2\mathbf{x} + (1 - m_2)\mathbf{y}) && (16)\\
&= \beta f(a) + (1 - \beta)f(b) && (17)
\end{aligned}$$

In $(13) - (14)$, I use the fact that $(\beta m_1 + (1 - \beta)m_2) \in (m_1, m_2)$ since $\beta \in (0, 1)$ and $\beta(1 - m_1) + (1 - \beta)(1 - m_2) = 1 - (\beta m_1 + (1 - \beta)m_2)$, therefore the inequality holds using the fact that $F(\lambda) > 0$ for $\lambda \in (m_1, m_2)$. For $(15) - (16)$, I use the equalities in (9) and (10).

Thus, I managed to show (11), for any $\beta \in (0, 1)$. Using the equalities/inequalities $(12) - (17)$ we have:

$$f(b + \beta(a - b)) - \beta f(a) - (1 - \beta)f(b) > 0 \tag{18}$$

From (18), we get by integrating over β, $(0, 1)$:

$$\int_0^1 f(b + \beta(a - b))\, d\beta > \int_0^1 (\beta f(a) + (1 - \beta)f(b))\, d\beta = \frac{f(a) + f(b)}{2} \tag{19}$$

However (19) contradicts the condition in b) of this problem: i.e., there are $a, b \in \mathbb{R}^d$ such that the inequality in b) does not hold. We therefore conclude that our assumption that f is not convex is rejected. Finally, I showed the other direction $(b) \rightarrow (a)$, i.e., if the inequality in b) holds then f is convex.

Exercise 3: Gradient Descent with Inexact Gradient Oracle

Consider an unconstrained optimization problem $min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Assume f is μ-strongly convex and L-Lipschitz smooth. Now we only have access to an inexact gradient $g(\mathbf{x})$ at each point $\mathbf{x}$ such that $\|g(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \delta$ with $\delta > 0$. Consider gradient descent with this inexact gradient:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma g(\mathbf{x}_t)$$

where $\gamma > 0$ is the step-size. Define $\mathbf{x}^* = arg\, min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ and $f^* = f(x^*)$.

a) Show that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{4}||\mathbf{x} - \mathbf{y}||^2 - \frac{\delta^2}{\mu}$$

and moreover,

$$\frac{1}{\mu}||g(\mathbf{y})||^2 \geq f(\mathbf{y}) - f^* - \frac{\delta^2}{\mu}$$

Solution a): The following series of equalities/inequalities hold:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{20}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{21}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2) - \tag{22}$$

$$- \frac{1}{2}(||(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})||^2) + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{23}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2) \tag{24}$$

$$- \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2 - 2\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle) + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{25}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{26}$$

$$\geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - ||\mathbf{x} - \mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})|| + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{27}$$

$$\geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - \delta||\mathbf{x} - \mathbf{y}|| + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{28}$$

$$\geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - \frac{\delta^2}{\mu} + \frac{\mu}{4}||\mathbf{x} - \mathbf{y}||^2 \tag{29}$$

Explications about the equalities/inequalities (20) – (29):
In (20), I use the property that a μ-strongly convex function satisfies the inequality: $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2$.
In (21), I use the fact that $\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.
In (22) – (23), I use the following property: $\langle v, w \rangle = \frac{1}{2}(\|v\|^2 + \|w\|^2 - \|v - w\|^2)$, which suited to our case is: $\langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \frac{1}{2}(\|\nabla f(\mathbf{y}) - g(\mathbf{y})\|^2 +$
$+ \|\mathbf{x} - \mathbf{y}\|^2 - \|(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})\|^2)$.
In (24) – (25) – (26) I use the following: $||(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})||^2 = \langle (\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y}), (\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y}) \rangle = ||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2 -$
$- 2 \cdot \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle$.
In (27) I use the Cauchy-Schwarz Inequality that enables:
$-||\mathbf{x} - \mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})|| \leq \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle \leq ||\mathbf{x} - \mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})||$
The inequality in (28) is based on the condition of the problem $||g(\mathbf{y}) - \nabla f(\mathbf{y})|| \leq \delta$ from which we get $-||g(\mathbf{y}) - \nabla f(\mathbf{y})|| \geq -\delta$ and also since $||\mathbf{x} - \mathbf{y}|| \geq 0$ we get that

$-||\mathbf{x}-\mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})|| \geq -\delta||\mathbf{x}-\mathbf{y}||$.
In (29), I use the fact that $\frac{\delta^2}{\mu} + \frac{\mu||\mathbf{x}-\mathbf{y}||^2}{4} \geq \delta||\mathbf{x}-\mathbf{y}||$ since it is equivalent to $(\frac{\delta}{\sqrt{\mu}} - \frac{\sqrt{\mu}||\mathbf{x}-\mathbf{y}||}{2})^2 \geq 0$. (We know that $\mu \in \mathbb{R}_+$). From this we obtain that $-\delta||\mathbf{x}-\mathbf{y}|| \geq -\frac{\delta^2}{\mu} - \frac{\mu||\mathbf{x}-\mathbf{y}||^2}{4}$. We plug this result in (28) and get the inequality as in (29).

