Ranking and Risk-Aware Reinforcement Learning

—

PhD Defense

Mastane Achab

Télécom Paris, LTCI, Palaiseau

July 10, 2020

Agenda

Agenda

1. Introduction
   - Offline Minimization of the Empirical Risk
   - (Online) Reinforcement Learning
2. Beyond Ranking Aggregation
   - Dimensionality Reduction on Permutations
   - Learning Bucket Orders
3. Risk-Aware Bandits
   - Bandits for Credit Risk
   - Extreme Bandits Revisited
4. Distributional Reinforcement Learning
   - 1-Step Operators
   - Atomic Bellman Equations
5. Perspectives

Empirical Risk Minimization (ERM)

Many ML problems belong to the ERM paradigm [Devroye et al., 1996].

What we really want ...

- Minimize the *true risk*:

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmin}} \mathscr{R}_P(\theta) := \mathbb{E}_{Z \sim P}[\ell(\theta, Z)].$$

- Example - Classification: $Z = (X, Y)$, $\theta =$ "classifier".

What we can compute ...

- Minimize the *empirical risk*:

$$\widehat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}} \widehat{\mathscr{R}}_P(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i).$$

- Training dataset: n independent observations $Z_i \sim P$.

Classification *vs.* Ranking

Binary classification and *bipartite ranking* ([Agarwal et al., 2005]) are two ERM problems with the same type of supervised data:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \ , \qquad \text{valued in } \mathcal{X} \times \{-1, +1\}.$$

The optimal elements $\theta^* \in \Theta$ are given by the posterior probability $\eta(x) = \mathbb{P}\{Y = +1 | X = x\}$.

Binary classification: answer to "$\eta(x) > 0.5$?" for all x

- $\theta =$ classifier $g : \mathcal{X} \to \{-1, +1\}$
- Zero-one loss function: $\ell_{0/1}(g, (x, y)) = \mathbb{I}\{g(x) \neq y\}$

Bipartite ranking: answer to "$\eta(x) > \eta(x')$?" for all x, x'

- $\theta =$ scoring function $s : \mathcal{X} \to \mathbb{R}$
- Maximize the empirical $AUC(s) = \mathbb{P}\{s(X) < s(X') | Y = -1, Y' = +1\}$:

$$\widehat{AUC}_n(s) = \frac{1}{n_+ \cdot n_-} \sum_{i: Y_i = -1} \sum_{j: Y_j = +1} \mathbb{I}\{s(X_i) < s(X_j)\}.$$

Many Rankings for Many Labels

- Bipartite ranking: $Y \in \{$👍, 👎$\}$
- Multipartite ranking [Rajaram and Agarwal, 2005], [S. Clémençon and Vayatis, 2013]: $Y \in \{1\bigstar, \dots, 5\bigstar\}$

Continuous ranking [Clémençon and Achab, 2017]: $Y \in [0,1]$

- Application to implicit feedback [Radlinski and Joachims, 2005]:

$$Y = \frac{\text{listening time of song } X \text{ until skip}}{\text{total duration of song } X} \in [0,1].$$

- For threshold y: binary subproblem with $Z_y = 2\mathbb{I}\{Y > y\} - 1$.
- Continuum of binary subproblems: $IROC(s) = \int_{y=0}^{1} ROC_y(s) dF_Y(y)$, and $IAUC(s) = \int_{y=0}^{1} AUC_y(s) dF_Y(y)$.
- Empirical maximization of $\widehat{IAUC}_n$.

Ranking From Rankings

Ranking Aggregation [Korba et al., 2017]

Summarize a distribution P on the set of permutations $\mathfrak{S}_N$ by a single consensus/median ranking σ^*:

$$\sigma^* = \underset{\sigma \in \mathfrak{S}_N}{\operatorname{argmin}} L_P(\sigma) := \mathbb{E}[d_\tau(\sigma, \Sigma)] = \sum_{\sigma(i) < \sigma(j)} p_{j,i},$$

with Kendall's tau distance:

$$d_\tau(\sigma, \sigma') = \sum_{1 \le i < j \le N} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

and pairwise probabilities $p_{i,j} = \mathbb{P}_{\Sigma \sim P}\{\Sigma(i) < \Sigma(j)\}$.

Extension to Partial Orders

Definition (Bucket Order)

It is an ordered partition $\mathscr{C} = (\mathscr{C}_1, \dots, \mathscr{C}_K)$ of the N items $\{1, \dots, N\}$.

Figure : This bucket order constrains football teams to be preferred over hockey's. It has size $K = 2$ and shape $\lambda = (4, 2)$.

Learning bucket orders by ERM [Achab et al., 2018b]

Find the bucket order $\mathscr{C}^*$ (of given *size* K and *shape* λ) with minimal distortion measure:

$$\mathscr{C}^* = \underset{\mathscr{C} \in \mathbf{C}_{K,\lambda}}{\operatorname{argmin}} \Lambda_P(\mathscr{C}) := \min_{P' \in \mathbf{P}_{\mathscr{C}}} W_{d_\tau,1}(P, P') = \sum_{1 \le k < l \le K} \sum_{(i,j) \in \mathscr{C}_k \times \mathscr{C}_l} p_{j,i}.$$

1^{st} Relaxation of ERM: Biased Data

Question: What if the training dataset $\{Z'_1, \ldots, Z'_n\}$ is i.i.d. sampled from P' (training distrib.) $\neq P$ (testing distrib.) ?

