Online Decision-Making

0. References

This course is mainly based on the following references:

1. Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122;
2. Elad Hazan. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325;
3. Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

On some slides, specific references will be cited. Otherwise the full list of references used to prepare this presentation in addition to some reading suggestions are provided in Appendix.

1. Online decision-making

- facing a problem (environment, system, etc.): implement the best decision (control, action) to meet some objective ($\min$ or $\max$);
- **uncertainty:** unknown environment, subject to exogenous factors, limited models;
- **online**: the only information we have access to comes from the past, even the current problem is not well characterized (predictive aspect);
- **sequential** nature of the problem: consecutive decisions;
- today: **online (machine) learning** approaches.

Online decision process

In each round:

1. implement decision;
2. suffer losses & get new information;
3. compute next round decision.

Example: uncertain resource allocation in real-time.

- manage resources while learning their attributes.

Applications

- telecommunication: channel access, network resource allocation;
- recommender systems: preference learning;
- finance: rebalanced portfolio;
- sensing: target localization or tracking;
- power systems: demand response, real-time pricing, economic dispatch/optimal power flow, state estimation, etc.

Motivation for online decision-making – static setting

First, our motivation is to learn the optimal fixed decision when all information is revealed (hindsight). We call this context the static setting.

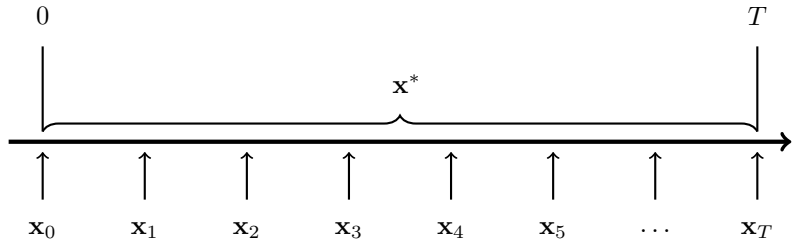


Figure 1: Sequence of decisions: static hindsight (top) and online (bottom)

Motivation for online decision-making – dynamic setting

Then, we will move to the dynamic setting where one wants to implement the round optimal decision at each round. This is of interest in many engineering contexts – but it is also a harder problem.

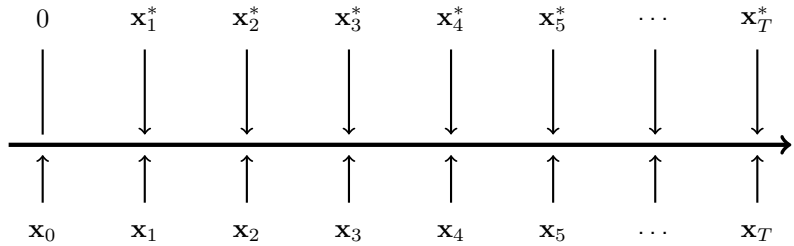


Figure 2: Sequence of decisions: dynamic hindsight (top) and online (bottom)

Overview

We will cover two important families of problems and their solution concepts:

1. Multi-armed bandit (MAB)
   - stochastic
   - adversarial
   - Markovian
2. Online convex optimization (OCO)

Their main advantages is that their simplicity allows for a thorough performance analysis and multiple extension tailored to the problem at hand.

Real-time online decision-making

We primarily focus on **real-time**, online decision-making. In other words, we want to design computationally efficient (time, CPU, memory) algorithms.

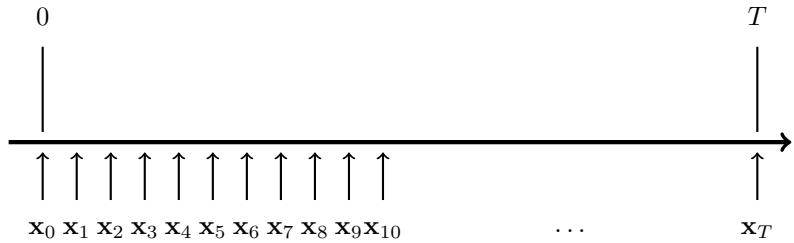


Figure 3: Fast-timescale decision making

If real-time is not our objective: we still get readily-implementable, hardware-compatible approaches.

2. Preliminaries

Notation:

- Consider a discretized time horizon $T \in \mathbb{N}$. We index rounds by t.
- We have access to $n \in \mathbb{N}$ resources.
- We can either pick $0 < m < n$ resources (binary decisions) or a combination of all resources (continuous decision);
- Let $\mathbf{x}_t \subseteq 2^n$ or $\mathbf{x}_t \in \mathbb{R}^n$ be our decision variable at time t.
- The problem can also be subject to context-specific constraints $\mathcal{X}$, e.g., m binary decision at the time: $\operatorname{card} \mathbf{x}_t = m$ or on the probability simplex: $\sum_{i=1}^{n} \mathbf{x}_t(i) = 1$.

Preliminaries – II

Regret:

- performance indicator, used to design our algorithm $\mathcal{A}$;
- definition (static):

$$\texttt{Regret}_{\mathcal{A}}(T) = \underbrace{\sum_{t=1}^{T} \text{Loss}(\mathbf{x}_{t,\mathcal{A}})}_{\text{loss we incurred}} - \underbrace{\min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} \text{Loss}(\mathbf{x})}_{\text{hindsight fixed minimum}}$$

- can be adapted to gain maximization instead of loss minimization.
- we wish to design $\mathcal{A}$ such that $\texttt{Regret}_{\mathcal{A}}(T) < O(T)$, i.e., regret is sublinear.
- sublinear regret: Hannan-consistent and $\texttt{Regret}_{\mathcal{A}}(T)/T \to 0$ as T grows meaning $\mathcal{A}$ performs as well as comparator, on average.

Preliminaries – III

- we will refer to this regret as the *static* regret;
- an interesting extension is the *dynamic* regret:

$$\texttt{Regret}_{\mathcal{A}}^{\text{dynamic}}(T) = \underbrace{\sum_{t=1}^{T} \text{Loss}(\mathbf{x}_{t,\mathcal{A}})}_{\text{loss we incurred}} - \underbrace{\sum_{t=1}^{T} \min_{\mathbf{x}_t \in \mathcal{X}} \text{Loss}(\mathbf{x}_t)}_{\text{round minima}}$$

- at this time, we will need to be more humble in our performance analysis;
- let V_T be the cumulative variation in optima, used to characterize the complexity of dynamic problems.

3. The multi-armed bandit problem

- from "American slang":

  one-armed bandit $=$ slot machine;

- each arm leads to a reward;
- **objective:** maximize the player's gain by determining the best sequence of arms (decisions) to play;
- *unknown* resources, *only* feedback is from the played arm;
- canonical example of exploration *vs.* exploitation problem.

Figure 4: Slot machines in Reno Airport, NV, USA.

Multi-armed bandits (MAB)

The three main family of bandits as characterized by the arm's reward process:

- stochastic bandits (S);
- adversarial bandits (A):
- Markovian bandits (M).

For each family, their exists an efficient, sublinear regret solution concepts. But that's only the tip of the iceberg, there are many more family of bandits, e.g., contextual [36] or infinite-armed [1] bandits.

Applications (MAB)

- channel access in cognitive radio network [13] (S);
- intelligent transport systems [48] (A);
- load curtailment [46, 47] (M) and [32] (S);
- curtailment with load fatigue [21] (S);
- curtailment of prosumers using [5] (S);
- learning load models [30] (A);
- vehicle-to-grid for load flattening [19] (S);
- setpoint tracking with flexible loads [37] (S);
- frequency regulation [45] (S);

3.1. Stochastic MAB

Setting:

- single arm can be played ($m = 1$), and $\mathbf{x}_t \in \{1, 2, \ldots, n\}$;
- let $X_{i,t}$ be arm's i reward at time t;
- the reward $X_{i,t}$ is distributed according to an unknown i.i.d. random variable;
- bounded reward: $0 \leq X_{i,t} \leq \overline{X}$, then normalized so that $X_{i,t} \in [0, 1]$.
- no prior information is known about $X_{i,t}$;
- if $\mathbf{x}_t = i$, i.e., arm i is played, then we observe $X_{i,t}$ and nothing more.

