Adversarial Bandit Linear Optimization

Chen-Yu Wei

Review: Online Linear Optimization

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

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

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

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

Bandit Linear Optimization

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

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

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

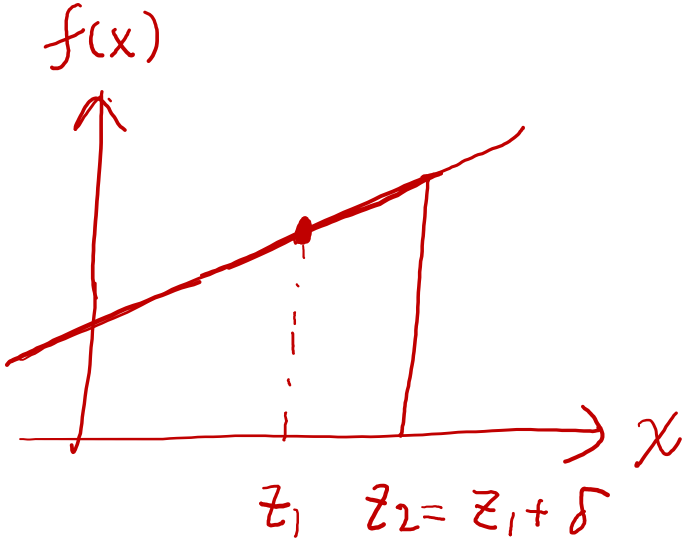
Learner chooses a point $w_t \in \Omega$

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 $f(z)$

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

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

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

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

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

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

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

Unbiased Gradient Estimator (1/3)

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

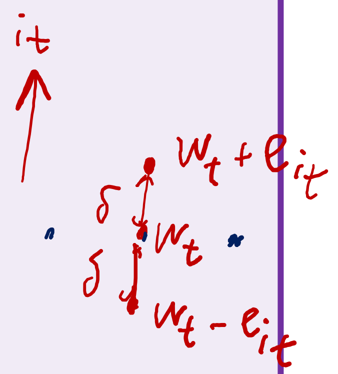
Uniformly randomly choose a direction $i_t \in \{1, 2, \dots, d\}$

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

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

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

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



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

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

$$\mathbb{E}[\hat{r}_t] = \mathbb{E}\Big[\underbrace{\frac{d}{\delta}\left(\langle w_t, r_t \rangle\right) \alpha_t e_{i_t}}_{0} + \frac{d}{\cancel{\delta}} \langle \cancel{\delta} \cancel{\alpha_t} e_{i_t}, r_t \rangle \cancel{\alpha_t} e_{i_t}\Big]$$

$$\mathbb{E}\left[d \langle e_{i_t}, r_t \rangle e_{i_t}\right]$$

$$= d \sum_{i=1}^{d} \frac{1}{d} \langle e_i, r_t \rangle e_i$$

$$= \sum_{i=1}^{d} \langle e_i, r_t \rangle e_i = r_t$$

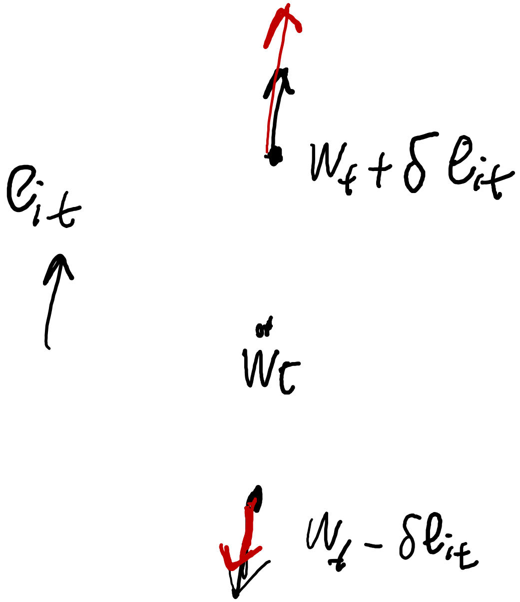
Unbiased Gradient Estimator (1/3)

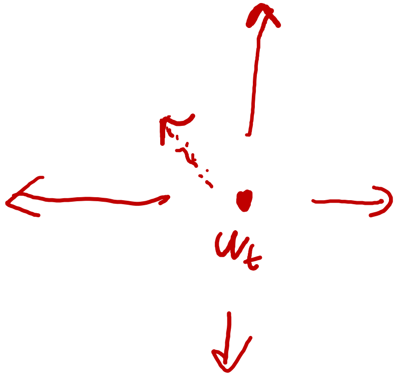
$$\hat{r}_t = \frac{d\, y_t}{\delta}\, \boxed{\alpha_t e_{i_t}}$$

