Reinforcement Learning
Adversarial Learning

Odalric-Ambrym Maillard

Spring 2022

Inria Scool

Aggregation of experts

- Compare to: Best arm, Best convex combinations, Best sequence, Best recurring sequence.

Adversarial bandits

- Exp3, Exp4
- Best of both world

Min-max games

- Bandits and Nash equilibrium

Risk-aversion

- CVaR, EVaR, etc.

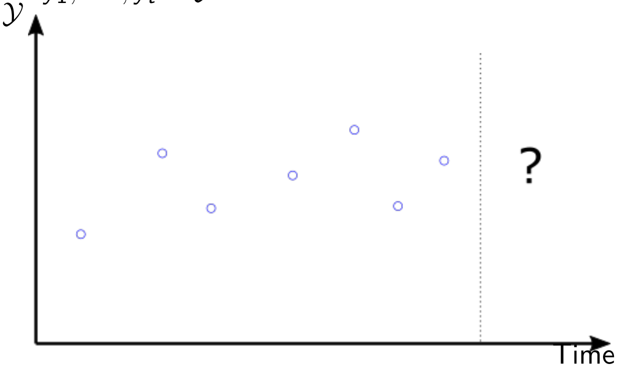
Robust planning

- Autonomous vehicles.

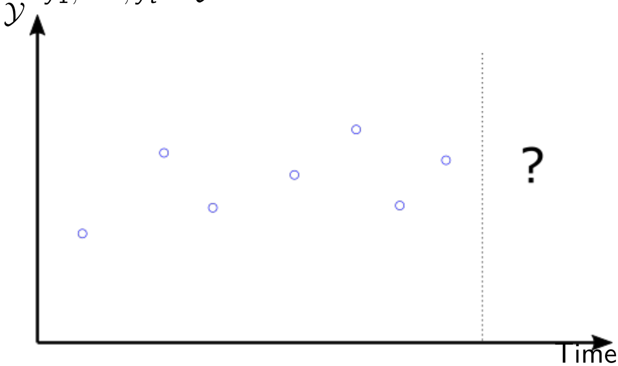
Table of contents

Many models?

▷ Observe a signal $y_1, \dots, y_t \in \mathcal{Y}$

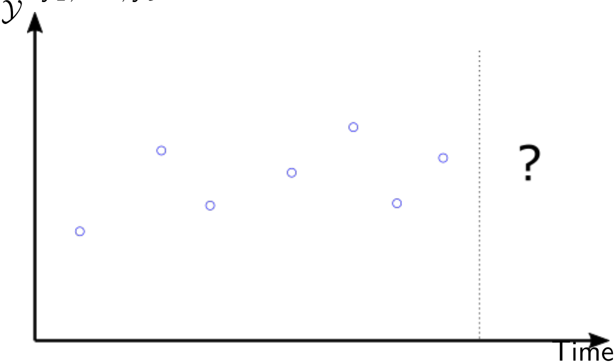


- Observe a signal $y_1, \dots, y_t \in \mathcal{Y}$



- Goal: Predict observation at time $t+1$?

- Observe a signal $y_1, \dots, y_t \in \mathcal{Y}$



- Goal: Predict observation at time $t+1$?
- **Many** available models:
  - *I.i.d.*: $[0,1]$-bounded ?
  - *Parametric*: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$ for φ: polynomials, wavelets, etc. ?
  - *Markov*: $y_t \sim P(\cdot | y_{t-1})$, *k-order Markov*: $y_t \sim P(\cdot | y_{t-1}, \dots, y_{t-k})$?
  - *States*: representation maps $\psi(h_t) = s_t$ for observation history h_t?

Which model is best?

Parametric: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$

Parametric: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$

▷ $\varphi(t) = (1, t, t^2, t^3)$

Parametric: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$

- $\varphi(t) = (1, t, t^2, t^3)$
- $\varphi(t) = (cos(t), cos(2t), cos(4t), \dots)$

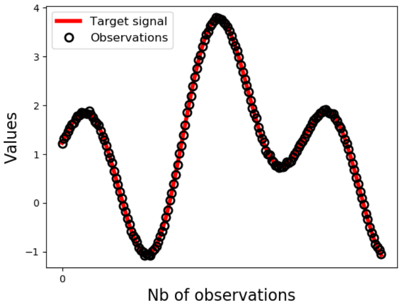
Many models?

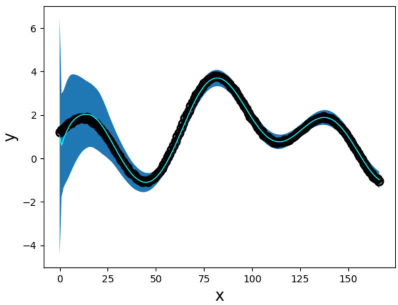
Parametric: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$

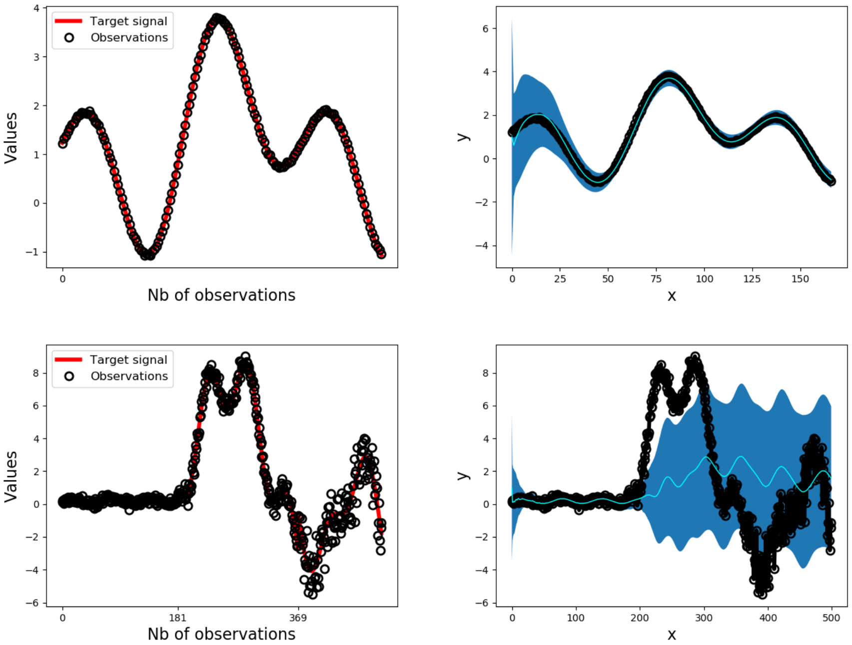
- $\varphi(t) = (1, t, t^2, t^3)$
- $\varphi(t) = (cos(t), cos(2t), cos(4t), \dots)$
- $\varphi(t) =$ wavelet basis

Parametric: $y_t = \langle \theta, \varphi(t) \rangle + \xi_t$

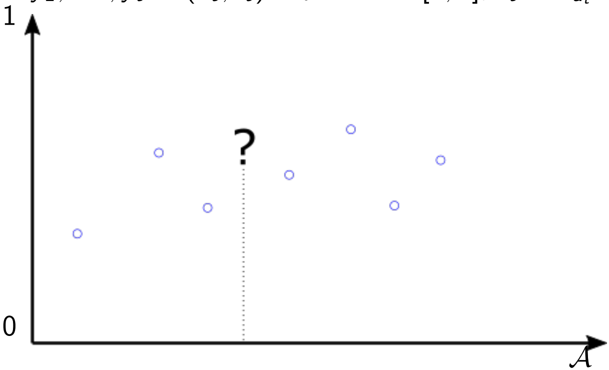
- $\varphi(t) = (1, t, t^2, t^3)$
- $\varphi(t) = (cos(t), cos(2t), cos(4t), \dots)$
- $\varphi(t) =$ wavelet basis
- ...


Target signal
Observations
Values
Nb of observations
4
3
2
1
0
−1
0

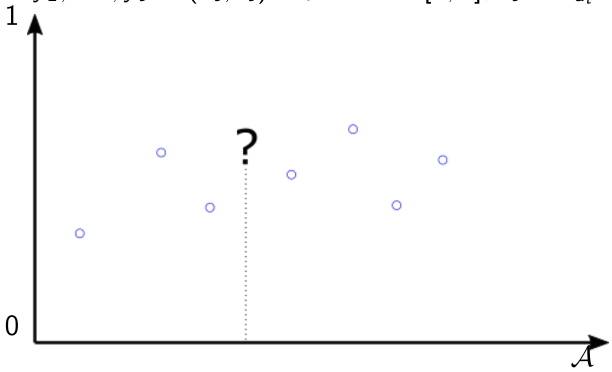

y
x
6
4
2
0
−2
−4
0
25
50
75
100
125
150


Target signal
Observations
Values
Nb of observations
y
x
Target signal
Observations
Values
Nb of observations
y
x

- Sample a signal $y_1, \ldots, y_t = (a_t, r_t) \in \mathcal{Y} = \mathcal{A} \times [0,1]$, $r_t \sim \nu_{a_t}$.

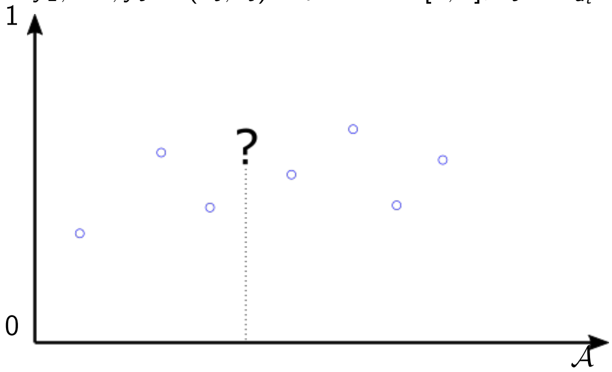


- Sample a signal $y_1, \ldots, y_t = (a_t, r_t) \in \mathcal{Y} = \mathcal{A} \times [0,1]$, $r_t \sim \nu_{a_t}$.



- Goal: choose $a_t \in \mathcal{A}$ to maximize rewards.

- Sample a signal $y_1, \ldots, y_t = (a_t, r_t) \in \mathcal{Y} = \mathcal{A} \times [0,1]$, $r_t \sim \nu_{a_t}$.



- Goal: choose $a_t \in \mathcal{A}$ to maximize rewards.
- Many available algorithms:
  - *Bandits*: UCB? UCB-V? KL-UCB? TS?
  - *Structured bandits*: OFUL, GP-UCB? IMED?
  - *MDPs*: UCRL? Q-learning? DQN?

Which algorithm is best?

Table of contents

▷ **Set of models** $\mathcal{M}$.

At each time step:

- **Set of models** $\mathcal{M}$.

At each time step:

- Each model $m \in \mathcal{M}$ outputs a **decision** $x_{t,m} \in \mathcal{X}$:
  - $\mathcal{X} = \mathcal{Y}$, $\mathcal{X} = \mathcal{P}(\mathcal{Y})$, $\mathcal{X} = \mathcal{A}$.

- **Set of models** $\mathcal{M}$.

At each time step:

- Each model $m \in \mathcal{M}$ outputs a **decision** $x_{t,m} \in \mathcal{X}$:
  - $\mathcal{X} = \mathcal{Y}$, $\qquad \mathcal{X} = \mathcal{P}(\mathcal{Y})$, $\qquad \mathcal{X} = \mathcal{A}$.
- We output **decision** $x_t \in \mathcal{X}$ based on $(x_{t,m})_{m \in \mathcal{M}}$.

Decisions and Losses

- **Set of models** $\mathcal{M}$.

At each time step:

- Each model $m \in \mathcal{M}$ outputs a **decision** $x_{t,m} \in \mathcal{X}$:
  - $\mathcal{X} = \mathcal{Y}$, $\mathcal{X} = \mathcal{P}(\mathcal{Y})$, $\mathcal{X} = \mathcal{A}$.
- We output **decision** $x_t \in \mathcal{X}$ based on $(x_{t,m})_{m\in\mathcal{M}}$.
- All decisions evaluated via a **loss** $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
  - Quadratic: $\ell(x, y) = \frac{(x-y)^2}{2}$,
  - Self-information: $\ell(x, y) = -\log(x(y))$,
  - Reward: $\ell(x, y) = 1 - y(x)$

- **Set of models** $\mathcal{M}$.

At each time step:

- Each model $m \in \mathcal{M}$ outputs a **decision** $x_{t,m} \in \mathcal{X}$:
  - $\mathcal{X} = \mathcal{Y}$, $\mathcal{X} = \mathcal{P}(\mathcal{Y})$, $\mathcal{X} = \mathcal{A}$.
- We output **decision** $x_t \in \mathcal{X}$ based on $(x_{t,m})_{m\in\mathcal{M}}$.
- All decisions evaluated via a **loss** $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
  - Quadratic: $\ell(x, y) = \frac{(x-y)^2}{2}$,
  - Self-information: $\ell(x, y) = -\log(x(y))$,
  - Reward: $\ell(x, y) = 1 - y(x)$
- We receive **observation** $y_t \in \mathcal{Y}$, and incur **loss** $\ell_t(x_t) := \ell(x_t, y_t)$.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

▷ **Set of models** $\mathcal{M}$.

At each time step:

▷ Each model $m \in \mathcal{M}$ outputs a **decision** $x_{t,m} \in \mathcal{X}$:
- ◇ $\mathcal{X} = \mathcal{Y}$, $\quad \mathcal{X} = \mathcal{P}(\mathcal{Y})$, $\quad \mathcal{X} = \mathcal{A}$.

▷ We output **decision** $x_t \in \mathcal{X}$ based on $(x_{t,m})_{m\in\mathcal{M}}$.

▷ All decisions evaluated via a **loss** $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$
- ◇ Quadratic: $\ell(x,y) = \frac{(x-y)^2}{2}$,
- ◇ Self-information: $\ell(x,y) = -\log(x(y))$,
- ◇ Reward: $\ell(x,y) = 1 - y(x)$

▷ We receive **observation** $y_t \in \mathcal{Y}$, and incur **loss** $\ell_t(x_t) := \ell(x_t, y_t)$.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

▷ Q: in Expectation? High probability?

▷ Say after playing x_t, you **observe** y_t (more generally ℓ_t). Then, you can compute $\ell_t(x)$ for all other choice.

Full information

▷ In bandit, only $\ell_t(x_t)$ is observed, but ℓ_t is **unknown**:

Partial information: Bandit feedback

▷ Intermediate settings: e.g. Classification $\ell(x, y) = \mathbb{I}\{x \neq y\}$.
(Only) If I receive **loss** 0, then, I know y, hence I can compute $\ell(x, y)$ for all x.

Semi-bandit Feedback

In the sequel, we first consider **full information**.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \; \ldots$$

w.r.t.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

▷ Goal 1: **best model** (Model **selection**) ?

$$\min_{m \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(x_{t,m})$$

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- Goal 1: **best model** (Model **selection**) ?

$$\min_{m \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(x_{t,m})$$

- Goal 2: **best combination** of models (Model **aggregation**)?

$$\min_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \min_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m \in \mathcal{M}} q_m x_{t,m} \right)$$

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- Goal 1: **best model** (Model **selection**) ?

$$\min_{m \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(x_{t,m})$$

- Goal 2: **best combination** of models (Model **aggregation**)?

$$\min_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \min_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m \in \mathcal{M}} q_m x_{t,m} \right)$$

- Goal 3: **best sequence** of models (Model **tracking**)?

$$\sum_{t=1}^{T} \min_{m \in \mathcal{M}} \ell_t(x_{t,m})$$

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \dots$$

w.r.t.

▷ Goal 1: **best model** (Model **selection**) ?

$$\min_{m \in \mathcal{M}} \sum_{t=1}^{T} \ell_t(x_{t,m})$$

▷ Goal 2: **best combination** of models (Model **aggregation**)?

$$\min_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \min_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m \in \mathcal{M}} q_m x_{t,m} \right)$$

▷ Goal 3: **best sequence** of models (Model **tracking**)?

$$\sum_{t=1}^{T} \min_{m \in \mathcal{M}} \ell_t(x_{t,m})$$

▷ ... ?

- You are given a **set** $\mathcal{M}$ of models.

At each time step,

- You maintain some **distribution** $p_t \in \mathcal{P}(\mathcal{M})$ on the set of models.
- You receive **recommendation** $x_{t,m}$ from each model $m \in \mathcal{M}$.
- You use them in order to output some **decision** x_t.
- You incur the corresponding **loss** $\ell_t(x_t)$, an receive **feedback**.