MAB process

In each round $t = 1, 2, \ldots, T$:

1. play arm $\mathbf{x}_t \in \{0, 1, 2, \ldots, n\}$;
2. obtain reward $X_{\mathbf{x}_t, t}$;
3. compute $\mathbf{x}_{t+1}$, the next arm to play given additional knowledge.

Our objective: design a policy $\mathbf{x}_{t+1}$ given only observations as we go.

Regret for MAB

Regret: we make slight modification → pseudo-regret.

In MAB, the regret translates to:

$$\texttt{Regret}_{\mathcal{A}}(T) = \max_{i=1,2,\ldots,n} \sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{\mathbf{x}_t,t}$$

which is a random variable because the reward and potentially the policy $\mathbf{x}_t$ are stochastic.

We rather opt for the expected regret defined as:

$$\mathbb{E}\left[\texttt{Regret}_{\mathcal{A}}(T)\right] = \mathbb{E}\left[\max_{i=1,2,\ldots,n} \sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{\mathbf{x}_t,t}\right].$$

Note. The $\mathbb{E}$ is taken w.r.t. the random reward and the random decision-making policy.

Pseudo-regret for MAB

This is still a very strict performance indicator because the expectation is taken over $\max$..

We swap the $\mathbb{E}$ and $\max$ and use a weaker definition of the regret, the pseudo-regret:

$$\begin{aligned}
\mathbb{E}\left[\texttt{Regret}_{\mathcal{A}}(T)\right] &\geq \max_{i=1,2,\dots,n} \mathbb{E}\left[\sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{\mathbf{x}_t,t}\right] \\
&= \max_{i=1,2,\dots,n} \sum_{t=1}^{T} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{\mathbf{x}_t,t}\right] \\
&= T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\mathbf{x}_t,t}\right] \\
&= \overline{\texttt{Regret}}_{\mathcal{A}}(T),
\end{aligned}$$

where $\mathbb{E}\left[X_{i,t}\right] = \mu_i$ because i.i.d. random variable and $\mu^* = \max_{i=1,2,\dots,n} \mu_i$.

Now, how do we compute $\mathbf{x}_t$?

Optimism in the face of uncertainty

- **Policy #1**: greedy, i.e., sample each arm once, then play the one with highest mean.
  - no guarantee of sublinear regret, might be "stuck" on bad arm.
- **Policy #2**: ε-greedy, i.e., same but explore at random with probability ε [2];
  - constant non-zero probability of exploration leads to linear regret
  - if $\varepsilon_t \propto \frac{1}{td^2}$ where $0 < d < \min_{\mathbf{x} \neq \mathbf{x}^*} \mu^* - \mu_{\mathbf{x}}$, sublinear regret but needs prior knowledge for d.
- **Policy #3**: upper confidence bound (`UCB1`) [2], i.e., be optimistic about the reward and play the arm with the highest supposed reward. That also means don't ignore arms that poorly performed at some point.
  - sublinear regret bound, with no further assumption.

Upper confidence bound-1 (UCB1) algorithm

Let c_i^t be the number of time arm i has been played after t rounds.

Initialization: play each arm once and let current sample mean $\hat{\mu}_i = X_{i,1:n}$ and $c_i^n = 1 \; \forall \; i$.

In each round $t = 1, 2, \ldots, T$:

1. play arm with largest index, $\mathbf{x}_t = \arg\max_i \lambda_i$;
2. obtain reward $X_{\mathbf{x}_t, t}$;
3. update current sample mean $\hat{\mu}_{\mathbf{x}_t}$ and counter c_i^t;
4. update indices: $\lambda_i \leftarrow \underbrace{\hat{\mu}_i}_{\text{sample mean after } t} + \underbrace{\sqrt{\frac{\ln t + 1}{c_i^t}}}_{\text{upper confidence} \propto 1/\text{explored}}.$

Optimism in the face of uncertainty – II

Our sample average of arm i's is out of our confidence interval with a vanishing probability:

$$\begin{aligned}
\Pr\left[|\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{\ln t}{c_i^t}}\right] &= \Pr\left[\left|\sum_{t=1}^{c_i^t} X_{i,t} - c_i^t \mu_i\right| \geq c_i^t \sqrt{\frac{\ln t}{c_i^t}}\right] \\
&\leq 2\mathrm{e}^{-\frac{2}{c_i^t}\left(c_i^t\sqrt{\frac{\ln t}{c_i^t}}\right)^2} \qquad \text{(Hoeffding)} \\
&= \frac{2}{t^2}.
\end{aligned}$$

So let's trust our sample mean $\hat{\mu}_i, i = 1, 2, \ldots, n$.

Optimism in the face of uncertainty – III

More specifically, the upper confidence bound of arm i is bounded by the expected reward with high probability:

$$\begin{aligned}\Pr\left[\hat{\mu}_i + \sqrt{\frac{\ln t}{c_i^t}} \leq \mu_i\right] &= \Pr\left[\sum_{t=1}^{c_i^t} X_{i,t} - c_i^t \mu_i \leq -c_i^t \sqrt{\frac{\ln t}{c_i^t}}\right] \\ &\leq \mathrm{e}^{-\frac{2}{c_i^t}\left(c_i^t \sqrt{\frac{\ln t}{c_i^t}}\right)^2} \qquad \text{(Hoeffding)} \\ &= \frac{1}{t^2}.\end{aligned}$$

The UCB is not misleading with high probability, let's be optimistic and follow the most promising resource so far.

Regret analysis

Let:

- $\Delta_{\min} = \min_{i \neq \mathbf{x}^*} \mu^* - \mu_i$
- $\Delta_{\max} = \max_i \mu^* - \mu_i$

Theorem 1. ***(UBC1 regret bound)*** *The pseudo-regret of UCB1 is bounded above by:*

$$\overline{\texttt{Regret}}(T) \leq n\Delta_{\max}\left(\frac{4\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3}\right).$$

The pseudo-regret is at most $O(\ln T)$ and is, therefore, sublinear.

Proof*

1. The regret can be re-expressed as:

$$\begin{aligned}\overline{\texttt{Regret}}(T) &= T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{\mathbf{x}_t,t}\right] \\ &= \sum_{i=1}^{n} \left(\mu^* - \mu_i\right) \mathbb{E}\left[c_i^T\right] \\ &= \sum_{i=1}^{n} \Delta_i \, \mathbb{E}\left[c_i^T\right]\end{aligned}$$

where $\Delta_i = \mu^* - \mu_i \; \forall \; i$ and we recall that c_i^T is the number of time arm i was played after T rounds.

Then, we need to show that for $i \neq \mathbf{x}^*$, $\mathbb{E}\left[c_i^T\right]$ grows sublinearly in T.

Proof* – II

2. Selecting arm $i \neq \mathbf{x}^*$ at t occurs when:

$$\hat{\mu}_{\mathbf{x}^*} + \sqrt{\frac{\ln t}{c^t_{\mathbf{x}^*}}} \leq \hat{\mu}_i + \sqrt{\frac{\ln t}{c^t_i}}.$$

This in turns occur if:

- sample mean of the optimal arm is below our lower confidence bound (underestimate):

$$\hat{\mu}_{\mathbf{x}^*} \leq \mu^* - \sqrt{\frac{\ln t}{c^t_{\mathbf{x}^*}}} \tag{1}$$

- sample mean of arm i is above our upper confidence bound (overestimate):

$$\hat{\mu}_i > \mu_i + \sqrt{\frac{\ln t}{c^t_i}} \tag{2}$$

Proof* – III

- If $i \neq \mathbf{x}^*$ at t and (1) & (2) are false, then:

$$\mu^* < \mu_i + 2\sqrt{\frac{\ln t}{c_i^t}}. \tag{3}$$

  That is, the expected values are closed to each other and under insufficient sampling seem indistinguishable given our current upper and lower confidence bounds.

