

Adversarial Bandit Linear Optimization

Chen-Yu Wei

Review: Online Linear Optimization

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

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

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

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

Bandit Linear Optimization

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

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

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

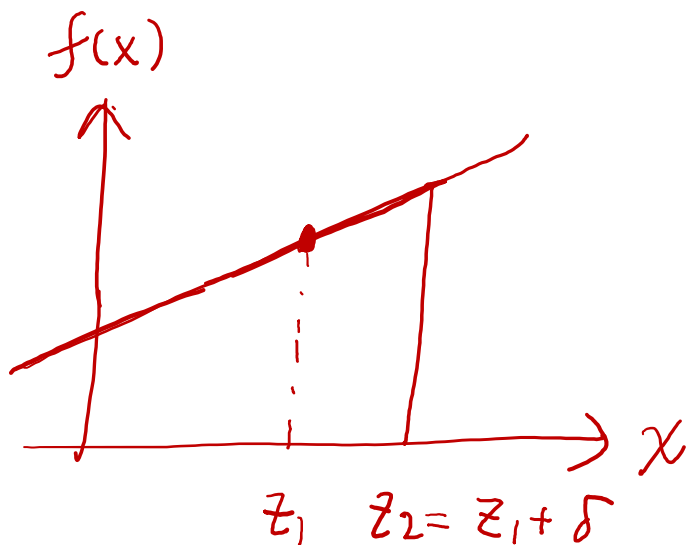
Learner chooses a point $w_t \in \Omega$

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 $\langle w_t, r_t \rangle + \epsilon_t$)



Env tells $f(z)$

$$f(x) = ax + b$$

$$\begin{cases} f(z_1) = az_1 + b \\ f(z_2) = az_2 + b \end{cases}$$

$$f(z_1) - f(z_2) = a(z_1 - z_2)$$

$$a = \frac{f(z_2) - f(z_1)}{z_2 - z_1} = \frac{f(z_2) - f(z_1)}{\delta}$$

$$z = \begin{cases} z_1 & \text{with prob } 1/2 \\ z_2 & \text{with prob } 1/2 \end{cases}$$

$$\hat{a} = \frac{2f(z_2)}{\delta} \mathbb{1}\{z=z_2\} - \frac{2f(z_1)}{\delta} \mathbb{1}\{z=z_1\}$$

$$\mathbb{E}[\hat{a}] = \frac{2f(z_2)}{\delta} \cdot \frac{1}{2} - \frac{2f(z_1)}{\delta} \cdot \frac{1}{2} = a$$

Unbiased Gradient Estimator (1/3)

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{-th entry}$$

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

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

Sample $\tilde{w}_t = w_t + \delta \alpha_t e_{i_t}$

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

Define $\hat{r}_t = \frac{dy_t}{d\delta} \alpha_t e_{i_t}$



$$\hat{r}_t = \frac{d}{d\delta} \left(\langle \tilde{w}_t, r_t \rangle + \epsilon_t \right) \alpha_t e_{i_t}$$

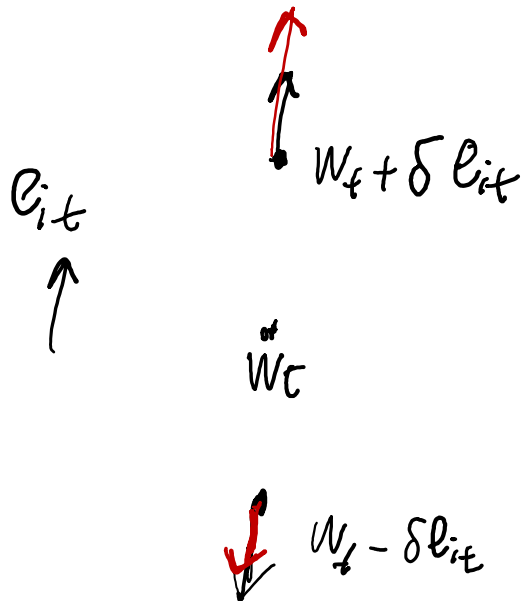
$$= \frac{d}{d\delta} \left(\langle w_t + \delta \alpha_t e_{i_t}, r_t \rangle + \epsilon_t \right) \alpha_t e_{i_t}$$

