ADAM Problem Set

December 3, 2020

1 Problems

1. Consider the data set

$$x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, y_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$x_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, y_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

   - What is w^*, assuming the bias is 0
   - Let $w = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}$, and bias $b = 0$. Using ℓ_2 loss $\ell_2(w) = \sum_{i=1}^{n} \|y_i - (wx_i + b)\|_2$ and w^* from above, calculate the regret R.

2. Note that a stochastic optimization is written as $\min_{x \in X} \mathbb{E}_{\xi}[(f(x, \xi)]$; however, it is hard to work on the this problem directly due to the expensiveness of taking the expectation. So instead of solving the optimization problem directly, we can consider another approach by considering a sequence of sample: ξ_1,ξ_2,...ξ_n and consider the following optimization problem: $\min_{x \in X} \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i)$. Give an intuitive explanation why does this approach make sense? This approach is actually part of the intuition behind regret analysis

3. In ADAM's paper, the **Lemma 10.3** can be formulated as follow: suppose $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ and let $\|\mathbf{x}\|_\infty \leq \mathbf{G}_\infty$,then we have:

$$\sum_{t=1}^{n} \sqrt{\frac{x_t^2}{t}} \leq 2\mathbf{G}_\infty \|\mathbf{x}\|_2 \tag{1}$$

   (a) Show that this lemma is wrong even when $n = 1$

   (b) Sow that Eq.(1) is true when $\mathbf{x} = (1, 1, \cdots, 1)$

4. Define $Reg(T, H) = \sum_{t=1}^{T} l_t(h_t) - \min_{h' \in H} \sum_{t=1}^{T} l_t(h')$. Show that Adam achieves regret $Reg(T, H) = R(T) = O(\sqrt{T})$.

5. Graduate Question: Let $h_t(x) = h(x_t, \theta_t)$. Let $l_t(h_t) = l(h(x_t, \theta_t), y_t)$. Assume l_t are i.i.d from D, define $L(h, D) = \mathbb{E}_{l_t \sim D} l_t(h)$ for any $t = 1, ..., T$. Show that the following is true for ADAM:

$$\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} L(h_t, D)] \leq \min_{h' \in H} \mathbb{E}[L(h', D)] + \frac{R(T)}{T}$$

2 Solutions

1.
   - $w^* = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$
   - $R = \ell_2(w) - \ell_2(w^*) = \| \begin{bmatrix} -1 \\ -2 \end{bmatrix} \| + \| \begin{bmatrix} 1 \\ 0 \end{bmatrix} \| - 0 \approx 3.2361$

2. By law of large number, $\mathbb{E}_\xi[(f(x,\xi)] \approx \frac{1}{n}\sum_{i=1}^n f(x,\xi_i)$.

3. (a) consider $x = \frac{1}{4}$

   (b) note that $\sum_{t=1}^n \frac{1}{\sqrt{t}} \leq \sum_{t=1}^n \frac{2}{\sqrt{t}+\sqrt{t-1}} = \sum_{t=1}^n 2(\sqrt{t} - \sqrt{t-1}) = 2\sqrt{n}$

4. Recall that Adam achieves $R(T) = \sum_{t=1}^T f_t(\theta_t) - \sum_{t=1}^T f_t(\theta^*) = O(\sqrt{T})$. Then let $f_t(\theta_t) = l(h(x_t,\theta_t),y_t)$, so $R(T) = \sum_{t=1}^T l(h(x_t,\theta_t),y_t) - l(h(x_t,\theta^*),y_t)$. On the other hand, $\sum_{t=1}^T l_t(h_t) = \sum_{t=1}^T l(h_t(x_t),y_t) = \sum_{t=1}^T l(h(x_t,\theta_t),y_t)$. So $Reg(T,H) = \sum_{t=1}^T l_t(h_t) - \sum_{t=1}^T l_t(h^*) = \sum_{t=1}^T l(h(x_t,\theta_t),y_t) - l(h(x_t,\theta^*),y_t) = R(T)$

5. Since $Reg(T,H) = R(T)$, we have: $\sum_{t=1}^T l_t(h_t) = \min_{h'\in H}\sum_{t=1}^T l_t(h') + R(T)$. Take expectation with respect to the algorithm as well as $l \sim D$, we have:$\mathbb{E}[\sum_{t=1}^T l_t(h_t)] = \mathbb{E}[\min_{h'\in H}\sum_{t=1}^T l_t(h')] + R(T) \leq \min_{h'\in H}\mathbb{E}[\sum_{t=1}^T l_t(h')] + R(T)$, where the inequality follows from applying Jensen's inequality with the fact that min is concave. Now, Denote $H_t = \{h_1, l_1, ..., h_{t-1}, l_{t-1}, h_t\}$ to be the entire history up to iterate t, by the tower property, we have: $\mathbb{E}[l_t(h_t)] = \mathbb{E}_{H_t}\mathbb{E}[l_t(h_t)|H_t]$. When we condition $l_t(h_t)$ with H_t, the only source of randomness is from $l_t \sim D$, so: $\mathbb{E}[l_t(h_t)|H_t] = \mathbb{E}_{l_t\sim D}[l_t(h_t)] = L(h_t,D)$. Similarly, $\mathbb{E}[\sum_{t=1}^T l_t(h')] = \sum_{t=1}^T \mathbb{E}[l_t(h')] = \sum_{t=1}^T L(h',D) = T\cdot L(h',D)$. Put everything together, we have: $\mathbb{E}[\sum_{t=1}^T L(h_t,D)] \leq T\cdot\min_{h'\in H}\mathbb{E}[L(h',D)] + R(T)$. Divide both sides by T, and we have what we wanted to show.