Examples of Sample Selection Bias

- censored data [Kaplan and Meier, 1958]
- Positive-Unlabeled learning [du Plessis et al., 2014]
- varying class probabilities, stratified data [Bekker and Davis, 2018]

Weighted ERM (WERM) [Vogel, Achab, et al., 2020]

Minimize the *weighted empirical risk*:

$$\widetilde{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \widetilde{\mathscr{R}}_{P'}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \underbrace{\widehat{\Phi}(Z'_i)}_{\approx \frac{dP}{dP'}(Z'_i)} \cdot \ell(\theta, Z'_i).$$

2^{nd} Relaxation: Non-I.I.D. Data

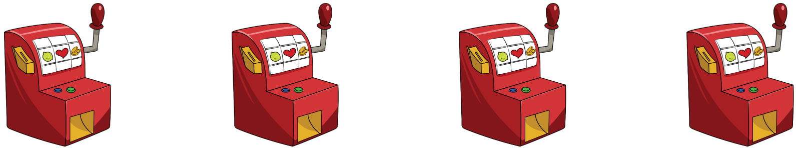
In *online learning*, the training data is collected through time, depending on the learner's decisions:

- *active learning* [Minsker, 2012], [Locatelli et al., 2017]: faster convergence rates than offline ERM,
- *multi-armed bandits* (MAB) [Bubeck et al., 2012],
- *reinforcement learning* (RL) [Sutton and Barto, 2018].

Agenda

1. Introduction
   - Offline Minimization of the Empirical Risk
   - (Online) Reinforcement Learning
2. Beyond Ranking Aggregation
   - Dimensionality Reduction on Permutations
   - Learning Bucket Orders
3. Risk-Aware Bandits
   - Bandits for Credit Risk
   - Extreme Bandits Revisited
4. Distributional Reinforcement Learning
   - 1-Step Operators
   - Atomic Bellman Equations
5. Perspectives

The Casino Dilemma



$\mathbb{E}[\nu_1] = \mu_1$ $\mathbb{E}[\nu_2] = \mu_2$ $\mathbb{E}[\nu_3] = \mu_3$ $\mathbb{E}[\nu_4] = \mu_4$

Figure : "Which slot machine should I choose?"

Stochastic Multi-Armed Bandit (MAB)

At each time $t \in \{1, \ldots, T\}$,

- pull an arm $A_t \in \{1, \ldots, K\}$,
- receive reward $X_{A_t,t} \sim \nu_{A_t}$.

Minimize the *regret*: $R_T = \sum_{a=1}^{K} \mathbb{E}[N_a(T)] \cdot (\mu_{a^*} - \mu_a)$.

Cautious Bandits

Risk-sensitive MAB

Mean reward $\mu_a = \mathbb{E}[\nu_a]$ replaced by alternative risk-measures such as:

- quantiles in [Szorenyi et al., 2015],
- the CVaR in [Galichet et al., 2013] and [Kolla et al., 2019],
- a mean-variance tradeoff in [Sani et al., 2012], generalized in [Maillard, 2013].

In environmental or financial applications, *extreme rewards* are sometimes more relevant than mean values [Beirlant et al., 2006].

Max K-Armed Bandits [Cicirello and Smith, 2005]

- Maximize: $\mathbb{E}[\max_{1\leq t\leq T} X_{A_t,t}]$.
- ... or equivalently, *minimize the extreme regret*:

$$R_T = \max_{1\leq a\leq K} \mathbb{E}\left[\max_{1\leq t\leq T} X_{a,t}\right] - \mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t}\right]$$

Max K-Armed Bandits for Pareto Tails

Max K-armed bandits for Pareto-like distributions in [Carpentier and Valko, 2014].

Contributions in [Achab et al., 2017]

- "Explore-Then-Commit" (ETC) variant of ExtremeHunter ([Carpentier and Valko, 2014]).
- For both ExtremeHunter and ExtremeETC: refined extreme regret analysis + tight lower bound.
- Reduction to MAB by truncating the rewards: $X_{\text{truncated}} = X \cdot \mathbb{I}\{X > u\}$.

Learning Distributions in a Dynamic Environment

Question: Why not learning the whole distribution, instead of just a risk-sensitive measure?
⟶ Distributional reinforcement learning (DRL) [Bellemare et al., 2017].

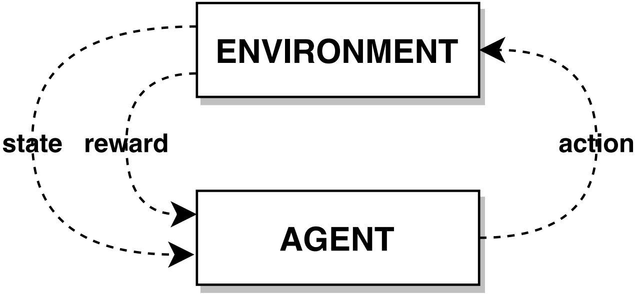


Figure : MAB is a particular case of RL.

The MDP Model of RL

Markov decision process (MDP)

A *Markov decision process (MDP)* is described by a tuple $(\mathscr{X}, \mathscr{A}, P, R)$

- countable state space $\mathscr{X}$,
- countable action space $\mathscr{A}$,
- transition kernel $P : \mathscr{X} \times \mathscr{A} \to \mathscr{P}(\mathscr{X})$,
- distributional reward function $R : \mathscr{X} \times \mathscr{A} \to \mathscr{P}(\mathbb{R})$.