$$\mathbb{E}[\hat{r}_t] = \mathbb{E} \left[\underbrace{\frac{d}{d\delta} \left(\langle w_t, r_t \rangle \right) \alpha_t e_{i_t}}_0 + \frac{d}{d\delta} \left(\langle \delta \alpha_t e_{i_t}, r_t \rangle \right) \alpha_t e_{i_t} \right]$$

$$\begin{aligned} \mathbb{E} \left[\frac{d}{d\delta} \left(\langle \delta \alpha_t e_{i_t}, r_t \rangle \right) \alpha_t e_{i_t} \right] &= \mathbb{E} \left[\sum_{i=1}^d \frac{1}{d} \langle e_i, r_t \rangle e_i \right] \\ &= \sum_{i=1}^d \langle e_i, r_t \rangle e_i = r_t \end{aligned}$$

Unbiased Gradient Estimator (1/3)

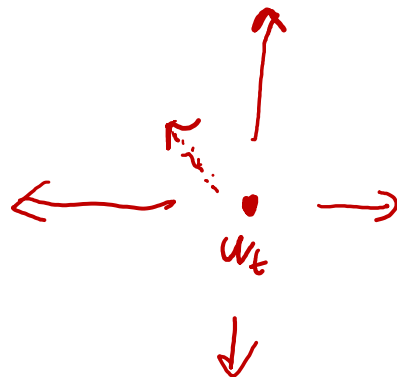
$$\hat{r}_t = \frac{d y_t}{\delta} \boxed{\alpha_t e_{it}}$$



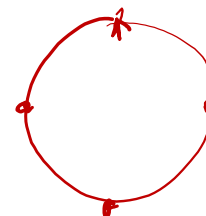
$$i_t \sim \text{uniform} \{1, \dots, d\} \quad \alpha_t \sim \text{unif} \{-1, 1\}$$

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

$$\tilde{w}_t = w_t + \delta \alpha_t e_{it}$$



$$\mathbb{E}[s_t s_t^T] = \sum_i \frac{1}{d} e_i e_i^T = \frac{1}{d} I$$



$\text{uniform}(\text{sphere}) \approx \text{uniform}(\text{orientation of standard basis})$
 \times perturbation

Unbiased Gradient Estimator (2/3)

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

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

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

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

$$\begin{aligned} \mathbb{E}[\hat{r}_t] &= \mathbb{E}\left[\frac{d(\langle \tilde{w}_t, r_t \rangle + \epsilon_t)}{\delta} s_t\right] = \mathbb{E}\left[\underbrace{\frac{d\langle w_t + \delta s_t, r_t \rangle}{\delta}}_{\textcircled{1} + \textcircled{2}} s_t\right] \\ &= \mathbb{E}\left[d s_t s_t^T r_t\right] = r_t \quad \text{where } \mathbb{E}[s_t s_t^T] = \frac{1}{d} \mathbf{I} \\ &\quad \textcircled{1} = \mathbb{E}\left[\frac{d\langle w_t, r_t \rangle}{\delta} s_t\right] = 0 \\ &\quad \textcircled{2} = \mathbb{E}\left[\frac{d\langle \delta s_t, r_t \rangle}{\delta} s_t\right] \end{aligned}$$

Unbiased Gradient Estimator (3/3)