A first approach

▷ Choose x_t as a convex combination of the $(x_{t,m})_{m\in\mathcal{M}}$? or sample $x_t \sim p_t$?

$$x_t = \sum_{m\in\mathcal{M}} p_t(m) x_{t,m} \text{ where } p_t \in \mathcal{P}(\mathcal{M})\,.$$

$\Longrightarrow$ Assuming that $\ell_t(\cdot) = \ell(\cdot, y_t)$ is **convex**, convex combination is better:

$$\ell_t(x_t) \leqslant \sum_{m\in\mathcal{M}} p_t(m)\ell_t(x_{t,m}) = \mathbb{E}_{M\sim p_t}[\ell_t(x_{t,M})]$$

Technical property (Hoeffing Lemma for bounded random variables)

Let r.v. X s.t. $a \leqslant X \leqslant b$ a.s. then

$$\forall \eta \in \mathbb{R}^+, \quad \mathbb{E}[X] \leqslant -\frac{1}{\eta}\log \mathbb{E}[\exp(-\eta X)] + \eta\frac{(b-a)^2}{8}\,.$$

$\Longrightarrow$ Assuming that ℓ is **bounded** by 1, then

$$\mathbb{E}_{M\sim p_t}[\ell_t(x_{t,M})] \leqslant -\frac{1}{\eta}\log \sum_{m\in\mathcal{M}} p_t(m) e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}\,.$$

For **Bounded, convex** loss:

$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\sum_{m\in\mathcal{M}} p_t(m)e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}$$

For **Bounded, convex** loss:

$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\sum_{m\in\mathcal{M}} p_t(m)e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}$$

▷ This suggests:

$$p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}, \qquad w_{t+1}(m) = w_t(m)e^{-\eta\ell_t(x_{t,m})}$$

For **Bounded, convex** loss:

- This suggests:

$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log \sum_{m\in\mathcal{M}} p_t(m) e^{-\eta \ell_t(x_{t,m})} + \frac{\eta}{8}$$

$$p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}, \qquad w_{t+1}(m) = w_t(m) e^{-\eta\ell_t(x_{t,m})}$$

- We get $\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{t+1}}{W_t}\right) + \frac{\eta}{8}$ where $W_t = \sum_{m\in\mathcal{M}} w_t(m)$

For **Bounded, convex** loss:

$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\sum_{m\in\mathcal{M}} p_t(m)e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}$$

- This suggests:

$$p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}, \qquad w_{t+1}(m) = w_t(m)e^{-\eta\ell_t(x_{t,m})}$$

- We get $\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{t+1}}{W_t}\right) + \frac{\eta}{8}$ where $W_t = \sum_{m\in\mathcal{M}} w_t(m)$
- Summing over t yields $\sum_{t=1}^{T}\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{T+1}}{W_1}\right) + \frac{\eta T}{8}$

For **Bounded, convex** loss:

$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\sum_{m\in\mathcal{M}} p_t(m)e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}$$

- This suggests:

$$p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}, \qquad w_{t+1}(m) = w_t(m)e^{-\eta\ell_t(x_{t,m})}$$

- We get $\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{t+1}}{W_t}\right) + \frac{\eta}{8}$ where $W_t = \sum_{m\in\mathcal{M}} w_t(m)$
- Summing over t yields $\sum_{t=1}^T \ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{T+1}}{W_1}\right) + \frac{\eta T}{8}$
- Finally, $W_1 = |\mathcal{M}|$ and for any $m^\star \in \mathcal{M}$,

$$W_{T+1} \geqslant w_{t+1}(m^\star) = \exp\Big(-\eta\sum_{t=1}^T \ell_t(x_{t,m^\star})\Big).$$

For **Bounded, convex** loss:

- This suggests:
$$\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\sum_{m\in\mathcal{M}} p_t(m)e^{-\eta\ell_t(x_{t,m})} + \frac{\eta}{8}$$
$$p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}, \qquad w_{t+1}(m) = w_t(m)e^{-\eta\ell_t(x_{t,m})}$$
- We get $\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{t+1}}{W_t}\right) + \frac{\eta}{8}$ where $W_t = \sum_{m\in\mathcal{M}} w_t(m)$
- Summing over t yields $\sum_{t=1}^{T}\ell_t(x_t) \leqslant -\frac{1}{\eta}\log\left(\frac{W_{T+1}}{W_1}\right) + \frac{\eta T}{8}$
- Finally, $W_1 = |\mathcal{M}|$ and for any $m^\star \in \mathcal{M}$,
$$W_{T+1} \geqslant w_{t+1}(m^\star) = \exp\left(-\eta\sum_{t=1}^{T}\ell_t(x_{t,m^\star})\right).$$
- Hence $$\sum_{t=1}^{T}\ell_t(x_t) \leqslant \sum_{t=1}^{T}\ell_t(x_{t,m^\star}) + \frac{\log(|\mathcal{M}|)}{\eta} + \frac{\eta T}{8}.$$

This leads to the following strategy

1: Let $\forall m \in \mathcal{M}, w_1(m) = 1$
2: **for** $t = 1, \ldots$ **do**
3: Receive $x_{t,m}$ from each model $m \in \mathcal{M}$.
4: Let $p_t(m) = \frac{w_t(m)}{\sum_{m \in \mathcal{M}} w_t(m)}$.
5: Choose $x_t = \sum_{m \in \mathcal{M}} p_t(m) x_{t,m}$
6: Receive loss function ℓ_t.
7: **Update** $w_{t+1}(m) = w_t(m) e^{-\eta \ell_t(x_{t,m})}$ for each m,
Equivalently, $w_{t+1}(m) = \exp(-\eta L_{t,m})$
8: **end for**

This leads to the following strategy

- Choose $x_t = \sum_{m\in\mathcal{M}} p_t(m)x_{t,m}$ where $p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}$,
  - $\forall m \in \mathcal{M}, w_1(m) = 1$ and $w_{t+1}(m) = w_t(m)e^{-\eta\ell_t(x_{t,m})}$.

Theorem (Cesa-Bianchi,Lugosi 2006)

Assume that ℓ_t is **convex** and **bounded** by 1, then this strategy satisfies:

$$\underbrace{\sum_{t=1}^{T}\ell_t(x_t)}_{L_T} - \min_{m\in\mathcal{M}}\underbrace{\sum_{t=1}^{T}\ell_t(x_{t,m})}_{L_{T,m}} \leqslant \frac{\log(|\mathcal{M}|)}{\eta} + \frac{\eta T}{8}$$

This leads to the following strategy

- Choose $x_t = \sum_{m\in\mathcal{M}} p_t(m) x_{t,m}$ where $p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}$,
  - $\forall m \in \mathcal{M}, w_1(m) = 1$ and $w_{t+1}(m) = w_t(m) e^{-\eta \ell_t(x_{t,m})}$.

Theorem (Cesa-Bianchi,Lugosi 2006)

Assume that ℓ_t is **convex** and **bounded** by 1, then this strategy satisfies:

$$\underbrace{\sum_{t=1}^{T} \ell_t(x_t)}_{L_T} - \min_{m\in\mathcal{M}} \underbrace{\sum_{t=1}^{T} \ell_t(x_{t,m})}_{L_{T,m}} \leqslant \frac{\log(|\mathcal{M}|)}{\eta} + \frac{\eta T}{8}$$

- In particular for the choice of parameter $\eta = \sqrt{8\log(|\mathcal{M}|)/T}$,

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T\log(|\mathcal{M}|)}{2}}$$

Aggregation with exponential weights?

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2} \log(|\mathcal{M}|)}$$

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2} \log(|\mathcal{M}|)}$$

▷ **No statistical assumption** on y_t: ℓ_t only convex and bounded!

Aggregation with exponential weights?

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2} \log(|\mathcal{M}|)}$$

- **No statistical assumption** on y_t: ℓ_t only convex and bounded!
- Logarithmic in $|\mathcal{M}|$: Can handle a large amount of models!

Questions

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2}\log(|\mathcal{M}|)}$$

- **No statistical assumption** on y_t: ℓ_t only convex and bounded!
- Logarithmic in $|\mathcal{M}|$: Can handle a large amount of models!

Questions

- Anytime **tuning** of η ($\eta = \eta_t$) ?
  Using $\eta_t = \sqrt{8\log(|\mathcal{M}|)/t}$ at time t, one can show (more involved):

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant 2\sqrt{\frac{T\log(|\mathcal{M}|)}{2}} + \sqrt{\frac{\log(|\mathcal{M}|)}{2}}$$

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2}\log(|\mathcal{M}|)}$$

- **No statistical assumption** on y_t: ℓ_t only convex and bounded!
- Logarithmic in $|\mathcal{M}|$: Can handle a large amount of models!

Questions

- Anytime **tuning** of η ($\eta = \eta_t$) ?
  Using $\eta_t = \sqrt{8\log(|\mathcal{M}|)/t}$ at time t, one can show (more involved):

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant 2\sqrt{\frac{T\log(|\mathcal{M}|)}{2}} + \sqrt{\frac{\log(|\mathcal{M}|)}{2}}$$

- Examples of convex/bounded losses?

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant \sqrt{\frac{T}{2}\log(|\mathcal{M}|)}$$

- **No statistical assumption** on y_t: ℓ_t only convex and bounded!
- Logarithmic in $|\mathcal{M}|$: Can handle a large amount of models!

Questions

- Anytime **tuning** of η ($\eta = \eta_t$) ?
  Using $\eta_t = \sqrt{8\log(|\mathcal{M}|)/t}$ at time t, one can show (more involved):

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant 2\sqrt{\frac{T\log(|\mathcal{M}|)}{2}} + \sqrt{\frac{\log(|\mathcal{M}|)}{2}}$$

- Examples of convex/bounded losses?
- Simplify this assumption, cf. Technical property ??

We only used this:

$$\ell_t\big(\underbrace{\mathbb{E}_{M\sim p_t}[x_{t,M}]}_{x_t}\big) \leqslant -\frac{1}{\eta}\log \mathbb{E}_{M\sim p_t}\exp\big(-\eta\ell_t(x_{t,M})\big) + \frac{\eta}{8}$$

We only used this:

$$\ell_t\big(\underbrace{\mathbb{E}_{M\sim p_t}[x_{t,M}]}_{x_t}\big) \leqslant -\frac{1}{\eta}\log\mathbb{E}_{M\sim p_t}\exp\big(-\eta\ell_t(x_{t,M})\big) + \frac{\eta}{8}$$

- Satisfied if convex, bounded by 1.
  Ok for **quadratic** loss, pb for **self-information**: not bounded when x small!

We only used this:

$$\ell_t\big(\underbrace{\mathbb{E}_{M\sim p_t}[x_{t,M}]}_{x_t}\big) \leqslant -\frac{1}{\eta}\log\mathbb{E}_{M\sim p_t}\exp\big(-\eta\ell_t(x_{t,M})\big)+\frac{\eta}{8}$$

- ▷ Satisfied if convex, bounded by 1.
  Ok for **quadratic** loss, pb for **self-information**: not bounded when x small!
- ▷ What about dropping $\eta/8$ term?
  Equivalent to $\exp(-\eta\ell_t(\cdot))$ is concave: **η-exp-concavity**.
  - ◇ **Self-information** loss is 1-exp-concave (with $=$ instead of $\leqslant$)
  - ◇ **Quadratic** loss is η-exp-concave for $\eta \leqslant \frac{1}{2(b-a)^2}$ on $\mathcal{X}=\mathcal{Y}\subset[a,b]$.
  - ◇ **Absolute** loss $\ell(x,y)=|x-y|$ is not exp-concave for any η.

A SECOND LOOK AT ASSUMPTIONS

- Interpretation of $-\frac{1}{\eta}\log \mathbb{E}_{M\sim p_t} \exp\big(-\eta \ell_t(x_{t,M})\big)$?
  Entropy formula:

$$-\frac{1}{\eta}\log \mathbb{E}_{M\sim p} \exp\big(-\eta X_M\big) = \inf_{q\in\mathcal{P}(\mathcal{M})} \mathbb{E}_{M\sim q}[X_M] + \frac{1}{\eta}\mathtt{KL}(q,p)\,.$$

▷ Interpretation of $-\frac{1}{\eta}\log\mathbb{E}_{M\sim p_t}\exp\big(-\eta\ell_t(x_{t,M})\big)$?
Entropy formula:

$$-\frac{1}{\eta}\log\mathbb{E}_{M\sim p}\exp\big(-\eta X_M\big) = \inf_{q\in\mathcal{P}(\mathcal{M})}\mathbb{E}_{M\sim q}[X_M] + \frac{1}{\eta}\mathtt{KL}(q,p)\,.$$

▷ Hence, η-exp-concavity becomes:

η-exp-concavity

A loss ℓ is η-exp-concave if $\forall\mathbf{x}\in\mathcal{X}^{\mathcal{M}}, p\in\mathcal{P}(\mathcal{M}), \forall y\in\mathcal{Y}$,

$$\ell(\mathbb{E}_{M\sim p}[\mathbf{x}_M], y) \leqslant \inf_{q\in\mathcal{P}(\mathcal{M})}\mathbb{E}_{M\sim q}[\ell(\mathbf{x}_M, y)] + \frac{1}{\eta}\mathtt{KL}(q,p)$$

▷ Interpretation of $-\frac{1}{\eta}\log \mathbb{E}_{M\sim p_t} \exp\big(-\eta \ell_t(x_{t,M})\big)$?
Entropy formula:

$$-\frac{1}{\eta}\log \mathbb{E}_{M\sim p} \exp\big(-\eta X_M\big) = \inf_{q\in\mathcal{P}(\mathcal{M})} \mathbb{E}_{M\sim q}[X_M] + \frac{1}{\eta}\mathtt{KL}(q,p)\,.$$

▷ Hence, η-exp-concavity becomes:

η-exp-concavity

A loss ℓ is η-exp-concave if $\forall \mathbf{x}\in\mathcal{X}^{\mathcal{M}}, p\in\mathcal{P}(\mathcal{M}), \forall y\in\mathcal{Y}$,

$$\ell(\mathbb{E}_{M\sim p}[\mathbf{x}_M], y) \leqslant \inf_{q\in\mathcal{P}(\mathcal{M})} \mathbb{E}_{M\sim q}[\ell(\mathbf{x}_M, y)] + \frac{1}{\eta}\mathtt{KL}(q,p)$$

▷ Further, infimum obtained for $q(m) = \frac{\exp(-\eta X_m)p(m)}{\sum_{m'\in\mathcal{M}}\exp(-\eta X_{m'})p(m')}$.

Generalization: we don't need that $x_t = \mathbb{E}_{M\sim p_t}[x_{t,M}]$.

η-mixability

A loss ℓ is η-mixable if $\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \forall y \in \mathcal{Y}$,

$$\ell(x_{\mathbf{x},\mathbf{p}}, y) \leqslant \inf_{q\in\mathcal{P}(\mathcal{M})} \mathbb{E}_{M\sim q}[\ell(\mathbf{x}_M, y)] + \frac{1}{\eta}\mathtt{KL}(q,p)$$

$[\mathbf{x}], \mathbf{p} \mapsto \mathbf{x}_{\mathbf{x},\mathbf{p}}$ is called the **substitution function**.

Generalization: we don't need that $x_t = \mathbb{E}_{M\sim p_t}[x_{t,M}]$.

η-mixability

A loss ℓ is η-mixable if $\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \forall y \in \mathcal{Y}$,

$$\ell(x_{\mathbf{x},\mathbf{p}}, y) \leqslant \inf_{q\in\mathcal{P}(\mathcal{M})} \mathbb{E}_{M\sim q}[\ell(\mathbf{x}_M, y)] + \frac{1}{\eta}\mathtt{KL}(q,p)$$

$[\mathbf{x}], \mathbf{p} \mapsto \mathbf{x}_{\mathbf{x},\mathbf{p}}$ is called the **substitution function**.

- η-exp-concave loss is η-mixable with $x_{\mathbf{x},\mathbf{p}} = \mathbb{E}_{M\sim p}\mathbf{x}_{\mathbf{M}}$.
- **Quadratic** loss is η-exp-concave for $\eta \leqslant \frac{1}{2}$ on $\mathcal{X} = \mathcal{Y} \subset [0,1]$, but η-mixable for η up to $\eta \leqslant 2$!

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.
- At time $t+1$, given $\mathbf{x}_t \in \mathcal{X}^{\mathcal{M}}$, and $p_t \in \mathcal{P}(\mathcal{M})$, output decision $x_t = x_{\mathbf{x}_t, p_t}$,

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.
- At time $t+1$, given $\mathbf{x}_t \in \mathcal{X}^{\mathcal{M}}$, and $p_t \in \mathcal{P}(\mathcal{M})$, output decision $x_t = x_{\mathbf{x}_t, p_t}$,
- Receive y_t and update

$$p_{t+1} = \operatorname*{argmin}_{q \in \mathcal{P}_M} \mathbb{E}_{M \sim q}[\underbrace{\ell(\mathbf{x}_{t,M}, y_t)}_{\ell_{t,M}}] + \frac{1}{\eta}\mathtt{KL}(q, p_t).$$

Theorem

Assume that ℓ_t is η-**mixable**, then after T time steps, this strategy satisfies:

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \frac{\log(|\mathcal{M}|)}{\eta}.$$

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.
- At time $t+1$, given $\mathbf{x}_t \in \mathcal{X}^{\mathcal{M}}$, and $p_t \in \mathcal{P}(\mathcal{M})$, output decision $x_t = x_{\mathbf{x}_t, p_t}$,
- Receive y_t and update

$$p_{t+1} = \operatorname*{argmin}_{q \in \mathcal{P}_M} \mathbb{E}_{M\sim q}[\underbrace{\ell(\mathbf{x}_{t,M}, y_t)}_{\ell_{t,M}}] + \frac{1}{\eta}\mathtt{KL}(q, p_t).$$

Theorem

Assume that ℓ_t is η-**mixable**, then after T time steps, this strategy satisfies:

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \frac{\log(|\mathcal{M}|)}{\eta} \,.$$

- Still for arbitrary $y_t \in \mathcal{Y}$.