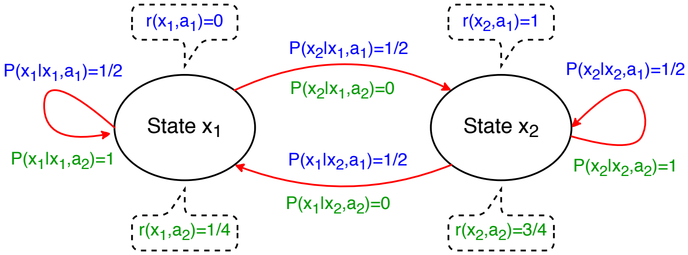


Figure : Example of MDP with deterministic rewards: $R(x,a) = \delta_{r(x,a)}$.

Average Performance of a Policy

Distributional Discounted Return

For a discount factor $\gamma \in [0,1)$, the distributional discounted return $Z^\pi(x,a)$ of a policy π is the *probability distribution* of:

$$\sum_{t=0}^{\infty} \gamma^t R_t \text{ given that } X_0 = x, A_0 = a,$$

and for all $t \in \mathbb{N}, R_t \sim R(X_t, A_t), X_{t+1} \sim P(\cdot|X_t, A_t), A_{t+1} \sim \pi(\cdot|X_{t+1})$.

How good (in expectation) is a policy π?

State-Action Value Function: for all $(x,a) \in \mathscr{X} \times \mathscr{A}$,

$$Q^\pi(x,a) = \mathbb{E}_{Z_0 \sim Z^\pi(x,a)}[Z_0] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \;\middle|\; X_0 = x, A_0 = a, \pi\right]$$

An Atomic Extension of Bellman's Equations

Atomic Bellman Equations (Chap. VII)

- The $N \geq 1$ atoms $\Theta_1^\pi(x,a), \ldots, \Theta_N^\pi(x,a)$ are "conditional expectations" summarizing the distribution $Z^\pi(x,a)$.
- They verify: for all x, a, for all $1 \leq i \leq N$,

$$\Theta_i^\pi(x,a) = \text{Function}\left(\left\{\Theta_j^\pi(x',a') : x', a', j\right\}\right).$$

- $\longrightarrow$ Atomic temporal difference algorithm.

Agenda

Bucket Ranking

Bucket Order

A bucket order $\mathcal{C} = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ is an ordered partition of $\{1, \dots, N\}$:

- $\mathcal{C}_k$'s disjoint non empty subsets of $\{1, \dots, N\}$
- $\bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \dots, N\}$

$\mathcal{C}$ is described by its *size* K, and its *shape* $\lambda = (\#\mathcal{C}_1, \dots, \#\mathcal{C}_K)$.

Question: How much does P violate the constraints of $\mathcal{C}$?

$$\longrightarrow \text{Distortion:} \quad \Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_\tau,1}(P, P') = \sum_{1 \le k < l \le K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} p_{j,i},$$

where any $P' \in \mathbf{P}_{\mathcal{C}}$ is described by $d_{\mathcal{C}} = \prod_{1 \le k \le K} \#\mathcal{C}_k! - 1 \le N! - 1$ parameters ($d_{\mathcal{C}}$ is the *dimensionality* of $\mathbf{P}_{\mathcal{C}}$).

Dimension-Distortion Tradeoff

The smaller the dimension, the larger the distortion ...

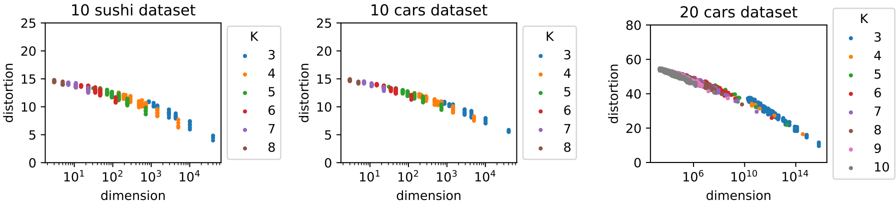


Figure : Dimension-Distortion plot for different bucket sizes on real-world preference datasets.

Agenda

Learning Buckets From Pairwise Comparisons

ERM Setting

Training sample: $\Sigma_1, \ldots, \Sigma_n$ i.i.d. from P.

- Empirical pairwise probabilities:

$$\widehat{p}_{i,j} = \frac{1}{n} \sum_{s=1}^{n} \mathbb{I}\{\Sigma_s(i) < \Sigma_s(j)\}.$$

- Empirical distortion of any bucket order $\mathscr{C}$:

$$\widehat{\Lambda}_n(\mathscr{C}) = \Lambda_{\widehat{P}_n}(\mathscr{C}) = \sum_{1 \leq k < l \leq K} \sum_{(i,j) \in \mathscr{C}_k \times \mathscr{C}_l} \widehat{p}_{j,i}. \quad (1)$$

- **Remark:** Alternatively, observe only pairwise comparisons.

Excess of Distortion for Given Shape

Empirical distortion minimizer $\widehat{C}_{K,\lambda}$ is solution of:

$$\min_{\mathscr{C}\in\mathbf{C}_{K,\lambda}} \widehat{\Lambda}_n(\mathscr{C}),$$

where $\mathbf{C}_{K,\lambda}$ set of bucket orders $\mathscr{C}$ of size K and shape λ (i.e. $\#\mathscr{C}_k = \lambda_k$ for all $1 \leq k \leq K$).