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

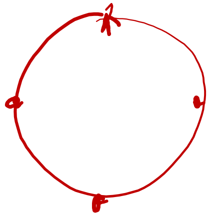
$$y_t = \langle \tilde{w}_t, r_t \rangle + \varepsilon_t$$

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





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



uniform(sphere) = uniform(orientation of standard basis)

× perturbation

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

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

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

$$\mathbb{E}[\hat{r}_t] = \mathbb{E}\left[\frac{d(\langle \tilde{w}_t, r_t\rangle + \cancel{\epsilon_t}^{\,0})}{\delta} s_t\right] = \mathbb{E}\left[\underbrace{\frac{d\langle w_t + \delta s_t, r_t\rangle}{\delta} s_t}\right]$$

$$= \mathbb{E}\left[d\, s_t s_t^T r_t\right] = r_t$$

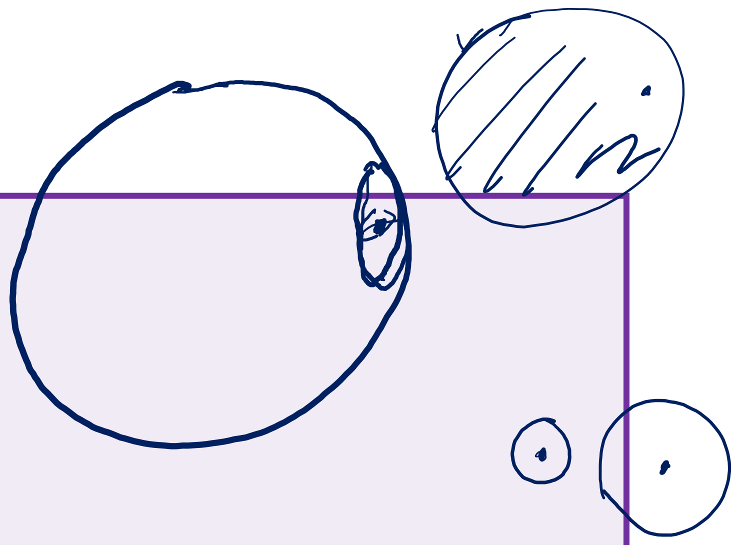
$$\mathbb{E}[s_t s_t^T] = \frac{1}{d} I$$

① + ②

$$① = \mathbb{E}\left[\frac{d\langle w_t, r_t\rangle}{\delta} s_t\right] = 0$$

$$② = \mathbb{E}\left[\frac{d\langle \cancel{\delta} s_t, r_t\rangle}{\cancel{\delta}} s_t\right]$$

Unbiased Gradient Estimator (3/3)



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 $H_t \coloneqq \mathbb{E}_{s\sim\mathcal{D}}[s s^\top]$

$$\mathbb{E}[\hat{r}_t] = \mathbb{E}\left[\left(\langle \cancelto{0}{w_t} + s_t, r_t \rangle + \cancelto{0}{\epsilon_t}\right) H_t^{-1} s_t\right] = \mathbb{E}\left[s_t^T r_t H_t^{-1} s_t \right] = \mathbb{E}\left[H_t^{-1} \underbrace{s_t s_t^T}_{\mathbb{E} = H_t} r_t \right]$$

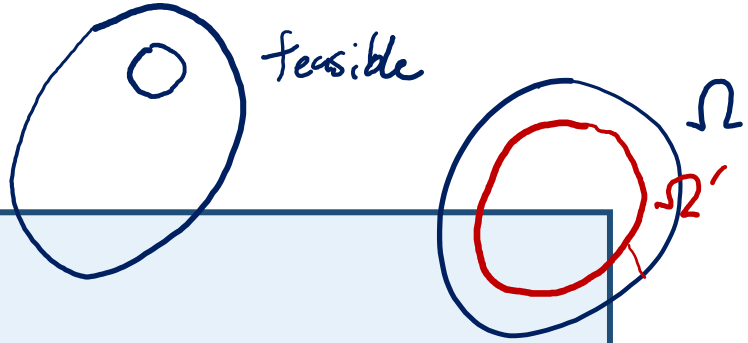
$$= r_t$$

Projected Gradient Descent for Bandit Linear Optimization

Assume the feasible set Ω contains a ball of radius δ

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

ball of radius δ
centered around w



Arbitrarily pick $\cancel{\tilde{w}_1}$ w_1 $\in \Omega'$

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

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

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

Update policy:

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

Regret Bound for Bandit Linear Optimization

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

$|y_t| \le DG$

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

$\hat{G} = \max_t \|\hat{r}_t\| \le \frac{d\, y_t}{\delta} \|s_t\| \le \frac{dDG}{\delta}$

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

w_*' is the regret benchmark in Ω'