Now we need to show that

$$\frac{1}{\mu}||g(\mathbf{y})||^2 \geq f(\mathbf{y}) - f^* - \frac{\delta^2}{\mu}$$

We already proved that:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{\mu}{4}||\mathbf{x}-\mathbf{y}||^2 - \frac{\delta^2}{\mu}$$

Equivalent to:

$$\langle g(\mathbf{y}), \mathbf{y}-\mathbf{x}\rangle - \frac{\mu}{4}||\mathbf{x}-\mathbf{y}||^2 \geq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\delta^2}{\mu}$$

Plug in $\mathbf{x} = \mathbf{x}^*$, we obtain:

$$\langle g(\mathbf{y}), \mathbf{y}-\mathbf{x}^*\rangle - \frac{\mu}{4}||\mathbf{x}^*-\mathbf{y}||^2 \geq f(\mathbf{y}) - f(\mathbf{x}^*) - \frac{\delta^2}{\mu} \tag{30}$$

If we achieve to show that:

$$\frac{1}{\mu}||g(\mathbf{y})||^2 \geq \langle g(\mathbf{y}), \mathbf{y}-\mathbf{x}^*\rangle - \frac{\mu}{4}||\mathbf{x}^*-\mathbf{y}||^2$$

or equivalently:

$$\frac{1}{\mu}||g(\mathbf{y})||^2 - \langle g(\mathbf{y}), \mathbf{y}-\mathbf{x}^*\rangle + \frac{\mu}{4}||\mathbf{x}^*-\mathbf{y}||^2 \geq 0$$

then we are done using (30). Indeed, the above inequality holds, since we have:

$$\frac{1}{\mu}||g(\mathbf{y})||^2 - \langle g(\mathbf{y}), \mathbf{y}-\mathbf{x}^*\rangle + \frac{\mu}{4}||\mathbf{x}^*-\mathbf{y}||^2 = ||\frac{1}{\sqrt{\mu}}g(\mathbf{y}) - \frac{\sqrt{\mu}}{2}(\mathbf{y}-\mathbf{x}^*)||^2 \geq 0$$

b) Show that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have:

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

Solution b): This inequality can be shown similarly as in a), using the definition of an L-smooth function. We have the following series of inequalities:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{31}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{32}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2) - \tag{33}$$

$$- \frac{1}{2}(||(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})||^2) + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{34}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2) \tag{35}$$

$$- \frac{1}{2}(||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2 - 2\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle) + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{36}$$

$$= f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{37}$$

$$\leq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + ||\mathbf{x} - \mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})|| + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{38}$$

$$\leq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \delta||\mathbf{x} - \mathbf{y}|| + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 \tag{39}$$

$$\leq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + L||\mathbf{x} - \mathbf{y}||^2 + \frac{\delta^2}{2L} \tag{40}$$

Explanations for all the steps (31) – (40):
In (31), I use the fact that an L-smooth function satisfies the inequality
$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2$.
Similar to the previous step in a), for (32), I make use of the trivial equality: $\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$. In (33) – (34) I use the equality: $\langle \nabla f(\mathbf{y}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \frac{1}{2}(\|\nabla f(\mathbf{y}) - g(\mathbf{y})\|^2 +$
$+ \|\mathbf{x} - \mathbf{y}\|^2 - \|(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})\|^2)$.
In (35) – (36) – (37) I use the property that $||(\nabla f(\mathbf{y}) - g(\mathbf{y})) - (\mathbf{x} - \mathbf{y})||^2 = ||\nabla f(\mathbf{y}) - g(\mathbf{y})||^2 + ||\mathbf{x} - \mathbf{y}||^2 - 2\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle$.
In (38) I apply the Cauchy-Schwarz inequality which states, in our example, that: $\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) - g(\mathbf{y}) \rangle \leq ||\mathbf{x} - \mathbf{y}|| \cdot ||\nabla f(\mathbf{y}) - g(\mathbf{y})||$. In (39), I use the condition of the problem, mainly that $||\nabla f(\mathbf{y}) - g(\mathbf{y})|| \leq \delta$.
In (40) I use the inequality $||\mathbf{x} - \mathbf{y}|| \cdot \delta \leq \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 + \frac{\delta^2}{2L}$ which is equivalent to $(\frac{\sqrt{L}}{\sqrt{2}}||\mathbf{x} - \mathbf{y}|| - \frac{\delta}{\sqrt{2L}})^2 \geq 0$ which of course holds. (We know that $L \in \mathbb{R}_+$). I apply the inequality $||\mathbf{x} - \mathbf{y}|| \cdot \delta \leq \frac{L}{2}||\mathbf{x} - \mathbf{y}||^2 + \frac{\delta^2}{2L}$ in the expression of (39) and obtain (40).