Theorem 1 in [Achab et al., 2018b]

For all $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\Lambda_P(\widehat{C}_{K,\lambda}) - \inf_{\mathscr{C}\in\mathbf{C}_{K,\lambda}} \Lambda_P(\mathscr{C}) \leq \beta(N,\lambda) \times \sqrt{\frac{\log(\frac{1}{\delta})}{n}}.$$

Balancing Dimension & Distortion

BuMeRank Algorithm

- Start with ranking aggregation:

$$\mathscr{C}(0) = \left(\{\sigma_P^{*-1}(1)\}, \ldots, \{\sigma_P^{*-1}(N)\}\right), \quad \text{dimension } d_{\mathscr{C}(0)} = 0.$$

- For step $j \geq 0$, merge two adjacent cells:

$$\mathscr{C}(j+1) = (\mathscr{C}_1(j), \ldots, \mathscr{C}_{k-1}(j), \mathscr{C}_k(j) \cup \mathscr{C}_{k+1}(j), \mathscr{C}_{k+2}(j), \ldots, \mathscr{C}_K(j)).$$

- The agglomerative stage $\mathscr{C}(j) \to \mathscr{C}(j+1)$ increases the dimension:

$$d_{\mathscr{C}(j+1)} = (d_{\mathscr{C}(j)} + 1) \times \binom{\#\mathscr{C}_k(j) + \#\mathscr{C}_{k+1}(j)}{\#\mathscr{C}_k(j)} - 1,$$

- while reducing the distortion by:
$\Lambda_P(\mathscr{C}(j)) - \Lambda_P(\mathscr{C}(j+1)) = \sum_{i \in \mathscr{C}_k(j), j \in \mathscr{C}_{k+1}(j)} p_{j,i}.$

Agenda

Credit Risk Management

Model: The population (of credit applicants) is stratified over $K \geq 1$ categories.

Iterations

At each time $1 \leq t \leq T$,

- a client of each category $a \in \{1, \ldots, K\}$ asks for a credit of amount τ_a,
- the bank chooses a subset $\mathscr{A}_t \subseteq \{1, \ldots, K\}$, and pays τ_a for each chosen category $a \in \mathscr{A}_t$,
- then, the bank receives the corresponding reimbursements: $X_{a,t} = (1 + \rho_a)\tau_a \cdot B_{a,t}$ with Bernoulli r.v. $B_{a,t} \sim \mathscr{B}(p_a)$.

Reimbursement ... or credit default!

- In case of credit default: $B_{a,t} = 0 \Longrightarrow X_{a,t} = 0$ (no refunding!).
- Otherwise, $B_{a,t} = 1$, i.e. the bank gets refunded $(1 + \rho_a)\tau_a$.
- Category a is "profitable" if: $\mathbb{E}[X_{a,t}] > \tau_a \Longleftrightarrow p_a > \frac{1}{1+\rho_a}$.

Make Profit, Not Reward

Profitable Bandits [Achab et al., 2018a]

At each time $t \in \{1, \dots, T\}$,

- pull a subset of arms $\mathscr{A}_t \subseteq \{1, \dots, K\}$,
- for all pulled arms $a \in \mathscr{A}_t$,
  - pay (known) price τ_a (e.g. loan financed by a bank),
  - receive reward $X_{a,t} \sim \nu_a$ (loan reimbursement + interest ... or default!).

Maximize expected profits: $\mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in \mathscr{A}_t} (X_{a,t} - \tau_a)\right]$.

Here, the regret is:

$$R_T = \sum_{a \in \mathscr{A}^*} \Delta_a \cdot (T - \mathbb{E}[N_a(T)]) - \sum_{a \notin \mathscr{A}^*} \Delta_a \cdot \mathbb{E}[N_a(T)],$$

with (unknown) expected profit $\Delta_a = \mu_a - \tau_a$, and set of profitable arms:

$$\mathscr{A}^* = \Big\{a \in \{1, \dots, K\} : \Delta_a > 0\Big\}.$$

$R_T \gtrsim Constant \times \log T$

Lower Bound: Theorem 1 in [Achab et al., 2018a]

Any *uniformly efficient* profitable bandits strategy produces a regret R_T asymptotically lower bounded as follows:

$$\liminf_{T \to \infty} \frac{R_T}{\log T} \geq \sum_{a \notin \mathscr{A}^*} \frac{|\Delta_a|}{\mathscr{K}_{\inf}(\nu_a, \tau_a, \mathscr{D}_a)},$$

where $\mathscr{K}_{\inf}(\nu_a, x, \mathscr{D}_a) = \inf\left\{\mathrm{KL}(\nu_a, \nu'_a) : \nu'_a \in \mathscr{D}_a \text{ and } \mathbb{E}_{X' \sim \nu'_a}[X'] > x\right\}.$

Pull Arm If Index Above Threshold

Algorithm 1 Profitable bandits index policy

Require: time horizon T, thresholds $(\tau_a)_{a\in\{1,\dots,K\}}$.

1: **Initialize:** Pull all arms: $\mathscr{A}_1 = \{1,\dots,K\}$.
2: **for** $t = 1$ **to** $T-1$ **do**
3: Compute index $u_a(t)$ for all arms $a \in \{1,\dots,K\}$.
4: Pull arms in $\mathscr{A}_{t+1} = \{a \in \{1,\dots,K\} : u_a(t) \geq \tau_a\}$.
5: **end for**

Asymptotically optimal algorithms ($R_T \lesssim \sum_{a\notin\mathscr{A}^*} \frac{|\Delta_a|}{\mathscr{K}_{\inf}(\nu_a,\tau_a,\mathscr{D}_a)} \log T$):

- the kl-UCB index [Garivier and Cappé, 2011]:

$$u_a(t) = \sup\left\{q > \widehat{\mu}_a(t) : N_a(t) d(\widehat{\mu}_a(t), q) \leq \log t + c \log\log t\right\},$$

- the Bayes-UCB index [Kaufmann et al., 2012]:

$$u_a(t) = Q(1 - 1/(t(\log t)^c), \pi_{a,t}),$$

- the Thompson Sampling index [Thompson, 1933]: $u_a(t) = \mu(\theta_a(t))$.

