Bandits

Vincent Evers

Seminar Online Learning

B.Sc. Statistics & Data Science

December 16, 2025

Outline

Multi-armed Bandit Problem

Round	**Arm 1**	**Arm 2**	**Arm 3**
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
⋮	⋮	⋮	⋮

Multi-armed Bandit Problem

Round	**Arm 1**	**Arm 2**	**Arm 3**
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
⋮	⋮	⋮	⋮

Multi-armed Bandit Problem

Round	Arm 1	Arm 2	Arm 3
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
⋮	⋮	⋮	⋮

Multi-armed Bandit Problem

Round	Arm 1	Arm 2	Arm 3
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
⋮	⋮	⋮	⋮

Multi-armed Bandit Problem

Round	Arm 1	Arm 2	Arm 3
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
⋮	⋮	⋮	⋮

Multi-armed Bandit Problem

Round	Arm 1	Arm 2	Arm 3
1	0.5	0.3	0.2
2	0.9	0.3	0.5
3	0.8	0.1	0.6
4	0.4	0.7	0.3
$\vdots$	$\vdots$	$\vdots$	$\vdots$

How can we choose an arm to minimize our cumulative cost?

Multi-armed Bandit Problem

d: number of arms
T: number of rounds played
t: current round
p_t: arm chosen
$\mathbf{w}_t \in S$: weights

$S = \{\mathbf{w} : ||\mathbf{w}||_1 = 1 \wedge \mathbf{w} \geq 0\}$
$\mathbb{P}[p_t = i] = w_t[i]$
$\mathbf{y}_t \in [0,1]^d$: cost vector
$f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{y}_t \rangle$: loss function
$\nabla f_t(\mathbf{w}) = \mathbf{y}_t$

Regret

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T} y_t[p_t]\right]}_{\substack{\text{cumulative cost} \\ \text{of the algorithm}}} - \underbrace{\min_i \sum_{t=1}^{T} y_t[i]}_{\substack{\text{cumulative cost of} \\ \text{the best arm in hindsight}}}$$

The expected value is taken with respect to the randomness in choosing an arm.

Estimating the Gradient

Since only one arm is picked, just $y_t[p_t]$ is known.

Unbiased gradient estimate

$\mathbf{z}_t$ is an unbiased estimate of $\mathbf{y}_t$ and defined as

$$z_t[j] = \begin{cases} \dfrac{y_t[j]}{w_t[j]} & \text{if } j = p_t, \\ 0 & \text{else.} \end{cases}$$

$$\mathbb{E}[z_t[j] | \mathbf{z}_{t-1}, \ldots, \mathbf{z}_1] = \sum_{i=1}^{d} \mathbb{P}[p_t = i] z_t^{(i)}[j] = w_t[j] \frac{y_t[j]}{w_t[j]} = y_t[j]$$

Sub-gradient definition

- Since $\mathbf{y}_t$ is the gradient of f_t at $\mathbf{w}_t$, it is also the sub-gradient at that point.
- This can be written as $\mathbb{E}[\mathbf{z}_t] = \mathbf{y}_t \in \partial f_t(\mathbf{w}_t)$.

Sub-gradient

Let f be a convex function.
$\mathbf{y}$ is a sub-gradient of f at w if it satisfies the inequality

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{y} \rangle$$

.

Stochastic Bandits

- We assume a fixed iid distribution of costs for every arm.
- This assumption can be difficult to verify.
- Goal: Learning the arm with the lowest cost.
- Applications:
  - A/B testing
  - Choosing ads for consumers

Adversarial Bandits

- Costs are chosen by an adversary, which may be oblivious or adaptive.
- Oblivious adversary: The adversary chooses all the costs before the first round.
- Adaptive adversary: The adversary chooses the cost before each round. The adversary can adapt to our prior decisions.
- Randomization in choosing an arm is essential when dealing with an adaptive adversary.
- Goal: Perform well even in worst-case scenarios.
- Applications:
  - Auction bidding
  - Spam detection / Cybersecurity

Learning with Expert Advice

- We randomly choose one out of d experts every round. Every expert receives a cost, but we get to see every experts cost.
- Since we have full information in every round, learning is faster.
- Costs could be generated by a fixed iid distribution or chosen by an adversary, just like with bandits.
- Applications:
  - Portfolio selection
  - weather forecasting

