Adversarial Bandit Linear Optimization

Review: Online Linear Optimization

Given: Convex feasible set $\Omega \subseteq \mathbb{R}^d$

For time $t = 1, 2, \ldots, T$:

Learner chooses a point $w_t \in \Omega$

Environment reveals a reward vector $r_t \in \mathbb{R}^d$

$$\text{Regret} = \max_{w \in \Omega} \sum_{t=1}^{T} \langle w, r_t \rangle - \sum_{t=1}^{T} \langle w_t, r_t \rangle$$

Projected Gradient Descent

Arbitrary $w_1 \in \Omega$

$w_{t+1} = \Pi_\Omega(w_t + \eta r_t)$

Review: Online Linear Optimization

Theorem. Projected Online Gradient Descent ensures

$$\text{Regret} = \max_{w^\star \in \Omega} \sum_{t=1}^{T} \langle w^\star - w_t, r_t \rangle \leq \frac{\max_{w \in \Omega} \|w\|_2^2}{\eta} + \eta \sum_{t=1}^{T} \|r_t\|_2^2$$

Bandit Linear Optimization

Given: Convex feasible set $\Omega \subseteq \mathbb{R}^d$

For time $t = 1, 2, \ldots, T$:

Environment decides the reward vector $r_t \in \mathbb{R}^d$

Learner chooses a point $w_t \in \Omega$

Environment reveals $\langle w_t, r_t \rangle + \epsilon_t$, where ϵ_t is a zero-mean noise

$$\text{Regret} = \max_{w \in \Omega} \sum_{t=1}^{T} \langle w, r_t \rangle - \sum_{t=1}^{T} \langle w_t, r_t \rangle$$

Unbiased Gradient Estimator

Goal: construct a $\hat{r}_t \in \mathbb{R}^d$ with $\mathbb{E}[\hat{r}_t] = r_t$ (using only the feedback $\langle w_t, r_t \rangle + \epsilon_t$)

Unbiased Gradient Estimator (1/3)

Uniformly randomly choose a direction $i_t \in \{1, 2, \ldots, d\}$

Uniformly randomly choose $\alpha_t \in \{1, -1\}$

Sample $\widetilde{w}_t = w_t + \delta \alpha_t \mathrm{e}_{i_t}$

Observe $y_t = \langle \widetilde{w}_t, r_t \rangle + \epsilon_t$

Define $\hat{r}_t = \frac{d y_t}{\delta} \alpha_t \mathrm{e}_{i_t}$

Property 1: $\mathbb{E}[\hat{r}_t] = r_t$

Property 2: $\mathbb{E}[\widetilde{w}_t] = w_t$

Unbiased Gradient Estimator (2/3)

Uniformly randomly choose s_t from the unit sphere $\mathbb{S}_d = \{s \in \mathbb{R}^d : \|s\|_2 = 1\}$

Sample $\widetilde{w}_t = w_t + \delta s_t$

Observe $y_t = \langle \widetilde{w}_t, r_t \rangle + \epsilon_t$

Define $\hat{r}_t = \frac{d y_t}{\delta} s_t$

Property 1: $\mathbb{E}[\hat{r}_t] = r_t$

Property 2: $\mathbb{E}[\widetilde{w}_t] = w_t$

Unbiased Gradient Estimator (3/3)

Uniformly randomly choose $s_t \sim \mathcal{D}$ with $\mathbb{E}_{s\sim\mathcal{D}}[s] = 0$

Sample $\widetilde{w}_t = w_t + s_t$

Observe $y_t = \langle \widetilde{w}_t, r_t \rangle + \epsilon_t$

Define $\hat{r}_t = y_t H_t^{-1} s_t$ where $H_t \coloneqq \mathbb{E}_{s\sim\mathcal{D}}[ss^\top]$

Property 1: $\mathbb{E}[\hat{r}_t] = r_t$

Property 2: $\mathbb{E}[\widetilde{w}_t] = w_t$

Projected Gradient Descent for Bandit Linear Optimization

Assume the feasible set Ω contains a ball of radius δ

Define $\Omega' = \{w \in \Omega:\ \mathcal{B}(w, \delta) \subset \Omega\}$

Arbitrarily pick $\widetilde{w}_1 \in \Omega'$

For $t = 1,2, \ldots, T$:

Let $\widetilde{w}_t = w_t + \delta s_t$ where $s_t \in \mathbb{R}^d$ is uniformly sampled from unit sphere

Receive $y_t = \langle \widetilde{w}_t, r_t \rangle + \epsilon_t$

Define

$$\hat{r}_t = \frac{d y_t}{\delta} s_t$$

Update policy:

$$w_{t+1} = \Pi_{\Omega'} \left(w_t + \eta \hat{r}_t \right)$$

Regret Bound for Bandit Linear Optimization

Theorem. Suppose $\max_{w \in \Omega} \|w\| \leq D$, $\max_{t} \|r_t\| \leq G$. Then the BLO algorithm ensures

$$\text{Regret} = \max_{w^\star \in \Omega} \mathbb{E}\left[\sum_{t=1}^{T} \langle w^\star - w_t, r_t \rangle\right] \leq O\left(\frac{D^2}{\eta} + \eta \frac{d^2 D^2 G^2}{\delta^2} T + \delta G T\right) = O\left(DG\sqrt{d}T^{3/4}\right)$$

Bandit Optimization / Zeroth-Order Optimization

For time $t = 1, 2, \ldots, T$:

Learner chooses a point w_t

Environment reveals $R_t(w_t) + \epsilon_t$, where ϵ_t is a zero-mean noise