Experiment - Profitable Bandits

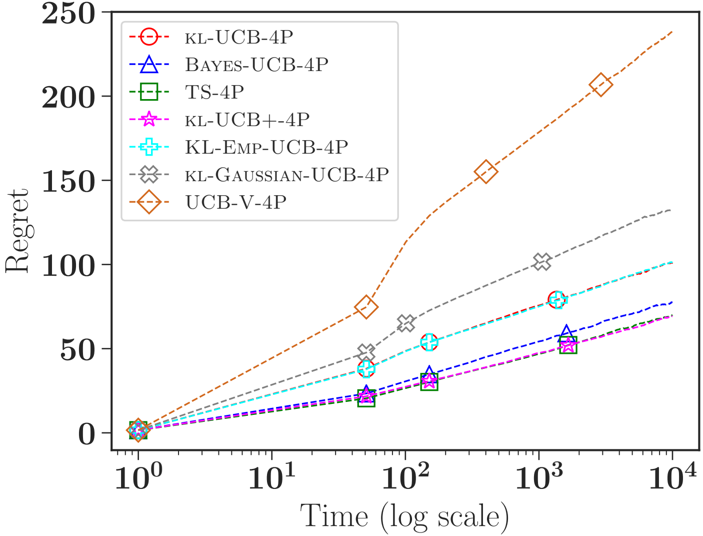


Figure : Regret as a function of time in the Bernoulli scenario.

Agenda

Small Tail Index α Means "Heavy-Tailed"

Definition (2^{nd}-order Pareto Distributions)

It is a distribution with c.d.f. F that satisfies: $\forall x \geq 0$,

$$|1 - Cx^{-\alpha} - F(x)| \leq C' x^{-\alpha(1+\beta)}.$$

Assumptions [Carpentier and Valko, 2014]

- The distributions $\nu_1, \ldots, \nu_K$ of the K arms are 2^{nd}-order Pareto.
- For any arm $1 \leq a \leq K$,
  - tail index $\alpha_a > 1$ (finite mean),
  - $\beta_a \geq b > 0$.

Property

For T large enough, *the optimal arm* has the smallest tail index:

$$a^* = \operatorname*{argmin}_{1 \leq a \leq K} \alpha_a = \operatorname*{argmax}_{1 \leq a \leq K} \mathbb{E}\left[\max_{1 \leq t \leq T} X_{a,t}\right].$$

On the Hunt of Extremes

ExtremeHunter [Carpentier and Valko, 2014]

- Main idea: UCB for $\widehat{1/\alpha_a}$.
- Upper bound for the extreme regret: $R_T = O\left(T^{\frac{1}{(1+b)\alpha_{a^*}}}\right)$.

Our contribution [Achab et al., 2017]

- Refined upper bound for ExtremeHunter and ExtremeETC:

$$R_T = O\left(\log(T)^{\frac{2(2b+1)}{b}} \cdot T^{-\left(1-\frac{1}{\alpha_{a^*}}\right)} + T^{-\left(b-\frac{1}{\alpha_{a^*}}\right)}\right).$$

- Lower bound (tight if $b \geq 1$):

$$R_T = \Omega\left(\log(T)^{\frac{2(2b+1)}{b}} \cdot T^{-\left(1-\frac{1}{\alpha_{a^*}}\right)}\right).$$

Reduction to MAB with Truncated Rewards

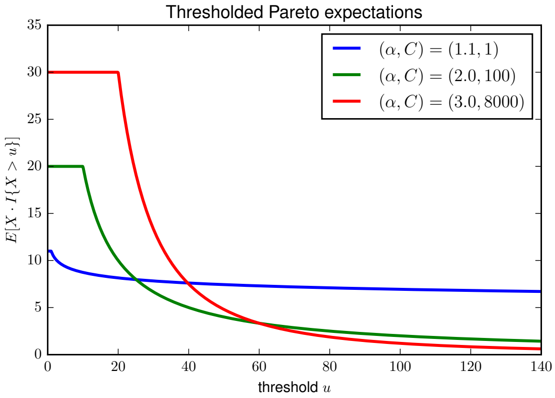


Figure : Expected truncated rewards $\mathbb{E}[X_a \mathbb{I}\{X_a > u\}]$ as a function of threshold u.

- Lemma 6 in [Achab et al., 2017]: For threshold u large enough,

$$a^* = \underset{1 \le a \le K}{\operatorname{argmin}} \, \alpha_a = \underset{1 \le a \le K}{\operatorname{argmax}} \, \mathbb{E}[X_a \cdot \mathbb{I}\{X_a > u\}].$$

Truncating *vs.* ExtremeETC

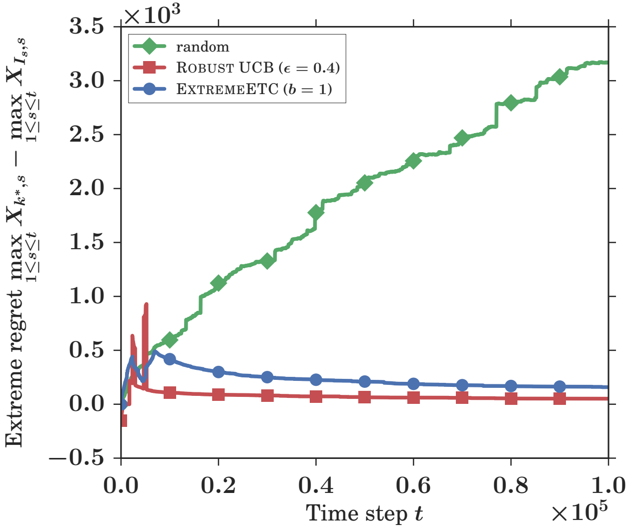


Figure : Extreme regret across time for different strategies.

