Theory of Fair Reinforcement Learning

Pratik Gajane

Tutorial on Advances in Fairness-aware Reinforcement Learning: Theory and Applications,
at the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024),
Aug 5, 2024.

Slides available at https://fair-rl.github.io/

Learning Objectives

- Fairness notions.
- Key ideas used in fair-RL solutions.
- Mathematical performance guarantees.

Introduction to Reinforcement Learning

Multi-armed Bandits

- Stateless (or single-state) reinforcement learning.
- **Classical bandits**: In each round $t = 1, 2, \ldots, T$,
  - the algorithm selects an action i_t from the available actions, and
  - the algorithm receives as feedback a reward r_t according to $\mathrm{RewardFunction}_t(i_t)$.
- **Contextual bandits**: In each round $t = 1, 2, \ldots, T$,
  - the algorithm observes a context vector $x_t \in C \subseteq \mathbb{R}^d$,
  - the algorithm selects an action i_t from the available actions, and
  - the algorithm receives as feedback a reward r_t according to $\mathrm{RewardFunction}_t(x_t, i_t)$.
- **Notations**
  - Total number of actions = A.
  - Total number of rounds = T.
  - Number of dimensions in a context/feature vector = d (wherever applicable).
  - Number of actions that can be selected in each round = m (wherever applicable).

Markov Decision Processes

- Multi-state reinforcement learning.
- Episodic MDP: Learning proceeds in episodes of length H.
- In each episode, at $h = 1, 2, \ldots, H$,
  - at the beginning of the round, the algorithm is in state s_h,
  - the algorithm selects an action i_h from the available actions,
  - the algorithm receives a reward according to $\mathrm{RewardFunction}_h(s_h, i_h)$, and
  - the environment transitions to the next state s_{h+1} according to $\mathrm{TransitionFunction}_h(s_h, i_h)$.
- **Notations**
  - Total number of states = S.
  - Total number of actions = A.
  - Episode length = H.
  - Total number of episodes = E.
  - Total number of agents = N (wherever applicable).

Performance Measure: Regret

- Algorithm's objective is to maximize its cumulative reward (aka, *return*).
- Regret: Difference between the optimal return and the algorithm's return.
- Maximizing return is equivalent to minimizing regret.
- Throughout this tutorial, we will see regret bounds using
  - Big-Oh notation, and
  - $\tilde{O}$: $O(X \cdot \text{log terms}) = \tilde{O}(X)$.
- Goal: Sublinear (in $\#$rounds T or in $\#$episodes E) regret bound.

Common Solution Strategies

- **Upper Confidence Bound (UCB) Algorithms**
  - Compute an estimate of the relevant quantity (i.e. reward or transition probabilities).
  - Build confidence intervals around these estimates.
  - "Optimism in the face of uncertainty": Selection favours the choice with the highest upper confidence bound.
- **Thompson Sampling (aka Posterior Sampling) Algorithms**
  - Maintain belief (*prior*) distribution(s) about relevant quantities.
  - Sample a set of parameters from prior distribution(s).
  - Selection based on samples.
  - Update belief (*posterior*) distributions.

Fairness Notions and Corresponding RL Solutions

Fairness Notions

- Group Fairness
- Distance/Metric/Similarity-based Fairness
- Minimum Selection Criteria
- Counterfactual Fairness
- Nash Social Welfare
- Maxi-min Welfare
- Generalized Gini Welfare

Group Fairness

(Parity across subgroups)

Group Fairness in Multi-armed Bandits

- **Fairness notion**: Parity in expected mean reward for subgroups [1].
- **Key idea**: Adjusted Upper Confidence Bound (UCB) = UCB + fairness penalty, where fairness penalty = linear function of the disparity in observed mean rewards.
- Actions showing high disparity $\Rightarrow$ Decreased adjusted UCB.
- **Performance guarantee**: [1] prove an upper bound of $\tilde{O}(d\sqrt{T})$ on cumulative regret.
  (Same order as the corresponding bound for fairness-unaware RL solutions, albeit with a larger constant.)

⚠ Concern: Assumption that rewards for the decision-maker are aligned with rewards for subgroups.
(Not always the case.)
For example, in credit lending scenario:

- the decision-maker's reward $\equiv$ maximize profits via loan repayments, and
- any subgroup's reward $\equiv$ get more loans.