c) Show that by running gradient descent and setting $\gamma = \frac{1}{2L}$, we have

$$f(\mathbf{x}_{t+1}) - f^* \leq \left(1 - \frac{\mu}{4L}\right)(f(\mathbf{x}_t) - f^*) + \frac{3\delta^2}{4L}$$

This directly implies:

$$f(\mathbf{x}_T) - f^* \leq \left(1 - \frac{\mu}{4L}\right)^T (f(\mathbf{x}_0) - f^*) + \frac{3\delta^2}{\mu}$$

Solution 3c): I start with the first inequality:

$$f(\mathbf{x}_{t+1}) - f^* \leq \left(1 - \frac{\mu}{4L}\right) (f(\mathbf{x}_t) - f^*) + \frac{3\delta^2}{4L}$$

which is equivalent to the inequality:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \frac{\mu}{4L}(f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \frac{3\delta^2}{4L} \tag{41}$$

Thus, if we manage to show (41), then we are done. Using 3a) and 3b) I obtain the following series of inequalities:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq \langle g(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + L||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 + \frac{\delta^2}{2L} && (42)\\
&= \langle g(\mathbf{x}_t), -\frac{1}{2L} g(\mathbf{x}_t) \rangle + L||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 + \frac{\delta^2}{2L} && (43)\\
&= -\frac{1}{2L}\langle g(\mathbf{x}_t), g(\mathbf{x}_t) \rangle + L||-\frac{1}{2L} g(\mathbf{x}_t)||^2 + \frac{\delta^2}{2L} && (44)\\
&= -\frac{1}{2L}\langle g(\mathbf{x}_t), g(\mathbf{x}_t) \rangle + \frac{1}{4L}||g(\mathbf{x}_t)||^2 + \frac{\delta^2}{2L} && (45)\\
&= -\frac{1}{2L}||g(\mathbf{x}_t)||^2 + \frac{1}{4L}||g(\mathbf{x}_t)||^2 + \frac{\delta^2}{2L} && (46)\\
&= -\frac{1}{4L}||g(\mathbf{x}_t)||^2 + \frac{\delta^2}{2L} && (47)\\
&\leq \frac{1}{4L}[\mu(f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \delta^2] + \frac{\delta^2}{2L} && (48)\\
&= \frac{\mu}{4L}(f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \frac{3\delta^2}{4L} && (49)
\end{aligned}$$

Explanation for all the equalities/inequalities in (42) – (49):
In (42), I use the inequality proved in 3b).
In (43) – (44) – (45) – (46) – (47) I make use of the equality (with the indicated step size): $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{2L} g(\mathbf{x}_t)$.
In (48), I consider the second inequality proved in 3a), mainly for $\mathbf{y} = \mathbf{x}_t$: $\frac{1}{\mu}||g(\mathbf{x}_t)||^2 \geq f(\mathbf{x}_t) - f(\mathbf{x}^*) - \frac{\delta^2}{\mu}$, which yields that $-||g(\mathbf{x}_t)||^2 \leq -\mu f(\mathbf{x}_t) + \mu f(\mathbf{x}^*) + \delta^2$. We plug it in and reach at the end the result in (49), which is what we wanted to show mainly the inequality (41).

We now show the second inequality:

$$f(\mathbf{x}_T) - f^* \leq \left(1 - \frac{\mu}{4L}\right)^T (f(\mathbf{x}_0) - f^*) + \frac{3\delta^2}{\mu}$$

by the following:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})(f(\mathbf{x}_{T-1}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L} \tag{50}$$

$$\leq (1 - \frac{\mu}{4L})((1 - \frac{\mu}{4L})(f(\mathbf{x}_{T-2}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L}) + \frac{3\delta^2}{4L} \tag{51}$$

$$\leq ..... \tag{52}$$

$$\leq (1 - \frac{\mu}{4L})^T(f(\mathbf{x}_0) - f(x^*)) + \sum_{i=0}^{T-1}(1 - \frac{\mu}{4L})^i \cdot (\frac{3\delta^2}{4L}) \tag{53}$$

$$= (1 - \frac{\mu}{4L})^T(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{3\delta^2}{\mu}[1 - (1 - \frac{\mu}{4L})^T] \tag{54}$$

$$\leq (1 - \frac{\mu}{4L})^T(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{3\delta^2}{\mu} \tag{55}$$