In fact, we have:

$$\text{(1) and (2) are false} \implies \text{(3) is true,}$$

and the contrapositive

$$\text{(3) is false} \implies \text{(1) or (2) is true.}$$

Proof* – IV

Assuming (3) holds, we get:

$$\mu^* \leq \mu_i + 2\sqrt{\frac{\ln t}{c_i^t}} \iff c_i^t \leq \frac{4 \ln t}{(\mu^* - \mu_i)^2}$$

Hence, if $c_i^t > \left\lceil \frac{4 \ln t}{\Delta_i^2} \right\rceil$, then inequalities (1) or (2) must be true.

Proof* – V

3. Back to upper bounding $\mathbb{E}\left[c_i^T\right]$ for $i \neq \mathbf{x}^*$.

$$\begin{aligned}
\mathbb{E}\left[c_i^T\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \text{pick arm } i \text{ at } t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_{i=\arg\max \lambda_i \text{ at } t}\right] \\
&\leq \left\lceil \frac{4\ln T}{\Delta_i^2} \right\rceil + \mathbb{E}\left[\sum_{t=\left\lceil \frac{4\ln T}{\Delta_i^2} \right\rceil + 1}^{T} \mathbb{I}_{i=\arg\max \lambda_i \text{ at } t \cap (3) \text{ is false}}\right] \\
&\leq \frac{4\ln T}{\Delta_i^2} + 1 + \mathbb{E}\left[\sum_{t=\left\lceil \frac{4\ln T}{\Delta_i^2} \right\rceil + 1}^{T} \mathbb{I}_{(1) \text{ is true} \cup (2) \text{ is true}}\right]
\end{aligned}$$

Proof* – VI

$$\begin{aligned}
\mathbb{E}\left[c_i^T\right] &\leq \frac{4\ln T}{\Delta_i^2} + 1 + \sum_{t=\left\lceil \frac{4\ln T}{\Delta_i^2} \right\rceil + 1}^{T} \Pr\left[(1) \text{ is true} \cup (2) \text{ is true}\right] \\
&\leq \frac{4\ln T}{\Delta_i^2} + 1 + \sum_{t=\left\lceil \frac{4\ln T}{\Delta_i^2} \right\rceil + 1}^{T} \Pr\left[(1) \text{ is true}\right] + \Pr\left[(2) \text{ is true}\right] \quad \text{(union bound)} \\
&\leq \frac{4\ln T}{\Delta_i^2} + 1 + \sum_{t=1}^{+\infty} \frac{2}{t^2} \\
&= \frac{4\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3},
\end{aligned}$$

and we have our bound the the number of time a non-optimal arm is picked.

Proof* – VII

4. Altogether. We finally obtain

$$\begin{aligned}
\overline{\texttt{Regret}}(T) &= \sum_{i=1}^{n} \Delta_i \, \mathbb{E}\left[c_i^T\right] \\
&\leq \sum_{i=1}^{n} \Delta_i \left(\frac{4 \ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \right) \quad \text{(previous result)} \\
&\leq \sum_{i=1}^{n} \Delta_{\max} \left(\frac{4 \ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} \right) \\
&\leq n \Delta_{\max} \left(\frac{4 \ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} \right)
\end{aligned}$$

which completes the proof.

3.2. Adversarial MAB

Setting:

- single arm can be played ($m = 1$), and $\mathbf{x}_t \in \{1, 2, \ldots, n\}$;
- historically, in terms of losses instead of reward;
- let $\ell_{i,t}$ be arm's i loss at time t;
- the loss is set by an adversary (nature/environment/system) when the decision is taken;
- bounded reward: $0 \leq \ell_{i,t} \leq \overline{X}$, then normalized so that $\ell_{i,t} \in [0, 1]$.
- no prior information is known about $\ell_{i,t}$;
- if $\mathbf{x}_t = i$, i.e., arm i is played, then we observe $\ell_{i,t}$ and nothing more.

Adversarial MAB process

In each round $t = 1, 2, \ldots, T$:

1. play arm $\mathbf{x}_t \in \{0, 1, 2, \ldots, n\}$ / simultaneously adversary sets $\ell_{i,t} \forall\ i$;
2. suffer loss $\ell_{\mathbf{x}_t, t}$;
3. compute $\mathbf{x}_{t+1}$, the next arm to play given additional knowledge.

Regret: the pseudo-regret in the loss-referential is

$$\overline{\texttt{Regret}}(T) = \sum_{t=1}^{T} \mathbb{E}\left[\ell_{\mathbf{x}_t, t}\right] - \min_{i=1,2,\ldots,n} \sum_{t=1}^{T} \mathbb{E}\left[\ell_{i,t}\right]$$

Note. The $\mathbb{E}$ is taken w.r.t. the decision maker and adversary random policy. To be continued.

How do we compute $\mathbf{x}_t$ to play against an (non-oblivious) adversary?

Randomized decision policy

- **Policy #1**: deterministic.
  - linear regret, adversary can constructed a strategy against us.
  - need a randomized policy.
- **Policy #2**: exponential weights for exploration and exploitation (Exp3) [3], i.e., randomly select an arm i according to a probability mass function (pmf) computed by the exponential weighting approach.
  - sublinear regret bound;

Exponential weights for exploration and exploitation (Exp3) algorithm

Let $\mathbf{p} = \begin{pmatrix} p_1 & p_2 & \dots & p_n \end{pmatrix}^\top$ the vector of arm probabilities.

Initialization: $p_i = 1/n \ \forall \ i$.

In each round $t = 1, 2, \dots, T$:

1. play arm randomly $\mathbf{x}_t$ according the distribution $\mathbf{p}$
2. suffer loss $\ell_{\mathbf{x}_t, t}$;
3. update estimated cumulative loss of arm $\mathbf{x}_t$: $\hat{L}_{\mathbf{x}_t} \leftarrow \hat{L}_{\mathbf{x}_t} + \frac{\ell_{\mathbf{x}_t, t}}{p_{\mathbf{x}_t, t}}$
4. update probability distribution $\forall \ i$:

$$p_i \leftarrow \frac{\mathrm{e}^{-\eta_t \hat{L}_i}}{\sum_{i=1}^{n} \mathrm{e}^{-\eta_t \hat{L}_i}}$$

Regret analysis

Theorem 2. ***(`Exp3` regret bound)*** *Let $\eta_t = \sqrt{\frac{\ln n}{tn}}$. Then, the pseudo-regret of `Exp3` is upper bounded by:*

$$\overline{\texttt{Regret}}(T) \leq 2\sqrt{Tn \ln n}.$$

The pseudo-regret is at most $O(\sqrt{T})$ and is, therefore, sublinear.

- looser regret bound, but arguably harder setting (less constrained setting);
- interested readers are referred to [3, 6] for the proof.

3.3. Markovian MAB

Setting:

- single arm can be played ($m = 1$), and $\mathbf{x}_t \in \{1, 2, \ldots, n\}$;
- consider the state of the arm i at time t: $s_{i,t} \in \mathcal{S}$, where $\mathcal{S}$ is the state space.
- the process is Markovian;
- if arm selected, the state evolves according to $\Pr[\, s_{t+1,\mathbf{x}_t} |\, s_{t,\mathbf{x}_t}]$, otherwise stay unchanged.
- back to reward, and we consider a discount factor γ;
- let $\underline{r}_i \leq r_i(s_{i,t}) \leq \overline{r}_i$ be arm's i bounded reward at time t, can be negative (e.g., overused);
- transition probability $\Pr$ & reward function r_i are known prior to decision process;
- states $s_{i,t}$ are all observed at t.