Outline

Online Mirror Descent with Estimated Gradient

Algorithm for OMD with estimated Gradient

parameter: a link function $g : \mathbb{R}^d \to S$
initialize: $\boldsymbol{\theta}_1 = 0$
for $t = 1, 2, \ldots$
predict $\mathbf{w}_t = g(\boldsymbol{\theta}_t)$
pick $\mathbf{z}_t$ at random such that $\mathbb{E}[\mathbf{z}_t | \mathbf{z}_{t-1}, \ldots, \mathbf{z}_1] \in \partial f_t(\mathbf{w}_t)$
update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$

g links / mirrors the primal space S, where $\mathbf{w}$ lives and the dual space $\mathbb{R}^d$, where θ lives.

OMD Bound

Theorem 1

If the subgradient is chosen such that with Probability 1 we have

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \leq B(\mathbf{u}) + \sum_{t=1}^{T} ||\mathbf{z}_t||_t^2$$

B is some function and $|| \cdot ||_t$ depends on $\mathbf{w}_t$. From this follows

$$\mathbb{E}\left[\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u}))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle\right] \leq B(\mathbf{u}) + \sum_{t=1}^{T} \mathbb{E}[||\mathbf{z}_t||_t^2]$$

The expected value is being taken with respect to the randomness of the learner in choosing the gradient $\mathbf{z}_1, ..., \mathbf{z}_T$.

Outline

Normalized Exponentiated Gradient Result

Consider the OMD with estimated Gradient Algorithm. If we choose $S = \{\mathbf{w} : ||\mathbf{w}||_1 = 1 \wedge \mathbf{w} \geq 0\}$ and g, such that the i-th element of g is

$$g_i(\boldsymbol{\theta}) = \frac{\exp(\eta\theta[i])}{\sum_j \exp(\eta\theta[j])},$$

we get the Normalized Exponentiated Gradient Algorithm.

For this Algorithm we have the following Result:

Theorem 2

For a sequence of linear loss functions such that for all t, i we have $\eta z_t[i] \geq -1$, then

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \leq \frac{\log(d)}{\eta} + \eta \sum_{t=1}^{T} \sum_i w_t[i] z_t[i]^2$$

Bandit Algorithm

Bandit Exponentiated Gradient Algorithm

parameters: $\eta \in (0, 1)$

initialize: $\mathbf{w}_1 = (\frac{1}{d}, ..., \frac{1}{d})$

for $t = 1, 2, ...$

choose $p_t \sim \mathbf{w}_t$

receive $y_t[p_t] \in [0, 1]$

update

$$\tilde{w}[p_t] = w_t[p_t] \exp\left(\frac{-\eta y_t[p_t]}{w_t[p_t]}\right)$$

for $i \neq p_t, \tilde{w}[i] = w_t[i]$

$$\forall i, w_{t+1} = \frac{\tilde{w}[i]}{\sum_j \tilde{w}[j]}$$

Bandit Regret Bound Proof

Proof of a Regret Bound for the Bandit Algorithm:

Combining Theorem 1 and Theorem 2 and taking the expected value, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{u}))\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t\rangle\right] \\ &\leq \frac{\log(d)}{\eta} + \eta\sum_{t=1}^{T}\mathbb{E}\left[\sum_i w_t[i]z_t[i]^2\right].\end{aligned}$$

Bandit Regret Bound Proof

$$\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{u}))\right] \leq \frac{\log(d)}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_i w_t[i] z_t[i]^2\right]$$

We can bound the expected value in the last term.

$$\begin{aligned}
\mathbb{E}\left[\sum_i w_t[i] z_t^{(p_t)}[i]^2 | \mathbf{z}_{t-1}, .., \mathbf{z}_1\right] &= \sum_j \mathbb{P}[p_t = j] \sum_i w_t[i] z_t^{(j)}[i]^2 \\
&= \sum_j w_t[j] w_t[j] \frac{y_t[j]^2}{w_t[j]^2} \\
&= \sum_j \underbrace{y_t[j]^2}_{\leq 1, \text{ since } \mathbf{y} \in [0,1]^d} \\
&\leq d
\end{aligned}$$