[1] Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee and Neil Heffernan. Achieving User-Side Fairness in Contextual Bandits. In Human-Centric Intelligent Systems, 2022.

Group Fairness with Distinct Rewards in Multi-agent Episodic MDPs

- [2] make a distinction between decision-maker's rewards and subgroups' rewards.
- Agents belonging to different subgroups interact with the environment according to the sub-group specific transition functions.
- **Fairness-aware objective**: Maximize return of the decision-maker with the constraint that difference in returns of any two agents $\leq \alpha$ (fairness tolerance).
- Assumption: Access to a policy satisfying the fairness constraint with $\alpha_0 < \alpha$. (Allows for exploration without violating fairness guarantees.)
- **Key idea**: For a pair of subgroups, compute optimistic and pessimistic estimates.
- **Performance guarantees**:
  - sublinear cumulative regret (in #episodes E), and
  - fairness constraint is never violated with arbitrarily high probability.

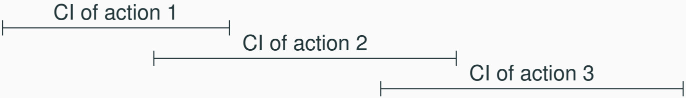
[2] Harsh Satija, Alessandro Lazaric, Matteo Pirotta and Joelle Pineau. Group Fairness in Reinforcement Learning, TMLR 2023.

Metric/Distance/Similarity-based Fairness

(“*Similar* individuals should be treated *similarly*.”)

Meritocratic Fairness

- [3] consider meritocratic fairness in contextual bandits.
- **Fairness constraint**: Given a merit function f,
  if $f(i) \geq f(j)$,
  selection probability of $i \geq$ selection probability of j [3].
- [3] consider merit function to be the expected reward.
- **Key idea**: Use confidence intervals (CI) to link actions.



- **Performance guarantee**: Cumulative regret bound of $\tilde{O}(dAm\sqrt{T})$.
  where m is the maximum #actions that can be selected at each round.

Concern 1: Allows a subgroup best by only a small margin to be selected all the time.

Concern 2: Does not constrain the algorithm in case one subgroup is much better.

[3] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel and Aaron Roth. Fair Algorithms for Infinite and Contextual Bandits. arXiv:1610.09559.

Smooth Fairness and Calibrated Fairness - I

[4] propose:

- **Smooth fairness** — Actions with similar reward distributions should be selected with similar probability (similarity determined by a given divergence function), and
- **Calibrated fairness** — Select each action with probability equal to its realized reward being the highest.

Illustrative Example: Bandit problem with two actions.

- Action 1: $\mathbb{P}(r_1 = 1) = 1$ i.e. $\mathbb{E}[r_1] = 1$.
- Action 2: $\mathbb{P}(r_2 = 0) = 0.52$ and $\mathbb{P}(r_2 = 2) = 0.48$ i.e. $\mathbb{E}[r_2] = 0.96$.
- Meritocratic Fairness: Always select action 1 over action 2.
- Smooth Fairness: In every round, probability of selecting action 1 is close to that of action 2.
- Calibrated Fairness: In every round, select action 1 with probability 0.52 and action 2 with probability 0.48.

[4] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal and David C. Parkes. Calibrated Fairness in Bandits. arXiv:1707.01875.

Smooth Fairness and Calibrated Fairness - II

- Fairness regret = Cumulative amount by which an algorithm is miscalibrated
  $=\sum_1^T \mathbb{E}\left[\sum_{i=1}^A \max\left(\mathbb{P}(\text{realized reward of i is highest}) - \mathbb{P}(\text{i is selected}), 0\right)\right]$.
- Objective: Devise a solution
  - adhering to smooth fairness in each round, and
  - minimizing fairness regret.
- [4] propose a solution based on Thompson sampling with an initial exploration phase which ensures that all actions have been sampled *enough*.
- **Performance guarantees**:
  - Smooth fairness in each round.
    (W.r.t. the divergence function of total variation distance.)
  - Fairness regret = $\tilde{O}(AT)^{2/3}$.