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

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

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

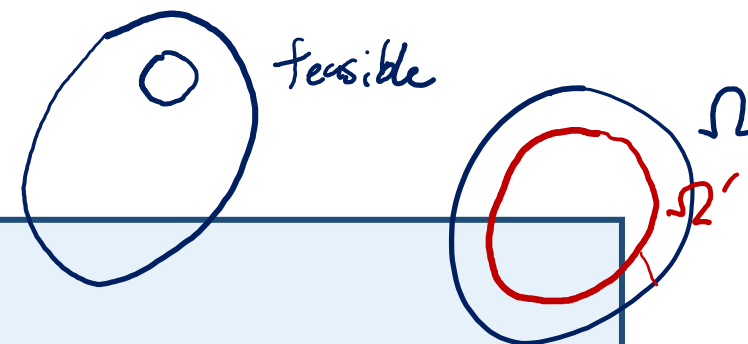
Define $\hat{r}_t = y_t H_t^{-1} s_t$ where $H_t := \mathbb{E}_{s \sim \mathcal{D}}[s s^\top]$

$$\begin{aligned} \mathbb{E}[\hat{r}_t] &= \mathbb{E} \left[\left(\langle \cancel{w_t} + \cancel{s_t}, r_t \rangle + \cancel{\epsilon_t} \right) H_t^{-1} s_t \right] = \mathbb{E} \left[\underbrace{s_t^\top r_t}_{\text{scalar}} H_t^{-1} s_t \right] = \mathbb{E} \left[H_t^{-1} \underbrace{s_t s_t^\top}_{\substack{\mathbb{E} \\ H_t}} r_t \right] \\ &= r_t \end{aligned}$$

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 $w_1 \in \Omega'$

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

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

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

Define

$$\hat{r}_t = \frac{dy_t}{\delta} s_t$$

Update policy:

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

Regret Bound for Bandit Linear Optimization

$$y_t = (\tilde{w}_t, r_t) + \xi_t$$

$$|y_t| \leq DG$$

Theorem. Suppose $\max_{w \in \Omega} \|w\| \leq D$, $\max_t \|r_t\| \leq G$. Then projected GD for BLO ensures

$$\hat{G} = \max_t \|\hat{r}_t\| \leq \frac{dy_t}{\delta} \|\xi_t\| \leq \frac{dDG}{\delta}$$

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

w'_* is the regret benchmark in Ω'

$$\mathbb{E} \left[\sum_{t=1}^T \langle w'_* - w_t, \hat{r}_t \rangle \right] \leq \frac{D^2}{\eta} + 2T \hat{G}^2$$

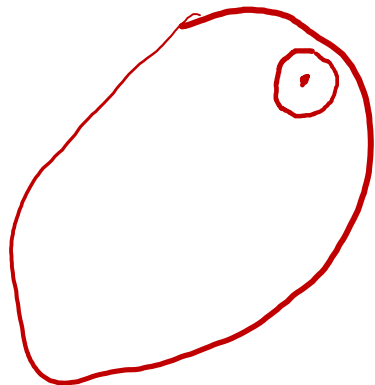
$$= \frac{D^2}{\eta} + \eta T \cdot \frac{d^2}{\delta^2} D^2 G^2$$

$$\Rightarrow \mathbb{E} \left[\sum_t \langle w'_* - w_t, r_t \rangle \right] \leq \frac{D^2}{\eta} + 2T \frac{d^2}{\delta^2} D^2 G^2$$

For any w^* in Ω , we can find a $w'_* \in \Omega'$ such that $\sum_t \langle w^* - w'_*, r_t \rangle \leq \sum_t \delta G$

①

PGD



Optimal



$$\|x - w_t\|_A \leq 1$$



②

$$w_{t+1} = \operatorname{argmin} \left\{ \langle w_t, \hat{r}_t \rangle - \underbrace{\frac{1}{\eta} D(w, w_t)}_{\approx \|w - w_t\|_A^2} \right\}$$

Abernethy, Hazan, and Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. 2008.

