Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 2

Student Name: Dania Sana

Exercise 1: CGD with Unknown Smooth Parameter

Suppose a function $f : \mathbb{R}^d \to \mathbb{R}$ (for $d \geq 2$) that is differentiable and coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, ..., L_d)$ where exactly one of L_j-s is equal to β ($\beta > 1$) and the others are equal to 1. Moreover, suppose that f is μ-strongly convex ($\mu \leq 1$). Now we know β and μ but we do not know which coordinate is β-smooth. We consider Algorithm 1: starting with a guess $\tilde{\mathcal{L}}^{(0)} = (\tilde{L}_1 = 1, \tilde{L}_2 = 1, ..., \tilde{L}_d = 1)$, when a coordinate-wise sufficient decrease criterion:

$$f(\mathbf{x}_t - \frac{1}{\tilde{L}_i^t}\nabla_i f(\mathbf{x}_t)\mathbf{e}_i) \leq f(\mathbf{x}_t) - \frac{1}{2\tilde{L}_i^{(t)}}||\nabla_i f(\mathbf{x}_t)||^2 \tag{1}$$

is not satisfied, we update our guess of the smoothness parameter of this coordinate to be β. In algorithm 1, $\mathcal{D}_{IS}(L_1, L_2, ..., L_d)$ represents the probability distribution with the following mass function for every $k \in [d]$:

$$\mathbb{P}[i = k] = \frac{L_k}{\sum_{j=1}^d L_j}$$

Prove the following for Algorithm 1.

a) Show that when Algorithm 1 stops, it queries at most $O(\frac{d\bar{L}}{\mu} ln\frac{1}{\epsilon})$ numbers of partial derivatives of f with $\bar{L} = \frac{1}{d}\sum_{j=1}^d L_j$

Solution 1a):

Algorithm 1 consists of an "if" and "else" part. It is given that the function f is coordinate-wise smooth with true smoothness parameter $\mathcal{L} = (L_1, L_2, ..., L_d)$. Lemma 5.5 of the lecture notes, page 131 states that for such a function and $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L_i}\nabla_i f(\mathbf{x}_t)\mathbf{e}_i$ the following holds:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2$$

where $i \in \{1, 2, ..., d\}$ is the active coordinate. However, if one of the L_i is not correct (suppose L_k is not the correct one with $k \in \{1, 2, ..., d\}$) then the inequality $f(\mathbf{x}_t - \frac{1}{L_k}\nabla_k f(\mathbf{x}_t)\mathbf{e}_k) \leq f(\mathbf{x}_t) - \frac{1}{2L_k}|\nabla_k f(\mathbf{x}_t)|^2$ is not guaranteed to hold. According to the condition of this problem $\mathcal{L} = (L_1, L_2, ..., L_d)$ where exactly one of L_j ($j \in \{1, 2, ..., d\}$) is β (with $\beta > 1$) and the others equal to 1. Now, the "if" part is run if for i sampled from $\mathcal{D}_{IS}(\tilde{L}_1^{(t)}, \tilde{L}_2^{(t)}, ..., \tilde{L}_d^{(t)})$ the inequality $f(\mathbf{x}_t - \frac{1}{\tilde{L}_i^{(t)}}\nabla_i f(\mathbf{x}_t)\mathbf{e}_i) \leq$

$f(\mathbf{x}_t) - \frac{1}{2\tilde{L}_i^{(t)}}|\nabla_i f(\mathbf{x}_t)|^2$ holds. On the other hand, the "else" part is run if the "if" part is not satisfied i.e., if for i sampled from $\mathcal{D}_{IS}(\tilde{L}_1^{(t)}, \tilde{L}_2^{(t)}, ..., \tilde{L}_d^{(t)})$, the inequality $f(\mathbf{x}_t - \frac{1}{\tilde{L}_i^{(t)}}\nabla_i f(\mathbf{x}_t)\mathbf{e}_i) \leq f(\mathbf{x}_t) - \frac{1}{2\tilde{L}_i^{(t)}}|\nabla_i f(\mathbf{x}_t)|^2$ does not hold. This means that the corresponding $\tilde{L}_i^{(t)}$ was not the correct one and thus it should be assigned to β. Since all the other values of L_i are 1, we are done with the finding of the true smoothness parameter, and therefore only the "if" part with be executed afterwards. In other words, the "else" part runs at most once (if potentially the sampled i does not have the correct corresponding smoothness parameter). Finally the "if" part is executed at most $T = \lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \rceil$ times and at each time it finds at most 3 (same) partial derivatives. If we optimize the code, this can be reduced to a single partial derivative, however for the matter of O-notation, is the same. The else part runs at most once from which we have at most $\lceil \frac{\beta+(d-1)}{\mu} ln2 \rceil$ number of distinct partial derivatives (because Algorithm 2 runs $\lceil \frac{\beta+(d-1)}{\mu} ln(2) \rceil$ times and each time it queries one partial derivative). To conclude, we query at most $T \cdot 3 + \lceil \frac{\beta+(d-1)}{\mu} ln2 \rceil$ numbers of partial derivatives. However we also know that $\lceil x \rceil \leq x + 1$. Moreover, if $x \geq 1$, it holds that $\lceil x \rceil \leq x + 1 \leq 2x$. Since $d \geq 2$, $\beta > 1$, and $\mu \leq 1$, it is true that $\frac{\beta+(d-1)}{\mu} > 2$. We therefore have that $\frac{\beta+(d-1)}{\mu} ln2 \geq 2ln2 \geq 1$. Based on this, we have that $\lceil \frac{\beta+(d-1)}{\mu} ln2 \rceil \leq 2 \cdot \frac{\beta+(d-1)}{\mu} ln2$.

Thus, since $\frac{1}{\epsilon} \geq max(2, e^{1/2}) = 2$, (announced on moodle, in Q & A on Graded Assigenements under the topic "GA2-Clarification about range of eps") then $ln\frac{1}{\epsilon} \geq ln(e^{\frac{1}{2}}) = \frac{1}{2}$. Thus for this ϵ it holds that $\frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \geq 1$ since $\frac{2(\beta+(d-1))}{\mu} \geq 2$. Therefrom we have $\lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \rceil \leq 2 \cdot \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon}$. Using these two relations we have that the number of queried partial derivatives is at most $3\lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \rceil + \lceil \frac{\beta+(d-1)}{\mu} ln2 \rceil$ which can be bounded as:

$$3\left\lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \right\rceil + \left\lceil \frac{\beta+(d-1)}{\mu} ln2 \right\rceil \leq 6 \cdot \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} + 2 \cdot \frac{\beta+(d-1)}{\mu} ln2$$

Moreover, since $\frac{1}{\epsilon} \geq 2 \to ln(2) \leq ln\frac{1}{\epsilon}$, we get the inequalities:

$$3\left\lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \right\rceil + \left\lceil \frac{\beta+(d-1)}{\mu} ln2 \right\rceil \leq 6 \cdot \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} + 2 \cdot \frac{\beta+(d-1)}{\mu} ln2$$

$$\leq 6 \cdot \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} + 2 \cdot \frac{\beta+(d-1)}{\mu} ln\frac{1}{\epsilon} = 14 \cdot \frac{(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon}$$

Substitute $\beta + (d-1) = d\bar{L}$ and we have that when Algorithm 1 stops, it queries at most $3\lceil \frac{2(\beta+(d-1))}{\mu} ln\frac{1}{\epsilon} \rceil + \lceil \frac{\beta+(d-1)}{\mu} ln2 \rceil = O(\frac{\beta+(d-1)}{\mu} ln\frac{1}{\epsilon}) = O(\frac{d\bar{L}}{\mu} ln\frac{1}{\epsilon})$ partial derivatives.