Markovian MAB process

In each round $t = 1, 2, \ldots$:

1. play arm $\mathbf{x}_t \in \{0, 1, 2, \ldots, n\}$;
2. observe new state $s_{\mathbf{x}_t, t}$
3. receive reward $r_{\mathbf{x}_t}(s_{\mathbf{x}_t, t})$;
4. compute $\mathbf{x}_{t+1}$ given the new state.

- we have more information and we can performed optimally in the expected sense;
- problem translates to the following program:

$$V(\mathbf{s}_0) = \max_{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots} \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^t r_{\mathbf{x}_t}(s_{\mathbf{x}_t, t}) \middle| \mathbf{s}_0 \right].$$

Index policy

- **Policy #1**: solve the dynamic program via Bellman equation.
  - curse of dimensionality, problem dimension's exponential in n, computationally intractable.
- **Policy #2**: Gittins index [14], i.e., select the arm i possessing the largest index $\nu_i\left(s_{i,t}\right)$.
  - optimal policy for the $m = 1$ case;
  - n indices to compute, each independent of other arms;
  - based on optimal stopping problems (i.e., when to stop using an arm).

Gittins index algorithm

In each round $t = 1, 2, \ldots$:

1. play the arm $\mathbf{x}_t$ with the largest index $\nu_i \left(s_{i,t-1} \right)$
2. observe new state $s_{\mathbf{x}_t,t}$
3. obtain reward $r_{\mathbf{x}_t}(s_{\mathbf{x}_t,t})$;
4. update Gittins index:

$$\nu_{\mathbf{x}_t} = \sup_{T>0} \frac{\mathbb{E}\left[\sum_{\tau=0}^{T} \gamma^\tau r_{\mathbf{x}_\tau} \left(s_{\mathbf{x}_\tau,\tau} \right) \middle| s_{\mathbf{x}_\tau,0} = s_{\mathbf{x}_t,t} \right]}{\mathbb{E}\left[\sum_{\tau=0}^{T} \gamma^\tau \middle| s_{\mathbf{x}_\tau,0} = s_{\mathbf{x}_t,t} \right]}$$

- no regret analysis – optimal decisions;
- multiple play at each round, $m > 1$ (suboptimal policy).

Extension: restless MAB

An interesting extension is the restless MAB [49] in which:

1. multiple play at each round ($m > 1$) and the decision $\mathbf{x}_t \in 2^n$, where $\operatorname{card} \mathbf{x}_t = m$;
2. all states evolves $\sim \Pr$;
3. reward also obtained from unselected arms, new definition: $r_i\left(s_{i,t}, \mathbb{I}_{i \in \mathbf{x}_t}\right)$;

Whittle index: suboptimal heuristic to solve this problem, pick m largest indices defined as

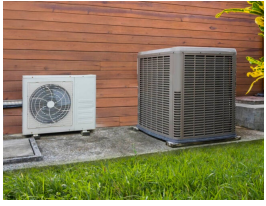
$$\nu_i(s_{i,t}) = \inf_{\lambda} \left\{ \lambda \mid r_i\left(s_{i,t}, \mathbb{I}_{i \in \mathbf{x}_t}\right) = r_i\left(s_{i,t}, \mathbb{I}_{i \notin \mathbf{x}_t}\right) + \lambda \mathbb{I}_{i \notin \mathbf{x}_t} \right\}.$$

Interpretation: subsidiary λ such that is it equally good to play or not arm i.

3.4. Example: Stochastic-MAB for load curtailment [32]

- demand response: aggregator wishes to curtail load power consumption to assist the system operator, e.g., peak-shaving or frequency regulation;
- at each time step, m_t loads must be curtailed;
- consider a set of n unknown, uncertain loads ($\rightarrow$ arms);
- the potential for power reduction of loads is characterized by a i.i.d. bounded random variables.

(a) Load aggregation *Source: Pecan Street*



(b) Heat pump *Source: iStockPhoto*

Figure 5: DR of flexible loads

MAB with stochastic plays

Theorem 3. (UBC-SP regret bound) *Let $m_t \sim$ wide-sense stochastic process. Then pseudo-regret of UCB1 where the m_t largest indices λ_i are selected is bounded above by:*

$$\overline{\texttt{Regret}}(T) \leq n\Delta_{\max}\left(\frac{6\left(\sigma_\kappa^2+\kappa^2\right)\ln T}{\Delta_{\min}^2}+1+\frac{\kappa\pi^2}{3}\right).$$

where κ and σ_κ^2 are the mean and variance of m_t.

In this case, the pseudo-regret has been modified to account for multiple plays:

$$\overline{\texttt{Regret}}(T)=\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i\in\mathbf{x}_t^*}\mu_i-\sum_{i\in\mathbf{x}_t}\mu_i\right)\right],$$

where $\mathbf{x}_t^*$ is the set collecting the m_t largest μ_i, i.e., $\mathbf{x}_t \in \underset{\mathbf{x}\in 2^n}{\arg\max}\sum_{i\in\mathbf{x}}|\mu_i|$ s.t. $\operatorname{card}\mathbf{x}=m_t$.

Numerical example

Consider a frequency regulation setting where power deficit is mitigated.

- $n = 20$ loads;
- curtailment $X_{i,t} \sim$ i.i.d. Uniform;
- Deficit signal $s^t \sim \mathrm{N}_{>0}(\mu_{\mathsf{ACE}}, \sigma^2_{\mathsf{ACE}})$;
- Load to deploy to mitigate imbalance m_t: min number of loads that summed are $\geq s^t$ (using d-moving average of $X_{i,t}$);
- $T = 10^5$, i.e., 112 hours for 4 second regulation time steps;
- Naive = Policy #1.

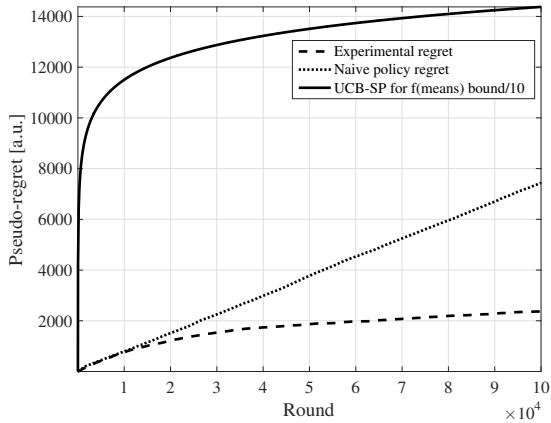


Figure 6: Experimental pseudo-regret

Conclusion: multi-armed bandit

- low information, sequential decision-making framework;
- setting & solution concepts for the three main types of bandits:
  1. stochastic;
  2. adversarial;
  3. Markovian (bonus, restless).
- simple assumption means that it can be extended to many problems;
- MAB with stochastic plays for load curtailment in power systems.

4. The online convex optimization problem

- optimization as a process [16];
- **objective:** iteratively minimize objective function ($\approx$ learn & adapt strategy);
- design very computationally efficient decision rules;
- we will be looking two types of algorithm and theirs respective regret analysis:
  1. static
  2. dynamic
- rich performance analysis.

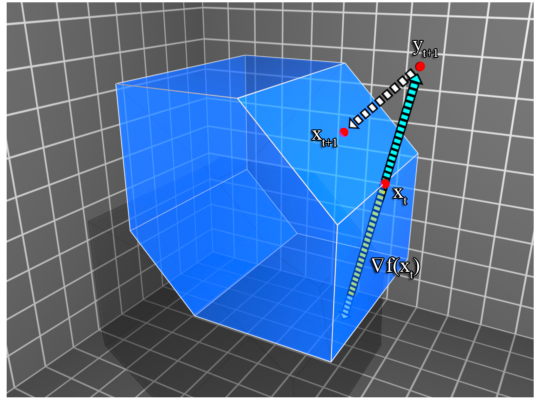


Figure 7: Online gradient descent *Source: [16]*

Applications

1. Static:
   - spam filtering [16];
   - portfolio selection;
   - recommendation systems via matrix completion;
   - localization of fixed target or online regression of patrolling target;
   - computer breach detection via online support vector machine [25];
   - learning EV behaviour models [44];
   - pricing for EV charging [39];
   - economic dispatch [41, 4, 50];
   - state estimation [24];
   - optimal power flow [23].