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.
- At time $t+1$, given $\mathbf{x}_t \in \mathcal{X}^{\mathcal{M}}$, and $p_t \in \mathcal{P}(\mathcal{M})$, output decision $x_t = x_{\mathbf{x}_t, p_t}$,
- Receive y_t and update

$$p_{t+1} = \operatorname*{argmin}_{q \in \mathcal{P}_M} \mathbb{E}_{M\sim q}[\underbrace{\ell(\mathbf{x}_{t,M}, y_t)}_{\ell_{t,M}}] + \frac{1}{\eta}\mathtt{KL}(q, p_t).$$

Theorem

Assume that ℓ_t is η-**mixable**, then after T time steps, this strategy satisfies:

$$L_T - \min_{m\in\mathcal{M}} L_{T,m} \leqslant \frac{\log(|\mathcal{M}|)}{\eta}\,.$$

- Still for arbitrary $y_t \in \mathcal{Y}$.
- **Independent** on T !

Aggregation of experts revisited

- Consider an η-**mixable** loss ℓ, and let $p_1 = \text{Uniform}(\mathcal{M}) \in \mathcal{P}(\mathcal{M})$.
- At time $t+1$, given $\mathbf{x}_t \in \mathcal{X}^{\mathcal{M}}$, and $p_t \in \mathcal{P}(\mathcal{M})$, output decision $x_t = x_{\mathbf{x}_t, p_t}$,
- Receive y_t and update

$$p_{t+1} = \operatorname*{argmin}_{q \in \mathcal{P}_M} \mathbb{E}_{M \sim q}[\underbrace{\ell(\mathbf{x}_{t,M}, y_t)}_{\ell_{t,M}}] + \frac{1}{\eta}\mathtt{KL}(q, p_t).$$

Theorem

Assume that ℓ_t is η-**mixable**, then after T time steps, this strategy satisfies:

$$L_T - \min_{m \in \mathcal{M}} L_{T,m} \leqslant \frac{\log(|\mathcal{M}|)}{\eta}.$$

- Still for arbitrary $y_t \in \mathcal{Y}$.
- **Independent** on T !
- Only for **specific**, possibly small η (all $\eta' \leqslant \eta$, but not larger).

We can actually get a stronger result:

Theorem (Aggregation of experts)

Assume that ℓ_t is η-mixable, then after T time steps, the aggregation strategy with $p_1 = \pi$, satifies

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \mathbb{E}_{M\sim q}\Big[L_{T,M}\Big] \leqslant \frac{1}{\eta}\Big(\mathtt{KL}(q,\pi) - \mathtt{KL}(q, p_{T+1})\Big)\,.$$

We can actually get a stronger result:

Theorem (Aggregation of experts)

Assume that ℓ_t is η-mixable, then after T time steps, the aggregation strategy with $p_1 = \pi$, satifies

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \mathbb{E}_{M\sim q}\Big[L_{T,M}\Big] \leqslant \frac{1}{\eta}\Big(\mathtt{KL}(q,\pi) - \mathtt{KL}(q, p_{T+1})\Big)\,.$$

▷ Now, we compete against **convex combination** of loss of experts!

We can actually get a stronger result:

Theorem (Aggregation of experts)

Assume that ℓ_t is η-mixable, then after T time steps, the aggregation strategy with $p_1 = \pi$, satifies

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \mathbb{E}_{M\sim q}\Big[L_{T,M}\Big] \leqslant \frac{1}{\eta}\Big(\mathtt{KL}(q,\pi) - \mathtt{KL}(q,p_{T+1})\Big)\,.$$

- Now, we compete against **convex combination** of loss of experts!
- In particular for $q = \delta_{m^\star}$ (Dirac mass as $m^\star$), we deduce

$$L_T - L_{T,m^\star} \leqslant \frac{1}{\eta}\log\left(\frac{1}{\pi(m^\star)}\right)\,.$$

We can actually get a stronger result:

Theorem (Aggregation of experts)

Assume that ℓ_t is η-mixable, then after T time steps, the aggregation strategy with $p_1 = \pi$, satifies

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \mathbb{E}_{M\sim q}\Big[L_{T,M}\Big] \leqslant \frac{1}{\eta}\Big(\mathtt{KL}(q,\pi) - \mathtt{KL}(q, p_{T+1})\Big).$$

- Now, we compete against **convex combination** of loss of experts!
- In particular for $q = \delta_{m^\star}$ (Dirac mass as $m^\star$), we deduce

$$L_T - L_{T,m^\star} \leqslant \frac{1}{\eta}\log\left(\frac{1}{\pi(m^\star)}\right).$$

- We can move from finitely many to **countably** many experts:

$$\pi(m) = \frac{1}{m(m+1)}, \qquad \pi(m) = \log(2)\left(\frac{1}{\log(m+1)} - \frac{1}{\log(m+2)}\right).$$

▷ **Bregman** divergence generalizes KL:

$$\mathcal{B}(p, q) = \psi(p) - \psi(q) - \langle p - q, \nabla\psi(q)\rangle$$

($\psi(p) = \sum_i p_i \log(p_i)$ gives KL as a special case.)

▷ Assumption: ℓ is **η-Bregman-mixable** w.r.t. Bregman divergence $\mathcal{B}$:

$$\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \in \mathcal{X},\ \ell(x_{\mathbf{x},\mathbf{p}}) \leqslant \min_{q \in \mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x}} \rangle + \frac{1}{\eta}\mathcal{B}(q, p).$$

where $\ell_{\mathbf{x}}$ denotes the vector $(\ell(x_1), \ldots, \ell(x_M))$.

- **Bregman** divergence generalizes KL:

$$\mathcal{B}(p,q) = \psi(p) - \psi(q) - \langle p - q, \nabla\psi(q)\rangle$$

($\psi(p) = \sum_i p_i \log(p_i)$ gives KL as a special case.)

- Assumption: ℓ is η-**Bregman-mixable** w.r.t. Bregman divergence $\mathcal{B}$:

$$\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \in \mathcal{X},\ \ell(x_{\mathbf{x},\mathbf{p}}) \leqslant \min_{q\in\mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p).$$

where $\ell_{\mathbf{x}}$ denotes the vector $(\ell(x_1),\dots,\ell(x_M))$.

- **Strategy**: Play $x_{\mathbf{x_t},\mathbf{p_t}}$, update $p_{t+1} = \mathop{\mathrm{argmin}}_{q\in\mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x_t}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p_t)$.

Bregman aggregation

- **Bregman** divergence generalizes KL:

$$\mathcal{B}(p,q) = \psi(p) - \psi(q) - \langle p - q, \nabla\psi(q)\rangle$$

($\psi(p) = \sum_i p_i \log(p_i)$ gives KL as a special case.)

- Assumption: ℓ is **η-Bregman-mixable** w.r.t. Bregman divergence $\mathcal{B}$:

$$\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \in \mathcal{X},\ \ell(x_{\mathbf{x},\mathbf{p}}) \leqslant \min_{q \in \mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p).$$

where $\ell_{\mathbf{x}}$ denotes the vector $(\ell(x_1), \ldots, \ell(x_M))$.

- **Strategy**: Play $x_{\mathbf{x_t},\mathbf{p_t}}$, update $p_{t+1} = \operatorname{argmin}_{q \in \mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x_t}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p_t)$.
- **Performance**:

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \langle q, \mathbf{L}_T\rangle \leqslant \frac{1}{\eta}\Big(\mathcal{B}(q,\pi) - \mathcal{B}(q,p_{T+1})\Big).$$

- **Bregman** divergence generalizes KL:

$$\mathcal{B}(p,q) = \psi(p) - \psi(q) - \langle p-q, \nabla\psi(q)\rangle$$

($\psi(p) = \sum_i p_i \log(p_i)$ gives KL as a special case.)

- Assumption: ℓ is **η-Bregman-mixable** w.r.t. Bregman divergence $\mathcal{B}$:

$$\forall \mathbf{x} \in \mathcal{X}^{\mathcal{M}}, p \in \mathcal{P}(\mathcal{M}), \exists x_{\mathbf{x},\mathbf{p}} \in \mathcal{X},\ \ell(x_{\mathbf{x},\mathbf{p}}) \leqslant \min_{q\in\mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p).$$

where $\ell_{\mathbf{x}}$ denotes the vector $(\ell(x_1),\dots,\ell(x_M))$.

- **Strategy**: Play $x_{\mathbf{x_t},\mathbf{p_t}}$, update $p_{t+1} = \operatorname{argmin}_{q\in\mathcal{P}(\mathcal{M})} \langle q, \ell_{\mathbf{x_t}}\rangle + \frac{1}{\eta}\mathcal{B}(q,p_t)$.

- **Performance**:

$$\forall q \in \mathcal{P}(\mathcal{M}) \quad L_T - \langle q, \mathbf{L}_T\rangle \leqslant \frac{1}{\eta}\Big(\mathcal{B}(q,\pi) - \mathcal{B}(q,p_{T+1})\Big).$$

- Other interpretation: Use Legendre-Fenchel dual objective function, perform gradient descent!

When the best expert has **small loss**, we may prefer to express regret bounds on terms of this loss:

▷ Consider a loss **convex and bounded** in $[0,1]$, then:

$$L_T - L_T^\star \leqslant \left(\frac{\eta}{1-\exp(-\eta)} - 1\right) L_T^\star + \frac{\log(M)}{1-\exp(-\eta)}$$

where $L_T^\star = \min_{m\in\mathcal{M}} L_{t,m}$

Proof: One can show that any loss ℓ convex and bounded in $[0,1]$ satisfies the following extension of η-mixability property:

$$\ell(\mathbb{E}_{M\sim q}(x_M)) \leqslant -\frac{\eta}{1-\exp(-\eta)}\frac{1}{\eta}\ln\left(\mathbb{E}_{m\sim q}\exp(-\eta\ell(x_M))\right).$$

(almost η-mixable!) The rest is obtained by following the initial derivation.

Different objectives

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- best **combination** of models (Model aggregation)?

$$\inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \Big(\sum_{t=1}^{T} \ell_t(x_{t,m}) \Big) \quad \text{or} \quad \inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \Big(\sum_{m \in \mathcal{M}} q_m x_{t,m} \Big)$$

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- best **combination** of models (Model aggregation)?

$$\inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m \in \mathcal{M}} q_m x_{t,m} \right)$$

- Left: best **combination of losses** Right: **loss of best combination**.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- best **combination** of models (Model aggregation)?

$$\inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{m\in\mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m\in\mathcal{M}} q_m x_{t,m} \right)$$

- Left: best **combination of losses** Right: **loss of best combination**.
- Right is **harder**: $\ell_t(\mathbf{q}\cdot\mathbf{x}_t) \leqslant \mathbf{q}\cdot\boldsymbol{\ell}_t$ by convexity.

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \ldots$$

w.r.t.

- ▷ best **combination** of models (Model aggregation)?

$$\inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{m \in \mathcal{M}} q_m \left(\sum_{t=1}^{T} \ell_t(x_{t,m}) \right) \quad \text{or} \quad \inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t \left(\sum_{m \in \mathcal{M}} q_m x_{t,m} \right)$$

- ▷ Left: best **combination of losses** Right: **loss of best combination**.
- ▷ Right is **harder**: $\ell_t(\mathbf{q} \cdot \mathbf{x}_t) \leqslant \mathbf{q} \cdot \boldsymbol{\ell}_t$ by convexity.
- ▷ ! From set of experts $\mathcal{M}$ (finite) to set of experts $\mathcal{P}(\mathcal{M})$ (continuous) !

$$\text{Minimize} \quad \sum_{t=1}^{T} \ell_t(x_t) \ \dots$$

w.r.t.

- best **combination** of models (Model aggregation)?

$$\inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{m\in\mathcal{M}} q_m \bigg(\sum_{t=1}^{T} \ell_t(x_{t,m})\bigg) \quad \text{or} \quad \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t\bigg(\sum_{m\in\mathcal{M}} q_m x_{t,m}\bigg)$$

- Left: best **combination of losses** Right: **loss of best combination**.
- Right is **harder**: $\ell_t(\mathbf{q}\cdot\mathbf{x}_t) \leqslant \mathbf{q}\cdot\boldsymbol{\ell}_t$ by convexity.
- ! From set of experts $\mathcal{M}$ (finite) to set of experts $\mathcal{P}(\mathcal{M})$ (continuous) !
- If ℓ is η-exp-concave on $\mathcal{X}$, then $\overline{\ell} : q \to \ell_t(\mathbf{q}\cdot\mathbf{x}_t)$ is η-exp-concave on $\mathcal{P}(\mathcal{M})$.

- $\overline{p}_1(q) = \frac{1}{\mathrm{vol}(\mathcal{P}(\mathcal{M}))} = M!$, $p_1 = \frac{1}{|\mathcal{M}|}\mathbf{1}$.

Aggregation over $\mathcal{P}(\mathcal{M})$: Strategy

- $\overline{p}_1(q) = \frac{1}{\mathrm{vol}(\mathcal{P}(\mathcal{M}))} = M!$, $p_1 = \frac{1}{|\mathcal{M}|}\mathbf{1}$.
- For given$(x_{t,m})_{m\in\mathcal{M}}$, choose $x_t = \sum_{m\in\mathcal{M}} p_t(m)x_{t,m}$, where $p_t = \mathbb{E}_{q\sim\overline{p}_t}[q]$.

- $\overline{p}_1(q) = \frac{1}{\mathrm{vol}(\mathcal{P}(\mathcal{M}))} = M!$, $p_1 = \frac{1}{|\mathcal{M}|}\mathbf{1}$.
- For given$(x_{t,m})_{m\in\mathcal{M}}$, choose $x_t = \sum_{m\in\mathcal{M}} p_t(m)x_{t,m}$, where $p_t = \mathbb{E}_{q\sim\overline{p}_t}[q]$.
- Get the next observation y_t from which we can compute $\ell_t(\mathbf{q}\cdot\mathbf{x}_t)$ for all $\mathbf{q}$.

$$\overline{p}_{t+1}(q) = \frac{\overline{p}_t(q)\exp(-\eta\overline{\ell}_t(q))}{\int_{\mathcal{P}(\mathcal{M})}\overline{p}_t(u)\exp(-\eta\overline{\ell}_t(q))du}$$

- $\overline{p}_1(q) = \frac{1}{\mathrm{vol}(\mathcal{P}(\mathcal{M}))} = M!$, $p_1 = \frac{1}{|\mathcal{M}|}\mathbf{1}$.
- For given$(x_{t,m})_{m\in\mathcal{M}}$, choose $x_t = \sum_{m\in\mathcal{M}} p_t(m)x_{t,m}$, where $p_t = \mathbb{E}_{q\sim\overline{p}_t}[q]$.
- Get the next observation y_t from which we can compute $\ell_t(\mathbf{q}\cdot\mathbf{x}_t)$ for all $\mathbf{q}$.

$$\overline{p}_{t+1}(q) = \frac{\overline{p}_t(q)\exp(-\eta\overline{\ell}_t(q))}{\int_{\mathcal{P}(\mathcal{M})}\overline{p}_t(u)\exp(-\eta\overline{\ell}_t(q))du}$$

- Update $p_{t+1} = \mathbb{E}_{q\sim\overline{p}_{t+1}}[q]$.

$$L_T - \inf_{q \in \mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \overline{\ell}_t(q) \leqslant \frac{M}{\eta}\left(1 + \log\left(1 + \frac{T}{M}\right)\right).$$

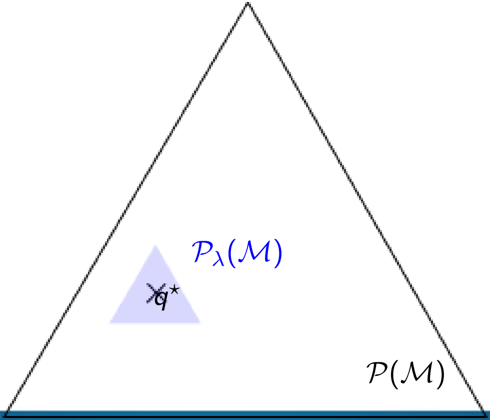
$$L_T - \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \overline{\ell}_t(q) \leqslant \frac{M}{\eta}\Big(1+\log\Big(1+\frac{T}{M}\Big)\Big).$$

- For comparison we had: $L_T - \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_m q(m) L_{T,m} \leqslant \frac{\log(M)}{\eta}$.

Aggregation over $\mathcal{P}(\mathcal{M})$:Performance

$$L_T - \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \overline{\ell}_t(q) \leqslant \frac{M}{\eta}\Big(1+\log\Big(1+\frac{T}{M}\Big)\Big).$$

- For comparison we had: $L_T - \inf_{q\in\mathcal{P}(\mathcal{M})} \sum_m q(m) L_{T,m} \leqslant \frac{\log(M)}{\eta}$.
- Proof technique: Similar +



- Consider Binary prediction and self-information loss ℓ.

- Consider Binary prediction and self-information loss ℓ.
- ▷ Aggregation over all Bernoulli $\mathcal{B}(\theta)$, $\theta \in [0,1]$.

- Consider Binary prediction and self-information loss ℓ.
  - Aggregation over all Bernoulli $\mathcal{B}(\theta)$, $\theta \in [0, 1]$.
  - KT-predictor: Use prior $g(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}$ on each parameter.

