Optimization for Machine Learning HW 5

Shuyue Jia
BUID: U62343813

Due: 10/27/2023

All parts of each question are equally weighted. When solving one question/part, you may assume the results of all previous questions/parts. You may also assume all previous homework results and results from class or lecture notes, but please explain which result you are using when you use it.

In this homework, you will extend the deterministic accelerated algorithm to a stochastic setting. The goal is to obtain a convergence rate like:

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)\right] \leq O\left(\frac{H\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{T^2} + \frac{\sigma\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}\right)$$

Thus, when σ is very small the convergence rate is nearly $O(1/T^2)$, but when σ is larger it decays to the ordinary $O(1/\sqrt{T})$. Obtaining this result in an adaptive way (i.e. via an algorithm that does not know H or σ ahead of time) is rather difficult, although some progress has been made recently. The state-of-the art here is currently this ICML 2020 paper: `http://proceedings.mlr.press/v119/joulani20a.html`.

Throughout this problem, assume that $\mathcal{L}$ is a convex, H-smooth function, and that $\ell(\mathbf{w}, z)$ is such that $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z) - \nabla\mathcal{L}(\mathbf{w})\|^2] \leq \sigma^2$ for all $\mathbf{w}$. Recall that by bias-variance decomposition this also implies $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z)\|^2] \leq \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w})\|^2 + \sigma^2]$ for all (possibly random) $\mathbf{w}$.

Algorithm 1 Accelerated Gradient Descent

Input: Initial Point $\mathbf{w}_1$, smoothness constant H, time horizon T, learning rate η
Set $\mathbf{y}_1 = \mathbf{w}_1$
Set $\alpha_0 = 0$, $\alpha_1 = 1$.
for $t = 1 \dots T$ **do**
  Set $\tau_t = \frac{\alpha_t}{\sum_{i=1}^t \alpha_t}$
  Set $\mathbf{x}_t = (1 - \tau_t)\mathbf{w}_t + \tau_t \mathbf{y}_t$
  Set $\mathbf{g}_t = \alpha_t \nabla\ell(\mathbf{x}_t, z_t)$.
  Set $\mathbf{y}_{t+1} = \mathbf{y}_t - \eta\mathbf{g}_t$.
  Set $\mathbf{w}_{t+1} = \mathbf{x}_t - \eta\nabla\ell(\mathbf{x}_t, z_t)$
  Set α_{t+1} to satisfy $\alpha_{t+1}^2 - \alpha_{t+1} = \sum_{i=1}^t \alpha_i$.
end for

1. Show that Algorithm 1 satisfies:

$$\mathbb{E}\left[\sum_{t=1}^T \alpha_t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star))\right] \leq \mathbb{E}\left[\sum_{t=1}^T \langle\nabla\mathcal{L}(\mathbf{x}_t), \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle + \sum_{t=1}^T \langle\mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle\right]$$

Solution:

Proof. By convexity, we have $\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star) \le \langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star \rangle$, so,

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t (\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) &\le \sum_{t=1}^{T} \alpha_t \langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star \rangle \\
&= \sum_{t=1}^{T} \alpha_t \langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t \rangle + \sum_{t=1}^{T} \alpha_t \langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{y}_t - \mathbf{w}_\star \rangle \\
&= \sum_{t=1}^{T} \langle \nabla \mathcal{L}(\mathbf{x}_t), \alpha_t (\mathbf{x}_t - \mathbf{y}_t) \rangle + \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle.
\end{aligned} \tag{1}$$

Thus, by taking expectations, we will have:

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t (\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star))\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla \mathcal{L}(\mathbf{x}_t), \alpha_t (\mathbf{x}_t - \mathbf{y}_t) \rangle + \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle\right]. \tag{2}$$

□

2. Show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle\right] \le \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2 \eta \sum_{t=1}^{T} \alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t^2 \|\nabla \mathcal{L}(\mathbf{x}_t)\|^2\right].$$

Solution:
Proof. First, we will have:

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{w}_\star\|^2 &= \|\mathbf{y}_t - \eta \mathbf{g}_t - \mathbf{w}_\star\|^2 \\
&= \|\mathbf{y}_t - \mathbf{w}_\star\|^2 - 2\eta \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle + \eta^2 \|\mathbf{g}_t\|^2.
\end{aligned} \tag{1}$$

Thus, we will have:

$$\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle = \frac{\|\mathbf{y}_t - \mathbf{w}_\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\|\mathbf{g}_t\|^2. \tag{2}$$

Sum over t, and telescope:

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle &= \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2 - \|\mathbf{y}_{T+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|^2 \\
&= \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} - \frac{\|\mathbf{y}_{T+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|^2 \\
&\le \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|^2.
\end{aligned} \tag{3}$$

By taking the expectation:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star \rangle\right] \le \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{g}_t\|^2\right]. \tag{4}$$

Since $\mathbf{g}_t = \alpha_t \nabla \ell(\mathbf{x}_t, z_t)$, we will have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle\right] &\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_t \nabla \ell(\mathbf{x}_t, z_t)\|^2\right] \\
&\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla \ell(\mathbf{x}_t, z_t)\|^2\right] \\
&\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\|\nabla \ell(\mathbf{x}_t, z_t)\|^2\right].
\end{aligned} \tag{5}$$

From the problem description, we know that $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z)\|^2] \leq \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w})\|^2 + \sigma^2]$, thus we will have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle\right] &\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\|\nabla \ell(\mathbf{x}_t, z_t)\|^2\right] \\
&\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2(\|\nabla \ell(\mathbf{x}_t)\|^2 + \sigma^2)\right] \\
&\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\|\nabla \ell(\mathbf{x}_t)\|^2\right] + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\sigma^2\right] \\
&\leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\|\nabla \ell(\mathbf{x}_t)\|^2\right] + \frac{\sigma^2\eta}{2}\sum_{t=1}^{T}\alpha_t^2.
\end{aligned} \tag{6}$$

Thus, we will have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle\right] \leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla \mathcal{L}(\mathbf{x}_t)\|^2\right]. \tag{7}$$

$\square$

3. Show that

$$\begin{aligned}
-\,\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t \mathcal{L}(\mathbf{w}_\star)\right] \leq\,& \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \\
&+ \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla \mathcal{L}(\mathbf{x}_t)\|^2\right]
\end{aligned}$$

Solution:
Proof. From the **Problem 1 - Equation 1**, we have:

$$\sum_{t=1}^{T}\alpha_t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) = \sum_{t=1}^{T}\langle \nabla \mathcal{L}(\mathbf{x}_t), \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle + \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle. \tag{1}$$

From the **Problem 2 - Equation 3**, we have:

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle \leq \frac{\|\mathbf{y}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2. \tag{2}$$

Since $\mathbf{x}_t = (1-\tau_t)\mathbf{w}_t + \tau_t\mathbf{y}_t$, then we will have:

$$\mathbf{x}_t = (1-\tau_t)\,\mathbf{w}_t + \tau_t\mathbf{y}_t = \left(1 - \frac{\alpha_t}{\sum_{i=1}^{t}\alpha_i}\right)\mathbf{w}_t + \frac{\alpha_t}{\sum_{i=1}^{t}\alpha_i}\mathbf{y}_t. \tag{3}$$

and

$$\left(\sum_{i=1}^{t}\alpha_i\right)\mathbf{x}_t = \left(\left(\sum_{i=1}^{t}\alpha_i\right) - \alpha_t\right)\mathbf{w}_t + \alpha_t\mathbf{y}_t. \tag{4}$$

By subtracting $\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathbf{x}_t$ and $\alpha_t\mathbf{y}_t$ frombothboth sides:

$$\begin{aligned}\alpha_t\mathbf{x}_t - \alpha_t\mathbf{y}_t &= \left(\left(\sum_{i=1}^{t}\alpha_i\right) - \alpha_t\right)\mathbf{w}_t - \left(\sum_{i=1}^{t}\alpha_i\right)\mathbf{x}_t \\ &= \left(\sum_{i=1}^{t-1}\alpha_i\right)(\mathbf{w}_t - \mathbf{x}_t)\,.\end{aligned} \tag{5}$$