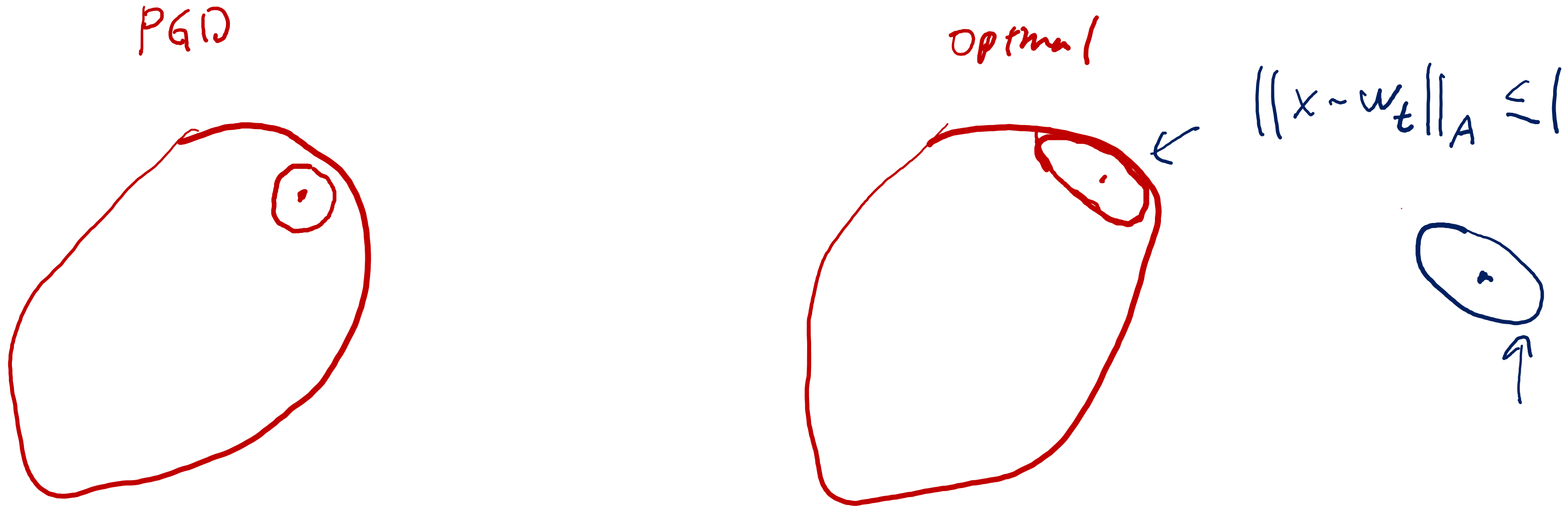
$$\mathbb{E}\left[\sum_{t=1}^{T} \langle w_*' - w_t, \hat{r}_t \rangle\right] \le \frac{D^2}{\eta} + \eta T \hat{G}^2$$

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

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

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

1



2

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

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

Bandit Optimization / Zeroth-Order Optimization

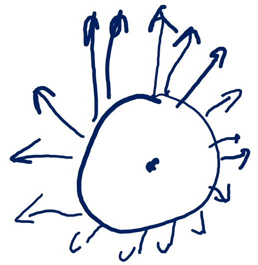
For time $t = 1, 2, \dots, T$:

Learner chooses a point w_t

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

$\left(\nabla R_t(w_t)\right)$

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

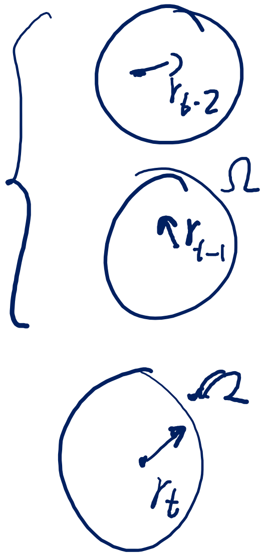


Doubly Robust Estimator

Unbiased Estimator vs. Regression Estimator

$$\boxed{f_t(w) = w^T r_t}$$

$\hat{\theta}_t$



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

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

Unbiased High variance

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

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

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

$\uparrow$

$\phi(x, a)$

Biased Low variance

$\downarrow$

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

An estimator that maintains the unbiasedness but with reduced variance?

Doubly Robust Estimator

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

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

$\rightarrow \left(\langle \tilde{w}_t, r_t \rangle + \varepsilon_t - \langle \tilde{w}, \hat{\theta}_t \rangle\right) H_t^{-1} s_t + \hat{\theta}_t$

$= \left(\langle \tilde{w}_t, r_t - \hat{\theta}_t \rangle + \varepsilon_t\right) H_t^{-1} s_t + \hat{\theta}_t$

$\mathbb{E}$

$\Rightarrow r_t$

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

Dudik, Langford, and Li. Doubly Robust Policy Evaluation and Learning. 2011.

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)