Exercise 2: Normalized GD for Nonconvex Optimization

Consider a L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$, which could be nonconvex. In addition, we assume the function is differentiable and has a global minimum $\mathbf{x}^*$. Our goal is to find a stationary point $\mathbf{x}$ such that $||\nabla f(\mathbf{x})||$ is small. Instead of using conventional gradient descent, we consider a normalized version as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \frac{\nabla f(\mathbf{x}_t)}{||\nabla f(\mathbf{x}_t)|| + \beta_t},$$

where $\eta_t, \beta_t > 0$.

(a) In this part, we consider fixed η_t and β_t, i.e., $\eta_t = \eta$ and $\beta_t = \beta$ for $t \geq 0$. Find a stepsize η such that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ for all iterations.

Solution 2a):

Choose $\eta = \frac{\beta}{L}$. Using the iterative condition $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \frac{\nabla f(\mathbf{x}_t)}{||\nabla f(\mathbf{x}_t)||+\beta}$ and substituting $\eta = \frac{\beta}{L}$ we get

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\beta}{L} \cdot \frac{\nabla f(\mathbf{x}_t)}{||\nabla f(\mathbf{x}_t)|| + \beta} \tag{2}$$

From (2) we obtain that $\mathbf{x}_{t+1} - \mathbf{x}_t = -\frac{\beta \nabla f(\mathbf{x}_t)}{L(||\nabla f(\mathbf{x}_t)||+\beta)}$. Using this fact and the property that the function f is L-smooth, we obtain the following series of inequalities/equalities:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 && (3)\\
&= f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), -\frac{\beta \nabla f(\mathbf{x}_t)}{L(||\nabla f(\mathbf{x}_t)|| + \beta)} \rangle + \frac{L}{2}||(-\frac{\beta \nabla f(\mathbf{x}_t)}{L(||\nabla f(\mathbf{x}_t)|| + \beta)})||^2 && (4)\\
&= f(\mathbf{x}_t) - \frac{\beta ||\nabla f(\mathbf{x}_t)||^2}{L(||\nabla f(\mathbf{x}_t)|| + \beta)} + \frac{L}{2}\frac{\beta^2 ||\nabla f(\mathbf{x}_t)||^2}{L^2(||\nabla f(\mathbf{x}_t)|| + \beta)^2} && (5)\\
&= f(\mathbf{x}_t) - \frac{1}{2L}\beta ||\nabla f(\mathbf{x}_t)||^2 \frac{(\beta + 2||\nabla f(\mathbf{x}_t)||)}{(||\nabla f(\mathbf{x}_t)|| + \beta)^2} && (6)
\end{aligned}$$

With this choosen stepsize $\eta = \frac{\beta}{L}$ we achieve that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ since $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\beta||\nabla f(\mathbf{x}_t)||^2 \frac{(\beta+2||\nabla f(\mathbf{x}_t)||)}{(||\nabla f(\mathbf{x}_t)||+\beta)^2}$ from (6) holds and $-\frac{1}{2L}\beta||\nabla f(\mathbf{x}_t)||^2 \frac{(\beta+2||\nabla f(\mathbf{x}_t)||)}{(||\nabla f(\mathbf{x}_t)||+\beta))^2} \leq 0$ because $\frac{1}{2L}\beta||\nabla f(\mathbf{x}_t)||^2 \frac{(\beta+2||\nabla f(\mathbf{x}_t)||)}{(||\nabla f(\mathbf{x}_t)||+\beta))^2} \geq 0$, given $\beta > 0$, $L \geq 0$, and $||\nabla f(\mathbf{x}_t)|| \geq 0$.

(b) Under the same setting as part (a), show that the algorithm converges with rate $\mathcal{O}(T^{-1})$, i.e.,

$$\frac{1}{T}\sum_{t=0}^{T-1}||\nabla f(\mathbf{x}_t)||^2 = \mathcal{O}(T^{-1})$$

Solution 2b):

We consider the same $\eta = \frac{\beta}{L}$ as in 2a). Moreover, inequality (6) in 2a), gives the following series of inequalities/equalities:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{2L}\beta||\nabla f(\mathbf{x}_t)||^2\frac{(\beta + 2||\nabla f(\mathbf{x}_t)||)}{(||\nabla f(\mathbf{x}_t)|| + \beta)^2} && (7)\\
&= f(\mathbf{x}_t) - \frac{\beta^2||\nabla f(\mathbf{x}_t)||^2}{2L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} - \frac{\beta||\nabla f(\mathbf{x}_t)||^3}{L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} && (8)\\
&\leq f(\mathbf{x}_t) - \frac{\beta||\nabla f(\mathbf{x}_t)||^3}{L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} && (9)
\end{aligned}
$$

From (9) we obtain the following inequality:

$$\frac{||\nabla f(\mathbf{x}_t)||^3}{(||\nabla f(\mathbf{x}_t)|| + \beta)^2} \leq (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}))\frac{L}{\beta} \tag{10}$$

Moreover it holds that:

$$||\nabla f(\mathbf{x}_t)|| - 2\beta \leq \frac{||\nabla f(\mathbf{x}_t)||^3}{(||\nabla f(\mathbf{x}_t)|| + \beta)^2} \tag{11}$$

We can prove (11) by showing that its equivalent inequality holds as well, mainly that the following holds:

$$(||\nabla f(\mathbf{x}_t)|| - 2\beta)(||\nabla f(\mathbf{x}_t)|| + \beta)^2 \leq ||\nabla f(\mathbf{x_t})||^3 \tag{12}$$

Indeed (12) holds from the following series of inequalities/equalities:

$$
\begin{aligned}
(||\nabla f(\mathbf{x}_t)|| - 2\beta)(||\nabla f(\mathbf{x}_t)|| + \beta)^2 &= (||\nabla f(\mathbf{x}_t)|| - 2\beta)(||\nabla f(\mathbf{x}_t)||^2 + \beta^2 + 2\beta||\nabla f(\mathbf{x}_t)||) && (13)\\
&= ||\nabla f(\mathbf{x}_t)||^3 - 2\beta||\nabla f(\mathbf{x}_t)||^2 + \beta^2||\nabla f(\mathbf{x}_t)|| - && (14)\\
&- 2\beta^3 + 2\beta||\nabla f(\mathbf{x}_t)||^2 - 4\beta^2||\nabla f(\mathbf{x}_t)|| && (15)\\
&= ||\nabla f(\mathbf{x}_t)||^3 - 3\beta^2||\nabla f(\mathbf{x}_t)|| - 2\beta^3 && (16)\\
&\leq ||\nabla f(\mathbf{x}_t)||^3 && (17)
\end{aligned}
$$

Therefore, (11) holds and using it together with (10), we obtain that for every t:

$$||\nabla f(\mathbf{x}_t)|| - 2\beta \leq (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}))\frac{L}{\beta} \tag{18}$$

In other words, $||\nabla f(\mathbf{x}_t)||$ is bounded by $(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}))\frac{L}{\beta} + 2\beta$. However, since we have chosen a stepsize such that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$, we have that $f(\mathbf{x}_t) \leq f(\mathbf{x}_0)$ for every $t \in \{0, 1, 2, ..., T-1\}$. Moreover, f attains its global minimum at $\mathbf{x} = \mathbf{x}^*$ thus $f(\mathbf{x}_t) \geq f(\mathbf{x}^*)$ yielding the inequality $-f(\mathbf{x}_t) \leq -f(\mathbf{x}^*)$. Therefore we can further upper bound $(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}))\frac{L}{\beta} + 2\beta$ by $(f(\mathbf{x}_0) - f(\mathbf{x}^*))\frac{L}{\beta} + 2\beta$. In other words, we obtain the following inequality for $t \in \{0, 1, 2, ..., T-1\}$:

$$||\nabla f(\mathbf{x}_t)|| \leq (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}))\frac{L}{\beta} + 2\beta \leq (f(\mathbf{x}_0) - f(\mathbf{x}^*))\frac{L}{\beta} + 2\beta = M \tag{19}$$

