

# **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  $r_t \in \mathbb{R}^d$

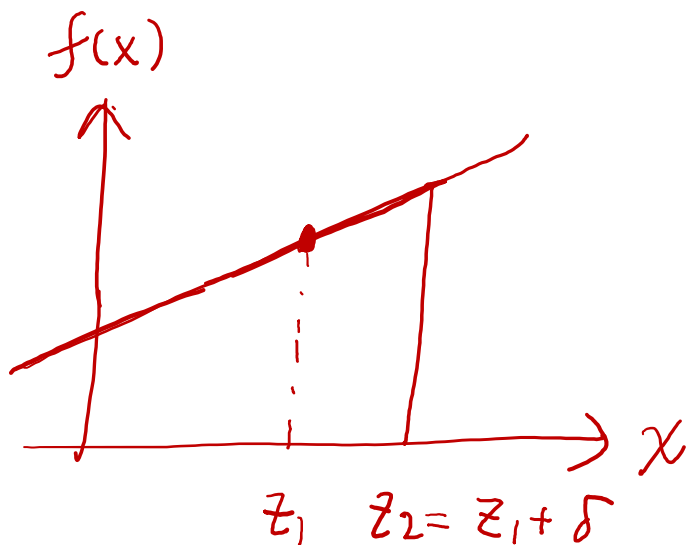
Learner chooses a point  $w_t \in \Omega$

Environment reveals  $\langle w_t, r_t \rangle + \epsilon_t$ , where  $\epsilon_t$  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  $\hat{r}_t = \frac{dy_t}{d\delta} \alpha_t e_{i_t}$



**Property 1:**  $\mathbb{E}[\hat{r}_t] = r_t$   $\hat{r}_t = \frac{d}{d\delta} \left( \langle \tilde{w}_t, r_t \rangle + \epsilon_t \right) \alpha_t e_{i_t}$

**Property 2:**  $\mathbb{E}[\tilde{w}_t] = w_t$   $\tilde{w}_t = \frac{d}{d\delta} \left( \langle w_t + \delta \alpha_t e_{i_t}, r_t \rangle + \epsilon_t \right) \alpha_t e_{i_t}$

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

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

# Unbiased Gradient Estimator (1/3)

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

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

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



# Unbiased Gradient Estimator (3/3)

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

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

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

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

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

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

# Projected Gradient Descent for Bandit Linear Optimization

Assume the feasible set  $\Omega$  contains a ball of radius  $\delta$

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

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

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

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

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

# 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  $\epsilon_t$  is a zero-mean noise