⚠ Might be difficult to specify a suitable divergence function (or distance/similar metric) for individuals.

[4] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal and David C. Parkes. Calibrated Fairness in Bandits. arXiv:1707.01875.

Individual Fairness with Unknown Distance Metric

- [5] consider contextual bandits with the **fairness constraint**: $|\mathrm{SelectionProbability}(i) - \mathrm{SelectionProbability}(j)| \leq \mathrm{DistanceMetric}(\text{context}_i, \text{context}_j)$.
- Unknown $\mathrm{DistanceMetric}$.
- Oracle assumption: Selection rule $\xrightarrow{\text{input}}$ Oracle $\xrightarrow{\text{output}}$ Pairs of actions for which fairness constraint is violated.
- Objectives:
  - Minimize regret w.r.t. the best fair policy.
  - Minimize number of fairness violations.
- Solution based on upper confidence bound and *optimism* principle.
- **Performance guarantees**:
  - Regret w.r.t. the best fair policy = $\tilde{O}\left(A^2 d^2 \log(T) + d\sqrt{T}\right)$
  - Fairness constraint violations of more than ϵ on at most $O(A^2 d^2 \log(d/\epsilon))$ rounds.

[5] Stephen Gillen, Christopher Jung, Michael Kearns, Aaron Roth. Online Learning with an Unknown Fairness Metric. NeurIPS 2018.

Minimum Selection Criteria

Fairness in Wireless Scheduling using Multi-armed Bandits

- Clients (symbolized as actions) compete for a shared wireless channel.
- Multiple actions (up to m) can be selected in each round.
- Some actions can be "sleeping" (i.e. unavailable) and the set of available actions is revealed to the algorithm at the beginning of each round.
- **Asymptotic fairness criteria**: Selection fraction of action $i \geq v_i$ asymptotically. $\liminf_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}[\mathrm{IndicatorFunction}(\text{i is selected at t})] \geq v_i$.
- **Performance guarantees**: Cumulative regret w.r.t. the best fair policy is $O(\sqrt{mAT \log T} + A)$.

⚠ Concern: Fairness guarantees not anytime but only asymptotic.

[6] Fengjiao Li, Jia Liu and Bo Ji. Combinatorial Sleeping Bandits with Fairness Constraints. IEEE Conference on Computer Communications 2019.

Anytime Fairness Guarantees with Minimum Selection Criteria

- Robot-human collaboration where each human teammate is represented by an action and selecting an action corresponds to assigning resources.
- Motivation: Vastly unequal resource assignment leads to loss of trust.
- **Fairness criteria**: Minimum selection rate for every action is at least v (either anytime from 1 to T, or in expectation).
- Proposed UCB-based solutions for above fairness criteria.
- **Performance guarantee**: Cumulative regret w.r.t. the best fair policy $O(\sqrt{AT \log T} + A \log T)$.
- Characterization of regret in terms of the minimum selection rate v is also possible.(Not always tight, bound can sometimes become trivial i.e. linear in T.)

[7] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung and Stefanos Nikolaidis. Multi-Armed Bandits with Fairness Constraints for Distributing Resources to Human Teammates. ACM/IEEE International Conference on Human-Robot Interaction, 2020.

Cost of Fairness with Minimum Selection Criteria

- **Anytime fairness criteria**: Selection fraction of action $i \geq v_i - \alpha$.
- [8] propose a meta-algorithm that can use any suitable bandit algorithm as a black-box.
- **Performance guarantees**:
  - Cumulative regret w.r.t. *the best fair policy* is $O\left(\sqrt{AT \log T}\right)$.
  - Also proved a problem-dependent regret bound which grows as $\log T$.
    (consistent with classical fairness-unaware bandits literature).
- **Cost of fairness** (Regret w.r.t. *the best policy*):
  - When fairness tolerance α is *high*,
    (i.e. $\alpha > v_i - \frac{8 \log T}{T\Delta_i^2}$ for all suboptimal i, where Δ_i is the suboptimlaity gap),
    $\Rightarrow$ regret bound grows as $\log T$.
  - When fairness tolerance α is *low*,
    $\Rightarrow$ regret bound grows as T.

[8] Vishakha Patil, Ganesh Ghalme, Vineet Nair, Y. Narahari. Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. JMLR, 2021.

Counterfactual Fairness

Counterfactual Fairness

- Contextual Bandits for recommender system.
  - Each item (symbolized as action) has a feature vector $y \in \mathcal{Y}$.
  - User arriving at round t has a feature vector $x_t \in \mathcal{X}$.
  - The algorithm recommends an item based on (x_t, y).
- **Fairness constraint**: Expected reward for a user remains within α if their protected attribute were changed to its counterpart.
  Fairness tolerance: α.
- Causal graph

$$\begin{tikzcd} & \mathcal{I} \arrow[dr] & \\ \mathcal{Y} \arrow[ur] \arrow[rr] & & \mathcal{R} \\ & \mathcal{X} \arrow[ul] \arrow[ur] \arrow[uu] & \end{tikzcd}$$

  where $\mathcal{R}$ represents reward and $\mathcal{I}$ represents intermediate features between $\mathcal{Y}$ and $\mathcal{R}$.
- **Key idea**: Find $\mathcal{W} \subseteq \mathcal{Y} \cup \mathcal{X} \cup \mathcal{I}$ that d-separates reward R from features $(\mathcal{Y} \cup \mathcal{X}) \setminus \mathcal{W}$.
- [9] propose an upper confidence bound algorithm based on $\mathcal{W}$.
- **Performance guarantee**: Regret bound of $O\left(\frac{\sqrt{\mathcal{W}}T}{\mathrm{LinearFunction}(\alpha)}\right)$.

[9] Wen Huang, Lu Zhang and Xintao Wu. Achieving Counterfactual Fairness for Causal Bandit, AAAI, 2022.

Welfare-based Notions

Nash Social Welfare in Multi-agent Multi-armed Bandits

- N agents, A actions.
- When agent i selects action j, reward $r \sim$ with mean $\mu_{i,j}$.
- Policy π: Select action j with probability π_j.
- Nash social welfare (NSW): Product of the expected reward of the agents
  i.e. $NSW(\pi) = \prod_{i=1}^{N} \left(\sum_{j=1}^{A} \pi_j \cdot \mu_{i,j} \right)$.
- **Fairness-aware objective**: Minimize regret = $\sum_{t=1}^{T} [NSW(\pi^*) - NSW(\pi_t)]$,
  where $\pi^* \in \arg\max NSW(\pi)$ and π_t is the policy being followed at round t.
- **Key idea**: Use upper confidence bound for NSW and *optimism* principle.
- **Performance guarantee**: Regret bound of $\tilde{O}\left(\sqrt{T} \min\left(\sqrt{N}K^{3/2}, NK\right)\right)$.
- **Caveat**: Exact implementation involves a NP-hard optimization problem.
  Polynomial-time approximation scheme is available $\xRightarrow{\text{unresolved}}$ Regret?

[10] Safwan Hossain, Evi Micha and Nisarg Shah. Fair Algorithms for Multi-Agent Multi-Armed Bandits. NeurIPS, 2021.

Nash Social Welfare in Multi-agent Markov Decision Processes

- N agents.
- In each episode of length H, at $h = 1, 2, \ldots, H$,
  - reward for agent i for action j_h in state s_h is according to $\mathrm{RewardFunction}_i(s_h, j_h)$;
    (separate reward function for each agent)
  - the environment transitions to the next state s_{h+1} according to $\mathrm{TransitionFunction}(s_h, j_h)$.
- Value of policy π corresponding to agent i
  $= \mathrm{Value}_\pi(i) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathrm{RewardFunction}_i(s_h, j_h) \right]$.
- **Nash social welfare (NSW)**: Product of the values received by all the agents
  i.e. $NSW(\pi) = \prod_{i=1}^{N} \mathrm{Value}_\pi(i)$.
- **Fairness-aware objective**: Minimize regret $\sum_{e=1}^{E} NSW(\pi^*) - NSW(\pi_e)$,
  where $\pi^* \in \arg\max NSW(\pi)$ and π_e is the policy being followed in episode e.
- **Key idea**: Upper confidence bound for NSW and *optimism* principle.
- **Performance guarantee**: Regret bound of $\tilde{O}(NH^{N+1} S \sqrt{AE})$.

[11] Debmalya Mandal and Jiarui Gan. Socially Fair Reinforcement Learning. arXiv:2208.12584.

Maxi-Min Welfare in Multi-agent Markov Decision Processes

- Same problem formulation as previous slide.
- In each episode of length H, at $h = 1, 2, \ldots, H$,
  - reward for agent i for action j_h in state s_h is according to $\mathrm{RewardFunction}_i(s_h, j_h)$; (separate reward function for each agent)
  - the environment transitions to the next state s_{h+1} according to $\mathrm{TransitionFunction}(s_h, j_h)$.
- Return or Value of policy π corresponding to agent i $= \mathrm{Value}_\pi(i) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathrm{RewardFunction}_i(s_h, j_h)\right]$.
- **Minimum welfare (MW)**: $MW(\pi) = \min_{i=1,2,\ldots,N} \mathrm{Value}_\pi(i)$
- **Fairness-aware objective**: Minimize regret $\sum_{e=1}^{E} MW(\pi^*) - MW(\pi_e)$, where $\pi^* \in \arg\max MW(\pi)$ and π_e is the policy being followed in episode e.
- **Key idea**: Upper confidence bound for MW and *optimism*.
- **Performance guarantee**: Regret bound of $\tilde{O}(H^2 S \sqrt{AE})$. (Independent of the number of agents N, unlike Nash social welfare bound).

[11] Debmalya Mandal and Jiarui Gan. Socially Fair Reinforcement Learning. arXiv:2208.12584.

Generalized Gini Welfare in Multi-agent Markov Decision Processes

- Same problem formulation as previous slide.
- In each episode of length H, at $h = 1, 2, \ldots, H$,
  - reward for agent i for action j_h in state s_h is according to $\mathrm{RewardFunction}_i(s_h, j_h)$; (separate reward function for each agent)
  - the environment transitions to the next state s_{h+1} according to $\mathrm{TransitionFunction}(s_h, j_h)$.
- Return or Value of policy π corresponding to agent i $= \mathrm{Value}_\pi(i) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathrm{RewardFunction}_i(s_h, j_h) \right]$.
- **Generalized Gini welfare (GGW)** (generalization of Maxi-min welfare).
  - Given weight vector w with $w_i \geq 0$, $\sum_i w_i = 1$ and $w_1 \geq w_2 \cdots \geq w_N$ (descending order).
  - $i_1, i_2, \ldots, i_N$: An ordering with $\mathrm{Value}_\pi(i_1) \leq \mathrm{Value}_\pi(i_2) \cdots \leq \mathrm{Value}_\pi(i_N)$ (ascending order).
  - Generalized Gini Welfare: Weighted sum of values received by all the agents i.e. $GGW(\pi) = \sum_{k=1}^{N} w_k \mathrm{Value}_\pi(i_k)$.
    (Agent receiving lowest value has highest weight, ...)
  - When $w_1 = 1$, generalized Gini welfare reduces to minimum welfare.
- **Fairness-aware Objective**: Minimize regret $\sum_{e=1}^{E} GGW(\pi^*) - GGW(\pi_e)$, where $\pi^* \in \arg\max GGW(\pi)$ and π_e is the policy being followed in episode e.
- **Performance guarantee**: Regret bound of $\tilde{O}(H^2 S \sqrt{AE})$.
  (Independent of the number of agents N, unlike Nash Social Welfare bound).

[11] Debmalya Mandal and Jiarui Gan. Socially Fair Reinforcement Learning. arXiv:2208.12584.

Summary

- Group fairness: Parity across subgroups
  - Contextual bandits.
  - Multi-agent episodic MDPs.
- Distance/metric/similarity-based fairness: “Similar individuals treated similarly.”
  - Meritocratic fairness.
  - Smooth fairness and calibrated fairness.
  - Individual fairness with unknown distance metric.
- Minimum Selection Criteria
  - Asymptotic fairness guarantees.
  - Anytime fairness guarantees.
  - Cost of achieving minimum selection criteria.
- Counterfactual Fairness: Causal approach to fairness
- Nash Social Welfare
- Maxi-min Welfare
- Generalized Gini Welfare

Thank you.