Example of Universal prediction

- Consider Binary prediction and self-information loss ℓ.
  - Aggregation over all Bernoulli $\mathcal{B}(\theta)$, $\theta \in [0, 1]$.
  - KT-predictor: Use prior $g(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}$ on each parameter.
  - Yields a **fully explicit** solution:

$$q_t(1) = \frac{t\widehat{\theta}_t + 1/2}{t + 1}$$

Efficient computation despite aggregation of continuum of models.

Example of Universal prediction

- Consider Binary prediction and self-information loss ℓ.
- Aggregation over all Bernoulli $\mathcal{B}(\theta)$, $\theta \in [0,1]$.
- KT-predictor: Use prior $g(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}$ on each parameter.
- Yields a **fully explicit** solution:

$$q_t(1) = \frac{t\widehat{\theta}_t + 1/2}{t+1}$$

Efficient computation despite aggregation of continuum of models.

- Called "Universal prediction". Extends to all Markov models of arbitrary order.

- So far, we only considered **fixed** experts:

$$\min_{m\in\mathcal{M}}\sum_{t=1}^{T}\ell_t(x_{t,m}),\qquad \min_{q\in\mathcal{P}(\mathcal{M})}\sum_{m\in\mathcal{M}}q(m)L_{T,m}\qquad \min_{q\in\mathcal{P}(\mathcal{M})}\sum_{t=1}^{T}\ell_t(\sum_{m\in\mathcal{M}}q(m)x_{t,m})$$

- So far, we only considered **fixed** experts:

$$\min_{m\in\mathcal{M}} \sum_{t=1}^{T} \ell_t(x_{t,m}), \qquad \min_{q\in\mathcal{P}(\mathcal{M})} \sum_{m\in\mathcal{M}} q(m)L_{T,m} \qquad \min_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^{T} \ell_t\Big(\sum_{m\in\mathcal{M}} q(m)x_{t,m}\Big)$$

- What about best **sequence** of experts:

$$\min_{m_1,\dots,m_T\in\mathcal{S}_k(\mathcal{M})} \sum_{t=1}^{T} \ell_t(x_{t,m_t}) \text{ where } \mathcal{S}_k(\mathcal{M}) : \text{at most } k \text{ switches.}$$

  - Difficulty: Concentrating mass **exponentially fast** to a single expert means putting near 0 on others.
  - When switching to other best expert, **need to catch-up**!
  - from $\mathcal{M}$ to $\mathcal{M}^T$ many experts??

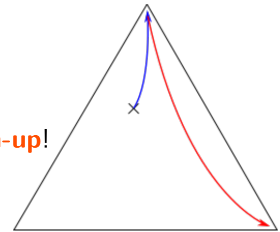
- So far, we only considered **fixed** experts:

$$\min_{m\in\mathcal{M}} \sum_{t=1}^T \ell_t(x_{t,m}), \qquad \min_{q\in\mathcal{P}(\mathcal{M})} \sum_{m\in\mathcal{M}} q(m)L_{T,m} \qquad \min_{q\in\mathcal{P}(\mathcal{M})} \sum_{t=1}^T \ell_t\Big(\sum_{m\in\mathcal{M}} q(m)x_{t,m}\Big)$$

- What about best **sequence** of experts:

$$\min_{m_1,\dots,m_T\in\mathcal{S}_k(\mathcal{M})} \sum_{t=1}^T \ell_t(x_{t,m_t}) \text{ where } \mathcal{S}_k(\mathcal{M}) : \text{at most } k \text{ switches.}$$

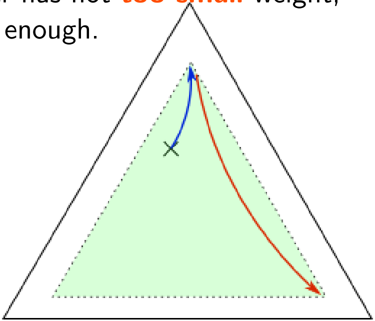
  - Difficulty: Concentrating mass **exponentially fast** to a single expert means putting near 0 on others.
  - When switching to other best expert, **need to catch-up**!
  - from $\mathcal{M}$ to $\mathcal{M}^T$ many experts??



Fixed-share solution

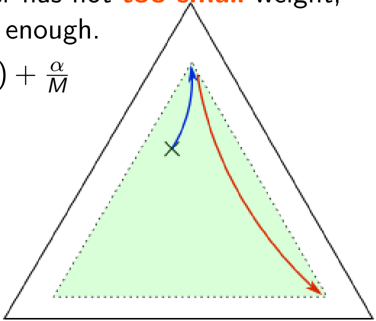
Fixed-share solution

▷ Guarantees each m never has not **too small** weight, hence can catch-up fast enough.



Fixed-share solution

- Guarantees each m never has not **too small** weight, hence can catch-up fast enough.
- $\tilde{p}_{t+1}(\cdot) = (1-\alpha)p_{t+1}(\cdot) + \frac{\alpha}{M}$



Fixed-share performance

For all sequence $q_1, \dots, q_T \in \mathcal{P}(\mathcal{M})$ with at most k switches,

$$L_T - \sum_{t=1}^{T} q_t \ell_t \leqslant \frac{\log(M)}{\eta} + \frac{k}{\eta} \log\left(\frac{M}{\alpha}\right) + \frac{T-k-1}{\eta} \log\left(\frac{1}{1-\alpha}\right).$$

For all sequence $q_1, \dots, q_T \in \mathcal{P}(\mathcal{M})$ with at most k switches,

$$L_T - \sum_{t=1}^{T} q_t \ell_t \leqslant \frac{\log(M)}{\eta} + \frac{k}{\eta} \log\left(\frac{M}{\alpha}\right) + \frac{T-k-1}{\eta} \log\left(\frac{1}{1-\alpha}\right).$$

- Choosing $\alpha = k/(T-1)$ yields

$$L_T - \sum_{t=1}^{T} q_t \ell_t \leqslant \frac{\log(M)}{\eta} + \frac{k}{\eta} \log\left(\frac{M(T-1)}{k}\right) + \frac{k}{\eta}.$$

For all sequence $q_1, \dots, q_T \in \mathcal{P}(\mathcal{M})$ with at most k switches,

$$L_T - \sum_{t=1}^{T} q_t \ell_t \leqslant \frac{\log(M)}{\eta} + \frac{k}{\eta} \log\left(\frac{M}{\alpha}\right) + \frac{T-k-1}{\eta} \log\left(\frac{1}{1-\alpha}\right).$$

- Choosing $\alpha = k/(T-1)$ yields

$$L_T - \sum_{t=1}^{T} q_t \ell_t \leqslant \frac{\log(M)}{\eta} + \frac{k}{\eta} \log\left(\frac{M(T-1)}{k}\right) + \frac{k}{\eta}.$$

- α going to 0 but not exponentially fast.

Let us consider $\tilde{p}_t$ obtained from p_t as $\tilde{p}_{t+1}(\cdot) = \sum_{m' \in \mathcal{M}} \theta(\cdot|m') p_{t+1}(m')$, from a Markov chain with initial low ω and **transition matrix** θ.
For all sequence $m_1, \dots, m_T \in \mathcal{M}$ with at most k switches

$$L_T - \sum_{t=1}^{T} \ell_{t,m_t} \leqslant \frac{1}{\eta} \log\left(\frac{1}{\omega(m_1)}\right) + \frac{1}{\eta} \sum_{t=2}^{T} \log\left(\frac{1}{\theta_t(m_t|m_{t-1})}\right).$$

Let us consider $\tilde{p}_t$ obtained from p_t as $\tilde{p}_{t+1}(\cdot) = \sum_{m' \in \mathcal{M}} \theta(\cdot|m')p_{t+1}(m')$, from a Markov chain with initial low ω and **transition matrix** θ.
For all sequence $m_1, \dots, m_T \in \mathcal{M}$ with at most k switches

$$L_T - \sum_{t=1}^{T} \ell_{t,m_t} \leqslant \frac{1}{\eta} \log\left(\frac{1}{\omega(m_1)}\right) + \frac{1}{\eta} \sum_{t=2}^{T} \log\left(\frac{1}{\theta_t(m_t|m_{t-1})}\right).$$

- Fixed share: $\theta(m'|m) = (1-\alpha)\mathbb{I}\{m = m'\} + \alpha/M$.

Let us consider $\tilde{p}_t$ obtained from p_t as $\tilde{p}_{t+1}(\cdot) = \sum_{m' \in \mathcal{M}} \theta(\cdot|m')p_{t+1}(m')$, from a Markov chain with initial low ω and **transition matrix** θ.
For all sequence $m_1, \dots, m_T \in \mathcal{M}$ with at most k switches

$$L_T - \sum_{t=1}^{T} \ell_{t,m_t} \leqslant \frac{1}{\eta} \log\left(\frac{1}{\omega(m_1)}\right) + \frac{1}{\eta}\sum_{t=2}^{T} \log\left(\frac{1}{\theta_t(m_t|m_{t-1})}\right).$$

- Fixed share: $\theta(m'|m) = (1-\alpha)\mathbb{I}\{m = m'\} + \alpha/M$.
- Variable share, sleeping experts, etc.

Note: even though huge amount of experts $O(M^T)$ they share a **rich structure**.
This enables to have an efficient strategy maintaining only few quantities $O(MT)$.

- Best **sequence** of experts:

$$\min_{m_1,\dots,m_T \in \mathcal{S}_k(\mathcal{M})} \sum_{t=1}^{T} \ell_t(x_{t,m_t}) \text{ where } \mathcal{S}_k(\mathcal{M}) : \text{at most } k \text{ switches.}$$

- Best **sequence** of experts:

$$\min_{m_1,\dots,m_T \in \mathcal{S}_k(\mathcal{M})} \sum_{t=1}^{T} \ell_t(x_{t,m_t}) \text{ where } \mathcal{S}_k(\mathcal{M}) : \text{at most } k \text{ switches.}$$

- Best sequence of experts with **few good** experts:

$$\min_{m_1,\dots,m_T \in \mathcal{S}_k(\mathcal{M}_0)} \sum_{t=1}^{T} \ell_t(x_{t,m_t}) \text{ where } \mathcal{M}_0 \subset \mathcal{M} \text{ unknown but small}\,.$$

  - Intuition: the good experts should be good in the recent past.

- Ensure that experts good in the recent past have large enough weight and catch-up.

- Ensure that experts good in the recent past have large enough weight and catch-up.
- **Mixing past posterior** $\tilde{p}_{t+1}(\cdot) = \sum_{s=0}^{t} \beta_{t+1}(s) p_s(\cdot)$

- Ensure that experts good in the recent past have large enough weight and catch-up.
- **Mixing past posterior** $\tilde{p}_{t+1}(\cdot) = \sum_{s=0}^{t} \beta_{t+1}(s) p_s(\cdot)$
- In particular:
  - Hedge: $\beta_{t+1}(t') = \begin{cases} 1 & \text{if } t' = t \\ 0 & \text{else} \end{cases}$
  - Fixed share: $\beta_{t+1}(t') = \begin{cases} 1-\alpha & \text{if } t' = t \\ \alpha & \text{if } t' = 0 \\ 0 & \text{else} \end{cases}$
  - ...

Assume ℓ is η-mixable. For all sequence $(q_t)_{t\in\mathcal{T}}$ with k switches between at most n values,

$$L_T - \sum_{t=1}^{T} q_t \cdot \ell_t \leqslant \frac{n}{\eta}\log\left(|\mathcal{M}|\right) + \frac{1}{\eta}\sum_{t=1}^{T}\log\left(\frac{1}{\beta_t(\tau_t)}\right).$$

where τ_t is last $\tau < t$ such that $q_\tau = q_t$ (or 0 if first occurrence).

- **Sleeping experts** (Koolen et al. 2012): When experts are not available at all rounds.

- **Sleeping experts** (Koolen et al. 2012): When experts are not available at all rounds.
- **Growing experts** (Mourtada&M. 2017): When set of base experts $\mathcal{M}$ is no longer fixed but may increase with time; Especially useful to handle **non-stationarity**.

- **Sleeping experts** (Koolen et al. 2012): When experts are not available at all rounds.
- **Growing experts** (Mourtada&M. 2017): When set of base experts $\mathcal{M}$ is no longer fixed but may increase with time; Especially useful to handle **non-stationarity**.
- ...

Most results are minimax-optimal, valid for any input sequence.
This contrasts with typical results for bandits: instance-optimal, for stochastic sequence.

Full information

- Powerful: Handle large number of experts

Full information

- Powerful: Handle large number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)

Full information

- Powerful: Handle large number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts

Full information

- Powerful: Handle large number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Table of contents

Adjusting for the differences:

Adjusting for the differences:

- Decision are arms $\mathcal{X} = \mathcal{A}$. Consider one expert per arm $\mathcal{M} = \mathcal{A}$.

Adjusting for the differences:

- Decision are arms $\mathcal{X} = \mathcal{A}$. Consider one expert per arm $\mathcal{M} = \mathcal{A}$.
- Losses $(\ell_{t,m})_{m\in\mathcal{M}}$ become rewards $(r_{t,a})_{a\in\mathcal{A}}$

Adjusting for the differences:

- Decision are arms $\mathcal{X} = \mathcal{A}$. Consider one expert per arm $\mathcal{M} = \mathcal{A}$.
- Losses $(\ell_{t,m})_{m\in\mathcal{M}}$ become rewards $(r_{t,a})_{a\in\mathcal{A}}$
- Can only output an arm $A_t \in \mathcal{A}$ (not a combination):
  $x_t = \sum_{m\in\mathcal{M}} p_{t,m} x_{t,m}$ becomes $x_t = x_{t,m_t}$ with $m_t \sim p_t$.
  - Less good, but ok as long as $\mathbb{E}$ performance.

Problem: we only observe the reward of A_t (i.e., only r_{t,A_t}) !!
Partial information: We don't observe $r_{t,a}$ for all arms.

Terminology: Adversarial setup. We want guarantees against arbitrary (bounded) sequence of rewards/losses.

The Exponentially Weighted Average Forecaster

- Output $m_t \sim p_t$ where $p_t(m) = \frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}$,
  - $\forall m \in \mathcal{M}, w_1(m) = 1$ and $w_{t+1}(m) = w_t(m)\exp(-\eta\ell_{t,m})$.

$\ell_{t,m}$ is not available for all arms!

$\ell_{t,m} = 1 - r_{t,a}$?

We can use **importance sampling**

$$\widehat{\ell}_{t,m} = \begin{cases} \frac{\ell_{t,m}}{p_t(m)} & \text{if } m = m_t \\ 0 & \text{otherwise} \end{cases}$$

We can use **importance sampling**

$$\widehat{\ell}_{t,m} = \begin{cases} \frac{\ell_{t,m}}{p_t(m)} & \text{if } m = m_t \\ 0 & \text{otherwise} \end{cases}$$

Why it is a good idea:

We can use **importance sampling**

$$\widehat{\ell}_{t,m} = \begin{cases} \frac{\ell_{t,m}}{p_t(m)} & \text{if } m = m_t \\ 0 & \text{otherwise} \end{cases}$$

Why it is a good idea:

- $\widehat{\ell}_{t,m}$ is an **unbiased** estimator of $\ell_{t,m}$:

$$\mathbb{E}[\widehat{\ell}_{t,m}] = \frac{\ell_{t,m}}{p_t(m)} p_t(m) + 0(1 - p_t(m)) = \ell_{t,m}$$

We can use **importance sampling**

$$\widehat{\ell}_{t,m} = \begin{cases} \frac{\ell_{t,m}}{p_t(m)} & \text{if } m = m_t \\ 0 & \text{otherwise} \end{cases}$$

Why it is a good idea:

▷ $\widehat{\ell}_{t,m}$ is an **unbiased** estimator of $\ell_{t,m}$:

$$\mathbb{E}[\widehat{\ell}_{t,m}] = \frac{\ell_{t,m}}{p_t(m)} p_t(m) + 0(1 - p_t(m)) = \ell_{t,m}$$

Why it may be a bad idea:

We can use **importance sampling**

$$\widehat{\ell}_{t,m} = \begin{cases} \frac{\ell_{t,m}}{p_t(m)} & \text{if } m = m_t \\ 0 & \text{otherwise} \end{cases}$$

Why it is a good idea:

- $\widehat{\ell}_{t,m}$ is an **unbiased** estimator of $\ell_{t,m}$:

$$\mathbb{E}[\widehat{\ell}_{t,m}] = \frac{\ell_{t,m}}{p_t(m)} p_t(m) + 0(1 - p_t(m)) = \ell_{t,m}$$

Why it may be a bad idea:

- $p_{t,m}$ typically small for bad arms, hence this estimates has large variance for bad arms!

