Optimization for Machine Learning HW 5

name

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:

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:

3. Show that

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

Solution:

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:

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:

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: