Optimization for Machine Learning HW 5

SOLUTIONS

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:

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

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 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`, which can adapt to an unknown values of σ and H.

Throughout this problem, assume that $\mathcal{L}$ is a convex, H-smooth function, and that $\ell(\mathbf{w}, z)$ is such that $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z) - \nabla\mathcal{L}(\mathbf{w})\|^2] \le \sigma^2$ for all $\mathbf{w}$. Recall that by bias-variance decomposition this also implies $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z)\|^2] \le \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w})\|^2 + \sigma^2]$ 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 $\eta$
  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 $\alpha_{t+1}$ to satisfy $\alpha_{t+1}^2 - \alpha_{t+1} = \sum_{i=1}^t \alpha_i$.
  end for
```

1. Show that Algorithm 1 satisfies:

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

Solution:

By convexity, we have:

$$\alpha_t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) \le \alpha_t\langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star\rangle$$

taking expectations:

$$\mathbb{E}[\alpha_t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star))] \le \mathbb{E}[\langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star\rangle]$$

now sum over r:

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

now, use the fact that z_t is independent of both $\mathbf{x}_t$ and $\mathbf{y}_t$ and $\mathbb{E}[\nabla\ell(\mathbf{x}_t, z_t)] = \nabla\mathcal{L}(\mathbf{x}_t)$:

$$\begin{aligned}
&= \left[\sum_{t=1}^{T} \alpha_t \nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{y}_t\rangle + \sum_{t=1}^{T} \alpha_t \langle \nabla\ell(\mathbf{x}_t, z_t), \mathbf{y}_t - \mathbf{w}_\star\rangle\right] \\
&= \left[\sum_{t=1}^{T} \langle \nabla\mathcal{L}(\mathbf{x}_t), \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle + \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{w}_\star\rangle\right]
\end{aligned}$$

2. Show that

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

Solution:

From Notes 8 Theorem 5 and the definition of $\mathbf{y}_t$, we have

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

now taking expectations allows us to conclude:

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

Finally, notice that by definition of $\mathbf{g}_t$ and bias-variance decomposition, we have:

$$\mathbb{E}[\|\mathbf{g}_t\|^2] = \mathbb{E}[\alpha_t^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 + \alpha_t^2\sigma^2]$$

so plugging this in yields:

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

distributing the sum over σ^2 now shows the desired result.

3. Show that

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

Solution:
First, using the same argument as in the proof of Theorem 6 in Notes 8, we rearrange the definition of $\mathbf{x}_t$ to obtain:

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

Further, by convexity, we have:

$$\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_t) \le \langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_t \rangle$$

Therefore:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla \ell(\mathbf{x}_t, z_t), \alpha_t(\mathbf{x}_t - \mathbf{y}_t)\rangle\right] &\le \mathbb{E}\left[\left(\sum_{i=1}^{t-1}\alpha_i\right)\langle \nabla \mathcal{L}(\mathbf{x}_t), \mathbf{w}_t - \mathbf{x}_t\rangle\right] \\ &\le \left(\sum_{i=1}^{t-1}\alpha_i\right)\mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{x}_t)\right]\end{aligned}$$

Now, plug this result and the result of question 2 into the result of question 1 to obtain:

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

Now subtract $\mathbb{E}[\sum_{t=1}^{T}\alpha_t\mathcal{L}(\mathbf{x}_t)]$ from both sides to obtain the desired result.

4. Show that for any $\eta \le \frac{1}{H}$:

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

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

Solution:

Observe that by our standard result for smooth losses:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \le \mathbb{E}\left[\mathcal{L}(\mathbf{x}_t) - \eta\langle\nabla\mathcal{L}(\mathbf{x}_t), \nabla\ell(\mathbf{x}_t, z_t)\rangle + \frac{\eta^2 H}{2}\|\nabla\ell(\mathbf{x}_t, z_t)\|^2\right]$$

from bias-variance decomposition:

$$\le \mathbb{E}\left[\mathcal{L}(\mathbf{x}_t) - \eta\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 + \frac{\eta^2 H}{2}(\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 + \sigma^2)\right]$$

using $\eta \le \frac{1}{H}$:

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

rearranging now yields the desired result.

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

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

Solution:
From the previous question, we have:

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

Substituting this into the result of the question 3:

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

Now, use the fact that $\alpha_t^2 = \sum_{i=1}^{t} \alpha_i$:

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

telescope the sum:

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

Rearrange to conclude the result.

6. Choose a value for η such that:

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

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:

From the previous question, we have for any $\eta \leq \frac{1}{H}$:

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

Further, by Proposition 15.3 in the notes, we have:

$$\frac{t^2}{9} \le \sum_{t=1}^{T} \alpha_t \le t^2$$

Further, $\alpha_t^2 = \sum_{i=1}^{t} \alpha_i \le t^2$. Therefore:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star)\right] &\le \frac{9\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2T^2\eta} + \frac{9\sigma^2\eta\sum_{t=1}^{T} t^2}{T^2} \\
&\le \frac{9\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2T^2\eta} + \frac{9\sigma^2\eta T^3}{T^2} \\
&= \frac{9\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2T^2\eta} + 9\sigma^2\eta T
\end{aligned}$$

Let us set $\eta = \min\left(\frac{1}{H}, \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}}\right)$. We consider two cases. First, if $\eta = \frac{1}{H}$, then we must have

$$\begin{aligned}
\frac{1}{H} &\le \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma\sqrt{T}} \\
\sigma^2 T &\le \frac{H^2\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{T^2}
\end{aligned}$$

Therefore,

$$\frac{9\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2T^2\eta} + 9\sigma^2\eta T \le \frac{13.5H\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{T^2}$$

Otherwise, we have $\eta = \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sigma T^{3/2}}$ so that:

$$\frac{9\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2T^2\eta} + 9\sigma^2\eta T \le \frac{13.5\sigma\|\mathbf{w}_\star - \mathbf{y}_1\|}{\sqrt{T}}$$

adding the two cases shows the final result.