Exp3: Exponential-weight algorithm for Exploration and Exploitation

Exp3: Exponential-weight algorithm for Exploration and Exploitation

▷ $\forall m \in \mathcal{M}, w_1(m) = 1.$

Exp3: Exponential-weight algorithm for Exploration and Exploitation

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $m_t \sim p_t$ where $p_t(m) = \dfrac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}$

Exp3: Exponential-weight algorithm for Exploration and Exploitation

- ▷ $\forall m \in \mathcal{M}, w_1(m) = 1.$
- ▷ Output $m_t \sim p_t$ where $p_t(m) = \dfrac{w_t(m)}{\sum_{m \in \mathcal{M}} w_t(m)}$
- ▷ Receive r_{t,m_t}

Exp3: Exponential-weight algorithm for Exploration and Exploitation

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $m_t \sim p_t$ where $p_t(m) = \dfrac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)}$
- Receive r_{t,m_t}
- Update $\forall m \in \mathcal{M}, w_{t+1}(m) = w_t(m)\exp(-\eta \widehat{\ell}_{t,m}).$

Question: is this enough? is this algorithm actually exploring enough?

Question: is this enough? is this algorithm actually exploring enough?
Answer: more or less...

Question: is this enough? is this algorithm actually exploring enough?
Answer: more or less...

- ▷ Exp3 has a small regret **in expectation**

Question: is this enough? is this algorithm actually exploring enough?
Answer: more or less...

- Exp3 has a small regret **in expectation**
- Exp3 might have large deviations with **high probability** (ie, from time to time it may **concentrate $\widehat{p}_t$ on the wrong arm** for too long and then incur a large regret)

Fix: add some extra uniform exploration

Fix: add some extra uniform exploration

▷ $\forall m \in \mathcal{M}, w_1(m) = 1.$

Fix: add some extra uniform exploration

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $m_t \sim p_t$ where $p_t(m) = (1-\gamma)\frac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)} + \frac{\gamma}{|\mathcal{M}|}$

Fix: add some extra uniform exploration

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $m_t \sim p_t$ where $p_t(m) = (1-\gamma)\dfrac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)} + \dfrac{\gamma}{|\mathcal{M}|}$
- Receive r_{t,m_t}

Fix: add some extra uniform exploration

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $m_t \sim p_t$ where $p_t(m) = (1-\gamma)\dfrac{w_t(m)}{\sum_{m\in\mathcal{M}} w_t(m)} + \dfrac{\gamma}{|\mathcal{M}|}$
- Receive r_{t,m_t}
- Update $\forall m \in \mathcal{M}, w_{t+1}(m) = w_t(m)\exp(-\eta\widehat{\ell}_{t,m})$.

Theorem

If Exp3 is run with $\gamma = \eta$, then it achieves a regret

$$R_T = \max_{a\in\mathcal{A}} \sum_{t=1}^{T} r_{t,a} - \mathbb{E}\Big[\sum_{t=1}^{T} r_{t,A_t}\Big] \leqslant (e-1)\gamma G_{\max} + \frac{A\log A}{\gamma}$$

with $G_{\max} = \max_{a\in\mathcal{A}} \sum_{t=1}^{T} r_{t,a}$.

Theorem

If Exp3 is run with

$$\gamma = \eta = \sqrt{\frac{A\log A}{(e-1)T}}$$

then it achieves a regret

$$R_T \leqslant O(\sqrt{TA\log A})$$

Comparison with online learning (convex, bounded):

$$R_T(Exp3) \leqslant O(\sqrt{TA\log A})$$

$$R_T(EWA) \leqslant O(\sqrt{T\log A})$$

Comparison with online learning (convex, bounded):

$$R_T(Exp3) \leqslant O(\sqrt{TA\log A})$$

$$R_T(EWA) \leqslant O(\sqrt{T\log A})$$

Intuition: in online learning at each round we obtain A feedbacks, while in bandits we receive 1 feedback.

$$R_T(Exp3) = \mathbb{E}\Big(\sum_{t=1}^{T} r_{t,a} - r_{t,a_t}\Big) \leqslant \frac{\log(A)}{\eta} + \frac{A}{2}\eta T\,.$$

Further, For any non-increasing sequence $(\eta_t)_t$:

$$R_T(Exp3) = \mathbb{E}\Big(\sum_{t=1}^{T} r_{t,a} - r_{t,a_t}\Big) \leqslant \frac{\log(A)}{\eta_T} + \frac{A}{2}\sum_{t=1}^{T}\eta_t\,.$$

Step 1. $\mathbb{E}_{a\sim p_{t,\eta}}\tilde{\ell}_t(a) = 1 - r_{t,a_t}$ and $\mathbb{E}_{a_t\sim p_{t,\eta}}\tilde{\ell}_t(a) = 1 - r_{t,a}$. Thus:

$$\forall a \in \mathcal{A}, \quad \sum_{t=1}^{T} r_{t,a} - r_{t,a_t} = \sum_{t=1}^{T} \mathbb{E}_{a\sim p_{t,\eta}}\tilde{\ell}_t(a) - \sum_{t=1}^{T} \mathbb{E}_{a_t\sim p_{t,\eta}}\tilde{\ell}_t(a)\,.$$

Step 2. The random variable $X = \tilde{\ell}_t(a)$, is positive. By Hoeffding's lemma,

$$\begin{aligned}
\mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)) &\leqslant -\frac{1}{\eta}\log\left(\mathbb{E}_{a\sim p_{t,\eta}}\left[\exp(-\eta\tilde{\ell}_t(a))\right]\right) + \frac{\eta}{2}\mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)^2) \\
&= -\frac{1}{\eta}\log\left(\frac{\sum_{a\in\mathcal{A}} e^{-\sum_{s=1}^{t}\eta\tilde{\ell}_s(a)}}{\sum_{a\in\mathcal{A}} e^{-\sum_{s=1}^{t-1}\eta\tilde{\ell}_s(a)}}\right) + \frac{\eta}{2}\mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)^2)\,.
\end{aligned}$$

Step 3. Thus,

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)) \leqslant -\frac{1}{\eta}\log\Big(\frac{1}{A}\sum_b \exp\big(-\sum_{t=1}^{T}\eta\tilde{\ell}_t(b)\big)\Big) + \sum_{t=1}^{T}\frac{\eta}{2}\mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)^2)\,.$$

Since the reward function is bounded by 1 we have:

$$\mathbb{E}_{a\sim p_{t,\eta}}(\tilde{\ell}_t(a)^2) = \mathbb{E}_{a\sim p_{t,\eta}}\Big(\frac{(1-r_{t,A_t})^2}{p_t^2(A_t)}\mathbb{I}\{A_t = a\}\Big) \leqslant \frac{1}{p_t(a_t)}.$$

Step 4. Using the fact that the sum of positive terms is bigger than any of its term,

$$-\frac{1}{\eta}\log\Big(\sum_b \exp\big(-\sum_{t=1}^{T}\eta\tilde{\ell}_t(b)\big)\Big) \quad\leqslant\quad \sum_{t=1}^{T}\tilde{\ell}_t(a) \text{ for each } a\in\mathcal{A}\,.$$

Taking expectations, it comes for all $a\in\mathcal{A}$,

$$\mathbb{E}\Big[\sum_{t=1}^{T} r_{t,a} - r_{t,a_t}\Big] \leqslant \frac{\log(A)}{\eta} + \sum_{t=1}^{T}\frac{\eta}{2}\underbrace{\mathbb{E}\Big[\frac{1}{p_t(a_t)}\Big]}_{A}\,.$$

The Improved-Exp3 Algorithm

Using importance sampling is bad as generates large variance, especially for arms with low probability of being chosen (bad arms).

Using importance sampling is bad as generates large variance, especially for arms with low probability of being chosen (bad arms).

▷ Exp3.P (Auer et al. 2002): $\tilde{r}_{t,a} = r_{t,a} + \frac{\beta}{p_{t,a}}$

Using importance sampling is bad as generates large variance, especially for arms with low probability of being chosen (bad arms).

- Exp3.P (Auer et al. 2002): $\tilde{r}_{t,a} = r_{t,a} + \frac{\beta}{p_{t,a}}$
- Exp3-IX (Kocak et al, 2014; Neu 2015): $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}}{p_{t,a} + \gamma}$.

The Improved-Exp3 Algorithm

Using importance sampling is bad as generates large variance, especially for arms with low probability of being chosen (bad arms).

- Exp3.P (Auer et al. 2002): $\tilde{r}_{t,a} = r_{t,a} + \dfrac{\beta}{p_{t,a}}$
- Exp3-IX (Kocak et al, 2014; Neu 2015): $\tilde{\ell}_{t,a} = \dfrac{\ell_{t,a}}{p_{t,a} + \gamma}$.
- Many other variants.

- Decisions are **distributions** on arms $\mathcal{X} = \mathcal{P}(\mathcal{A})$.

- Decisions are **distributions** on arms $\mathcal{X} = \mathcal{P}(\mathcal{A})$.
- One expert outputs $\xi_{t,m} \in \mathcal{P}(\mathcal{A})$ at time t.

- Decisions are **distributions** on arms $\mathcal{X} = \mathcal{P}(\mathcal{A})$.
- One expert outputs $\xi_{t,m} \in \mathcal{P}(\mathcal{A})$ at time t.
- **Loss** of expert $m \in \mathcal{M}$: $\ell_{t,m} = \sum_{a\in\mathcal{A}} \xi_{t,m}(a) r_t(a)$ (Instead of reward)

- Decisions are **distributions** on arms $\mathcal{X} = \mathcal{P}(\mathcal{A})$.
- One expert outputs $\xi_{t,m} \in \mathcal{P}(\mathcal{A})$ at time t.
- **Loss** of expert $m \in \mathcal{M}$: $\ell_{t,m} = \sum_{a\in\mathcal{A}} \xi_{t,m}(a) r_t(a)$ (Instead of reward)
- Case when $|\mathcal{M}| \gg |\mathcal{A}|$?

Exponential-weight algorithm for exploration and exploitation using expert advice.

Exponential-weight algorithm for exploration and exploitation using expert advice.

▷ $\forall m \in \mathcal{M}, w_1(m) = 1.$

Exp4

Exponential-weight algorithm for exploration and exploitation using expert advice.

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $a_t \sim p_t \in \mathcal{P}(\mathcal{A})$ where $p_t(a) = (1-\gamma)\frac{w_t(m)\xi_{t,m}(a)}{\sum_{m\in\mathcal{M}} w_t(m)} + \frac{\gamma}{|\mathcal{A}|}$

Exp4

Exponential-weight algorithm for exploration and exploitation using expert advice.

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $a_t \sim p_t \in \mathcal{P}(\mathcal{A})$ where $p_t(a) = (1-\gamma)\frac{w_t(m)\xi_{t,m}(a)}{\sum_{m\in\mathcal{M}} w_t(m)} + \frac{\gamma}{|\mathcal{A}|}$
- Receive r_{t,a_t}, build $\widehat{\ell}_t(a) = \begin{cases} \frac{1-r_t(a)}{p_t(a)} & \text{if } a = a_t \\ 0 & \text{else} \end{cases}$

Exponential-weight algorithm for exploration and exploitation using expert advice.

- $\forall m \in \mathcal{M}, w_1(m) = 1.$
- Output $a_t \sim p_t \in \mathcal{P}(\mathcal{A})$ where $p_t(a) = (1-\gamma)\frac{w_t(m)\xi_{t,m}(a)}{\sum_{m\in\mathcal{M}} w_t(m)} + \frac{\gamma}{|\mathcal{A}|}$
- Receive r_{t,a_t}, build $\widehat{\ell}_t(a) = \begin{cases} \frac{1-r_t(a)}{p_t(a)} & \text{if } a = a_t \\ 0 & \text{else} \end{cases}$
- Update $\forall m \in \mathcal{M}, w_{t+1}(m) = w_t(m)\exp(-\eta\widehat{\ell}_{t,m})$. where $\widehat{\ell}_{t,m} = \sum_{a\in\mathcal{A}} \xi_{t,m}(a)\widehat{\ell}_t(a)$.

Theorem

If Exp4 is run with $\gamma \in [0,1]$, then it achieves a regret

$$R_T = \max_{a\in\mathcal{A}} \sum_{t=1}^{T} r_{t,a} - \mathbb{E}\Big[\sum_{t=1}^{T} r_{t,A_t}\Big] \leqslant (e-1)\gamma G_{\max} + \frac{A\log M}{\gamma}$$

with $G_{\max} = \max_{a\in\mathcal{A}} \sum_{t=1}^{T} r_{t,a}$.

- $\Phi : \mathcal{H} \rightarrow \mathcal{D}$, mapping from set of histories to some set $\mathcal{D}$, such that $h_1 \sim h_2$ iff $\Phi(h_1) = \Phi(h_2)$ defines **equivalence relation**; let $[h]$ the equivalence class of h.

Constrained opponent/comparison class

- $\Phi : \mathcal{H} \to \mathcal{D}$, mapping from set of histories to some set $\mathcal{D}$, such that $h_1 \sim h_2$ iff $\Phi(h_1) = \Phi(h_2)$ defines **equivalence relation**; let $[h]$ the equivalence class of h.
- Φ-constrained policy is $\pi : \mathcal{H}/\Phi \to \mathcal{A}$.

- $\Phi : \mathcal{H} \to \mathcal{D}$, mapping from set of histories to some set $\mathcal{D}$, such that $h_1 \sim h_2$ iff $\Phi(h_1) = \Phi(h_2)$ defines **equivalence relation**; let $[h]$ the equivalence class of h.
- Φ-constrained policy is $\pi : \mathcal{H}/\Phi \to \mathcal{A}$.
- Examples:
  - $\Phi(h) = 1$ gives constant experts.
  - $\Phi(h) = (a_{-1}, \dots, a_{-m})$ last m actions, gives experts depending on last m actions only.
  - $\Phi(h) = |h| \mod k$ gives periodic experts.

- $\Phi : \mathcal{H} \to \mathcal{D}$, mapping from set of histories to some set $\mathcal{D}$, such that $h_1 \sim h_2$ iff $\Phi(h_1) = \Phi(h_2)$ defines **equivalence relation**; let $[h]$ the equivalence class of h.
- Φ-constrained policy is $\pi : \mathcal{H}/\Phi \to \mathcal{A}$.
- Examples:
  - $\Phi(h) = 1$ gives constant experts.
  - $\Phi(h) = (a_{-1}, \dots, a_{-m})$ last m actions, gives experts depending on last m actions only.
  - $\Phi(h) = |h| \bmod k$ gives periodic experts.
- We define the **Φ-constrained** regret:

$$\mathcal{R}_T^{\Phi} = \sup_{\pi:\mathcal{H}/\Phi\to\mathcal{A}} \mathbb{E}\Big[\sum_{t=1}^{T} r_{t,\pi([h_t])}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} r_{t,a_t}\Big]$$

More challenging than best constant expert.

- We can define a version of Exp4 for Φ-constrained policies.

- We can define a version of Exp4 for Φ-constrained policies.
- We simply **contextualize** Exp4 by indexing losses, weights, parameters η by the equivalence classes, and computing the current active class $c_t = \Phi(h_t)$.

- We can define a version of Exp4 for Φ-constrained policies.
- We simply **contextualize** Exp4 by indexing losses, weights, parameters η by the equivalence classes, and computing the current active class $c_t = \Phi(h_t)$.
- Result (M. Munos, 2011)

$$\mathcal{R}_T^{\Phi} \leqslant \sum_{c \in \mathcal{H}/\Phi} \mathbb{E}\left[\frac{A\eta_c}{2} T_c + \frac{\log(A)}{\eta_c}\right].$$

where T_c is number of activation times of class c until time T.

- We consider we have a **set** $(\Phi_\theta)_{\theta\in\Theta}$ of **constrained strategies**.

- We consider we have a **set** $(\Phi_\theta)_{\theta\in\Theta}$ of **constrained strategies**.
- One Φ_θ-Exp3 strategy for each θ: see them as different **experts**?

POOL OF CONSTRAINED STRATEGIES?

- We consider we have a **set** $(\Phi_\theta)_{\theta\in\Theta}$ of **constrained strategies**.
- One Φ_θ-Exp3 strategy for each θ: see them as different **experts**?
- Run Exp4 with all these base experts: Φ_1-Exp3, ..., Φ_P-Exp3 ?

Difficulty: The experts are **learning** algorithms. Their performance depends on the observations they received.
We are in **partial feedback**: When Φ_p-Exp3 recommends to play action a, Exp4 may **instead** play (and received reward from) action b. Hence Φ_p-Exp3 not only faces **partial feedback**, but also it does **not** observe the reward corresponding to what it decides.

Double-bandit feedback.

Theorem (M. Munos, 2011)

In the double-bandit feedback setup, Exp4, run on $(\Phi_\theta\text{-Exp3})_{\theta\in\Theta}$ strategies with appropriate parameter tuning satisfies

$$\mathcal{R}_T = O\Big(T^{2/3}(A\log(A)C)^{1/3}\log(|\Theta|)^{1/2}\Big) \text{ with } C = \max_{\theta\in\theta}|\mathcal{H}/\Phi_\theta|.$$

- Strategies for **Stochastic** bandits: UCB, KL-UCB, etc.
  $\log(T)$ regret bounds when stochastic model, but strong assumptions on signal.

- Strategies for **Stochastic** bandits: UCB, KL-UCB, etc.
  $\log(T)$ regret bounds when stochastic model, but strong assumptions on signal.
- Strategies for **Adversarial** bandits: Exp3, Exp4, etc.
  $\sqrt{T}$ regret bounds with little assumption on model, but perhaps too conservative.

Can we have the best of both worlds?

Several works on the topic

- Bubeck&Slivkins 2012, Auer&Chiang, 2016.
- Zimmert-Seldin 2018.