We now reuse the inequality in (5), shown in 2a) and get the following series of equalities/inequalities:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\beta}{L(||\nabla f(\mathbf{x}_t)|| + \beta)}||\nabla f(\mathbf{x}_t)||^2 + ||\nabla f(\mathbf{x}_t)||^2 \frac{\beta^2}{2L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} \tag{20}$$

$$= f(\mathbf{x}_t) + ||\nabla f(\mathbf{x}_t)||^2 \left(\frac{\beta^2}{2L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} - \frac{\beta}{L(||\nabla f(\mathbf{x}_t)|| + \beta)} \right) \tag{21}$$

Our goal is to upper bound $\frac{\beta^2}{2L(||\nabla f(\mathbf{x}_t)||+\beta)^2} - \frac{\beta}{L(||\nabla f(\mathbf{x}_t)||+\beta)}$ by a negative expression. Let $s = \frac{1}{||\nabla f(\mathbf{x}_t)||+\beta}$. Using (19), we get that $0 \leq ||\nabla f(\mathbf{x}_t)|| \leq M$ where $M = (f(\mathbf{x}_0) - f(\mathbf{x}^*))\frac{L}{\beta} + 2\beta$ as shown in (19), is a constant (M is strictly positive since $f(\mathbf{x}_0) - f(\mathbf{x}^*) \geq 0$ and $\beta > 0$). From this, we get that $\frac{1}{M+\beta} \leq s \leq \frac{1}{\beta}$. Observe the function $g(s) = \frac{\beta^2 s^2}{2L} - \frac{\beta}{L}s$ whose derivative $g'(s) = \frac{\beta^2 s}{L} - \frac{\beta}{L}$ is 0 at $s = \frac{1}{\beta}$ and it is negative for $s \in [\frac{1}{M+\beta}, \frac{1}{\beta})$. Therefore $g(s)$ is decreasing for $s \in [\frac{1}{M+\beta}, \frac{1}{\beta}]$ (meaning that $g(s_1) \leq g(s_2)$ if $s_1 \geq s_2$) and thus obtains its maximun in this interval for $s = \frac{1}{\beta+M}$. Thus for all $s \in [\frac{1}{M+\beta}, \frac{1}{\beta}]$ we have $g(s) \leq g(\frac{1}{M+\beta}) = \frac{\beta^2}{2L}\left(\frac{1}{M+\beta}\right)^2 - \frac{\beta}{L}\left(\frac{1}{M+\beta}\right) = -\frac{2\beta M+\beta^2}{2L(M+\beta)^2}$. Therefore, we go back to the inequality (21) and upper bound the expression on brackets as follows:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + ||\nabla f(\mathbf{x}_t)||^2 \left(\frac{\beta^2}{2L(||\nabla f(\mathbf{x}_t)|| + \beta)^2} - \frac{\beta}{L(||\nabla f(\mathbf{x}_t)|| + \beta)} \right) \tag{22}$$

$$\leq f(\mathbf{x}_t) - \frac{(2\beta M + \beta^2)}{2L(M+\beta)^2}||\nabla f(\mathbf{x}_t)||^2 \tag{23}$$

From (23) we get:

$$||\nabla f(\mathbf{x}_t)||^2 \leq \frac{2L(M+\beta)^2}{2\beta M+\beta^2}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) \tag{24}$$

We first sum up for $t \in \{0, 1, 2, ..., T-1\}$ both sides of the inequality in (24) as shown in (25) and then divide by T as in (26)-(28). By using the fact that $\frac{2L(M+\beta)^2}{2\beta M+\beta^2}$ is a constant (≥ 0) not depending on t we get:

$$\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)||^2 \leq \frac{2L(M+\beta)^2}{2\beta M+\beta^2} \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) \tag{25}$$

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)||^2 &\leq \frac{1}{T}\frac{2L(M+\beta)^2}{2\beta M+\beta^2} \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) &(26)\\ &= \frac{1}{T}\frac{2L(M+\beta)^2}{(2\beta M+\beta^2)}(f(\mathbf{x}_0) - f(\mathbf{x}_T)) &(27)\\ &\leq \frac{1}{T}\frac{2L(M+\beta)^2}{(2\beta M+\beta^2)}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) = \frac{1}{T}c &(28) \end{aligned}$$

where $c = \frac{2L(M+\beta)^2}{(2\beta M+\beta^2)}(f(\mathbf{x}_0) - f(\mathbf{x}^*))$ is a constant. Therefore, we are done with proving that

$$\frac{1}{T}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)||^2 = \mathcal{O}(T^{-1}).$$

(c) Now we consider the more general case where η_t and β_t are allowed to change over time. Design η_t and β_t such that without knowing the smoothness parameter i.e., η_t and β_t should not depend on L, the algorithm provides the following guarantee:

$$\frac{1}{T}\sum^{T-1} ||\nabla f(\mathbf{x}_t)|| = \tilde{\mathcal{O}}(T^{-\frac{1}{2}})$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms of T. Note that we bound the average gradient norms (instead of squared norms), and that is why the right-hand side depends on $\sqrt{T}$ rather than T.

Solution 2c):
Design 1-Solution 1:
Since f is L-smooth we have:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{29}$$

Take $\eta_t = \frac{1}{\sqrt{t+1}} > 0$ for every and $\beta_t = \frac{1}{t+1} > 0$ (both strictly positive for every $t \geq 0$). Then: $\mathbf{x}_{t+1} - \mathbf{x}_t = -\frac{\nabla f(\mathbf{x}_t)}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)||+\frac{1}{t+1})}$, wherefrom $||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 = \left\| \frac{\nabla f(\mathbf{x}_t)}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)||+\frac{1}{t+1})} \right\|^2 = \frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)||+\frac{1}{t+1})^2}$ We plug this result in the inequality (29) and obtain:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \left(-\frac{\nabla f(\mathbf{x}_t)}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1})} \right) + \frac{L}{2} \cdot \frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1})^2} \tag{30}$$

$$\leq f(\mathbf{x}_t) - \frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1})} + \frac{L}{2} \cdot \frac{1}{t+1} \tag{31}$$

In the transition (30) – (31), I use the fact that $\frac{||\nabla f(\mathbf{x}_t)||^2}{(||\nabla f(\mathbf{x}_t)||+\frac{1}{t+1})^2} \leq 1$ because it is equivalent to the trivial true inequality:
$||\nabla f(\mathbf{x}_t)||^2 \leq (||\nabla f(\mathbf{x}_t)||^2 + \frac{1}{(t+1)^2} + \frac{2||\nabla f(\mathbf{x}_t)||}{t+1})$. Using inequality (31), we have the following:

$$\frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1})} \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)} \tag{32}$$

The following inequality (which is equivalent to the trivial true inequality $||\nabla f(\mathbf{x}_t)||^2 - \frac{1}{(t+1)^2} \leq ||\nabla f(\mathbf{x}_t)||^2$) holds:

$$(||\nabla f(\mathbf{x}_t)|| - \frac{1}{t+1})(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1}) \leq ||\nabla f(\mathbf{x}_t)||^2$$

