Homework 4

Instructor: Haipeng Luo

1. (**Two-point Bandit**) In Lecture 18, we discussed the challenging Bandit Convex Optimization problem and a variant of SCRiBLe that obtains $\mathcal{O}(T^{3/4})$ regret for Lipschitz functions. In this exercise, you will see that the regret can be improved to $\mathcal{O}(\sqrt{T})$ if we are allowed to query the function value for just *one extra point*. This is called the two-point bandit problem. Specifically, consider Algorithm 1, which is similar to the algorithm we discussed in Lecture 18, except for the extra query of point $\widetilde{w}'_t$ and a different way of constructing the estimator $\widehat{g}_t$.

Algorithm 1: SCRiBLe for Two-point Bandit

Input: parameter $\delta \in (0, 1]$, learning rate $\eta > 0$, and a ν-self-concordant function ψ
for $t = 1, \ldots, T$ **do**
- compute $w_t = \operatorname{argmin}_{w \in \Omega} \sum_{\tau=1}^{t-1} w^\top \widehat{g}_\tau + \frac{1}{\eta}\psi(w)$
- compute Hessian $H_t = \nabla^2 \psi(w_t)$ and sample $s_t \in \mathbb{S}^d$ uniformly at random
- play $\widetilde{w}_t = w_t + \delta H_t^{-\frac{1}{2}} s_t$, suffer and observe $f_t(\widetilde{w}_t)$
- **extra step for two-point bandit**: query the function value at point $\widetilde{w}'_t = w_t - \delta H_t^{-\frac{1}{2}} s_t$
- construct estimator $\widehat{g}_t = \frac{d}{2\delta}(f_t(\widetilde{w}_t) - f_t(\widetilde{w}'_t)) H_t^{\frac{1}{2}} s_t$

(a) Prove that $\widehat{g}_t$ is still an unbiased estimator of the gradient $\nabla \widehat{f}_t(w_t)$ (recall $\widehat{f}_t(w) = \mathbb{E}_{b \sim \mathbb{B}^d}[f_t(w + \delta H_t^{-\frac{1}{2}} b)]$).

(b) Prove that if f_t is L-Lipschitz, that is, for all $w, w' \in \Omega$, $|f_t(w) - f_t(w')| \leq L \|w - w'\|_2$, then $\|\widehat{g}_t\|^\star_{w_t} \leq dLD$, where $D = \max_{w,w' \in \Omega} \|w - w'\|_2$ is the diameter Ω. (Note that this bound is independent of δ!)

(c) Prove that for sufficiently small δ and an appropriate choice of learning rate η, the expected regret of Algorithm 1 is $\widetilde{\mathcal{O}}\left(dLD\sqrt{\nu T}\right)$. (Hint: in Lecture 18 we decomposed the regret into five terms and analyzed them separately. See which terms need to be bounded differently here. Also, when setting the learning rate η you can assume that T is large enough, as we did in Lecture 18.)

2. (**Reducing Oracle Calls**) In Lecture 19, we discussed FTL and Epsilon-Greedy for the i.i.d. contextual bandit problem in the full information setting and bandit setting respectively. Both of these algorithms make (at most) one oracle call per round. In this exercise, we will analyze "lazy versions" of these algorithms that reduce the number of oracle calls significantly while ensuring the same regret guarantees.

Recall that in the i.i.d. contextual bandit setting, each (x_t, ℓ_t) is an i.i.d. sample of an unknown joint distribution $\mathcal{D}$. The expected loss of a policy π is denoted by $\bar{\ell}(\pi) = \mathbb{E}_{(x,\ell)\sim\mathcal{D}}[\ell(\pi(x))]$ and the policy with the smallest expected loss is denoted by $\pi^\star = \operatorname{argmin}_{\pi \in \Pi} \bar{\ell}(\pi)$. Regret is defined as $\mathcal{R}_T = \sum_{t=1}^T \left(\ell_t(a_t) - \bar{\ell}(\pi^\star)\right)$.

(a) First consider Algorithm 2 for the full information setting, which clearly only makes $\mathcal{O}(\ln T)$ oracle calls for T rounds.