Agenda

Distributional Bellman Operators

Question: Do we know DRL operators that are contractions?

Yes, for distributional policy evaluation [Bellemare et al., 2017]

- The *distributional Bellman operator* $\mathscr{T}^{\pi}$: for any $Z : \mathscr{X} \times \mathscr{A} \to \mathscr{P}(\mathbb{R})$,

$$\mathscr{T}^{\pi} Z(x,a) = \mathrm{Distrib}\big(R_0 + \gamma Z_1\big) \text{ with } R_0 \sim R(x,a), Z_1 \sim Z(X_1, A_1).$$

- Lemma 3 in [Bellemare et al., 2017]: $\mathscr{T}^{\pi}$ is a γ-contraction in sup-Wasserstein distance $\widetilde{W}_p$.
- Distributional Bellman equation: $Z^{\pi} = \mathscr{T}^{\pi} Z^{\pi}$.

... and for distributional control?

The answer is "No" in Proposition 1 in [Bellemare et al., 2017].

1-Step Distributional Bellman Operators (1/2)

Our contribution: We introduce 2 new DRL operators (1 for policy evaluation & 1 for control), that are both contractions.

Distributional policy evaluation

- The *1-Step Distributional Bellman Operator* $\mathbb{T}^{\pi}$:

$$\mathbb{T}^{\pi} Z(x,a) = \text{Distrib}(R_0 + \gamma \mathbb{E}[Z_1 | X_1, A_1]),$$

where $R_0 \sim R(x,a), X_1 \sim P(\cdot|x,a), A_1 \sim \pi(\cdot|X_1), Z_1 \sim Z(X_1,A_1)$.

- Lemma 1 in Chap. VII: $\mathbb{T}^{\pi}$ is a γ-contraction in $\widetilde{W}_p$.
- If deterministic rewards $R(x,a) = \delta_{r(x,a)}$, fixed point of $\mathbb{T}^{\pi}$:

$$(x,a) \mapsto \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} P(x'|x,a)\pi(a'|x')\delta_{r(x,a)+\gamma Q^{\pi}(x',a')}.$$

1-Step Distributional Bellman Operators (2/2)

... and our new DRL operator for control is ...

Distributional control

- The *1-Step Distributional Bellman Optimality Operator* $\mathbb{T}$: for all $Z : \mathscr{X} \times \mathscr{A} \to \mathscr{P}(\mathbb{R})$,

$$\mathbb{T}Z(x,a) = \text{Distrib}(R_0 + \gamma \max_{a' \in \mathscr{A}} \mathbb{E}[Z_{1,a'} | X_1]),$$

  where $R_0 \sim R(x,a), X_1 \sim P(\cdot|x,a), Z_{1,a'} \sim Z(X_1, a')$.
- Lemma 2 in Chap. VII: $\mathbb{T}$ is a γ-contraction in $\widetilde{W}_p$.
- If $R(x,a) = \delta_{r(x,a)}$, fixed point of $\mathbb{T}$:

$$(x,a) \mapsto \sum_{x' \in \mathscr{X}} P(x'|x,a) \delta_{r(x,a) + \gamma \max_{a'} Q^*(x',a')}.$$

Agenda

Projected Bellman Operators

Let's now focus on the (full) distributional Bellman operator $\mathscr{T}^\pi$ ...
Question: In practice, how to (approximately) compute $\mathscr{T}^\pi$?

Quantile regression approach in [Dabney et al., 2018]

- Projected Bellman operator $\Pi_{1,N} \circ \mathscr{T}^\pi$, with W_1-projection $\Pi_{1,N}$:

$$\Pi_{1,N} Z(x,a) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\Theta_i(x,a)}, \text{ with } \Theta_i(x,a) = F_{x,a}^{-1}\left(\frac{2i-1}{2N}\right).$$

- Prop. 2 in [Dabney et al., 2018]: $\Pi_{1,N} \circ \mathscr{T}^\pi$ is a γ-contraction in $\widetilde{W}_\infty$.

Our approach: W_2-projection $\Pi_{2,N}$

- The W_2-optimal atoms are *trimmed means*:
$\Theta_i(x,a) = N \int_{\tau=\frac{i-1}{N}}^{\frac{i}{N}} F_{x,a}^{-1}(\tau) d\tau \approx \mathbb{E}\left[Z_0 \middle| F_{x,a}^{-1}\left(\frac{i-1}{N}\right) \leq Z_0 \leq F_{x,a}^{-1}\left(\frac{i}{N}\right)\right].$
- Corollary 1 in Chap. VII: $\Pi_{2,N} \circ \mathscr{T}^\pi$ is a γ-contraction in $\widetilde{W}_\infty$.