We plug this inequality in (32) to offer a lower bound for $\frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)||+\frac{1}{t+1})}$ as below:

$$\frac{(||\nabla f(\mathbf{x}_t)|| - \frac{1}{t+1})}{\sqrt{t+1}} \leq \frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+1})} \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)}$$

Finally we obtain the following very important result:

$$\frac{(||\nabla f(\mathbf{x}_t)|| - \frac{1}{t+1})}{\sqrt{t+1}} \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)}$$

Sum up both sides of the above inequality w.r.t $t \in \{0, 1, ..., T-1\}$ and we get the inequality:

$$\sum_{t=0}^{T-1} \frac{||\nabla f(\mathbf{x}_t)|| - \frac{1}{t+1}}{\sqrt{t+1}} \leq \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) + \frac{L}{2}\sum_{t=0}^{T-1}\frac{1}{t+1} \tag{33}$$

$$= (f(\mathbf{x}_0) - f(\mathbf{x}_T)) + \frac{L}{2}\sum_{t=0}^{T-1}\frac{1}{t+1} \tag{34}$$

$$\leq (f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{L}{2}\sum_{t=0}^{T-1}\frac{1}{t+1} \tag{35}$$

For every $t \in \{0, 1, 2, ..., T-1\}$ we have that $\frac{1}{\sqrt{T}} \leq \frac{1}{\sqrt{t+1}}$. and $-1 \leq -\frac{1}{\sqrt{t+1}}$ Therefore, using the inequality (35), we get that:

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| - \sum_{t=0}^{T-1}\frac{1}{t+1} \leq \sum_{t=0}^{T-1}\frac{(||\nabla f(\mathbf{x}_t)|| - \frac{1}{t+1})}{\sqrt{t+1}} \leq (f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{L}{2}\sum_{t=0}^{T-1}\frac{1}{t+1} \tag{36}$$

$$= f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{L}{2}\left[log(T) + \gamma + \frac{1}{2T} - \epsilon_T\right] \tag{37}$$

In the expression (37), I use the Euler-Maclaurin formula [1] that states

$$\sum_{k=0}^{n-1}\frac{1}{k+1} = \sum_{k=1}^{n}\frac{1}{k} = log(n) + \gamma + \frac{1}{2n} - \epsilon_n$$

where $\gamma \approx 0.5772$, $0 \leq \epsilon_n \leq \frac{1}{8n^2}$ and log is the natural logarithm (with base e). Furthermore, using this formula, we can expand the left-side expression of inequality (37), as following:

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| - \sum_{t=0}^{T-1}\frac{1}{t+1} = \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| - (log(T) + \gamma + \frac{1}{2T} - \epsilon_T)$$

Since $T \geq 1$, $\frac{1}{2T} \leq \frac{1}{2}$. Moreover, $-\epsilon_T \leq 0$. Therefore, we can use these properties together with the previous equality and results in (36)-(37) to show that:

$$\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| \leq \tag{38}$$

$$\leq \sqrt{T}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sqrt{T}\frac{L}{2}\left[log(T) + \gamma + \frac{1}{2T} - \epsilon_T\right] + \sqrt{T}(log(T) + \gamma + \frac{1}{2T} - \epsilon_T) \leq \tag{39}$$

$$\leq \sqrt{T}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \sqrt{T}\frac{L}{2}\left[log(T) + \gamma + \frac{1}{2}\right] + \sqrt{T}(log(T) + \gamma + \frac{1}{2}) \tag{40}$$

We divide both sides of (39) by T and we obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1}||\nabla f(\mathbf{x}_t)|| \leq \frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\sqrt{T}}+\frac{L}{2\sqrt{T}}(log(T)+\gamma+\frac{1}{2})+\frac{1}{\sqrt{T}}(log(T)+\gamma+\frac{1}{2})$$

Moreover, if $T \geq 2$, it is true that $\frac{log(T)}{log(2)} \geq 1$ and $\frac{1}{T} \leq \frac{1}{\sqrt{T}}$. We use these facts to obtain the following valid inequalities that hold for $T \geq 2$:

$$\frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}}\cdot\frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{log(2)}$$

$$\frac{\gamma L}{2\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}}\cdot\frac{\gamma L}{2log(2)}$$

$$\frac{L}{4\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}}\cdot\frac{L}{4log(2)}$$

$$\frac{\gamma}{\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}}\cdot\frac{\gamma}{log(2)}$$

$$\frac{1}{2\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}}\cdot\frac{1}{2log(2)}$$

We plug these inequalities in the desired inequality yielding the following result (for $T \geq 2$):

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}||\nabla f(\mathbf{x}_t)|| &\leq \frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\sqrt{T}}+\frac{L}{2\sqrt{T}}(log(T)+\gamma+\frac{1}{2})+\frac{1}{\sqrt{T}}(log(T)+\gamma+\frac{1}{2})\\ &\leq \frac{log(T)}{\sqrt{T}}\left(\frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{log(2)}+\frac{L}{2}+\frac{\gamma L}{2log(2)}+\frac{L}{4log(2)}+1+\frac{\gamma}{log(2)}+\frac{1}{2log(2)}\right)\\ &= \frac{log(T)}{\sqrt{T}}\cdot c'\end{aligned} \tag{41}$$

where $c' = \frac{(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{log(2)}+\frac{L}{2}+\frac{\gamma L}{2log(2)}+\frac{L}{4log(2)}+1+\frac{\gamma}{log(2)}+\frac{1}{2log(2)}$ is a constant. Result (40) shows that:

$$\frac{1}{T}\sum_{t=0}^{T-1}||\nabla f(\mathbf{x}_t)|| = O(log(T)\cdot T^{-1/2}) = \tilde{O}(T^{-1/2})$$

Design 2-Solution 2:
Choose $\beta_t = \frac{1}{(t+2)^2} > 0$ and $\eta_t = \frac{1}{\sqrt{t+1}} > 0$. The following results hold: $\mathbf{x}_{t+1}-\mathbf{x}_t = -\eta_t\frac{\nabla f(\mathbf{x}_t)}{||\nabla f(\mathbf{x}_t)||+\beta_t} = -\frac{1}{\sqrt{t+1}}\cdot\frac{\nabla f(\mathbf{x}_t)}{(||\nabla f(\mathbf{x}_t)||+\frac{1}{(t+2)^2})} \rightarrow ||\mathbf{x}_{t+1}-\mathbf{x}_t||^2 =$

$\frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)||+\frac{1}{(t+2)^2})^2}$ We now use the inequality (29) enabled by the smoothness property:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{(t+2)^2})} + \frac{L}{2} \cdot \frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)|| + \frac{1}{(t+2)^2})^2}$$

The below inequality follows:

$$\frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{t+1}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{(t+2)^2})} \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2} \cdot \frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)|| + \frac{1}{(t+2)^2})^2}$$

Again, we use the inequalities $\frac{L}{2}\frac{||\nabla f(\mathbf{x}_t)||^2}{(t+1)(||\nabla f(\mathbf{x}_t)||+\frac{1}{(t+2)^2})^2} \leq \frac{L}{2(t+1)}$ and $||\nabla f(\mathbf{x}_t)||^2 - (\frac{1}{(t+2)^2})^2 = (||\nabla f(\mathbf{x}_t)|| - \frac{1}{(t+2)^2})(||\nabla f(\mathbf{x}_t)|| + \frac{1}{(t+2)^2}) \leq ||\nabla f(\mathbf{x}_t)||^2$ to obtain:

$$\frac{1}{\sqrt{t+1}}(||\nabla f(\mathbf{x}_t)|| - \frac{1}{(t+2)^2}) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)}$$