Therefore, we have:

$$\langle\nabla\mathcal{L}(\mathbf{x}_t)\,, \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle = \left(\sum_{i=1}^{t-1}\alpha_i\right)\langle\nabla\mathcal{L}(\mathbf{x}_t)\,, \mathbf{w}_t - \mathbf{x}_t\rangle\,. \tag{6}$$

Now, let's use convexity again: we have $\mathcal{L}(\mathbf{w}_t) \geq \mathcal{L}(\mathbf{x}_t) + \langle\nabla\mathcal{L}(\mathbf{x}_t)\,, \mathbf{w}_t - \mathbf{x}_t\rangle$, so:

$$\langle\nabla\mathcal{L}(\mathbf{x}_t)\,, \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle \leq \left(\sum_{i=1}^{t-1}\alpha_i\right)(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{x}_t))\,. \tag{7}$$

Going back and putting this all together, we have shown:

$$\sum_{t=1}^{T}\alpha_t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 + \sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{x}_t))\,. \tag{8}$$

Now, eventually we are going to want the last sum to telescope. So far there are two obstacles. First, there is a $\mathbf{w}$ instead of a $\mathbf{x}$, and second the coefficients on the $\mathcal{L}(\mathbf{w}_t)$ and $\mathcal{L}(\mathbf{x}_t)$ are the same. Let's fix the second problem first: subtract $\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{x}_t)$ from both sides to get,

$$-\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{w}_\star) \leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 + \sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\,. \tag{9}$$

Taking the expectation, we will have:

$$-\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{w}_\star)\right] \leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right]. \tag{10}$$

Since $\mathbf{g}_t = \alpha_t\nabla\ell(\mathbf{x}_t, z_t)$, we will have:

$$\begin{aligned}-\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{w}_\star)\right] &\leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\alpha_t\nabla\ell(\mathbf{x}_t, z_t)\|^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \\ &\leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\alpha_t^2\|\nabla\ell(\mathbf{x}_t, z_t)\|^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right].\end{aligned} \tag{11}$$

From the problem description, we know that $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z)\|^2] \leq \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w})\|^2 + \sigma^2]$ and similar to **Problem 2**, we will have:

$$\begin{aligned} -\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \mathcal{L}(\mathbf{w}_\star)\right] \leq & \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}\left[\alpha_t^2 \|\nabla \ell(\mathbf{x}_t, z_t)\|^2\right] \\ & + \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}\left(\mathbf{w}_t\right) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}\left(\mathbf{x}_t\right)\right] \\ \leq & \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}\left[\alpha_t^2 \|\nabla \ell(\mathbf{x}_t)\|^2\right] + \frac{\sigma^2\eta}{2}\sum_{t=1}^{T}\alpha_t^2 \\ & + \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}\left(\mathbf{w}_t\right) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}\left(\mathbf{x}_t\right)\right]. \end{aligned} \tag{12}$$

Thus, we will have:

$$\begin{aligned} -\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \mathcal{L}(\mathbf{w}_\star)\right] \leq & \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \\ & + \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right]. \end{aligned} \tag{13}$$

□

4. Show that for any $\eta \leq \frac{1}{H}$, for all t:

$$\mathbb{E}\left[-\mathcal{L}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[-\mathcal{L}(\mathbf{w}_{t+1}) - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 + \frac{\eta\sigma^2}{2}\right]$$

(note the η instead of η^2 in the last term!)