Applications – II

2. Dynamic:
   - moving target tracking [34];
   - network resource allocation [10];
   - internet of things [9];
   - flexible/controllable and nominal load disaggregation in distribution feeder [26];
   - real-time pricing in power systems [22];
   - demand response for frequency regulation [31];
   - multi-energy building management [35];
   - economic dispatch [8].

Online convex optimization (OCO)

Setting: in each round $t = 1, 2, \ldots, T$, the decision maker solves:

$$\begin{aligned} \min_{\mathbf{x}_t} \quad & f_t(\mathbf{x}_t) \\ \text{s.t.} \quad & \mathbf{x}_t \in \mathcal{X} \end{aligned}$$

- where $\mathbf{x}_t$: decision at t, $\mathcal{X} \subseteq \mathbb{R}^n$: compact & convex decision set;
- $f_t : \mathbb{R}^n \mapsto \mathbb{R}$ convex, (sub)differentiable, but only observed **after** the round;
- B-bounded function: $|f_t(\mathbf{x})| < B$ and is G-bounded gradient: $|\nabla f_t(\mathbf{x})| < G \ \forall t$;
- distribution of the sequence of f_t: stochastic, adversarial;
- (computational resources are **limited**)

4.1. OCO process

In each round $t = 1, 2, \ldots, T$:

1. implement decision $\mathbf{x}_t$;
2. suffer the loss $f_t(\mathbf{x}_t)$ and observe all online parameters of f_t;
3. compute next decision: $\mathbf{x}_{t+1} = \text{UpdateRule}(f_t, \mathbf{x}_t)$.

Objective: design an efficient update rule which leads to a bounded regret.

Regret

Regret: we adapt the regret to the OCO setting and obtain

$$\texttt{Regret}_T = \underbrace{\sum_{t=1}^{T} f_t(\mathbf{x}_t)}_{\text{our decisions}} - \underbrace{\sum_{t=1}^{T} f_t(\mathbf{x}_t^\star)}_{\text{comparators}}.$$

1. static regret: $\mathbf{x}_t^\star = \mathbf{x}^\star \; \forall \; t \rightarrow$ **best single decision**

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

- similar to MAB;
- application *online*: linear regression, localization/state estimation, portfolio rebalancing, etc.

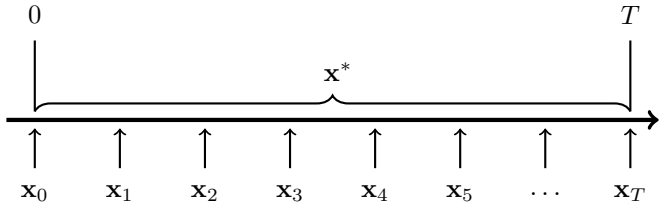
Regret – II

2. dynamic regret: $\mathbf{x}_t^\star \to$ round **optimum**

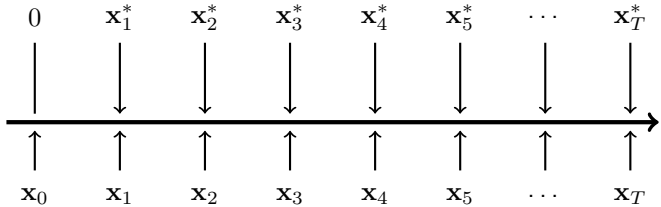
$$\mathbf{x}_t^\star \in \arg\min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x}_t)$$

- condition on $\{\mathbf{x}_t^\star\}_{t=1}^T$ via the **cumulative variation** V_T;
- measure *how dynamic is the problem*;
- $V_T = \sum_{t=1}^T \left\|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^\star\right\|$;
- we actually used a similar notion in MAB with stochastic plays;
- application *online*: moving-target localization, signal tracking, resource allocation, etc.

Regret – III



(a) Static regret



(b) Dynamic regret

Figure 8: Notions of regret

Step-by-step optimization

- next decision: $\mathbf{x}_{t+1}$ is computed using an **update rule**;
- idea: we use **single iteration** from a standard constrained convex optimization algorithm;
- **simple & efficient**, then establish performance guarantee;
- e.g., projected gradient descent [51].

Online gradient descent (OGD)

In each round $t = 1, 2, \ldots, T$:

1. implement decision $\mathbf{x}_t$;
2. suffer the loss $f_t(\mathbf{x}_t)$ and observe all online parameters of f_t;
3. compute next decision:

$$\mathbf{x}_{t+1} = \mathrm{proj}_{\mathcal{X}}\left(\mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t)\right)$$

where

$$\mathrm{proj}_{\mathcal{X}}(\mathbf{y}) = \operatorname*{arg\,min}_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - \mathbf{x}\|.$$

and $\eta_t > 0$ is a judiciously chosen descent step size.

Regret analysis

We recall our assumptions:

- $|f_t(\mathbf{x})| < B \ \forall \ \mathbf{x} \in \mathcal{X}, \forall \ t$ (bounded loss)
- $\mathcal{X}$ is compact: $\|\mathbf{x}\| < X \ \forall \ \mathbf{x} \in \mathcal{X}, \forall \ t$ (bounded decision)
- $\|\nabla f_t(\mathbf{x})\| < G \ \forall \ \mathbf{x} \in \mathcal{X}, \forall \ t$ (bounded gradient)

Theorem 4. (OGD static regret bound) *Let $\eta_t = \frac{X}{G\sqrt{t}}$ with $\frac{1}{\eta_0} = 0$, then OGD's static regret is bounded by:*

$$\texttt{Regret}^{\text{static}}(T) \leq 3GX\sqrt{T}$$

The static regret is $O(\sqrt{T})$ and, thus, sublinear.

Proof*

1. Convexity of f_t implies that:

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \leq \nabla f_t\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \mathbf{x}^*\right)$$

2. Upper bound on the gradient term: let's consider the update to which we substract to optimum:

$$\mathbf{x}_{t+1} - \mathbf{x}^* = \mathrm{proj}_{\mathcal{X}}\left(\mathbf{x}_t - \eta_t \nabla f_t\left(\mathbf{x}_t\right)\right) - \mathbf{x}^*.$$

Taking the $2-$norm on both sides yields:

$$\begin{aligned} &\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \left\|\mathbf{x}_t - \eta_t \nabla f_t\left(\mathbf{x}_t\right) - \mathbf{x}^*\right\|_2^2 \\ \iff &\nabla f_t\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \mathbf{x}^*\right) \leq \frac{1}{2\eta_t}\left(\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 - \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right) + \frac{\eta_t}{2}\left\|\nabla f_t\left(\mathbf{x}_t\right)\right\|_2^2. \end{aligned}$$

Proof* – II

3. Combining steps 1 & 2, and the regret definition:

$$\begin{aligned}\texttt{Regret}(T) &\leq \sum_{t=1}^{T} \nabla f_t\left(\mathbf{x}_t\right)^\top \left(\mathbf{x}_t - \mathbf{x}^*\right) \\ &\leq \sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right) + \frac{\eta_t}{2}\|\nabla f_t\left(\mathbf{x}_t\right)\|_2^2\end{aligned}$$

4. Rearranging the terms in the first parenthesis (and recalling that $\frac{1}{\eta_0} = 0$):

$$\texttt{Regret}(T) \leq -\frac{1}{2\eta_{T+1}}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 + \sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{\eta_t}{2}\|\nabla f_t\left(\mathbf{x}_t\right)\|_2^2$$

Proof* – III

5. By assumption, we have: $\|\mathbf{x}\|^2 \leq X$ and $\|\nabla f_t(\mathbf{x}_t)\|_2 \leq G$

$$\texttt{Regret}(T) \leq \sum_{t=1}^{T} 2X^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{\eta_t G^2}{2}.$$

6. Telescoping sum:

$$\texttt{Regret}(T) \leq \frac{2X^2}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t G^2}{2}.$$

Proof* – IV

7. Observing that:

$$\frac{1}{\sqrt{t}} \leq \frac{2}{\sqrt{t}+\sqrt{t-1}} = 2\left(\sqrt{t}-\sqrt{t-1}\right)$$

Then,

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{\sqrt{t}} &\leq \sum_{t=1}^{T} 2\left(\sqrt{t}-\sqrt{t-1}\right) \\ &= 2\sqrt{T} \qquad \text{(telescoping sum)}\end{aligned}$$

8. Using $\eta_t = \frac{X}{G\sqrt{t}}$ then completes the proof:

$$\texttt{Regret}^{\text{static}}(T) \leq 3XG\sqrt{T}.$$

Regret analysis – II

We extend the previous result to the dynamic case.