Idea: **Online Mirror Descent** regularized by **Tsallis Entropy**.
α-**Tsallis** entropy:

$$H_\alpha(x) = \frac{1}{1-\alpha}(1 - \sum_{a\in\mathcal{A}} x_a^\alpha)$$

◇ $\lim_{\alpha\to 1} H_\alpha(x) = \sum_{a\in\mathcal{A}} x_a \log(x_a)$
◇ $\lim_{\alpha\to 0} H_\alpha(x) = -\sum_{a\in\mathcal{A}} \log(x_a)$

Let us consider the potential:

$$\Psi_{t,\alpha}(q) = -\sum_{a\in\mathcal{A}} \frac{q^{\alpha}(a)}{\alpha}$$

Strategy:

Let us consider the potential:

$$\Psi_{t,\alpha}(q) = -\sum_{a\in\mathcal{A}} \frac{q^{\alpha}(a)}{\alpha}$$

Strategy:

▷ Choose

$$p_t = \operatorname*{argmin}_{q\in\mathcal{P}(\mathcal{A})} \langle q, \widehat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi_{\alpha}(q)$$

(This is gradient of dual of $\Psi_{t,\alpha}/\eta_t$ at position $\widehat{L}_{t-1}$)

Let us consider the potential:

$$\Psi_{t,\alpha}(q) = -\sum_{a\in\mathcal{A}} \frac{q^{\alpha}(a)}{\alpha}$$

Strategy:

- Choose

$$p_t = \operatorname*{argmin}_{q\in\mathcal{P}(\mathcal{A})} \langle q, \widehat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi_{\alpha}(q)$$

  (This is gradient of dual of $\Psi_{t,\alpha}/\eta_t$ at position $\widehat{L}_{t-1}$)
- Sample $a_t \sim p_t$

Let us consider the potential:

$$\Psi_{t,\alpha}(q) = -\sum_{a\in\mathcal{A}} \frac{q^\alpha(a)}{\alpha}$$

Strategy:

- Choose

$$p_t = \operatorname*{argmin}_{q\in\mathcal{P}(\mathcal{A})} \langle q, \widehat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi_\alpha(q)$$

  (This is gradient of dual of $\Psi_{t,\alpha}/\eta_t$ at position $\widehat{L}_{t-1}$)
- Sample $a_t \sim p_t$
- Observe ℓ_{t,a_t} then build $\widehat{\boldsymbol{\ell}}_t$ as unbiased estimate of $\boldsymbol{\ell}_t$, then $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\boldsymbol{\ell}}_t$.

	Regime	$\frac{\text{Upper bound}}{\text{Lower bound}}$	Learning rate
$\lim_{\alpha\to 0}$	Sto	$O(1)$	$\Theta(\Delta_a)$
	Adv	$O(\sqrt{\ln(T)})$	$\Theta\left(\frac{\ln(t)}{\sqrt{t}}\right)$
$\alpha = \frac{1}{2}$	Sto&Adv	$O(1)$	$\frac{1}{\sqrt{t}}$
$\lim_{\alpha\to 1}$	Sto	$O(\ln(T))$	$\Theta\left(\frac{\ln(t)}{\Delta_a t}\right)$
	Adv	$O(\sqrt{\ln(A)})$	$\Theta\left(\frac{1}{\sqrt{t}}\right)$.

Full information

- Powerful: Handle **large** number of experts

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts

Take home message

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

- Only output **one** arm, not a convex combination of arms.

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

- Only output **one** arm, not a convex combination of arms.
- Only receive reward on **one** arm.

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

- Only output **one** arm, not a convex combination of arms.
- Only receive reward on **one** arm.
- Difficulty to estimate reward/loss [Still not satisfactory]

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

- Only output **one** arm, not a convex combination of arms.
- Only receive reward on **one** arm.
- Difficulty to estimate reward/loss [Still not satisfactory]
- $\sqrt{A}$ factor in regret bounds.

Take home message

Full information

- Powerful: Handle **large** number of experts
- Increasingly challenging targets:
  - **Constant** expert, **combination of loss** of experts. Convex and bounded or η-mixable loss.
  - Constant **combination** of experts (Hedge)
  - Best **sequence of switching** experts
  - Best **sequence** of few **recurring** experts (Freund)
- Powerful results, log of number of experts
- Computationally efficient algorithms, leveraging structure of experts.

Bandit information

- Only output **one** arm, not a convex combination of arms.
- Only receive reward on **one** arm.
- Difficulty to estimate reward/loss [Still not satisfactory]
- $\sqrt{A}$ factor in regret bounds.
- Useful in **games**.

Open questions

- Bandit results for
  - Best sequence of experts?
  - Best sequence of few recurring experts ?
  - Sleeping, Growing experts ?
  - Beyond convex/bounded?

Open questions

- Bandit results for
  - Best sequence of experts?
  - Best sequence of few recurring experts ?
  - Sleeping, Growing experts ?
  - Beyond convex/bounded?
- Best of both world: Exact stochastic optimality? Estimation of loss?

Open questions

- Bandit results for
  - Best sequence of experts?
  - Best sequence of few recurring experts ?
  - Sleeping, Growing experts ?
  - Beyond convex/bounded?
- Best of both world: Exact stochastic optimality? Estimation of loss?
- Mixed world bandit: Some arms are stochastic, others are arbitrary bounded?

Table of contents

▷ A two–player **zero–sum** game

	A	B	C
1	30, -30	-10, 10	20, -20
2	10, -10	-20, 20	-20, 20

▷ A two–player **zero–sum** game

	A	B	C
1	30, -30	-10, 10	20, -20
2	10, -10	-20, 20	-20, 20

Nash equilibrium:
A set of strategies is a **Nash equilibrium** if **no player** can do better by **unilaterally changing** his strategy.

▷ A two–player **zero–sum** game

	A	B	C
1	30, -30	-10, 10	20, -20
2	10, -10	-20, 20	-20, 20

Nash equilibrium:
Red: take action 1 with **prob. 1**
Blue: take action B with **prob. 1**

▷ A two–player **zero–sum** game

	A	B	C
1	30, -30	-10, 10	20, -20
2	10, -10	-20, 20	-20, 20

Nash equilibrium:
Value of the game: $V = -10$ (reward of Red at the equilibrium)

A two–player zero–sum game

	A	B
1	-2, 2	3, -3
2	3, -3	-4, 4

A two–player zero–sum game

	A	B
1	-2, 2	3, -3
2	3, -3	-4, 4

Nash equilibrium:
A set of strategies is a Nash equilibrium if **no player** can do better by **unilaterally changing** his strategy.

A two–player zero–sum game

	A	B
1	-2, 2	3, -3
2	3, -3	-4, 4

Nash equilibrium:
Red: take action **1** with **prob. 7/12** and action **2** with **prob. 5/12**
Blue: take action **A** with **prob. 7/12** and action **B** with **prob. 5/7**

A two–player zero–sum game

	A	B
1	-2, 2	3, -3
2	3, -3	-4, 4

Nash equilibrium:
Value of the game: $V = 1/12$ (reward of Red at the equilibrium)

At each round t

- Row player computes a mixed strategy $\widehat{\mathbf{p}}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{N,t})$
- Column player computes a mixed strategy $\widehat{\mathbf{q}}_t = (\widehat{q}_{1,t}, \ldots, \widehat{q}_{M,t})$

At each round t

- Row player computes a mixed strategy $\widehat{\mathbf{p}}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{N,t})$
- Column player computes a mixed strategy $\widehat{\mathbf{q}}_t = (\widehat{q}_{1,t}, \ldots, \widehat{q}_{M,t})$
- Row player selects action $I_t \in \{1, \ldots, N\}$
- Column player selects action $J_t \in \{1, \ldots, M\}$

At each round t

- Row player computes a mixed strategy $\widehat{\mathbf{p}}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{N,t})$
- Column player computes a mixed strategy $\widehat{\mathbf{q}}_t = (\widehat{q}_{1,t}, \ldots, \widehat{q}_{M,t})$
- Row player selects action $I_t \in \{1, \ldots, N\}$
- Column player selects action $J_t \in \{1, \ldots, M\}$
- Row player suffers $\ell(I_t, J_t)$
- Column player suffers $-\ell(I_t, J_t)$

At each round t

- Row player computes a mixed strategy $\widehat{\mathbf{p}}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{N,t})$
- Column player computes a mixed strategy $\widehat{\mathbf{q}}_t = (\widehat{q}_{1,t}, \ldots, \widehat{q}_{M,t})$
- Row player selects action $I_t \in \{1, \ldots, N\}$
- Column player selects action $J_t \in \{1, \ldots, M\}$
- Row player suffers $\ell(I_t, J_t)$
- Column player suffers $-\ell(I_t, J_t)$

Value of the game

$$V = \max_{\mathbf{q}} \min_{\mathbf{p}} \bar{\ell}(\mathbf{p}, \mathbf{q})$$

with

$$\bar{\ell}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j \ell(i, j)$$

Question: what if the two players are both bandit algorithms (e.g., Exp3)?

Question: what if the two players are both bandit algorithms (e.g., Exp3)?

Row player: a bandit algorithm is able to minimize

$$R_n(\text{row}) = \sum_{t=1}^{n} \ell_{I_t,J_t} - \min_{i=1,\dots,N} \sum_{t=1}^{n} \ell_{i,J_t}$$

Question: what if the two players are both bandit algorithms (e.g., Exp3)?

Row player: a bandit algorithm is able to minimize

$$R_n(\text{row}) = \sum_{t=1}^{n} \ell_{I_t,J_t} - \min_{i=1,\dots,N} \sum_{t=1}^{n} \ell_{i,J_t}$$

Col player: a bandit algorithm is able to minimize

$$R_n(\text{col}) = \sum_{t=1}^{n} \ell_{I_t,J_t} - \min_{j=1,\dots,M} \sum_{t=1}^{n} \ell_{I_t,j}$$

Theorem

If both the row and column players play according to an **Hannan-consistent** strategy, then

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) = V$$

Theorem

The **empirical distribution** of plays

$$\widehat{p}_{i,n} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I} I_t = i \quad \widehat{q}_{j,n} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I} J_t = j$$

induces a product distribution $\widehat{\mathbf{p}}_n \times \widehat{\mathbf{q}}_n$ which converges to the **set of Nash equilibria** $\mathbf{p} \times \mathbf{q}$.

Since $\bar{\ell}(\mathbf{p}, J_t)$ is linear, over the simplex, the minimum is at one of the corners

[math]

$$\min_{i=1,\ldots,N} \frac{1}{N} \sum_{t=1}^{n} \ell(i, J_t) = \min_{\mathbf{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(\mathbf{p}, J_t)$$

Since $\bar{\ell}(\mathbf{p}, J_t)$ is linear, over the simplex, the minimum is at one of the corners **[math]**

$$\min_{i=1,\ldots,N} \frac{1}{N} \sum_{t=1}^{n} \ell(i, J_t) = \min_{\mathbf{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(\mathbf{p}, J_t)$$

We consider the empirical probability of the row player **[def]**

$$\widehat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I} J_t = j$$

Since $\bar{\ell}(\mathbf{p}, J_t)$ is linear, over the simplex, the minimum is at one of the corners **[math]**

$$\min_{i=1,\dots,N} \frac{1}{N} \sum_{t=1}^{n} \ell(i, J_t) = \min_{\mathbf{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(\mathbf{p}, J_t)$$

We consider the empirical probability of the row player **[def]**

$$\widehat{q}_{j,n} = \frac{1}{n} \sum_{t=1}^{n} \mathbb{I} J_t = j$$

Elaborating on it **[math]**

$$\begin{aligned} \min_{\mathbf{p}} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell}(\mathbf{p}, J_t) &= \min_{\mathbf{p}} \sum_{j=1}^{M} \widehat{q}_{j,n} \bar{\ell}(\mathbf{p}, j) \\ &= \min_{\mathbf{p}} \bar{\ell}(\mathbf{p}, \widehat{\mathbf{q}}_n) \\ &\leqslant \max_{\mathbf{q}} \min_{\mathbf{p}} \bar{\ell}(\mathbf{p}, \mathbf{q}) = V \end{aligned}$$

By definition of Hannan's consistent strategy **[def]**

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) = \min_{i=1,\dots,N} \frac{1}{n} \sum_{t=1}^{n} \ell(i, J_t)$$

By definition of Hannan's consistent strategy [def]

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) = \min_{i=1,\dots,N} \frac{1}{n}\sum_{t=1}^{n} \ell(i, J_t)$$

Then

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) \leqslant V$$

By definition of Hannan's consistent strategy **[def]**

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) = \min_{i=1,\dots,N} \frac{1}{n}\sum_{t=1}^{n} \ell(i, J_t)$$

Then

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) \leqslant V$$

If we do the same for the other player **[zero–sum game]**

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, J_t) \geqslant V$$

Question: how fast do they converge to the Nash equilibrium?

Question: how fast do they converge to the Nash equilibrium?
Answer: it depends on the specific algorithm. For EWA(η), we now that

$$\sum_{t=1}^{n} \ell(I_t, J_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{n} \ell(i, J_t) \leqslant \frac{\log N}{\eta} + \frac{n\eta}{8} + \sqrt{\frac{n}{2} \log \frac{1}{\delta}}$$

Generality of the results

- Players do not know the payoff matrix

Generality of the results

- Players do not know the payoff matrix
- Players do not observe the loss of the other player

Generality of the results

- Players do not know the payoff matrix
- Players do not observe the loss of the other player
- Players do not even observe the action of the other player

External (expected) regret

$$R_n = \sum_{t=1}^{n} \bar{\ell}(\widehat{\mathbf{p}}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{n} \ell(i, y_t)$$

$$= \max_{i=1,\ldots,N} \sum_{t=1}^{n} \sum_{j=1}^{N} \widehat{p}_{j,t} \left(\ell(j, y_t) - \ell(i, y_t)\right)$$

External (expected) regret

$$
\begin{aligned}
R_n &= \sum_{t=1}^{n} \bar{\ell}(\widehat{\mathbf{p}}_t, y_t) - \min_{i=1,\dots,N} \sum_{t=1}^{n} \ell(i, y_t) \\
&= \max_{i=1,\dots,N} \sum_{t=1}^{n} \sum_{j=1}^{N} \widehat{p}_{j,t} \big(\ell(j, y_t) - \ell(i, y_t)\big)
\end{aligned}
$$

Internal (expected) regret

$$
R_n^I = \max_{i,j=1,\dots,N} \sum_{t=1}^{n} \widehat{p}_{j,t} \big(\ell(i, y_t) - \ell(j, y_t)\big)
$$

Internal (expected) regret

$$R_n^I = \max_{i,j=1,\dots,N} \sum_{t=1}^{n} \widehat{p}_{j,t} \big(\ell(i, y_t) - \ell(j, y_t)\big)$$

Intuition: an algorithm has **small internal regret** if, for each pair of experts (i, j), the learner does not regret of not having followed expert j each time it followed expert i.

Theorem

Given a K–person game with a set of correlated equilibria $\mathcal{C}$. If all the players are internal–regret minimizers, then the **distance** *between the* **empirical distribution** *of plays and the set of* **correlated equilibria** $\mathcal{C}$ *converges to 0.*

A powerful model for **sequential** games

- Checkers / Chess / Go
- Poker
- Bargaining
- Monitoring
- Patrolling
- ...

Nash Equilibria in Extensive Form Games

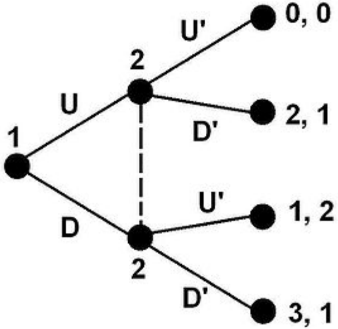
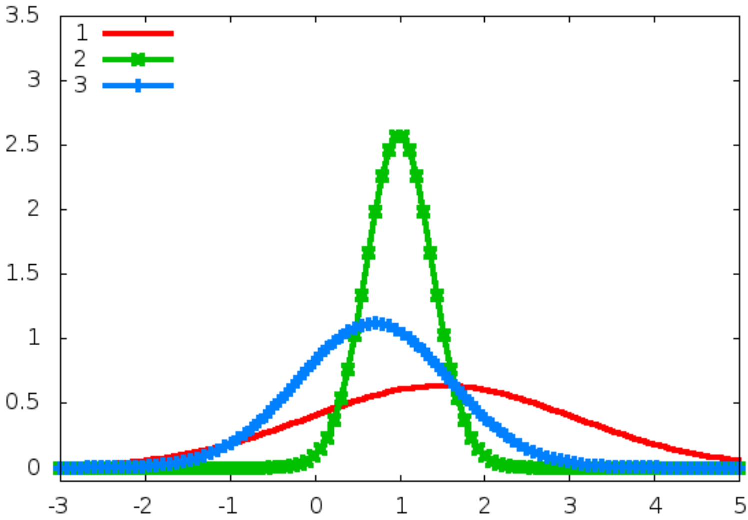


Table of contents

- We considered adversarial setup. One way to address risk.
- Other ways: **Risk-aversion** (model), **Robust** strategies (min-max).