Solution:

Proof. Let's use smoothness to relate $\mathcal{L}(\mathbf{x}_t)$ to $\mathcal{L}(\mathbf{w}_{t+1})$. So long as $\eta \leq \frac{1}{H}$,

$$\mathcal{L}\left(\mathbf{w}_{t+1}\right) \leq \mathcal{L}\left(\mathbf{x}_t\right) - \frac{\eta}{2}\left\|\nabla\mathcal{L}\left(\mathbf{x}_t\right)\right\|^2. \tag{1}$$

Since we know that $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z)\|^2] \leq \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w})\|^2 + \sigma^2]$, then we will have:

$$\begin{aligned} \mathcal{L}\left(\mathbf{w}_{t+1}\right) - \mathcal{L}\left(\mathbf{x}_t\right) & \leq -\frac{\eta}{2}\left\|\nabla\mathcal{L}\left(\mathbf{x}_t\right)\right\|^2 \\ & \leq -\frac{\eta}{2}\left[\|\nabla \ell(\mathbf{x}_t, \mathbf{z}_t)\|^2 - \sigma^2\right] \end{aligned} \tag{2}$$

By taking the expectation, we will have:

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}\left(\mathbf{w}_{t+1}\right) - \mathcal{L}\left(\mathbf{x}_t\right)\right] & \leq -\frac{\eta}{2}\mathbb{E}\left[\|\nabla \ell(\mathbf{x}_t, \mathbf{z}_t)\|^2 - \sigma^2\right] \\ & \leq -\frac{\eta}{2}\mathbb{E}\left[\|\nabla \ell(\mathbf{x}_t, \mathbf{z}_t)\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\sigma^2\right] \\ & \leq -\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 + \frac{\eta}{2}\mathbb{E}\left[\sigma^2\right]. \end{aligned} \tag{3}$$

Thus, we will have:

$$\mathbb{E}\left[-\mathcal{L}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[-\mathcal{L}(\mathbf{w}_{t+1}) - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 + \frac{\eta\sigma^2}{2}\right]. \tag{4}$$

□

5. Show that for any $\eta \leq \frac{1}{H}$:

$$\sum_{t=1}^{T} \alpha_t \, \mathbb{E}\left[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \sigma^2 \eta \sum_{t=1}^{T} \alpha_t^2$$

Solution:
Proof. From **Problem 4**, we have:

$$\mathbb{E}\left[-\mathcal{L}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[-\mathcal{L}(\mathbf{w}_{t+1}) - \frac{\eta}{2}\|\nabla \mathcal{L}(\mathbf{x}_t)\|^2 + \frac{\eta\sigma^2}{2}\right]. \tag{1}$$

By rearranging and multipling $\sum_{i=1}^{t-1} \alpha_i$ and adding $\mathcal{L}(\mathbf{w}_t)$, we will have:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \leq &\mathbb{E}[\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) \\ &- \left(\sum_{i=1}^{t}\alpha_i\right)\left(\mathcal{L}(\mathbf{w}_{t+1}) + \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 - \frac{\eta\sigma^2}{2}\right)].\end{aligned} \tag{2}$$

Then, by rearranging:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \leq &\mathbb{E}\left[\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{w}_{t+1})\right] \\ &- \mathbb{E}\left[\left(\sum_{i=1}^{t}\alpha_i\right)\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right] + \frac{\eta\sigma^2}{2}\sum_{i=1}^{t}\alpha_i.\end{aligned} \tag{3}$$

From **Problem 3 - Equation 13**, we have:

$$\begin{aligned}-\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}&\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{x}_t)\right] \\ &+ \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right].\end{aligned} \tag{4}$$

Thus, we then have:

$$\begin{aligned}-\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{w}_\star)\right] \leq &- \mathbb{E}\left[\left(\sum_{t=1}^{T}\alpha_t\right)\mathcal{L}(\mathbf{w}_{T+1})\right] \\ &+ \frac{\eta\sigma^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{t}\alpha_i - \sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{i=1}^{t}\alpha_i\right)\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right] \\ &+ \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right].\end{aligned} \tag{5}$$

By rearranging:

$$\begin{aligned}\left(\sum_{t=1}^{T}\alpha_t\right)\left(\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)\right) \leq &\frac{\eta\sigma^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{t}\alpha_i - \sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{i=1}^{t}\alpha_i\right)\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right] \\ &+ \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta} + \frac{\sigma^2\eta\sum_{t=1}^{T}\alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\right].\end{aligned} \tag{6}$$