Atomic Bellman Equation

- Proposition 2 in Chap. VII: For determinisic rewards $R(x,a) = \delta_{r(x,a)}$, the fixed point Z_{Θ^π} of the *atomic Bellman operator* $\Pi_{2,N} \circ \mathcal{T}^\pi$ solves the *atomic Bellman equation*: for all x, a, $1 \leq i \leq N$,

$$\Theta_i^\pi(x,a) = r(x,a) + \gamma N \sum_{x',a',j} \mu_i^\pi(\Theta^\pi, x, a, \Theta_j^\pi(x',a')) \cdot \Theta_j^\pi(x',a'),$$

- with "quantile level coefficients":

$$\mu_i^\pi(\Theta^\pi, x, a, \theta) = \text{Length}\left(\left[\frac{i-1}{N}, \frac{i}{N}\right] \bigcap \left[H_{x,a}^\pi(\theta), G_{x,a}^\pi(\theta)\right]\right),$$

- where $H_{x,a}^\pi(\theta) = G_{x,a}^\pi(\theta-)$ and $G_{x,a}^\pi$ is the c.d.f. of $Z_{\Theta^\pi}(X_1, A_1)$:

$$G_{x,a}^\pi(\theta) = \sum_{x',a'} P(x'|x,a)\pi(a'|x') \cdot \frac{1}{N}\sum_{j=1}^N \mathbb{I}\{\Theta_j^\pi(x',a') \leq \theta\}.$$

Atomic Dynamic Programming

Given known transition probabilities $P(\cdot|x,a)$, we recursively apply the atomic Bellman operator.

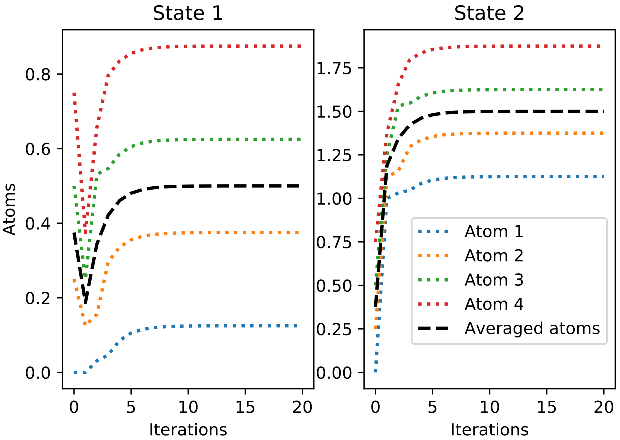


Figure : $\pi(a_1|x) \equiv 1$, $Z^\pi(x_1,a_1) = \text{Uniform}([0,1])$, $Z^\pi(x_2,a_1) = \text{Uniform}([1,2])$.

Atomic Approximation Error

How far is the atomic fixed point Z_{Θ^π} to the original fixed point Z^π?

W_∞-Approximation Error (Proposition 1 in Chap. VII)

$\sup_{x,a} W_\infty(Z^\pi(x,a), Z_{\Theta^\pi}(x,a)) = O\left(\frac{1}{N}\right)$

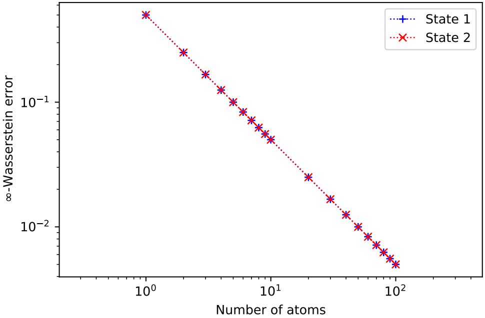


Figure : $W_\infty(Z^\pi(x,a_1), Z_{\Theta^\pi}(x,a_1))$ for the two states $x \in \{x_1, x_2\}$.

Atomic Temporal Difference

Consider a policy π and a single transition $x, a, r(x,a), X_1, A_1$ such that $X_1 \sim P(\cdot|x,a), A_1 \sim \pi(\cdot|X_1)$.

Atomic Temporal-Difference (ATD)

For all $x' \in \mathscr{X}$, $a' \in \mathscr{A}$, $j \in \{1, \dots, N\}$,

(a) $\theta \leftarrow \Theta_j(x', a')$,

(b) $G_{x,a}(\theta) \leftarrow (1-\beta) G_{x,a}(\theta) + \beta \cdot \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}\{\Theta_k(X_1, A_1) \leq \theta\}$,

(c) $H_{x,a}(\theta) \leftarrow (1-\beta) H_{x,a}(\theta) + \beta \cdot \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}\{\Theta_k(X_1, A_1) < \theta\}$,

(d) $\forall 1 \leq i \leq N$, $\mu_i(\Theta, x, a, \theta) \leftarrow \mathrm{Length}\left(\left[\frac{i-1}{N}, \frac{i}{N}\right] \cap [H_{x,a}(\theta), G_{x,a}(\theta)]\right)$.

Then, return the updated atoms in state-action (x,a): for $1 \leq i \leq N$,

$$\Theta_i(x,a) \leftarrow (1-\alpha)\Theta_i(x,a) + \alpha\Big(r(x,a) + \gamma N \sum_{\theta} \mu_i(\Theta, x, a, \theta) \cdot \theta\Big).$$

Experiment - Atomic TD

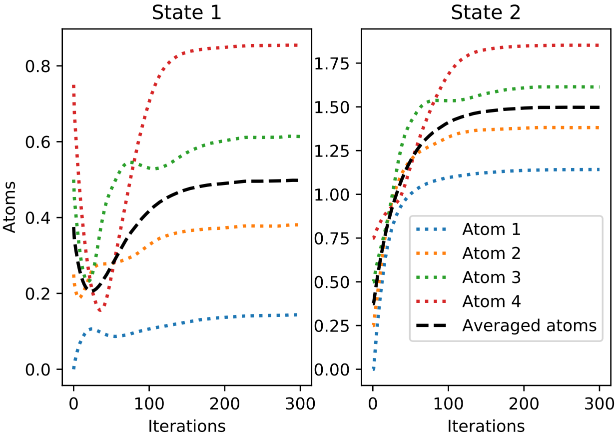


Figure : ATD with learning rates $\alpha = \beta = 0.1$.