Explanation for each of the steps $(50) - (55)$:
In $(50) - (51) - (52)$ I just apply recursively the inequality proved right before. i.e., $(f(\mathbf{x}_{t+1}) - f^* \leq \left(1 - \frac{\mu}{4L}\right)(f(\mathbf{x}_t) - f^*) + \frac{3\delta^2}{4L})$ for $\gamma = \frac{1}{2L}$. We prove the transition to (53) by induction. Let us start by $T = 1$. We have:

$$f(\mathbf{x}_1) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L}$$

which holds by (50). Suppose now that the inequality (53) holds for $T - 1$, and we have:

$$f(\mathbf{x}_{T-1}) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})^{T-1}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sum_{i=0}^{T-2}(1 - \frac{\mu}{4L})^i(\frac{3\delta^2}{4L})$$

We are only left to show the inequality (53) for T:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})^T(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sum_{i=0}^{T-1}(1 - \frac{\mu}{4L})^i(\frac{3\delta^2}{4L})$$

We know from (50) that:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})(f(\mathbf{x}_{T-1}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L}$$

Applying the inductive step for $T - 1$, we have:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})((1 - \frac{\mu}{4L})^{T-1}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sum_{i=0}^{T-2}(1 - \frac{\mu}{4L})^i(\frac{3\delta^2}{4L})) + \frac{3\delta^2}{4L} =$$

$$= (1 - \frac{\mu}{4L})^T(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sum_{i=0}^{T-1}(1 - \frac{\mu}{4L})^i(\frac{3\delta^2}{4L})$$

So, we just proved the transition to (53). In (53) − (54), the second term is the result of the sum of a geometric progression ($\sum_{i=0}^{T-1}(1-\frac{\mu}{4L})^i = \frac{4L}{\mu}[1-(1-\frac{\mu}{4L})^T]$). In (54) − (55) I use the property that $0 < \mu \leq L => 0 < \mu \leq 4L => 0 < \frac{\mu}{4L} \leq 1$ $=> 0 < [1-(1-\frac{\mu}{4L})^T] \leq 1$, i.e., $\frac{3\delta^2}{\mu}[1-(1-\frac{\mu}{4L})^T] \leq \frac{3\delta^2}{\mu}$

d) Find a function that is μ-strongly convex and show that the algorithm above can not guarantee to find a point $\mathbf{x}$ such that $f(\mathbf{x}) - f^* < \frac{\delta^2}{2\mu}$.

Solution 3d):
Consider $f(x) = \frac{\mu}{2}x^2$ where $x \in \mathbb{R}$ and $\mu \in \mathbb{R}_+$. Obviously, it is L-smooth for any $L \geq \mu$, and $x^* = 0$. (f is μ-strongly convex since $f(x) - \frac{\mu}{2}x^2 = 0$ is convex and it is L-smooth since $\frac{L}{2} - f(x) = (\frac{L-\mu}{2})x^2$ is convex whenever $L \geq \mu$: properties learned in the lecture notes page 100 and 109. I also use (without proof allowed: checked at QA in moodle) that $y = ax^2$ is convex whenever $a \geq 0$. f therefore reaches its minimum at the point when $f'(x^*) = \mu x^* = 0$, i.e., $x^* = 0$). Take $g(x) = \mu x - \delta$ ($|\nabla f(x) - g(x)| = |\mu x - (\mu x - \delta)| = \delta$, satisfies $|g(x) - \nabla f(x)| \leq \delta$) and $x_0 = \frac{\delta}{\mu}$. Then, based on $x_{t+1} = x_t - \frac{1}{2L}g(x_t) = x_t - \frac{\mu x_t}{2L} + \frac{\delta}{2L}$, we have:

$$x_1 = x_0 - \frac{\mu x_0}{2L} + \frac{\delta}{2L} = \frac{\delta}{\mu} - \frac{\delta}{2L} + \frac{\delta}{2L} = \frac{\delta}{\mu}$$

$$x_2 = x_1 - \frac{\mu x_1}{2L} + \frac{\delta}{2L} = \frac{\delta}{\mu}$$

And thus $x_T = \frac{\delta}{\mu}$ for every T. Therefrom, we get that $f(x_T) - f(x^*) = \frac{\mu}{2}(\frac{\delta}{\mu})^2 - 0 = \frac{\delta^2}{2\mu}$. Therefore we could find a μ-strongly convex and L-smooth ($L \geq \mu$) function $f(x) = \frac{\mu}{2}x^2$ such that $f(x_T) - f(x^*) < \frac{\delta^2}{2\mu}$ is not satisfied for any T where x_T is obtained by the gradient descent with an inexact gradient oracle.