Full-Information Online Learning with Adversarial Reward

Chen-Yu Wei

The Expert Problem

Given: set of experts $\mathcal{A} = \{1, \dots, A\}$

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

Learner chooses a distribution over experts $p_t \in \Delta_{\mathcal{A}}$

Environment reveals the reward vector $r_t = (r_t(1), \dots, r_t(A))$

Key difference from before: $r_1(a), \dots, r_T(a)$ do not have the same mean

$$\text{Regret} = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} \langle p_t, r_t \rangle$$

Strategies?

- Follow the leader

$$a_t = \max_{a \in \mathcal{A}} \left\{ \sum_{i=1}^{t-1} r_i(a) \right\}$$

Incremental Updates

Exponential weight updates:

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))}$$

Projected gradient ascent:

$$p_{t+1} = \Pi_{\Delta_{\mathcal{A}}}(p_t + \eta r_t)$$

Equivalent Forms of EWU

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))}$$

$$p_{t+1}(a) = \frac{\exp\left(\eta \sum_{i=1}^{t} r_i(a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\eta \sum_{i=1}^{t} r_i(a')\right)}$$

$$p_{t+1} = \underset{p \in \Delta_{\mathcal{A}}}{\operatorname{argmax}} \left\{ \langle p, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(p, p_t) \right\}$$

$$\mathrm{KL}(p, q) \coloneqq \sum_{a=1}^{A} p(a) \ln \frac{p(a)}{q(a)}$$ **(KL divergence)**

$$p_{t+1} = \underset{p \in \Delta_{\mathcal{A}}}{\operatorname{argmax}} \left\{ \left\langle p, \sum_{i=1}^{t} r_i \right\rangle + \frac{1}{\eta} H(p) \right\}$$

$$H(p) \coloneqq \sum_{a=1}^{A} p(a) \ln \frac{1}{p(a)}$$ **(Shannon entropy)**

Regret Bound for Exponential Weight Updates

Theorem.

Assume that $\eta r_t(a) \leq 1$ for all t, a. Then EWU

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))}$$

ensures

$$\text{Regret} = \max_{a^\star} \sum_{t=1}^{T} \left(r_t(a^\star) - \langle p_t, r_t \rangle \right) \leq \frac{\ln A}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{A} p_t(a) r_t(a)^2$$

Regret Bound Analysis

Online Mirror Descent

(Re-interpreting exponential weight updates)

Exponential Weight Updates

Exponential Weight Updates = KL divergence regularized policy updates

$$p_{t+1}(a) = \frac{p_t(a)\exp(\eta r_t(a))}{\sum_{a' \in \mathcal{A}} p_t(a')\exp(\eta r_t(a'))} = p_{t+1} = \underset{p \in \Delta_{\mathcal{A}}}{\operatorname{argmax}} \left\{ \langle p, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(p, p_t) \right\}$$

KL divergence regularized policy updates is the basis of many RL algorithms (e.g., PPO, SAC).

Projected Gradient Descent

Projected Gradient Descent = Euclidean norm regularized policy updates

$$p_{t+1} = \Pi_{\Delta_{\mathcal{A}}}(p_t + \eta r_t)$$

$$= \quad p_{t+1} = \underset{p \in \Delta_{\mathcal{A}}}{\operatorname{argmax}} \left\{ \langle p, r_t \rangle - \frac{1}{2\eta} \|p - p_t\|_2^2 \right\}$$

Why Regularized Updates?

Projected Gradient Descent

$$p_{t+1} = \Pi_{\Delta_{\mathcal{A}}}(p_t + \eta r_t)$$

$$p_{t+1} = \max_{p \in \Delta_{\mathcal{A}}} \left\{ \langle p, r_t \rangle - \frac{1}{2\eta} \|p - p_t\|_2^2 \right\}$$

Exponential Weight Updates

$$p_{t+1}(a) \propto p_t(a) \exp(\eta r_t(a))$$

$$p_{t+1} = \max_{p \in \Delta_{\mathcal{A}}} \left\{ \langle p, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(p, p_t) \right\}$$

- Adversarial reward
- Stochastic reward
- For non-linear functions, gradient only represent the function locally

Why Distance Measures Other than $\|\cdot\|_2$?