Algorithm 2: Lazy FTL

play $a_1 \in [K]$ uniformly at random for the first round
for $k = 0, 1, 2, \ldots$ **do**
  query the oracle once to obtain $\pi_k = \operatorname{argmin}_{\pi \in \Pi} \sum_{\tau=1}^{2^k} \ell_\tau(\pi(x_\tau))$
  for $t = 2^k + 1, \ldots, 2^{k+1}$ **do**
    observe x_t, play $a_t = \pi_k(x_t)$, and observe ℓ_t

Algorithm 3: Explore-then-exploit

Input: an integer T_0 between 1 and T
for $t = 1, \ldots, T_0$ **do**
  play $a_t \in [K]$ uniformly at random
  construct usual importance-weighted estimator $\widehat{\ell}_t$
query the oracle once to obtain $\widehat{\pi} = \operatorname{argmin}_{\pi \in \Pi} \sum_{\tau=1}^{T_0} \widehat{\ell}_\tau(\pi(x_\tau))$
for $t = T_0 + 1, \ldots, T$ **do**
  observe x_t and play $a_t = \widehat{\pi}(x_t)$

(i) Prove that for any T, with probability $1 - \delta/2$ we have

$$\sum_{t=1}^{T} \ell_t(a_t) \leq \sum_{t=1}^{T} \bar{\ell}(\pi_{\lceil \log_2 t \rceil - 1}) + \mathcal{O}\left(\sqrt{T \ln(1/\delta)}\right).$$

(ii) Prove that with probability $1 - \delta/2$, for all $k = 0, 1, \ldots, \lceil \log_2 T \rceil - 1$ we have

$$\bar{\ell}(\pi_k) \leq \bar{\ell}(\pi^\star) + \mathcal{O}\left(\sqrt{\frac{\ln((\ln T) N/\delta)}{2^k}}\right).$$

(iii) Conclude that for any T, with probability $1-\delta$ we have $\mathcal{R}_T = \mathcal{O}\left(\sqrt{T \ln((\ln T) N/\delta)}\right)$.

(b) Next consider Algorithm 3 for the bandit setting, which in total only makes *one* oracle call.

(i) With probability $1 - \delta/2$ we have

$$\bar{\ell}(\widehat{\pi}) \leq \bar{\ell}(\pi^\star) + \mathcal{O}\left(\sqrt{\frac{K \ln(N/\delta)}{T_0}} + \frac{K \ln(N/\delta)}{T_0}\right),$$

(ii) Conclude that with an appropriate choice of T_0, Algorithm 3 ensures that with probability $1 - \delta$, $\mathcal{R}_T = \mathcal{O}\left(T^{\frac{2}{3}} (K \ln(N/\delta))^{\frac{1}{3}} + \sqrt{TK \ln(N/\delta)}\right)$.

Algorithm 4: Exp4 with Randomized Policies

Input: learning rate η
for $t = 1, \dots, T$ **do**
  compute $P_t \in \Delta(\Pi)$ such that $P_t(\pi) \propto \exp\left(\eta \sum_{\tau=1}^{t-1} \left\langle \widehat{\ell}_\tau, \pi(x_\tau) \right\rangle\right)$
  play $a_t \sim p_t = \sum_{\pi \in \Pi} P_t(\pi)\pi(x_t)$
  construct usual importance-weighted estimator $\widehat{\ell}_t$

3. (**Randomized Policies**) In the lectures we only consider deterministic policies for contextual bandit, that is, each policy is a function from the context space $\mathcal{X}$ to $[K]$. However, it is also easy to generalize most of the results to the case of randomized policies where each policy is a function from the context space $\mathcal{X}$ to the simplex $\Delta(K)$. Exp4 can simply handle this setup as shown in Algorithm 4. Prove the following regret bound for this algorithm:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{\pi \in \Pi} \sum_{t=1}^{T} \langle \ell_t, \pi(x_t) \rangle \leq 2\sqrt{TM \ln N}$$

where $M = \max_t \sum_{a=1}^{K} \max_{\pi \in \Pi} \pi(x_t)(a)$, using the optimal choice of η (which for simplicity can depend on the unknown quantity M). Further argue that $M \leq \min\{K, N\}$ such that the bound is simply $\mathcal{O}(\sqrt{T \min\{K, N\} \ln N})$.