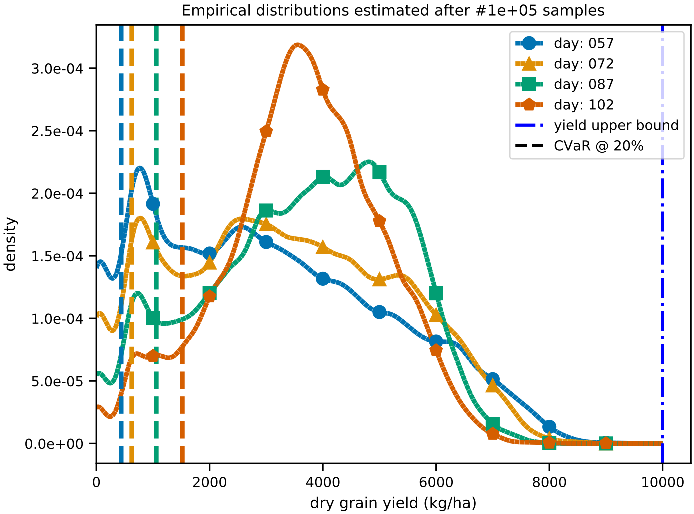
Illustration of risk-aversion.



▷ Choice for 1 sample ? For 1000 samples?

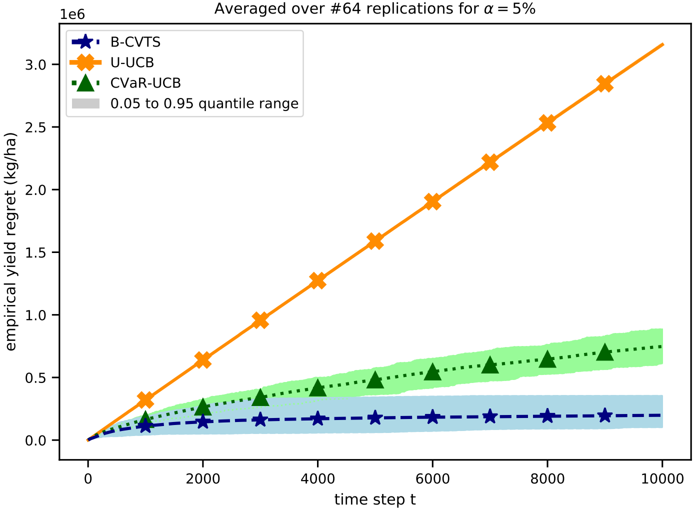
Bandit strategy for risk aversion

- DSSAT simulator: 30y of agronomy expertise, climate, ground, plant growth, etc.
- Distribution of yields for 4 different **planting date** (action) using **DSSAT**

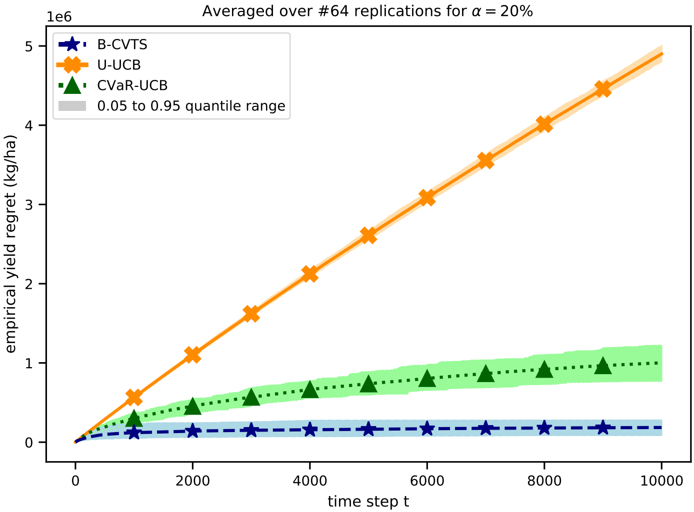


- May not want best expectation, but rather **risk-averse** criterion.

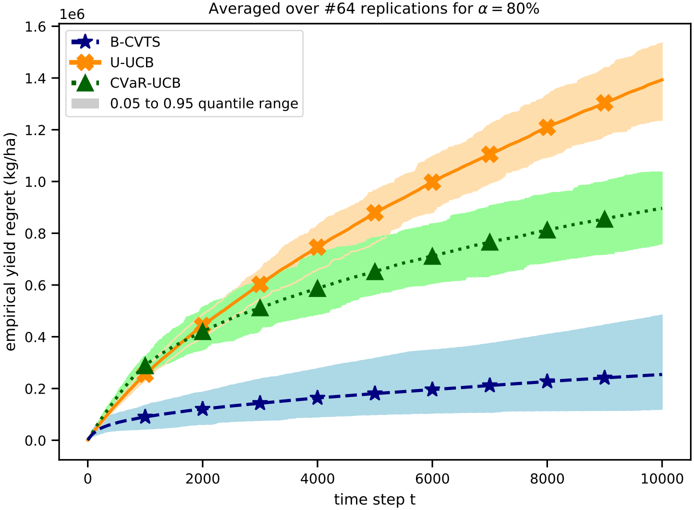
- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



Conditional Value at Risk

▷ Consider a **Gain** (reward): We are interested in **risky (low) gains**.

Maximize gain in worst-case situations

▷ Formally, for given $\alpha \in [0,1]$:

$$\mathtt{CVaR}_\alpha(\nu_k) = \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{\alpha}\mathbb{E}_{X\sim\nu_k}\left[(x-X)^+\right]\right\} . \quad (1)$$

▷ For **continuous** distributions, $\mathtt{CVaR}_\alpha(\nu_k) = \mathbb{E}_{X\sim\nu_k}[X|X \leqslant q_\alpha(\nu_k)]$, where $q_\alpha(\nu_k) = \inf\{x : \mathbb{P}(X \leqslant x) > \alpha\}$ is the quantile at level α.

- $\alpha = 1$ is the **expectation**, $\alpha = 0$ is very risk-averse (extreme).
- It is a **coherent** risk measure (Rockafellar, Acerbi et al.): many good properties.
- Rich litterature on CVaR in finance.
- Parameter α is **easy to interpret** for many practitioners.

CVaR Regret and bandits

- Unknown arm distributions $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$, given risk-level α.
- We write $c_k^\alpha = \mathtt{CVaR}_\alpha(\nu_k)$.
- Best arm is the one with the **largest** CVaR.
- The CVaR regret of a sequential sampling strategy $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T}\left(\max_k c_k^\alpha - c_{A_t}^\alpha\right)\right] = \sum_{k=1}^{K} \Delta_k^\alpha \mathbb{E}_{\boldsymbol{\nu}}[N_k(T)],$$

where $\Delta_k^\alpha = \max_{k'} c_{k'}^\alpha - c_k^\alpha$ is the CVaR gap.

- Algorithms: UCB style, we need upper confidence bounds for CVaR; TS, we need sampling scheme.

- **Concentration**: Brown 2007, Thomas and Learned-Miller 2019.
- **Bandits**: Agrawal et al. 2020, Galichet 2013, Tamkin et al. 2020, etc.
- **MDPs**:

Optimizing the CVaR via Sampling, Tamar et al. 2014

Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach, Chow et al. 2015

Definition

For any $\nu \in \mathcal{C}$ and $c \in \mathbb{R}$, we define

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{C}}(\nu, c) := \inf\left\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{C}, \mathtt{CVaR}_\alpha(\nu') \geqslant c\right\}.$$

Theorem (Regret Lower Bound in CVaR bandits)

Let $\alpha \in (0, 1]$. Let $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_K$ be a set of bandit models $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ where each ν_k belongs to the class of distribution $\mathcal{F}_k$. Let $\mathcal{A}$ be a strategy satisfying $\mathcal{R}_{\boldsymbol{\nu}}^{\alpha}(\mathcal{A}, T) = o(T^\beta)$ for any $\beta > 0$ and $\nu \in \mathcal{F}$. Then for any $\nu \in \mathcal{D}$, for any sub-optimal arm k, under the strategy $\mathcal{A}$ it holds that

$$\lim_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[N_k(T)]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{F}_k}(\nu_k, c^\star)},$$

where $c^\star = \max_{i \in [K]} \mathtt{CVaR}_\alpha(\nu_i)$.

▷ One can rewrite the $\texttt{CVaR}$ in terms of the **CDF** $F(x) = \mathbb{P}(X \leqslant x)$.

$$\texttt{CVaR}_\alpha(\nu) = \frac{1}{\alpha}\int g_\alpha(F_\nu(x)))dx$$

for some monotonic function g_α. Also, it holds

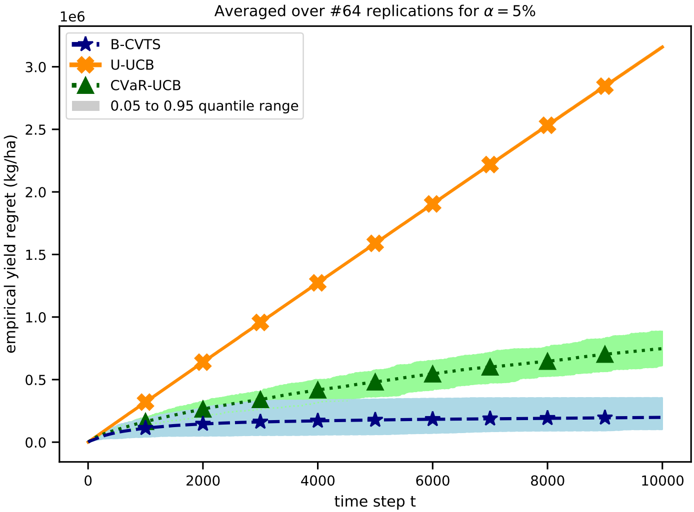
$$|\texttt{CVaR}_\alpha(\nu) - \texttt{CVaR}_\alpha(\nu')| \leqslant \frac{1}{\alpha}\|F_\nu - F_{\nu'}\|_\infty$$

▷ Main tool is Massart's version of Dvoretzky-Kiefer-Wolfowitz (DKW) inequality:

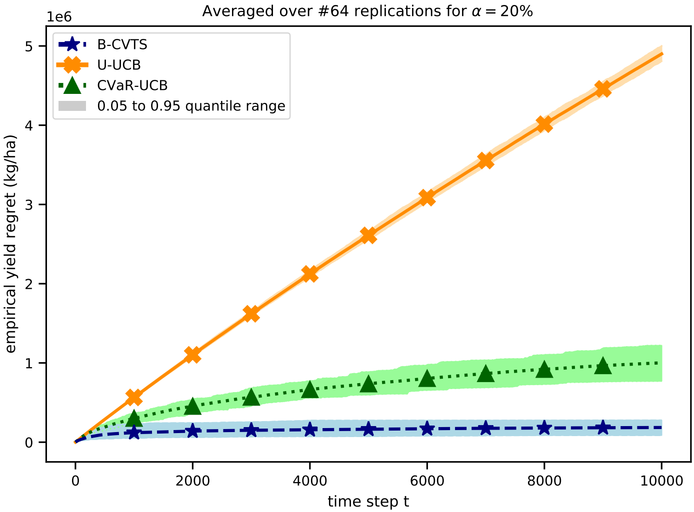
$$\forall \delta_0 \in [0, 0.5) \quad \mathbb{P}\bigg(\sup_{x\in\mathbb{R}} F_\nu(x) - F_n(x) > \sqrt{\frac{\ln(1/\delta_0)}{2n}}\bigg) \leqslant \delta_0\,.$$

where F_n is empirical CDF.

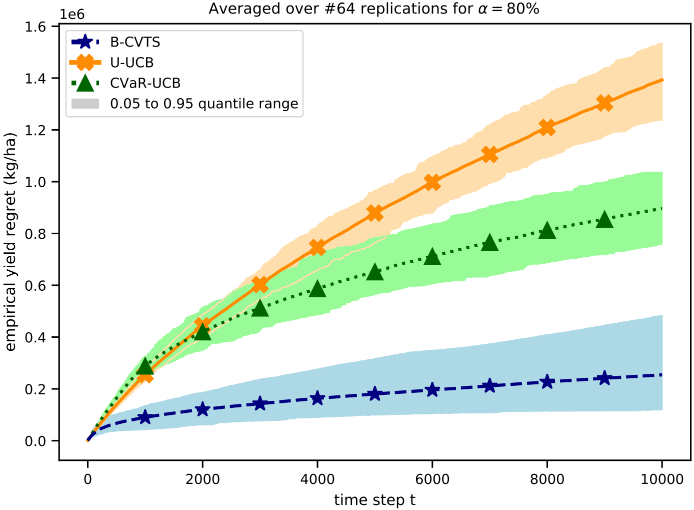
- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



- Novel bandit strategy for **Conditional Value at Risk** (CVaR)
- **Provably** optimal (regret bound matches lower bound).
- Based on novel statistical estimation tools. Performance (blue) against Sota:



- We want to **control** how big/small can a random variable be

$$\mathbb{P}\Big[X \geqslant \dots\Big] \leqslant \delta \tag{2}$$

$$\mathbb{P}\Big[X \leqslant \dots\Big] \leqslant \delta \tag{3}$$

- Quantiles, expectiles, expected shortfall, value at risk.

- We want to **control** how big/small can a random variable be

$$\mathbb{P}\Big[X \geqslant \inf_{\lambda>0}\Big\{\frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X) + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leqslant \delta \tag{4}$$

$$\mathbb{P}\Big[X \leqslant \sup_{\lambda>0}\Big\{-\frac{1}{\lambda}\log\mathbb{E}\exp(-\lambda X) - \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leqslant \delta \tag{5}$$

(by Markov's inequality, whenever $\log\mathbb{E}\exp$ is defined near 0)

For all $\lambda > 0$,

$$\begin{aligned}\mathbb{P}[X \geqslant \varepsilon] &= \mathbb{P}[\exp(\lambda X) \geqslant \exp(\lambda\varepsilon)] \\ &\leqslant \mathbb{E}[\exp(\lambda X)]\exp(-\lambda\varepsilon) \\ &= \exp\left(-\lambda\left(\varepsilon - \frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X)\right)\right)\end{aligned}$$

For $\varepsilon = \frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X) + \frac{\log(1/\delta)}{\lambda}$, we get

$$\mathbb{P}\left[X \geqslant \frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X) + \frac{\log(1/\delta)}{\lambda}\right] \leqslant \delta\,.$$

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda} \log \mathbb{E}_\nu \exp\Big(\lambda X\Big), \tag{6}$$

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda} \log \mathbb{E}_\nu \exp\Big(\lambda X\Big), \tag{6}$$

- more then one century year old,

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda} \log \mathbb{E}_\nu \exp\Big(\lambda X\Big), \tag{6}$$

- more then one century year old,
- at the heart of many key-results and tools in statistical theory (Cramer-Chernoff method, Chernoff transform, log-Laplace transform)

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda} \log \mathbb{E}_\nu \exp\Big(\lambda X\Big), \tag{6}$$

- more then one century year old,
- at the heart of many key-results and tools in statistical theory (Cramer-Chernoff method, Chernoff transform, log-Laplace transform)
- $\kappa_{-\lambda,\nu}$ is a key quantity to control the **probability that X is small**.

- Let $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.

Example: Gaussian distributions.

- Let $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.
- Let $X = \sum_{k=1}^{t} Z_k$ (thus $\mathcal{N}(\mu t, \sigma^2 t)$).

- Let $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from $\mathcal{N}(\mu, \sigma^2)$.
- Let $X = \sum_{k=1}^{t} Z_k$ (thus $\mathcal{N}(\mu t, \sigma^2 t)$).
- We recover in the **Gaussian** case the **mean-variance**

$$\kappa_{-\lambda,\nu} = \mu t - \frac{\lambda \sigma^2 t}{2}$$

Example: Gaussian distributions.

- Let $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from $\mathcal{N}(\mu,\sigma^2)$.
- Let $X = \sum_{k=1}^t Z_k$ (thus $\mathcal{N}(\mu t, \sigma^2 t)$).
- We recover in the **Gaussian** case the **mean-variance**

$$\kappa_{-\lambda,\nu} = \mu t - \frac{\lambda\sigma^2 t}{2}$$

- $\lambda = \sqrt{\frac{2\log(1/\delta)}{\sigma^2 t}}$ optimizes (4) and (5) and gives the familiar

$$\mathbb{P}\left(\frac{1}{t}\sum_{k=1}^t Z_k - \mu \geqslant \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\right) \leqslant \delta$$

$$\mathbb{P}\left(\mu - \frac{1}{t}\sum_{k=1}^t Z_k \geqslant \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\right) \leqslant \delta\,.$$

- We introduce the **mixability gaps** (always non negative):

$$m^+_{\lambda,\nu} = \kappa_{\lambda,\nu} - \mathbb{E}_\nu\Big[X\Big] \text{ and } m^-_{\lambda,\nu} = \mathbb{E}_\nu\Big[X\Big] - \kappa_{-\lambda,\nu}\,.$$

- We introduce the **mixability gaps** (always non negative):

$$m^{+}_{\lambda,\nu} = \kappa_{\lambda,\nu} - \mathbb{E}_{\nu}\Big[X\Big] \text{ and } m^{-}_{\lambda,\nu} = \mathbb{E}_{\nu}\Big[X\Big] - \kappa_{-\lambda,\nu} \,.$$

- Now equations (4) and (5) rewrite more compactly as

$$\mathbb{P}\Big[X - \mathbb{E}_{\nu}\Big[X\Big] \geqslant \inf_{\lambda>0}\Big\{m^{+}_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leqslant \delta \,, \tag{7}$$

$$\mathbb{P}\Big[\mathbb{E}_{\nu}\Big[X\Big] - X \geqslant \inf_{\lambda>0}\Big\{m^{-}_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leqslant \delta \,. \tag{8}$$

Entropic Value At Risk

- Control of the upper/lower tails involves $\kappa_{\lambda,\nu}/\kappa_{-\lambda,\nu}$.