General Framework: Mirror Descent

Projected Gradient Descent

$$p_{t+1} = \Pi_{\Delta_{\mathcal{A}}}(p_t + \eta r_t)$$

$$p_{t+1} = \max_{p \in \Delta_{\mathcal{A}}} \left\{ \langle p, r_t \rangle - \frac{1}{2\eta} \|p - p_t\|_2^2 \right\}$$

Exponential Weight Updates

$$p_{t+1}(a) \propto p_t(a) \exp(\eta r_t(a))$$

$$p_{t+1} = \max_{p \in \Delta_{\mathcal{A}}} \left\{ \langle p, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(p, p_t) \right\}$$

$$\psi(p) = \frac{1}{2} \|p\|_2^2$$

$$\psi(p) = \sum_{a=1}^{A} p(a) \ln p(a)$$

(Online) Mirror Descent

$$p_{t+1} = \max_{p \in \Omega} \left\{ \langle p, r_t \rangle - \frac{1}{\eta} D_\psi(p, p_t) \right\}$$

$$D_\psi(p, q) := \psi(p) - \psi(q) - \langle \nabla \psi(q), p - q \rangle$$

(Bregman divergence w.r.t. ψ)

Bregman Divergence

- Use a strictly convex function to define the distance on a space

Bregman Divergence

- Approximate the second-order derivative of ψ

- Provide local distance measure

Online Linear Optimization and Online Mirror Descent

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

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

Online Mirror Descent

Arbitrary $w_1 \in \Omega$

$$w_{t+1} = \max_{w \in \Omega} \left\{ \langle w, r_t \rangle - \frac{1}{\eta} D_\psi(w, w_t) \right\}$$

Regret Bound of Online Mirror Descent

Theorem. Online Mirror Descent ensures

$$\sum_{t=1}^{T}\langle u, r_t\rangle - \sum_{t=1}^{T}\langle w_t, r_t\rangle \leq \frac{D_\psi(u, w_1)}{\eta} + \sum_{t=1}^{T}\left(\langle w_{t+1} - w_t, r_t\rangle - \frac{1}{\eta} D_\psi(w_{t+1}, w_t)\right)$$

Recover the Bound of Exponential Weights

Mirror Descent under Matrix Norm

Corollary. Online Mirror Descent with $\psi(x) = \frac{1}{2}\|x\|_M^2$ ensures

$$\sum_{t=1}^{T}\langle u, r_t\rangle - \sum_{t=1}^{T}\langle w_t, r_t\rangle \leq \frac{\|u - w_1\|_M^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|r_t\|_{M^{-1}}^2$$

Linear Optimization → Convex 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 **convex** function $f_t \colon \mathbb{R}^d \to \mathbb{R}$

Algorithm

Run OMD with $r_t = -\nabla f_t(w_t)$

$$\text{Regret} = \sum_{t=1}^{T} \big(f_t(w_t) - f_t(w^\star)\big) \leq \sum_{t=1}^{T} \nabla f_t(w_t)^\top (w_t - w^\star) = \sum_{t=1}^{T} (w^\star - w_t)^\top r_t \leq \cdots$$

Recap

- Mirror Descent
  - Gradient update + distance regularization
  - There is flexibility to choose the distance measure: use a strictly convex function to define distances – **Bregman divergence**
  - A good choice of the potential would depend on
    1) the range of the feasible region, 2) the range of gradients
  - Can recover exponential weights and project gradient descent
- Mirror Descent is used in
  - RL algorithms such as NPG, PPO, SAC (covered later)
  - (online, stochastic) convex optimization

Lemmas about Bregman Divergence

Lemma 1. (Unaffected by adding a linear function)
If $G(w) = F(w) + w^\top c_1 + c_0$, then $D_G = D_F$.

Lemma 2. (Linear scaling)
If $G(w) = cF(w)$, then $D_G = cD_F$.

Lemmas about Bregman Divergence

Lemma 3.

Let F be a strictly convex function over a convex feasible set Ω.

If $w^\star \in \operatorname*{argmin}_{w \in \Omega} F(w)$, then for any $w \in \Omega$, $F(w) \geq F(w^\star) + D_F(w, w^\star)$.

Online Mirror Descent Regret Analysis