Bandit Optimization / Zeroth-Order Optimization

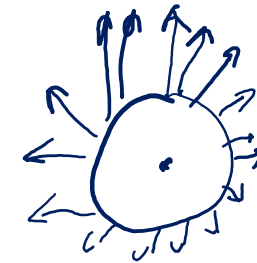
For time $t = 1, 2, \dots, T$:

Learner chooses a point w_t

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

$$(\nabla R_t(w_t))$$

$$\left. \begin{array}{l} \tilde{w}_t = w_t + s_t \\ \text{get } y_t \end{array} \right\} \rightarrow \text{estimated gradient } \underline{\hat{\nabla} R_t(w_t)}$$



Doubly Robust Estimator

Unbiased Estimator vs. Regression Estimator

$$\hat{r}_t = y_t H_t^{-1} s_t \text{ where } H_t := \mathbb{E}[s_t s_t^\top]$$

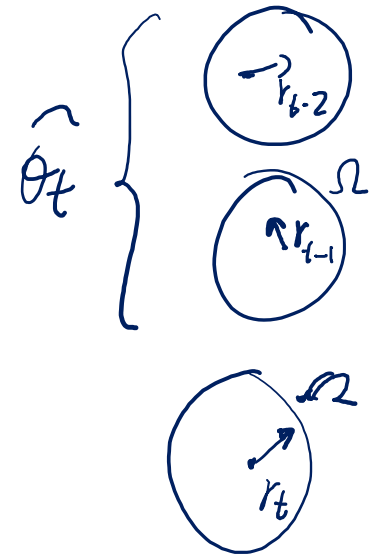
$$\hat{r}_t(a) = \frac{r_t(a) \mathbb{I}\{a_t = a\}}{p_t(a)}$$

$$\hat{\theta}_t = \operatorname{argmin}_{\theta} \sum_{i=1}^{t-1} (w_i^\top \theta - r_i)^2 + \|\theta\|^2$$

$$\hat{\theta}_t(a) = \frac{\sum_{i=1}^{t-1} r_i(a) \mathbb{I}\{a_i = a\}}{N_t(a)}$$

$$f_t(w) = w^\top r_t$$

Unbiased High variance



$$f(w) = w^\top \theta^*$$

↑
 $\phi(x, a)$

Biased Low variance

$$\mathbb{E}[\hat{\theta}_t] \neq r_t$$

An estimator that maintains the unbiasedness but with reduced variance?

Doubly Robust Estimator

$$\hat{r}_t = (y_t - \langle \tilde{w}_t, \hat{\theta}_t \rangle) H_t^{-1} s_t + \hat{\theta}_t$$

$$\hat{r}_t(a) = \frac{(r_t(a) - \hat{\theta}_t(a)) \mathbb{I}\{a_t = a\}}{p_t(a)} + \hat{\theta}_t(a)$$

$$\begin{aligned} &\rightarrow (\langle \tilde{w}_t, r_t \rangle + \varepsilon_t - \langle \tilde{w}, \hat{\theta}_t \rangle) H_t^{-1} s_t + \hat{\theta}_t \\ &= (\langle \tilde{w}_t, r_t - \hat{\theta}_t \rangle + \varepsilon_t) H_t^{-1} s_t + \hat{\theta}_t \end{aligned}$$

$$\begin{aligned} &\mathbb{E} \\ &\Rightarrow r_t \end{aligned}$$

$$\frac{\|\hat{r}_t\|^2}{\|r_t\|^2}$$

Summary for Bandits

- Value-based approach
 - Basic idea: **Regression**
 - Exploration strategies
 - Randomization based on $\hat{R}_t(x_t, a)$ (BE, IGW)
 - Adding uniform exploration (EG)
 - (Randomized) exploration bonus (UCB, TS)
- Policy-based approach
 - Basic idea: **Gradient updates subject to distance regularization**
 - Exploration strategies:
 - Intrinsic randomization (Exp3, IGW)
 - Adding extra uniform distribution (Exp3-1)
 - High baseline (Exp3-2)
 - Perturbed policy (PGD)
 - Exploration bonus is also used in policy-based approach (my talk at [AI/ML seminar](#))