Theorem 5. ***(OGD static regret bound)*** *Let $\eta_t = \frac{X}{G\sqrt{T}}$, then OGD's regret is bounded by:*

$$\texttt{Regret}^{\text{dynamic}}(T) \leq \frac{5}{2} G X \sqrt{T} + G\sqrt{T} V_T$$

The dynamic regret is $O(\sqrt{T}\,(V_T + 1))$ and, thus, sublinear if $V_T < O(\sqrt{T})$.

- stricter condition on the problem via V_T, but better guarantee for some applications.

Note: $V_T = \sum_{t=1}^{T} \left\| \mathbf{x}_{t+1}^\star - \mathbf{x}_t^\star \right\|$, the cumulative variation.

Proof

1. Same as static regret proof but with $\mathbf{x}^* \to \mathbf{x}_t^*$ (+ bounding the gradient term):

$$\texttt{Regret}(T) \leq \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\mathbf{x}_t - \mathbf{x}_t^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|_2^2 \right) + \frac{\eta_t G^2}{2}$$

2. Noting that $\eta_t = \eta$: we expand the sum

$$\begin{aligned}
\texttt{Regret}(T) &\leq \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\mathbf{x}_t\|_2^2 + \|\mathbf{x}_t^*\|_2^2 - 2\mathbf{x}_t^\top \mathbf{x}_t^* - \|\mathbf{x}_{t+1}\|_2^2 - \|\mathbf{x}_t^*\|_2^2 + 2\mathbf{x}_{t+1}^\top \mathbf{x}_t^* \right) + \frac{\eta G^2}{2} \\
&= \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\mathbf{x}_t\|_2^2 - \|\mathbf{x}_{t+1}\|_2^2 \right) + \frac{1}{\eta} \left(\mathbf{x}_{t+1} - \mathbf{x}_t \right)^\top \mathbf{x}_t^* + \frac{\eta G^2}{2}
\end{aligned}$$

Proof* – II

3. Re-expressing the second term in terms of of $V_T = \sum_{t=1}^{T} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|_2$:

$$\begin{aligned}
\texttt{Regret}(T) &\leq \frac{1}{\eta} \mathbf{x}_T^{*\top} \mathbf{x}_{T+1} - \frac{1}{\eta} \mathbf{x}_1^{*\top} \mathbf{x}_1 + \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\mathbf{x}_t\|_2^2 - \|\mathbf{x}_{t+1}\|_2^2 \right) + \frac{1}{\eta} \left(\mathbf{x}_t^* - \mathbf{x}_{t+1}^* \right)^\top \mathbf{x}_{t+1} \\
&\quad + \frac{\eta G^2}{2} \\
&\leq \frac{1}{\eta} \mathbf{x}_T^{*\top} \mathbf{x}_{T+1} + \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\mathbf{x}_t\|_2^2 - \|\mathbf{x}_{t+1}\|_2^2 \right) + \frac{\eta G^2}{2} + \sum_{t=2}^{T} \frac{1}{\eta} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\|_2 \|\mathbf{x}_{t+1}\|_2 \\
&\leq \frac{X^2}{\eta} + \frac{1}{\eta} V_T + \frac{T\eta G^2}{2} + \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\mathbf{x}_t\|_2^2 - \|\mathbf{x}_{t+1}\|_2^2 \right)
\end{aligned}$$

Proof* – III

3. Telescoping sum:

$$\begin{aligned}\texttt{Regret}(T) &\leq \frac{X^2}{\eta} + \frac{1}{\eta}V_T + \frac{T\eta G^2}{2} + \frac{1}{2\eta}\left(\|\mathbf{x}_1\|_2^2 - \|\mathbf{x}_{T+1}\|_2^2\right) \\ &\leq \frac{X^2}{\eta} + \frac{1}{\eta}V_T + \frac{T\eta G^2}{2} + \frac{X^2}{2\eta}.\end{aligned}$$

4. Setting $\eta = \frac{X}{G\sqrt{T}}$ concludes the proof:

$$\texttt{Regret}^{\text{dynamic}}(T) \leq \frac{5}{2}GX\sqrt{T} + G\sqrt{T}V_T.$$

Extensions

There are many possible extension (beauty of a simple framework):

- strongly convex function (tighter regret bounds) [17, 40];
- feedback types: bandit OCO (only value of $f_t(\mathbf{x}_t)$ is available) [12, 31];
- distributed OCO: decisions are computed locally [42, 25, 27];
- time-varying constraints: $\mathcal{X} \to \mathcal{X}_t$ [10, 7];
- second-order update: based on the Newton's step [34, 38];
- binary decisions: submodular function or using randomization [20, 18, 33, 28];
- predictive/rolling horizon [4, 29];

and many more.

4.2. Example: Demand response (DR) of flexible loads [31]

- **modulate power consumption** (continuous) of flexible loads in exchange of a reward.
- **low** infrastructure investment & renewable;
- applications: frequency regulation.

Figure 9: Residential load aggregation *Source: CBC/Hayward*

(a) Electric vehicles *Source: City of Ventura*

(b) Air conditioners/heat pumps *Source: LG*

(c) Water heater *Source: Homedepot*

Figure 10: Flexible loads

Demand response model

Online convex optimization for DR

$$\min_{\mathbf{x}_t \in [-1,1]^n} \underbrace{\left(s_t - p_{0,t} - \mathbf{p}_t^\top \mathbf{x}_t\right)^2}_{\text{Tracking error}} + \underbrace{\mathrm{R}\left(\mathbf{x}_t\right)}_{\text{Regularizer/side objective}}$$

- s_t: signal to track (**unknown**);
- $\mathbf{x}_t \in [-1,1]^N$: adjustment signal (**decision/control**);
- $p_{0,t}$: nominal load consumption (**uncertain**);
- $\mathbf{p}_t$: load response to signal (**uncertain**);
  - $\mathbf{p}_t^\top \mathbf{x}_t$: *total* power adjustment of flexible loads.

Composite objective gradient descent

Composite objective gradient descent (COGD)

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in[-1,1]} \eta \nabla f_t(\mathbf{x}_t)^\top \mathbf{x} + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2 + \eta \mathrm{R}(\mathbf{x})$$

- Specially tailored to round-independent regularizers [11, 15, 31];

Theorem 6. *Let $\eta = \frac{\delta}{\sqrt{T}}$ with $a > 0$ and $\mathbf{x}_0 = \mathbf{0}$. Then, COGD's regret is bounded by*

$$\texttt{Regret}^{\text{dynamic}}(T) \leq \sqrt{T}\left(\frac{X}{\delta} + \frac{G^2\delta}{2} + \frac{4X}{\delta}V_T\right)$$

- $\texttt{Regret}^{\text{dynamic}}(T) < O(\sqrt{T}(V_T + 1)$ similarly to OGD.

Extension: feedback level

We consider four types of feedback from the loads:

1. **full information**;
2. **bandit or limited:** only $f_t(\mathbf{x}_t)$ is observed, e.g., power measurement at the feeder level;
3. **partial bandit:** some loads have full & some bandit, e.g., opt-out for privacy reasons;
4. **Bernoulli:** rounds are either full or bandit for every loads, reduce communication burden.