Moreover the inequalities $\frac{1}{(t+2)^2} \leq \frac{1}{(t+1)(t+2)} = \frac{1}{t+1} - \frac{1}{t+2} \rightarrow -\frac{1}{(t+2)^2} \geq \frac{1}{t+2} - \frac{1}{t+1}$ hold. We plug this inequality in the above result and get:

$$\frac{1}{\sqrt{t+1}}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+2} - \frac{1}{t+1}) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)}$$

We sum over t ($t \in \{0, 1, ..., T-1\}$) the both sides of the previous inequality and using the fact that $\sqrt{t+1} \leq \sqrt{T}$ for every $t \in \{0, 1, ..., T-1\}$ and $-\frac{1}{\sqrt{t+1}} \geq -1$ we get:

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| - \frac{T}{(T+1)} = \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| + \sum_{t=0}^{T-1}(\frac{1}{t+2} - \frac{1}{t+1})$$

$$\leq \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}}(||\nabla f(\mathbf{x}_t)|| + \frac{1}{t+2} - \frac{1}{t+1}) \leq \sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2(t+1)}$$

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

We now use the Euler-Maclaurin formula $\sum_{t=0}^{T-1}\frac{1}{t+1} = log(T) + \gamma + \frac{1}{2T} - \epsilon_T$ (also the facts that $-\epsilon_T \leq 0$ and $\frac{1}{2T} \leq \frac{1}{2}$) and plug it in the previous result to finally get:

$$\frac{1}{T}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| \leq \frac{1}{\sqrt{T}}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \frac{L}{2\sqrt{T}}(log(T) + \gamma + \frac{1}{2}) + \frac{\sqrt{T}}{T+1}$$

Based on our previous argument, for $T \geq 2$ the following inequalities hold:

$$\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}} \cdot \frac{(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{log(2)}$$

$$\frac{L\gamma}{2\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}} \cdot \frac{L\gamma}{2log(2)}$$

$$\frac{L}{4\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}} \cdot \frac{L}{4log(2)}$$

$$\frac{\sqrt{T}}{T+1} \leq \frac{\sqrt{T}}{T} = \frac{1}{\sqrt{T}} \leq \frac{log(T)}{\sqrt{T}} \cdot \frac{1}{log(2)}$$

Thus the following inequality holds that for $T \geq 2$:

$$\frac{1}{T}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| \leq \frac{log(T)}{\sqrt{T}}\left(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{log(2)} + \frac{L}{2} + \frac{L\gamma}{2log(2)} + \frac{L}{4log(2)} + \frac{1}{log(2)}\right) = \frac{log(T)}{\sqrt{T}} \cdot c''$$

with $c'' = \frac{f(\mathbf{x}_0)-f(\mathbf{x}^*)}{log(2)} + \frac{L}{2} + \frac{L\gamma}{2log(2)} + \frac{L}{4log(2)} + \frac{1}{log(2)}$ is a constant. Thus, we just showed that also for this second design of η_t, β_t it holds that $\frac{1}{T}\sum_{t=0}^{T-1} ||\nabla f(\mathbf{x}_t)|| = O(log(T) \cdot T^{-1/2}) = \tilde{O}(T^{-1/2})$.

Exercise 3: Frank-Wolfe with an Approximation Oracle

Recall that in the Frank-Wolfe algorithm, we assume there is a linear minimization oracle (LMO) that can return the exact minimizer. However, this minimization problem can itself be challenging to solve. In this exercise, we analyze the convergence of a variant of the Frank-Wolfe algorithm with an approximation LMO.
We consider the optimization problem $min_{\mathbf{x} \in X} f(\mathbf{x})$ for a convex funcion $f : \mathbb{R}^d \to \mathbb{R}$ with smooth paramter l, and for X equals to $[-1/2, 1/2]^d$ (i.e., X is a unit d-dimensional cube). We assume a minimizer $\mathbf{x}^* \in X$ exists. For any accuracy $\alpha \geq 0$, let APPROX-LMO$_X^\alpha(\nabla f(\mathbf{x}_t))$ be an oracle that computes a vector in X such that

$$\nabla f(\mathbf{x}_t)^\top APPROX - LMO_X^\alpha(\nabla f(\mathbf{x}_t)) \leq min_{z \in X} \nabla f(\mathbf{x_t})^\top \mathbf{z} + \alpha C_{f,X} \qquad (42)$$

where $C_{f,X}$ is the curvature constant. So for α being zero, $APPROX - LMO_X^0$ is an exact oracle.

a) Show that

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 d$$

where $g(\mathbf{x}_t)$ is the duality gap that is defined in the lectures.

Solution 3a):

Given that the function is 1-smooth we have:

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

We now assign $\mathbf{y} = \mathbf{x}_{t+1} = (1 - \gamma_t)\mathbf{x}_t + \gamma_t \mathbf{s}_t = \mathbf{x}_t + \gamma_t(\mathbf{s}_t - \mathbf{x}_t)$ (with $\mathbf{s}_t$ as defined in Algorithm 3: $\mathbf{s}_t = APPROX - LMO_X^{\gamma_t}(\nabla f(\mathbf{x}_t)))$ and $\mathbf{x} = \mathbf{x}_t$ and using the previous inequality we obtain the following:

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t + \gamma_t(\mathbf{s}_t - \mathbf{x}_t)) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \gamma_t(\mathbf{s}_t - \mathbf{x}_t) + \frac{\gamma_t^2}{2}||\mathbf{s}_t - \mathbf{x}_t||^2 \tag{43}$$

Given that $\mathbf{s}_t, \mathbf{x}_t \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d$ (since it is given that $\mathbf{s}_t = APPROX - LMO_X^{\gamma_t}(\nabla f(\mathbf{x}_t))$ computes a vector in X) it is true that $||\mathbf{s}_t - \mathbf{x}_t||^2 = \sum_{i=1}^d (\mathbf{s}_{t_i} - \mathbf{x}_{t_i})^2 \leq \sum_{i=1}^d \left(\frac{1}{2} + \frac{1}{2}\right)^2 = d$. Therefore, using this fact and inequality (41), I can further bound the inequality in (42) by the following series:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \gamma_t(\mathbf{s}_t - \mathbf{x}_t) + \frac{\gamma_t^2}{2}d = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \gamma_t \mathbf{s}_t - \nabla f(\mathbf{x}_t)^\top \gamma_t \mathbf{x}_t + \frac{\gamma_t^2}{2}d \tag{44}$$

$$= f(\mathbf{x}_t) + \gamma_t[\nabla f(\mathbf{x}_t)^\top \mathbf{s}_t] - \nabla f(\mathbf{x}_t)^\top \gamma_t \mathbf{x}_t + \frac{\gamma_t^2}{2}d \tag{45}$$

$$\leq f(\mathbf{x}_t) + \gamma_t[min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top \mathbf{z} + \gamma_t C_{(f,X)}] - \gamma_t \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t + \frac{\gamma_t^2}{2}d \tag{46}$$

$$= f(\mathbf{x}_t) + \gamma_t[min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top \mathbf{z} - \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] + \gamma_t^2 C_{(f,X)} + \frac{\gamma_t^2}{2}d \tag{47}$$

In the transition of inequalities (44)-(45) I use inequality (41). In the lecture notes, Lemma 7.6, we have that $C_{(f,X)} \leq \frac{L}{2}diam(X)^2$. In our setting, $L = 1$ and as shown previously $diam(X)^2 = max_{\mathbf{x},\mathbf{y} \in [-\frac{1}{2},\frac{1}{2}]^d}||\mathbf{x} - \mathbf{y}||^2 = d$. We then finally get $C_{(f,X)} \leq \frac{L}{2}diam(X)^2 = \frac{d}{2}$. (This lemma is already proven in Homework 6 solutions, Exercise 2, so I am skipping the proof of it.)

We now go back to the inequality (46) and using the above inequality, we get:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \gamma_t[min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top z - \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] + \gamma_t^2 C_{(f,X)} + \frac{\gamma_t^2}{2}d \tag{48}$$

$$\leq f(\mathbf{x}_t) + \gamma_t[min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top \mathbf{z} - \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] + \gamma_t^2 d \tag{49}$$

Based on the lecture notes the duality gap is defined as $g(\mathbf{x}) = \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{s})$ where $s = argmin_{\mathbf{z}\in X} \nabla f(\mathbf{x})^\top \mathbf{z}$. Therefrom, we get

$min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top \mathbf{z} - \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t = -g(\mathbf{x}_t)$. We plug this in the inequality (48) and get the desired inequality:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \gamma_t[min_{\mathbf{z}\in X} \nabla f(\mathbf{x}_t)^\top \mathbf{z} - \nabla f(\mathbf{x}_t)^\top \mathbf{x}_t] + \gamma_t^2 d \tag{50}$$

$$\leq f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 d \tag{51}$$

We just showed (50): $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 d$.

b) Show that for any $\epsilon > 0$, and for any $T \geq 4d/\epsilon$, we have:

$$f(\mathbf{x}_T) - f(x^*) \leq \epsilon.$$

Solution 3b):

I claim that for every $T \geq 1$, $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{4d}{T}$. Inspired by the lecture notes, I define $h(\mathbf{x}_T) = f(\mathbf{x}_T) - f(\mathbf{x}^*)$. Lemma 7.2 of the lecture notes (proved in class) states that $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq g(\mathbf{x}_T)$. This implies that $h(\mathbf{x}_T) \leq g(\mathbf{x}_T)$. In 3a) we proved the inequality: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 d$. We substract $f(\mathbf{x}^*)$ from both sides of this inequality, let $t = T-1$ and obtain:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq f(\mathbf{x}_{T-1}) - f(\mathbf{x}^*) - \gamma_{T-1} g(\mathbf{x}_{T-1}) + \gamma_{T-1}^2 d$$

Using the above inequality we obtain the below inequalities by using the followings $h(\mathbf{x}_T) = f(\mathbf{x}_T) - f(\mathbf{x}^*)$, $h(\mathbf{x}_{T-1}) = f(\mathbf{x}_{T-1}) - f(\mathbf{x}^*)$, $-g(\mathbf{x}_{T-1}) \leq -h(\mathbf{x}_{T-1})$ and $\gamma_{T-1} = \frac{2}{T+1}$ (since $\gamma_t = \frac{2}{t+2}$):

$$h(\mathbf{x}_T) \leq h(\mathbf{x}_{T-1}) - \gamma_{T-1} h(\mathbf{x}_{T-1}) + \gamma_{T-1}^2 d = (1 - \gamma_{T-1}) h(\mathbf{x}_{T-1}) + \gamma_{T-1}^2 d \tag{52}$$

$$= \left(1 - \frac{2}{T+1}\right) h(\mathbf{x}_{T-1}) + \left(\frac{2}{T+1}\right)^2 d = \left(\frac{T-1}{T+1}\right) h(\mathbf{x}_{T-1}) + \left(\frac{2}{T+1}\right)^2 d \tag{53}$$

We use the inequality (52) to prove our claim: $h(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{4d}{k}$ by induction. Let us first prove it for $k_0 = 1$. Using (52), we have:

$$h(\mathbf{x}_1) \leq (\frac{1-1}{1+1}) h(\mathbf{x}_0) + (\frac{2}{1+1}) d = d \leq \frac{4d}{k_0} = \frac{4d}{1} = 4d$$

Let us now suppose it is true for all $k \leq m$, with $m > 1$ and we shall show it holds for $k = m+1$. The inequality holds:

$$h(\mathbf{x}_{m+1}) \leq \left(\frac{m}{m+2}\right) h(\mathbf{x}_m) + \left(\frac{2}{m+2}\right)^2 d \tag{54}$$

We apply the inductive reasoning which states that $h(\mathbf{x}_m) \leq \frac{4d}{m}$ and induce an inequality resulting from (53):

$$h(\mathbf{x}_{m+1}) \leq \left(\frac{m}{m+2}\right)h(\mathbf{x}_m) + \left(\frac{2}{m+2}\right)^2 d \leq \left(\frac{m}{m+2}\right)\frac{4d}{m} + \left(\frac{2}{m+2}\right)^2 d \tag{55}$$

$$= \frac{4d}{m+2} + \frac{4d}{(m+2)^2} = 4d\frac{m+3}{(m+2)^2} \leq \frac{4d}{m+1} \tag{56}$$

where the last inequality in (55) comes from the fact that $\frac{m+3}{(m+2)^2} \leq \frac{1}{m+1} \Leftrightarrow (m+3)(m+1) \leq (m+2)^2 \Leftrightarrow (m^2+4m+3) \leq (m^2+4m+4)$ which surely holds. We just showed that $h(\mathbf{x}_{m+1}) \leq \frac{4d}{m+1}$. Thus I already proved my claim i.e., that $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{4d}{T}$ holds for every $T \geq 1$. Now, we come back to the problem condition which states that if $T \geq 4d/\epsilon$ then $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \epsilon$. First, the inequality $T \geq 4d/\epsilon$ is equivalent to $\epsilon \geq \frac{4d}{T}$. Therefore (if $T \geq 4d/\epsilon$) we obtain (using the claim) the following:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{4d}{T} \leq \epsilon$$

Finally, I proved that for any $T \geq 4d/\epsilon$ we have $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \epsilon$. Thus, the problem statement is proven!

Exercise 4: Modified Newton's Method

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex twice-differentiable function with Lipschitz Hessian. In order to minimize the given function we consider a modified version of Newton's Method. Assume the following for this exercise:
1. There is a $\mathbf{x}^* \in \mathbb{R}^d$ such that $f(\mathbf{x}^*) = min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.
2. There is a constant $B \in \mathbb{R}$ such that for all $\mathbf{x} \in \mathbb{R}^d$, if $f(\mathbf{x}) \leq f(\mathbf{x}_0)$, then $||\mathbf{x} - \mathbf{x}^*|| \leq B$.
3. There is a constant $H > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{H}{3}||\mathbf{y} - \mathbf{x}||^3$$

and

$$||\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})|| \leq H||\mathbf{y} - \mathbf{x}||^2.$$

In the following steps, we derive a convergence result for Algorithm 4 given the three assumptions above on f.

a) Show that the following relations holds for all λ_t in Algorithm 4:

$$\lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) = -\nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t).$$

Solution 4a):

The next iterate $\mathbf{x}_{t+1}$ of the algorithm is as follows $\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1} \nabla f(\mathbf{x}_t)$. This enables the equivalent expressions:

$$\mathbf{x}_{t+1} - \mathbf{x}_t = -(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1} \nabla f(\mathbf{x}_t) \Leftrightarrow -\nabla f(\mathbf{x}_t) = (\nabla^2 f(\mathbf{x}_t) + \lambda_t I)(\mathbf{x}_{t+1} - \mathbf{x}_t)$$

For this above equivalence, we just multiplied the both sides of the first equality by the matrix $\nabla^2 f(\mathbf{x}_t) + \lambda_t I$. To the previous two expressions, we have the below equivalence:

$$\Leftrightarrow -\nabla f(\mathbf{x}_t) = \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) + \lambda_t I(\mathbf{x}_{t+1} - \mathbf{x}_t) \Leftrightarrow$$

$$\Leftrightarrow \lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) = -\nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)$$

I proved that $\lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) = -\nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)$ holds for all λ_t.