Entropic Value At Risk

- Control of the upper/lower tails involves $\kappa_{\lambda,\nu}/\kappa_{-\lambda,\nu}$.
- Coincides with mean-variance for Gaussian.

Entropic Value At Risk

- Control of the upper/lower tails involves $\kappa_{\lambda,\nu}/\kappa_{-\lambda,\nu}$.
- Coincides with mean-variance for Gaussian.
- Coherence.

Entropic Value At Risk

- Control of the upper/lower tails involves $\kappa_{\lambda,\nu}/\kappa_{-\lambda,\nu}$.
- Coincides with mean-variance for Gaussian.
- Coherence.
- General interpretation as **penalty**

$$\kappa_{-\lambda,\nu} = \inf_{\nu' \in \mathcal{M}(\mathbb{R})} \left\{ \mathbb{E}_{\nu'}(X) + \frac{1}{\lambda}\mathtt{KL}(\nu'||\nu) \right\} \leqslant \mathbb{E}_{\nu}\left[X\right] . \tag{9}$$

Entropic Value At Risk

- Control of the upper/lower tails involves $\kappa_{\lambda,\nu}/\kappa_{-\lambda,\nu}$.
- Coincides with mean-variance for Gaussian.
- Coherence.
- General interpretation as **penalty**

$$\kappa_{-\lambda,\nu} = \inf_{\nu' \in \mathcal{M}(\mathbb{R})} \left\{ \mathbb{E}_{\nu'}(X) + \frac{1}{\lambda}\mathtt{KL}(\nu'||\nu) \right\} \leqslant \mathbb{E}_\nu[X] \,. \tag{9}$$

- Natural measure of risk-aversion.

Setting: Unknown real-valued distributions $\{\nu_a\}_{a=1,\dots,A}$. At each t, we choose $A_t \in \{1,\dots,A\}$, receive reward $Y_t \sim \nu_{A_t}$.

The **expected regret** $\overline{\mathcal{R}}_T$ gives **no information on the risk** of the strategy and of pulling one arm (no control on the tails):

$$\overline{\mathcal{R}}_T = \sum_{a' \in \mathcal{A}} \Big(\max_{a \in \mathcal{A}} \mathbb{E}_{\nu_a}[X] - \mathbb{E}_{\nu_{a'}}[X] \Big) \mathbb{E}\Big[N_{T,a'} \Big],$$

where $N_{T,a'} = \sum_{t=1}^{A} \mathbb{I}\{A_t = a'\}$.

We are given some λ.
We define the optimal arm $a^\star$ as the one maximizing the **risk aversion** at level λ

$$a^\star \in \operatorname*{argmax}_{a=1,\dots,A} \kappa_{-\lambda,\nu_{a^\star}}.$$

Example: For $\mathcal{N}(\mu,\sigma^2)$ distributions $\kappa_{-\lambda,\nu_{a^\star}} = \mu_a - \frac{\lambda\sigma_a^2}{2}$.
In general it holds $\kappa_{-\lambda,\nu_{a^\star}} \leqslant \mathbb{E}_{\nu_a}[X]$.

- The empirical regret $\mathcal{R}_T(\lambda)$ of π with respect to the strategy $\star$ that constantly pulls arm $a^\star$ is:

$$\mathcal{R}_T(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{T} X_{i,a^\star} - \sum_{a=1}^{A} \sum_{i=1}^{N^\pi_{T,a}} X_{i,a}\,, \qquad (10)$$

where $X_{i,a}$ denotes the i^{th} (i.i.d) sample from arm a.

- The empirical regret $\mathcal{R}_T(\lambda)$ of π with respect to the strategy $\star$ that constantly pulls arm $a^\star$ is:

$$\mathcal{R}_T(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{T} X_{i,a^\star} - \sum_{a=1}^{A} \sum_{i=1}^{N^\pi_{T,a}} X_{i,a}\,, \qquad (10)$$

where $X_{i,a}$ denotes the i^{th} (i.i.d) sample from arm a.

- The risk-averse regret $\overline{\mathcal{R}}_T(\lambda)$ is defined by

$$\begin{aligned} \overline{\mathcal{R}}_T(\lambda) &= \sum_{a\in\mathcal{A}} \Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}\Big)\mathbb{E}\Big[N_{T,a}\Big] \qquad (11)\\ &= \sum_{a\in\mathcal{A}} \boldsymbol{\Delta}_{\mathbf{a}}\mathbb{E}\Big[N_{T,a}\Big] \end{aligned}$$

- The empirical regret $\mathcal{R}_T(\lambda)$ of π with respect to the strategy $\star$ that constantly pulls arm $a^\star$ is:

$$\mathcal{R}_T(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{T} X_{i,a^\star} - \sum_{a=1}^{A} \sum_{i=1}^{N^\pi_{T,a}} X_{i,a} \,, \tag{10}$$

where $X_{i,a}$ denotes the i^{th} (i.i.d) sample from arm a.

- The risk-averse regret $\overline{\mathcal{R}}_T(\lambda)$ is defined by

$$\overline{\mathcal{R}}_T(\lambda) = \sum_{a\in\mathcal{A}} \Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}\Big)\mathbb{E}\Big[N_{T,a}\Big] \tag{11}$$

$$= \sum_{a\in\mathcal{A}} \boldsymbol{\Delta}_{\mathbf{a}}\mathbb{E}\Big[N_{T,a}\Big]$$

- We study both (10) and (12) since they offer interesting and easy interpretations.

Tradeoff between
being risk-averse versus **targeting high reward**:

Tradeoff between
being risk-averse versus **targeting high reward**:

- If **"not enough"** risk-averse (protect against light lower tails only but arms have fat lower tails),
  $\implies$ we may get **high-regret**.

Tradeoff between
being risk-averse versus **targeting high reward**:

- If **“not enough”** risk-averse (protect against light lower tails only but arms have fat lower tails),
  $\Longrightarrow$ we may get **high-regret**.
- If **“too much”** risk-averse (protect against fat lower tails but all arms have light lower tails),
  $\Longrightarrow$ a less cautious algorithm can (e.g. UCB) get better rewards.

λ defines the **risk-aversion** of the problem, irrespectively of the actual distributions of the environment.

If we design an optimal algorithm for risk-averse level λ,

- Some environments will be **"simpler"** (a less risk-averse algorithm, e.g. UCB, gets better rewards),

λ defines the **risk-aversion** of the problem, irrespectively of the actual distributions of the environment.

If we design an optimal algorithm for risk-averse level λ,

- Some environments will be **"simpler"** (a less risk-averse algorithm, e.g. UCB, gets better rewards),
- Others will be **"harder"** (a more risk-averse algorithm, e.g. Exp3, gets better rewards).

λ defines the **risk-aversion** of the problem, irrespectively of the actual distributions of the environment.

If we design an optimal algorithm for risk-averse level λ,

- Some environments will be **"simpler"** (a less risk-averse algorithm, e.g. UCB, gets better rewards),
- Others will be **"harder"** (a more risk-averse algorithm, e.g. Exp3, gets better rewards).

λ defines the **risk-aversion** of the problem, irrespectively of the actual distributions of the environment.

If we design an optimal algorithm for risk-averse level λ,

- Some environments will be **"simpler"** (a less risk-averse algorithm, e.g. UCB, gets better rewards),
- Others will be **"harder"** (a more risk-averse algorithm, e.g. Exp3, gets better rewards).

We want to defeat e.g. not-enough cautious algorithms in hard environments.

Risk-aversion for a fixed λ (often justified in practical applications).

1. Decompose empirical regret with number of pulls of sub-optimal arms (allows **robust** analysis).
2. Introduce RAUCB for risk-aversion.
3. Get numerically efficient dual formulation.
4. Control both risk-averse and empirical regret.

Theorem (Generic decomposition of the empirical regret)

Let the event that strategy π **does not pull sub-optimal arms too often** be (for some non-negative constants $\{u_a\}_{a=1,\dots,A}$):

$$\Omega \stackrel{\text{def}}{=} \Big\{\exists a \neq a^\star \;:\; N_{T,a} > u_a\Big\}.$$

For all $\delta \in (0,1)$, with probability higher than $1-\delta-\mathbb{P}(\Omega)$, the empirical regret of π is upper bounded by

$$\mathcal{R}_T(\lambda) \leqslant \sum_{a\neq a^\star} \Delta_a u_a + \Big(\dots \mathbf{m}^-_{\lambda,\nu_{a^\star}} + \frac{\dots}{\lambda}\Big) + \inf_{\lambda'>0}\Big\{\dots \mathbf{m}^+_{\lambda',\nu_{a^\star}} + \frac{\dots}{\lambda'}\Big\}.$$

- First term: essentially risk-averse regret.
- Other terms: tails.

- Consider all rewards are upper bounded by B (known).

- Consider all rewards are upper bounded by B (known).
- RAUCB selects at time $t+1$ arm $A_{t+1} = \mathrm{argmax}_{a\in\mathcal{A}}\, U_t(a)$,

- Consider all rewards are upper bounded by B (known).
- RAUCB selects at time $t+1$ arm $A_{t+1} = \mathrm{argmax}_{a\in\mathcal{A}} U_t(a)$, where $U_t(a)$ is an **upper confidence bound** on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\textbf{def}}{=} \sup_{\nu\in\mathcal{P}(\mathbb{R}_B)} \left\{ \kappa_{-\lambda,\nu} \ : \ \mathbf{K}(\widehat{\nu}_{\mathbf{t}}(\mathbf{a}), \kappa_{-\lambda,\nu}) \leqslant \frac{f(t)}{N_{t,a}} \right\},$$

- Consider all rewards are upper bounded by B (known).
- RAUCB selects at time $t+1$ arm $A_{t+1} = \mathrm{argmax}_{a\in\mathcal{A}} U_t(a)$, where $U_t(a)$ is an **upper confidence bound** on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\textbf{def}}{=} \sup_{\nu\in\mathcal{P}(\mathbb{R}_B)} \left\{ \kappa_{-\lambda,\nu} \;:\; \mathbf{K}(\widehat{\nu}_{\mathbf{t}}(\mathbf{a}), \kappa_{-\lambda,\nu}) \leqslant \frac{f(t)}{N_{t,a}} \right\},$$

with parameter $f(t) \simeq \log(t)$ and where we introduced

$$\mathrm{K}(\widehat{\nu}_{\mathrm{t}}(\mathrm{a}), r) \stackrel{\mathrm{def}}{=} \inf_{\nu\in\mathcal{M}(\mathbb{R}_B)} \left\{ \mathrm{KL}(\widehat{\nu}_t(a)||\nu) \;:\; \kappa_{-\lambda,\nu} \geqslant r \right\}.$$

- Consider all rewards are upper bounded by B (known).
- RAUCB selects at time $t+1$ arm $A_{t+1} = \mathrm{argmax}_{a \in \mathcal{A}} U_t(a)$, where $U_t(a)$ is an **upper confidence bound** on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\textbf{def}}{=} \sup_{\nu \in \mathcal{P}(\mathbb{R}_B)} \left\{ \kappa_{-\lambda,\nu} \ : \ \mathbf{K}(\widehat{\nu}_\mathbf{t}(\mathbf{a}), \kappa_{-\lambda,\nu}) \leqslant \frac{f(t)}{N_{t,a}} \right\},$$

with parameter $f(t) \simeq \log(t)$ and where we introduced

$$\mathrm{K}(\widehat{\nu}_\mathrm{t}(\mathrm{a}), r) \stackrel{\mathrm{def}}{=} \inf_{\nu \in \mathcal{M}(\mathbb{R}_B)} \left\{ \mathrm{KL}(\widehat{\nu}_t(a) || \nu) \ : \ \kappa_{-\lambda,\nu} \geqslant r \right\}.$$

- **Note 1:** Using mean-based confidence bounds is useless here.

2. Risk-averse algorithm: RAUCB

- Consider all rewards are upper bounded by B (known).
- RAUCB selects at time $t+1$ arm $A_{t+1} = \mathrm{argmax}_{a\in\mathcal{A}} U_t(a)$, where $U_t(a)$ is an **upper confidence bound** on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\text{def}}{=} \sup_{\nu\in\mathcal{P}(\mathbb{R}_B)} \left\{\kappa_{-\lambda,\nu} \,:\, \mathbf{K}(\widehat{\nu}_{\mathbf{t}}(\mathbf{a}), \kappa_{-\lambda,\nu}) \leqslant \frac{f(t)}{N_{t,a}}\right\},$$

with parameter $f(t) \simeq \log(t)$ and where we introduced

$$\mathrm{K}(\widehat{\nu}_{\mathrm{t}}(\mathrm{a}), r) \stackrel{\mathrm{def}}{=} \inf_{\nu\in\mathcal{M}(\mathbb{R}_B)} \left\{\mathrm{KL}(\widehat{\nu}_t(a)||\nu) \,:\, \kappa_{-\lambda,\nu} \geqslant r\right\}.$$

- **Note 1:** Using mean-based confidence bounds is useless here.
- **Note 2:** Similarly to bandits, we do not need to estimate $\kappa_{-\lambda,\nu_a}$ (+ it would too loose here).

$\mathbf{K}(\widehat{\nu}_t(a), r)$ and then $U_t(a)$ can be **solved numerically** (deeply linked to numerically efficient dual formulation considered for standard MAB, e.g. Borwein-Lewis, 91, Harari-Kermadec, 06):

Lemma (Dual formulation)

Let $\widehat{\nu}_n$ be an empirical distribution built with n atoms $\{x_i\}_{1\leqslant i\leqslant n}$. Then the following dual formulation holds

$$\mathbf{K}(\widehat{\nu}_n, r) = \max_{0\leqslant\gamma^\star\leqslant\frac{\lambda}{1-e^{-\lambda(B-r)}}} \left\{\frac{1}{n}\sum_{i=1}^{n}\log\left(1-\frac{\gamma^\star}{\lambda}\left(1-e^{-\lambda(x_i-r)}\right)\right)\right\}.$$

Theorem (Regret of RAUCB)

The expected regret of RAUCB (for suitable f), is bounded by

$$\overline{\mathcal{R}}_T(\lambda) \leqslant 5 \sum_{a \neq a^\star} \frac{(1+\varepsilon_a)\Delta_a}{\mathbf{K}_a} \log(T) + O\Big(1\Big).$$

The empirical regret of RAUCB is bounded with high probability, for sub-Gaussians distributions of rewards (includes bounded as special case) and risk-aversion $\lambda = \Theta(\log(T)^{-1/2})$ as

$$\mathcal{R}_T(\lambda) \leqslant 5 \sum_{a \neq a^\star} \frac{(1+\varepsilon_a)\Delta_a}{\mathbf{K}_a} \log(T) + O\Big(\sqrt{\log(T)}\Big).$$

- `RAUCB` tuned with **known horizon** T (not anytime).

Comments about the result

- RAUCB tuned with **known horizon** T (not anytime).
- Ratio $\frac{\Delta_a}{\mathbf{K}_a}\log(T)$ similar to best known bounds for the expected regret Burnetas-Katehakis, 96; Cappe et al, 2013,

- RAUCB tuned with **known horizon** T (not anytime).
- Ratio $\frac{\Delta_a}{\mathbf{K}_a}\log(T)$ similar to best known bounds for the expected regret Burnetas-Katehakis, 96; Cappe et al, 2013,
- Choice of λ not too critical: still get $O(\log(T))$ for any λ not depending on T.

Further references

- Sani, A., Lazaric, A., Munos, R. *Risk-aversion in multi-armed bandits*. In NIPS 2012 (pp. 3275-3283).
- Maillard, O-A. *Robust risk-averse stochastic multi-armed bandits* ICML 2013. Springer, Berlin, Heidelberg.
- Galichet, N. PhD. Thesis, Torossian L., PhD. Thesis : Several risk measures (quantiles, expectiles, etc.)
- Baudry, Dorian, et al. *Optimal Thompson Sampling strategies for support-aware CVaR bandits.* International Conference on Machine Learning, 2021.

Table of contents

Slides Edouard:

`https://eleurent.github.io/robust-beyond-quadratic/paper/oral`

Self-check

- **Full** information vs **Partial** information (Bandit, semi-bandit)
- Objectives: Best model vs Best combination of losses vs Best combination of models vs Best sequence
- Convex losses
- Exponential weights, Hedge strategy
- Self-information loss
- Exp-concavity, mixability
- Entropy formula
- Bregman agggregation
- Fixed-share strategy
- Markov-Hedge
- Mixing past posteriors
- Exp3, Importance sampling, Exp3-P, Exp3-IX, Exp4
- Tsallis entropy
- Nash equilibria, Hannan consistency
- Conditional value at risk, mixability gap
- Robust learning

"The more applied you go, the stronger theory you need"

MERCI

odalricambrym.maillard@inria.fr

odalricambrymmaillard.wordpress.com