Since we know that $\alpha_{t+1}^2 - \alpha_{t+1} = \sum_{i=1}^{t} \alpha_i$ and $\alpha_t^2 - \alpha_t = \sum_{i=1}^{t-1} \alpha_i$, we will have $\forall t \geq 1, \alpha_t^2 = \sum_{i=1}^{t} \alpha_i$

$$\begin{aligned}\left(\sum_{t=1}^{T} \alpha_t\right) \left(\mathcal{L}\left(\mathbf{w}_{T+1}\right) - \mathcal{L}\left(\mathbf{w}_\star\right)\right) \leq & \frac{\eta\sigma^2}{2} \sum_{t=1}^{T} \alpha_t^2 - \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[\alpha_t^2 \left\|\nabla \mathcal{L}\left(\mathbf{x}_t\right)\right\|^2\right] \\ & + \frac{\left\|\mathbf{w}_\star - \mathbf{y}_1\right\|^2}{2\eta} + \frac{\eta\sigma^2 \sum_{t=1}^{T} \alpha_t^2}{2} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t^2 \left\|\nabla \mathcal{L}\left(\mathbf{x}_t\right)\right\|^2\right].\end{aligned} \tag{7}$$

Then, we will have:

$$\left(\sum_{t=1}^{T} \alpha_t\right) \left(\mathcal{L}\left(\mathbf{w}_{T+1}\right) - \mathcal{L}\left(\mathbf{w}_\star\right)\right) \leq \eta\sigma^2 \sum_{t=1}^{T} \alpha_t^2 + \frac{\left\|\mathbf{w}_\star - \mathbf{y}_1\right\|^2}{2\eta}. \tag{8}$$

□

6. Choose a value for η such that:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)] \leq O\left(\frac{H\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{T^2} + \frac{\sigma\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}\right)$$

Your choice for η may depend on values unknown in practice, such as $\|\mathbf{w}_\star - \mathbf{y}_1\|$. You would normally have to tune the learning rate to obtain this result without this knowledge.

Solution:

Proof. In the **Proposition 15.3**, for all $t \geq 1$, we have

$$\frac{t^2}{9} \leq \sum_{i=1}^{t} \alpha_i \leq t^2. \tag{1}$$

Since we know $\alpha_t^2 = \sum_{i=1}^{t} \alpha_i$, we will have:

$$\sum_{t=1}^{T} \alpha_t^2 = \sum_{t=1}^{T} \sum_{i=1}^{t} \alpha_i \leq \sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}. \tag{2}$$

From **Problem 5**, we know that:

$$\frac{T^2}{9} \mathbb{E}\left[\mathcal{L}\left(\mathbf{w}_{T+1}\right) - \mathcal{L}\left(\mathbf{w}_\star\right)\right] \leq \frac{\left\|\mathbf{w}_\star - \mathbf{y}_1\right\|^2}{2\eta} + \sigma^2\eta \frac{T(T+1)(2T+1)}{6}. \tag{3}$$

Here, we will use:

$$\eta = \min\left(\frac{1}{H}, c\frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}}\right), \tag{4}$$

where c is a constant.

Thus, we will have:

(1) If $\eta = \frac{1}{H} \leq c\frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}}$, we will have:

$$\frac{\left\|\mathbf{w}_\star - \mathbf{y}_1\right\|^2}{\eta T^2} = \frac{H\left\|\mathbf{w}_\star - \mathbf{y}_1\right\|^2}{T^2}. \tag{5}$$

As a result, we will have:

$$\sigma^2\eta T \le \sigma^2 T c \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}} = c\frac{\sigma\, \|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}. \tag{6}$$

(2) If $\eta = c\frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}}$, we will have:

$$\frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{\eta T^2} = \frac{\sigma\, \|\mathbf{w}_\star - \mathbf{y}_1\|^2}{c\sqrt{T}}. \tag{7}$$

As a result, we will have:

$$\sigma^2\eta T = c\frac{\sigma\, \|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}. \tag{8}$$

Finally, we will have:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)] \le O\left(\frac{H\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{T^2} + \frac{\sigma\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}\right). \tag{9}$$

□