b) Show that for all iterations, the following hold:

$$H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t,$$

$$||\nabla f(\mathbf{x}_{t+1})|| \leq 2\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq 2||\nabla f(\mathbf{x}_t)||.$$

Solution 4b):

i) Let us first show that $H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t$ holds. From 4a) $\mathbf{x}_{t+1} - \mathbf{x}_t = -(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1} \nabla f(\mathbf{x}_t)$ holds. We substitute this result and obtain the following series of inequalities/equalities:

$$H||\mathbf{x}_{t+1} - \mathbf{x}_t|| = H||-(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1} \nabla f(\mathbf{x}_t)|| \leq H||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}|| \cdot ||\nabla f(\mathbf{x}_t)|| \tag{57}$$

The inequality (56) holds because given a matrix $A \in \mathbb{R}^{d \times d}$, a vector $\mathbf{x} \in \mathbb{R}^d$ and

$|| \cdot ||$-l2-norm (applied to a vector) the inequality $||A\mathbf{x}|| \leq ||A|| \cdot ||\mathbf{x}||$ holds, where $||A|| = max_{||\mathbf{x}||=1} ||A\mathbf{x}||$. Moreover, according to algorithm 4, $\lambda_t = \sqrt{H||\nabla f(\mathbf{x}_t)||}$ therefore $H||\nabla f(\mathbf{x}_t)|| = \lambda_t^2$. Using (56) and this result we get:

$$H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t^2 ||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}|| \tag{58}$$

We are now left with estimating $||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}||$. It holds that $||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}|| = max_{||\mathbf{x}||=1} ||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}\mathbf{x}||$. Since the function f is convex, and twice differentiable the second order characterization of convexity (Lemma 2.18) implies that $\nabla^2 f(\mathbf{x})$ is positive semidefinite for every $\mathbf{x} \in \mathbb{R}^d$. This further indicates that the eigenvalues of $\nabla^2 f(\mathbf{x})$ are nonegative

for every $\mathbf{x} \in \mathbb{R}^d$. Since f is twice differentiable, it is continuously differentiable meaning that its hessian matrix $\nabla^2 f(\mathbf{x})$ is symmetric. This implies that the matrix
$\nabla^2 f(\mathbf{x}_t) + \lambda_t I$ is symmetric as well and moreover $(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}$ is symmetric too (since the inverse of a symmetric matrix is symmetric). Moreover, the eigenvalues of $\nabla^2 f(\mathbf{x}_t) + \lambda_t I$ are $\geq \lambda_t$ (since, as said before the eigenvalues of $\nabla^2 f(\mathbf{x})$ are non-negative). In other words the eigenvalues of the matrix $\nabla^2 f(\mathbf{x}_t) + \lambda_t I$ with have the form $\alpha + \lambda_t$ where $\alpha \geq 0$ is any eigenvalue of $\nabla^2 f(\mathbf{x}_t)$. Let α_{max} and α_{min} be the greatest and smallest eigenvalues of $\nabla^2 f(\mathbf{x}_t)$ respectively. Then if $\mathbf{x}$ is any unit vector, the following holds:

$$\alpha_{min} + \lambda_t \leq ||(\nabla^2 f(\mathbf{x}_t) + \lambda_t)\mathbf{x}|| \leq \alpha_{max} + \lambda_t$$

Now, using the fact that if m is the eigenvalue of matrix A then $\frac{1}{m}$ is the eigenvalue of of A^{-1}, we get that $\frac{1}{\alpha_{max}+\lambda_t}$ and $\frac{1}{\alpha_{min}+\lambda_t}$ are the smallest and greatest eigenvalues of the matrix $(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}$ respectively. Thus, given $\mathbf{x}$ is a unit vector, we have:

$$\frac{1}{\alpha_{max} + \lambda_t} \leq ||(\nabla^2 f(\mathbf{x}_t) + \lambda_t)^{-1}\mathbf{x}|| \leq \frac{1}{\alpha_{min} + \lambda_t}$$

Finally, we have $||(\nabla^2 f(\mathbf{x}_t) + \lambda_t)^{-1}|| = \frac{1}{\alpha_{min}+\lambda_t} \leq \frac{1}{\lambda_t}$, since $\alpha_{min} \geq 0$.

We plug this inequality in (57) and get the desired result:

$$H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t^2 ||(\nabla^2 f(\mathbf{x}_t) + \lambda_t I)^{-1}|| \leq \lambda_t^2 \frac{1}{\lambda_t} = \lambda_t \tag{59}$$

ii) Let us now prove the second inequality of this point. We start to show the left-hand side inequality:

$$||\nabla f(\mathbf{x}_{t+1})|| \leq 2\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t||$$

In order to prove this, I make use of the third assumption of this problem, mainly of the property that there is a constant $H > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have $||\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})|| \leq H||\mathbf{y} - \mathbf{x}||^2$. Plug in $\mathbf{y} = \mathbf{x}_{t+1}$ and $\mathbf{x} = \mathbf{x}_t$ and get:

$$||\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)|| \leq H||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{60}$$

We obtain the following series of equalities/inequalities:

$$\begin{aligned}
\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t|| \geq H||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 &\geq ||\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)|| && (61)\\
&\geq ||\nabla f(\mathbf{x}_{t+1})|| - ||\nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x_{t+1}} - \mathbf{x}_t)|| && (62)\\
&= ||\nabla f(\mathbf{x}_{t+1})|| - ||\lambda_t(\mathbf{x}_t - \mathbf{x}_{t+1})|| && (63)
\end{aligned}$$

Explications about the equalities/inequalities
In (60) we use the inequality proven previously: $H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t$ and the inequality in (59). In (61), I use the triangle inequality satisfied by the norm which states that: $||\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)|| + ||\nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x_{t+1}} - \mathbf{x}_t)|| \geq ||\nabla f(\mathbf{x}_{t+1})||$. In (62) we use the equality of 4a): $\lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) = -\nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)$. Finally we get the desired inequality from (62):

$$\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t|| \geq ||\nabla f(\mathbf{x}_{t+1})|| - ||\lambda_t(\mathbf{x}_t - \mathbf{x}_{t+1})|| \Leftrightarrow 2\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t|| \geq ||\nabla f(\mathbf{x}_{t+1})||$$

We just proved the left-hand side inequality: $||\nabla f(\mathbf{x}_{t+1})|| \leq 2\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||$
iii) Now we need to show that

$$2\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq 2||\nabla f(\mathbf{x}_t)||$$

We have the following series of equalities/inequalities:

$$2\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq 2\lambda_t\frac{\lambda_t}{H} = \frac{2\sqrt{H||\nabla f(\mathbf{x}_t)||}\sqrt{H||\nabla f(\mathbf{x}_t)||}}{H} = 2||\nabla f(\mathbf{x}_t)|| \tag{64}$$

Explanations for the step (63):
In the first inequality we make use of the inequality (58): $H||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq \lambda_t$ which implies $||\mathbf{x}_t - \mathbf{x}_{t+1}|| \leq \frac{\lambda_t}{H}$. In the next equality, we only substitute the value for λ_t as assigned by the algorithm 4: $\lambda_t = \sqrt{H||\nabla f(\mathbf{x}_t)||}$. Finally, we just proved the desired inequality: $2\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t|| \leq 2||\nabla f(\mathbf{x}_t)||$.

c) Prove the following descent lemma for Algorithm 4:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{2}{3}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2.$$

Solution 4c):