Numerical example

- thermostatically controlled loads with continuous decisions (e.g., HVAC);
- $s_t = 15\sin(0.1t)$;
- $N = 100$ loads;
- subject to noise: $\mathbf{p}_t = \overline{\mathbf{p}} + \mathbf{N}_{[-1,1]}[0, \frac{1}{2}]$;
- regularizer: sparsity & desired temperature;
- time horizon $T = 600$.

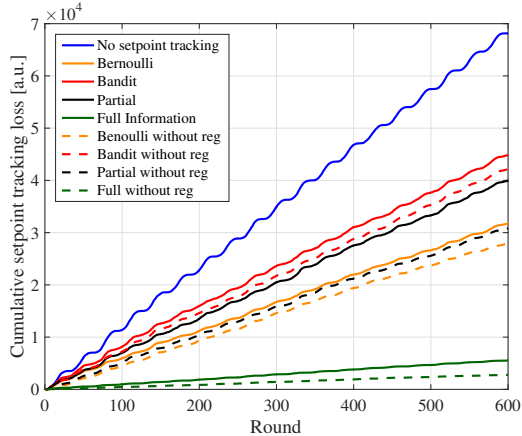
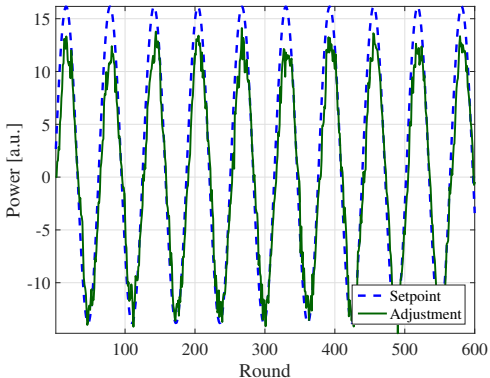
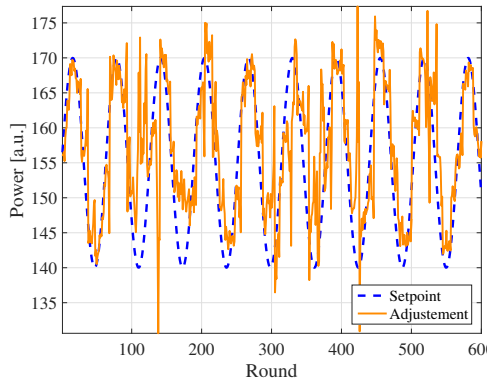


Figure 11: Setpoint tracking loss comparison (averaged over 100 simulations)

Setpoint tracking



(a) Full information setting



(b) Bernoulli feedback setting

Figure 12: Setpoint tracking with flexible loads

Conclusion: online convex optimization

- optimization problem decomposed as an **optimization process**;
- **real-time decision-making**: incorporate new data efficiently and mitigate uncertainty;
- next decision: **update rule** based single iteration of a convex optimization algorithm;
- performance **guarantee**: sublinear regret bound;
- large potential for extension **tailored** to the problem at hand.

5. Conclusion

- we discussed online decision-making algorithm & their provable performance analysis;
- two main families of approaches:
  - **multi-armed bandit** (stochastic, adversarial, and Markovian);
  - **online convex optimization**.
- set the basis, now time to adapt frameworks to your problems;
- a natural next step: reinforcement learning. To be continued!

The end.

Antoine Lesage-Landry

Department of Electrical Engineering

Polytechnique Montréal

antoine.lesage-landry@polymtl.ca

alesagelandry.github.io

Looking for PhD students & postdocs, always happy to host grad students for an external stay.

This work was funded by the National Science and Engineering Research Council of Canada (NSERC)

Hoeffding inequalities

Hoeffding inequalities

Let $X_{i,t} \in [0,1]$ be bounded i.i.d. random variables with expected value μ_i. Then, Hoeffding inequalities state that:

$$\Pr\left[\sum_{t=1}^{T} X_{i,t} - T\mu_i \geq \alpha\right] \leq \mathrm{e}^{-\frac{2\alpha^2}{T}}$$

$$\Pr\left[\sum_{t=1}^{T} X_{i,t} - T\mu_i \leq -\alpha\right] \leq \mathrm{e}^{-\frac{2\alpha^2}{T}}.$$

Thus, we also have:

$$\Pr\left[\left|\sum_{t=1}^{T} X_{i,t} - T\mu_i\right| \geq \alpha\right] \leq 2\mathrm{e}^{-\frac{2\alpha^2}{T}}.$$

Back to UCB.

OCO with time-varying constraints

OCO with time-varying constraints:

$$\begin{aligned} \min_{\mathbf{x}_t \in \mathcal{X}} \quad & f_t(\mathbf{x}_t) \\ \text{subject to} \quad & g_{t,j}(\mathbf{x}_t) \leq 0 \text{ for } j = 1, 2, \dots J \\ & h_{t,k}(\mathbf{x}_t) = 0 \text{ for } k = 1, 2, \dots K. \end{aligned}$$

Back to OCO extensions.

Lipschitz continuity

Let $0 < L < +\infty$. A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz continuous with respect to some norm if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| .$$

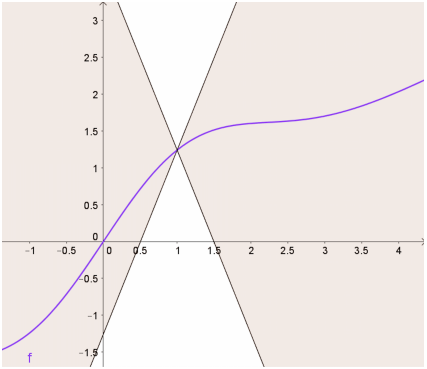


Figure 13: Lipschitz continuity *Source: Wikipedia*

Doubling trick

If the time horizon T is unknown, we can use the doubling trick [43] and use our algorithm successively with a pseudo-$T = 2^m$. This only increases the regret by a constant factor.

Let $0 \leq a < 1$. Consider a regret bound described by: $\texttt{Regret}(T) \leq \alpha T^a + \beta V_T + \gamma T^a V_T$.

Then, using $T = 2^m$ until the end of the process leads to:

$$\begin{aligned}
\texttt{Regret}(T) &\leq \sum_{m=1}^{\lceil \log_2 T \rceil} \alpha \left(2^m\right)^a + \beta V_{2^m} + \gamma \left(2^m\right)^a V_{2^m} \\
&\leq \frac{2^{a+1} 2^{a \log_2 T} - 1}{2^a - 1} \left(\alpha + \gamma V_T\right) + \beta V_T \\
&\leq \frac{2^{a+1}}{2^a - 1} T^a \left(\alpha + \gamma V_T\right) + \beta V_T.
\end{aligned}$$

Note. The regret must not have a constant term.

References I

[1] Rajeev Agrawal. "The continuum-armed bandit problem". In: *SIAM journal on control and optimization* 33.6 (1995), pp. 1926–1951.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47 (2002), pp. 235–256.

[3] Peter Auer et al. "The nonstochastic multiarmed bandit problem". In: *SIAM journal on computing* 32.1 (2002), pp. 48–77.

[4] Masoud Badiei, Na Li, and Adam Wierman. "Online convex optimization with ramp constraints". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE. 2015, pp. 6730–6736.

References II

[5] Sambaran Bandyopadhyay, Pratyush Kumar, and Vijay Arya. "Planning curtailment of renewable generation in power grids". In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 26. 2016, pp. 353–357.

[6] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *Foundations and Trends® in Machine Learning* 5.1 (2012), pp. 1–122.

[7] Xuanyu Cao, Junshan Zhang, and H Vincent Poor. "A virtual-queue-based algorithm for constrained online convex optimization with applications to data center resource allocation". In: *IEEE Journal of Selected Topics in Signal Processing* 12.4 (2018), pp. 703–716.