Perspectives

- Bucket ranking with Spearman ρ: $d_2(\sigma,\sigma') = \sqrt{\sum_{i=1}^{N}(\sigma(i)-\sigma'(i))^2}$. Proposition 16 in [Achab et al., 2018b]: alternative distortion measure $\Lambda'_P(\mathscr{C}) = \min_{P' \in \mathbf{P}_{\mathscr{C}}} W_{d_2,2}(P,P')$, whose explicit expression involves the *triplet-wise proabilities*:

$$p_{i,j,k} = \mathbb{P}_{\Sigma \sim P}\left\{\Sigma(i) < \Sigma(j) < \Sigma(k)\right\}.$$

- Atomic TD with function approximation for the c.d.f.'s $H_{x,a}(\theta)$ and $G_{x,a}(\theta)$.
- Also, Atomic Q-learning (Chap. VII) by projecting the 1-step distributional Bellman optimality operator.

Achab, M., Clémençon, S., and Garivier, A. (2018a).
Profitable bandits.
arXiv preprint arXiv:1805.02908.

Achab, M., Clémençon, S., Garivier, A., Sabourin, A., and Vernade, C. (2017).
Max k-armed bandit: On the extremehunter algorithm and beyond.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 389–404. Springer.

Achab, M., Korba, A., and Clémençon, S. (2018b).
Dimensionality reduction and (bucket) ranking: a mass transportation approach.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. (2005).
Generalization bounds for the area under the ROC curve.
J. Mach. Learn. Res., 6:393–425.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006).
Statistics of extremes: theory and applications.

John Wiley & Sons.

Bekker, J. and Davis, J. (2018).
Beyond the selected completely at random assumption for learning from positive and unlabeled data.
CoRR, abs/1809.03207.

Bellemare, M. G., Dabney, W., and Munos, R. (2017).
A distributional perspective on reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR. org.

Bubeck, S., Cesa-Bianchi, N., et al. (2012).
Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
Foundations and Trends® in Machine Learning, 5(1):1–122.

Carpentier, A. and Valko, M. (2014).
Extreme bandits.
In *Advances in Neural Information Processing Systems 27*, pages 1089–1097. Curran Associates, Inc.

Cicirello, V. A. and Smith, S. F. (2005).
The max k-armed bandit: A new model of exploration applied to search heuristic selection.
In *The Proceedings of the Twentieth National Conference on Artificial Intelligence*, volume 3, pages 1355–1361. AAAI Press.

Clémençon, S. and Achab, M. (2017).
Ranking data with continuous labels through oriented recursive partitions.
In *Advances in Neural Information Processing Systems*, pages 4600–4608.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. (2018).
Distributional reinforcement learning with quantile regression.
In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Devroye, L., Györfi, L., and Lugosi, G. (1996).
A Probabilistic Theory of Pattern Recognition.
Springer.

du Plessis, M. C., Niu, G., and Sugiyama, M. (2014).

Analysis of learning from positive and unlabeled data.
In *NIPS*, pages 703–711.

Galichet, N., Sebag, M., and Teytaud, O. (2013).
Exploration vs exploitation vs safety: Risk-aware multi-armed bandits.
In *Asian Conference on Machine Learning*, pages 245–260.

Garivier, A. and Cappé, O. (2011).
The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond.
ArXiv e-prints.

Kaplan, E. L. and Meier, P. (1958).
Nonparametric estimation from incomplete observations.
Journal of the American statistical association, 53(282):457–481.

Kaufmann, E., Cappé, O., and Garivier, A. (2012).
On bayesian upper confidence bounds for bandit problems.
In *Artificial intelligence and statistics*, pages 592–600.

Kolla, R. K., Jagannathan, K., et al. (2019).
Risk-aware multi-armed bandits using conditional value-at-risk.
arXiv preprint arXiv:1901.00997.

Korba, A., Clémençon, S., and Sibony, E. (2017).
A learning theory of ranking aggregation.
In *Proceeding of AISTATS 2017*.

Locatelli, A., Carpentier, A., and Kpotufe, S. (2017).
Adaptivity to noise parameters in nonparametric active learning.
arXiv preprint arXiv:1703.05841.

Maillard, O.-A. (2013).
Robust risk-averse stochastic multi-armed bandits.
In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer.

Minsker, S. (2012).
Plug-in approach to active learning.
The Journal of Machine Learning Research, 13(1):67–90.

Radlinski, F. and Joachims, T. (2005).
Query chains: learning to rank from implicit feedback.
In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248.

Rajaram, S. and Agarwal, S. (2005).
Generalization bounds for k-partite ranking.
In *NIPS 2005 Workshop on Learn to rank*.

S. Clémençon, S. R. and Vayatis, N. (2013).
Ranking data with ordinal labels: optimality and pairwise aggregation.
Machine Learning, 91(1):67–104.

Sani, A., Lazaric, A., and Munos, R. (2012).
Risk-aversion in multi-armed bandits.
In *Advances in Neural Information Processing Systems*, pages 3275–3283.

Sutton, R. S. and Barto, A. G. (2018).
Reinforcement learning: An introduction.
MIT press.

Szorenyi, B., Busa-Fekete, R., Weng, P., and Hüllermeier, E. (2015).
Qualitative multi-armed bandits: A quantile-based approach.

Thompson, W. (1933).

On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3/4):285–294.