We use the first inequality of assumption 3, substitute $\mathbf{y} = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$ and obtain the following series:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) + \tag{65}$$

$$+ \frac{H}{3}||\mathbf{x}_{t+1} - \mathbf{x}_t||^3 \tag{66}$$

$$= f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top [-\nabla f(\mathbf{x}_t) - \lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t)] + \tag{67}$$

$$+ \frac{H}{3}||\mathbf{x}_{t+1} - \mathbf{x}_t||^3 \tag{68}$$

$$= f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) - \frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) - \tag{69}$$

$$- \frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{H}{3}||\mathbf{x}_{t+1} - \mathbf{x}_t||^3 \tag{70}$$

$$= f(\mathbf{x}_t) + \frac{1}{2}\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) - \frac{1}{2}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 + \frac{H}{3}||\mathbf{x}_{t+1} - \mathbf{x}_t||^3 \tag{71}$$

$$\leq f(\mathbf{x}_t) - \frac{1}{2}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 - \frac{1}{2}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 + \frac{1}{3}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{72}$$

$$= f(\mathbf{x}_t) - \frac{2}{3}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{73}$$

Explanations about the steps (64)-(72):
In (64)-(65), I make use of the first inequality of assumption 3: $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{H}{3}||\mathbf{y}-\mathbf{x}||^3$. In (66)-(67) I use the result obtained from 4a), mainly that $\lambda_t(\mathbf{x}_{t+1}-\mathbf{x}_t) = -\nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)$. In (68)-(69) I use the property that since $(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t)$ is a scalar, then $(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t)$. In (70) I use that $(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \lambda_t (\mathbf{x}_{t+1} - \mathbf{x}_t) = \lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) = \lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$ since λ_t is a scalar. In the transition from (70) to (71), I use that $\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) \leq -\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$ because $\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) = (\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) = (\mathbf{x}_{t+1} - \mathbf{x}_t)^\top [-\lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t)] = -\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 - (\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) \leq -\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$ where I use the equality proven in 4a):
$\nabla f(\mathbf{x}_t) = -\lambda_t(\mathbf{x}_{t+1} - \mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)$ as well as the property that $\nabla^2 f(\mathbf{x}_t)$ is positive semi definite and therefore $-(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) \leq 0$. In (72) I proved the desired inequality: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{2}{3}\lambda_t||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$

d) Let $\mathcal{I}_\infty = \{i \in \mathbb{N} : ||\nabla f(\mathbf{x}_{i+1})|| \geq \frac{1}{4}||\nabla f(\mathbf{x}_i)||\}$ be the set of iterations at which the norm of gradient shrinks by at least a factor four. Show that for all

$t \in \mathcal{I}_\infty$, we have:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{96B^{3/2}\sqrt{H}}(f(\mathbf{x}_t) - f(\mathbf{x}^*))^{3/2}.$$

Solution 4d):

Guided by the hint, we first show that:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq B||\nabla f(\mathbf{x}_t)||$$

Since f is convex we make use of the first order characterization of convexity which enables:

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*)$$

Moreover $\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \leq ||\nabla f(\mathbf{x}_t)|| \cdot ||\mathbf{x}_t - \mathbf{x}^*||$ holds because of Cauchy-Schwarz inequality which states that $a^\top b \leq ||a|| \cdot ||b||$. Using the fact that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ (from the descent lemma in 4c)), it is true that $f(\mathbf{x}_t) \leq f(\mathbf{x}_0)$ therefore we could use assumption 2 of this exercise and get what we wanted to show:

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \leq ||\nabla f(\mathbf{x}_t)|| \cdot ||\mathbf{x}_t - \mathbf{x}^*|| \leq B||\nabla f(\mathbf{x}_t)|| \quad (74)$$

Next, we can show that for all $t \in \mathcal{I}_\infty$, the following inequality holds:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{96\sqrt{H}}||\nabla f(\mathbf{x}_t)||^{3/2} \quad (75)$$

Using 4c) (descent lemma), we have the following valid inequality:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{2}{3}\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$$

We now make use of one of the inequalities proven in 2b), mainly:
$||\nabla f(\mathbf{x}_{t+1})|| \leq 2\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t||$, equivalent to $||\mathbf{x}_{t+1} - \mathbf{x}_t|| \geq \frac{||\nabla f(\mathbf{x}_{t+1})||}{2\lambda_t}$ which further implies that $-||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \leq -\frac{||\nabla f(\mathbf{x}_{t+1})||^2}{4\lambda_t^2}$. We plug this result as an inequality in the above inequality and get:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{2}{3}\lambda_t ||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \leq -\frac{2}{3}\lambda_t \frac{||\nabla f(\mathbf{x}_{t+1})||^2}{4\lambda_t^2} = -\frac{||\nabla f(\mathbf{x}_{t+1})||^2}{6\lambda_t}$$

Moreover, since $t \in \mathcal{I}_\infty$, the property $||\nabla f(\mathbf{x}_{t+1})|| \geq \frac{1}{4}||\nabla f(\mathbf{x}_t)||$ equivalent to $||\nabla f(\mathbf{x}_{t+1})||^2 \geq \frac{1}{16}||\nabla f(\mathbf{x}_t)||^2$ is satisfied. This further implies that $-||\nabla f(\mathbf{x}_{t+1})||^2 \leq -\frac{1}{16}||\nabla f(\mathbf{x}_t)||^2$ We plug the last inequality in the above set of inequalities together with the equality $\lambda_t = \sqrt{H||\nabla f(\mathbf{x}_t)||}$ (defined in the algorithm 4) and finally obtain:

$$f(\mathbf{x}_{t+1})-f(\mathbf{x}_t) \leq -\frac{||\nabla f(\mathbf{x}_{t+1})||^2}{6\lambda_t} \leq -\frac{1}{16\cdot 6}\frac{||\nabla f(\mathbf{x}_t)||^2}{\lambda_t} = -\frac{1}{96}\frac{||\nabla f(\mathbf{x}_t)||^2}{\sqrt{H||\nabla f(\mathbf{x}_t)||}} = -\frac{||\nabla f(\mathbf{x}_t)||^{3/2}}{96\sqrt{H}}$$

We just show that (74) holds. We now make use of the proven inequality in (73): $f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq B||\nabla f(\mathbf{x}_t)||$ equivalent to $(f(\mathbf{x}_t) - f(\mathbf{x}^*))^{3/2} \leq (B||\nabla f(\mathbf{x}_t)||)^{3/2} = B^{3/2}||\nabla f(\mathbf{x}_t)||^{3/2}$ (raising into the respective power is eligible since both sides of the inequality are positive: we know that $f(\mathbf{x}_t) \geq f(\mathbf{x}^*)$). This further implies that $-||\nabla f(\mathbf{x}_t)||^{3/2} \leq -\frac{(f(\mathbf{x}_t)-f(\mathbf{x}^*))^{3/2}}{B^{3/2}}$. We plug this result as an inequality in the already proven result (74) and get:

$$f(\mathbf{x}_{t+1})-f(\mathbf{x}_t) \leq -\frac{1}{96\sqrt{H}}||\nabla f(\mathbf{x}_t)||^{3/2} \leq -\frac{1}{96\sqrt{H}}\cdot\frac{(f(\mathbf{x}_t) - f(\mathbf{x}^*))^{3/2}}{B^{3/2}} \quad (76)$$

Finally, (75) shows that $f(\mathbf{x}_{t+1})-f(\mathbf{x}_t) \leq -\frac{1}{96B^{3/2}\sqrt{H}}(f(\mathbf{x}_t)-f(\mathbf{x}^*))^{3/2}$ which is what we needed to prove!

References

[1] Wikipedia. *Harmonic series (mathematics)*. [Online; accessed 23-April-2023]. URL: `%5Curl%7Bhttps://en.wikipedia.org/wiki/Harmonic_series_(mathematics)%7D`.