References III

[8] Spiros Chadoulos and Iordanis Koutsopoulos. “Learning the Optimal Energy Supply Plan with Online Convex Optimization”. In: *ICC 2021-IEEE International Conference on Communications*. IEEE. 2021, pp. 1–6.

[9] Tianyi Chen and Georgios B Giannakis. “Bandit convex optimization for scalable and dynamic IoT management”. In: *IEEE Internet of Things Journal* 6.1 (2018), pp. 1276–1286.

[10] Tianyi Chen, Qing Ling, and Georgios B Giannakis. “An online convex optimization approach to proactive network resource allocation”. In: *IEEE Transactions on Signal Processing* 65.24 (2017), pp. 6350–6364.

References IV

[11] John C Duchi et al. "Composite objective mirror descent.". In: *COLT*. Vol. 10. Citeseer. 2010, pp. 14–26.

[12] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. 2005, pp. 385–394.

[13] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations". In: *IEEE/ACM Transactions on Networking* 20.5 (2012), pp. 1466–1478.

References V

[14] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[15] Eric C Hall and Rebecca M Willett. "Online convex optimization in dynamic environments". In: *IEEE Journal of Selected Topics in Signal Processing* 9.4 (2015), pp. 647–662.

[16] Elad Hazan. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325.

[17] Elad Hazan, Amit Agarwal, and Satyen Kale. "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3 (2007), pp. 169–192.

References VI

[18] Elad Hazan and Satyen Kale. "Online Submodular Minimization.". In: *Journal of Machine Learning Research* 13.10 (2012).

[19] Qinran Hu et al. "A user selection algorithm for aggregating electric vehicle demands based on a multi-armed bandit approach". In: *IET Energy Systems Integration* 3.3 (2021), pp. 295–305.

[20] Stefanie Jegelka and Jeff A Bilmes. "Online Submodular Minimization for Combinatorial Structures.". In: *ICML*. Citeseer. 2011, pp. 345–352.

[21] Dileep Kalathil and Ram Rajagopal. "Online learning for demand response". In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pp. 218–222.

References VII

[22] Seung-Jun Kim and Geogios B Giannakis. "An online convex optimization approach to real-time energy pricing for demand response". In: *IEEE Transactions on Smart Grid* 8.6 (2016), pp. 2784–2793.

[23] Seung-Jun Kim, Geogios B Giannakis, and Kwang Y Lee. "Online optimal power flow with renewables". In: *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE. 2014, pp. 355–360.

[24] Seung-Jun Kim, Gang Wang, and Geogios B Giannakis. "Online semidefinite programming for power system state estimation". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 6024–6027.

References VIII

[25] Alec Koppel, Felicia Y Jakubiec, and Alejandro Ribeiro. "A saddle point algorithm for networked online convex optimization". In: *IEEE Transactions on Signal Processing* 63.19 (2015), pp. 5149–5164.

[26] Gregory S Ledva, Laura Balzano, and Johanna L Mathieu. "Real-time energy disaggregation of a distribution feeder's demand using online learning". In: *IEEE Transactions on Power Systems* 33.5 (2018), pp. 4730–4740.

[27] Antoine Lesage-Landry and Duncan S Callaway. "Dynamic and distributed online convex optimization for demand response of commercial buildings". In: *IEEE Control Systems Letters* 4.3 (2020), pp. 632–637.

References IX

[28] Antoine Lesage-Landry and Julien Pallage. "Online dynamic submodular optimization". In: *IEEE Transactions on Automatic Control* (2023). Submitted.

[29] Antoine Lesage-Landry, Iman Shames, and Joshua A Taylor. "Predictive online convex optimization". In: *Automatica* 113 (2020), p. 108771.

[30] Antoine Lesage-Landry and Joshua A Taylor. "Learning to shift thermostatically controlled loads". In: *2017 50th Hawaii International Conference on System Science*. 2017, pp. 1–8.

[31] Antoine Lesage-Landry and Joshua A Taylor. "Setpoint tracking with partially observed loads". In: *IEEE Transactions on Power Systems* 33.5 (2018), pp. 5615–5627.

References X

[32] Antoine Lesage-Landry and Joshua A Taylor. "The multi-armed bandit with stochastic plays". In: *IEEE Transactions on Automatic Control* 63.7 (2017), pp. 2280–2286.

[33] Antoine Lesage-Landry, Joshua A Taylor, and Duncan S Callaway. "Online convex optimization with binary constraints". In: *IEEE Transactions on Automatic Control* 66.12 (2021), pp. 6164–6170.

[34] Antoine Lesage-Landry, Joshua A Taylor, and Iman Shames. "Second-order online nonconvex optimization". In: *IEEE Transactions on Automatic Control* 66.10 (2020), pp. 4866–4872.

References XI

[35] Antoine Lesage-Landry et al. "Online convex optimization of multi-energy building-to-grid ancillary services". In: *IEEE Transactions on Control Systems Technology* 28.6 (2019), pp. 2416–2431.

[36] Lihong Li et al. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 661–670.

[37] Yingying Li, Qinran Hu, and Na Li. "A reliability-aware multi-armed bandit approach to learn and select users in demand response". In: *Automatica* 119 (2020), p. 109015.

References XII

[38] Jean-Luc Lupien and Antoine Lesage-Landry. "An Online Newton's Method for Time-Varying Linear Equality Constraints". In: *IEEE Control Systems Letters* 7 (2023), pp. 1423–1428.

[39] Wann-Jiun Ma, Vijay Gupta, and Ufuk Topcu. "Distributed charging control of electric vehicles using online learning". In: *IEEE Transactions on Automatic Control* 62.10 (2016), pp. 5289–5295.

[40] Aryan Mokhtari et al. "Online optimization in dynamic environments: Improved regret rates for strongly convex problems". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE. 2016, pp. 7195–7201.

References XIII

[41] Balakrishnan Narayanaswamy, Vikas K Garg, and TS Jayram. "Online optimization for the smart (micro) grid". In: *Proceedings of the 3rd international conference on future energy systems: where energy, computing and communication meet*. 2012, pp. 1–10.

[42] Shahin Shahrampour and Ali Jadbabaie. "Distributed online optimization in dynamic environments using mirror descent". In: *IEEE Transactions on Automatic Control* 63.3 (2017), pp. 714–725.

[43] Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

References XIV

[44] Nasim Yahya Soltani, Seung-Jun Kim, and Georgios B Giannakis. "Real-time load elasticity tracking and pricing for electric vehicle charging". In: *IEEE Transactions on Smart Grid* 6.3 (2014), pp. 1303–1313.

[45] Jianfeng Sun et al. "A dynamic distributed energy storage control strategy for providing primary frequency regulation using multi-armed bandits method". In: *IET Generation, Transmission & Distribution* 16.4 (2022), pp. 669–679.

[46] Joshua A Taylor and Johanna L Mathieu. "Index policies for demand response". In: *IEEE Transactions on Power Systems* 29.3 (2013), pp. 1287–1295.

References XV

[47] Qingsi Wang, Mingyan Liu, and Johanna L Mathieu. "Adaptive demand response: Online learning of restless and controlled bandits". In: *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2014, pp. 752–757.

[48] Yiyang Wang and Neda Masoud. "Adversarial Online Learning With Variable Plays in the Pursuit-Evasion Game: Theoretical Foundations and Application in Connected and Automated Vehicle Cybersecurity". In: *IEEE Access* 9 (2021), pp. 142475–142488.

[49] Peter Whittle. "Restless bandits: Activity allocation in a changing world". In: *Journal of applied probability* 25.A (1988), pp. 287–298.

[50] Jianjun Yuan and Andrew Lamperski. "Online convex optimization for cumulative constraints". In: *Advances in Neural Information Processing Systems* 31 (2018).

References XVI

[51] Martin Zinkevich. "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 928–936.