Bandit Regret Bound Proof

$$\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\mathbf{w}_t) - f_t(\mathbf{u}))\right] \leq \frac{\log(d)}{\eta} + \eta T d$$

Let u be the decision to always pull the best arm in hindsight.

$$\begin{aligned}\sum_{t=1}^{T}\underbrace{\mathbb{E}[f_t(\mathbf{w}_t)]}_{=\mathbb{E}[y_t[p_t]]} - \sum_{t=1}^{T}\underbrace{f_t(\mathbf{u})}_{=\min_i y_t[i]} &= \sum_{t=1}^{T}\mathbb{E}[y_t[p_t]] - \min_i \sum_{t=1}^{T} y_t[i] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} y_t[p_t]\right] - \min_i \sum_{t=1}^{T} y_t[i]}_{=\text{Regret}}\end{aligned}$$

Bandit Regret Bound

Regret Bound

$$\mathbb{E}\left[\sum_{t=1}^{T} y_t[p_t]\right] - \min_i \sum_{t=1}^{T} y_t[i] \leq \frac{\log(d)}{\eta} + \eta T d$$

- Can be used to find optimal $\eta^* = \sqrt{\frac{\log(d)}{dT}}$.
- Plugging in η^* gives us sublinear regret, with $\text{Regret}_T = O(\sqrt{d \log(d) T})$.

Outline

Code Simulation

- We generate data, with $T = 500$, $d = 3$ and a fixed distribution for the costs of every arm. The costs are bounded, so $\mathbf{y}_t \in [0,1]^d$.

$$\text{Arm } 1 \sim N(0.3, 0.04),\ \text{Arm } 2 \sim N(0.5, 0.04),\ \text{Arm } 3 \sim N(0.7, 0.04)$$

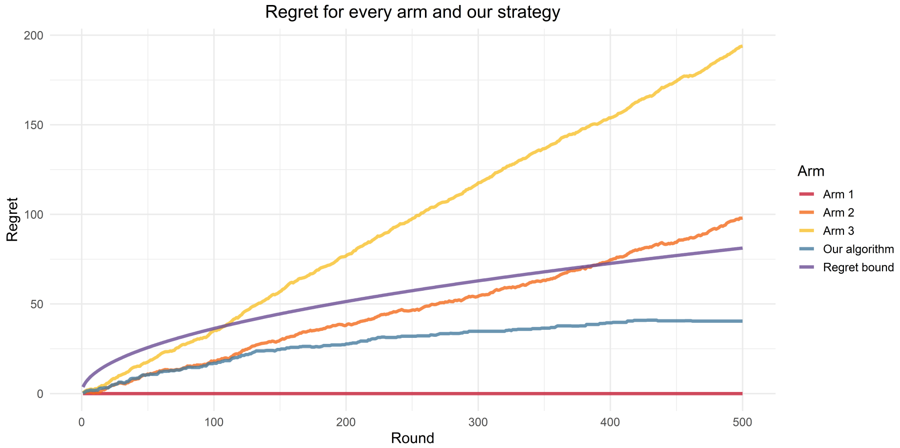
- But we make no assumptions about how the costs were generated.
- To minimize regret we use the bound and choose an optimal η.

$$\arg\min_{\eta} \frac{\log(d)}{\eta} + \eta T d \approx 0.0271$$

- We will look at how w_t changes and how our observed regret compares to the regret if we only pulled one arm.

Evolution of $\mathbf{w}_t$

Regret Comparison



Outline

Summary

- Goal: Optimizing sequential decision making under uncertainty.
- We can make assumptions about how the costs are generated (Stochastic, Adversarial Bandits).
- Regret is the central performance measure, where we compare the cumulative cost of our algorithm to the best possible fixed decision in hindsight.
- Through randomization in picking an arm we can achieve sublinear regret even if the costs are chosen by an adversary.

$$\lim_{T \to \infty} \frac{\text{Regret}_T}{T} = 0$$

Outline

References

[1] Shai Shalev-Shwartz, *Online Learning and Online Convex Optimization*, Foundations and Trends in Machine Learning, vol. 4, no. 2, pp. 107–194, 2011.

[2] Tor Lattimore and Csaba Szepesvári, *Bandit Algorithms*, July 2020

[3] Francesco Orabona, *A Modern Introduction to Online Learning*, May, 2025