Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning

Hengrui Cai

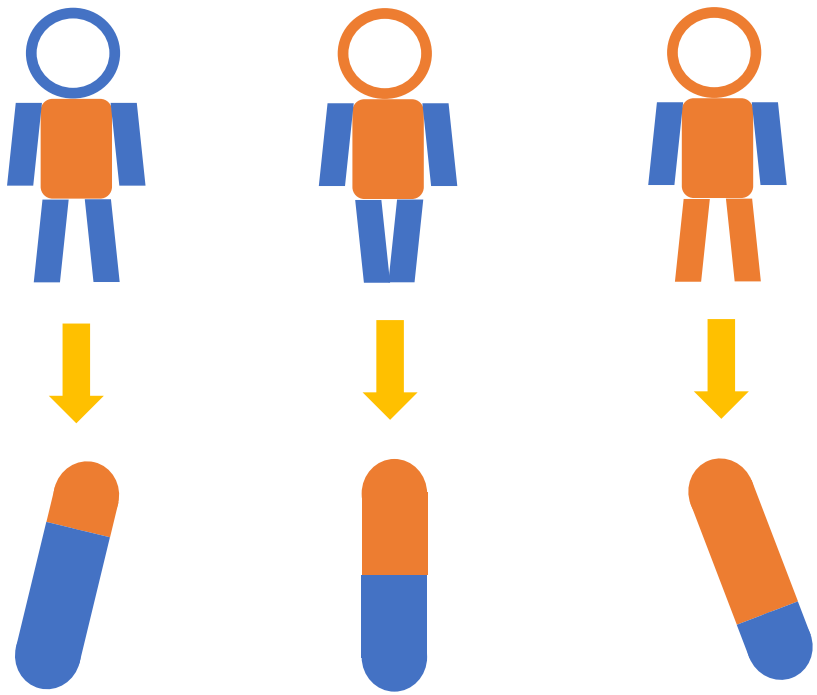
Joint work with Ye Shen, and Rui Song

North Carolina State University

EcoSta 2022

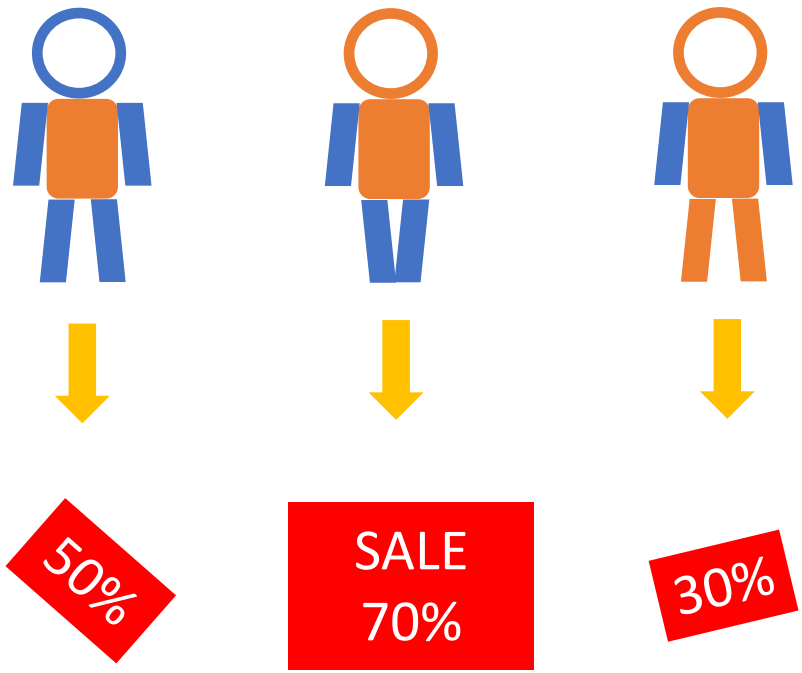
Personalized Sequential Decision Making

Precision Medicine (Lu et al. 2021): Developing an individualized dynamic treatment regime for patients to optimize expected clinical outcomes of interest;



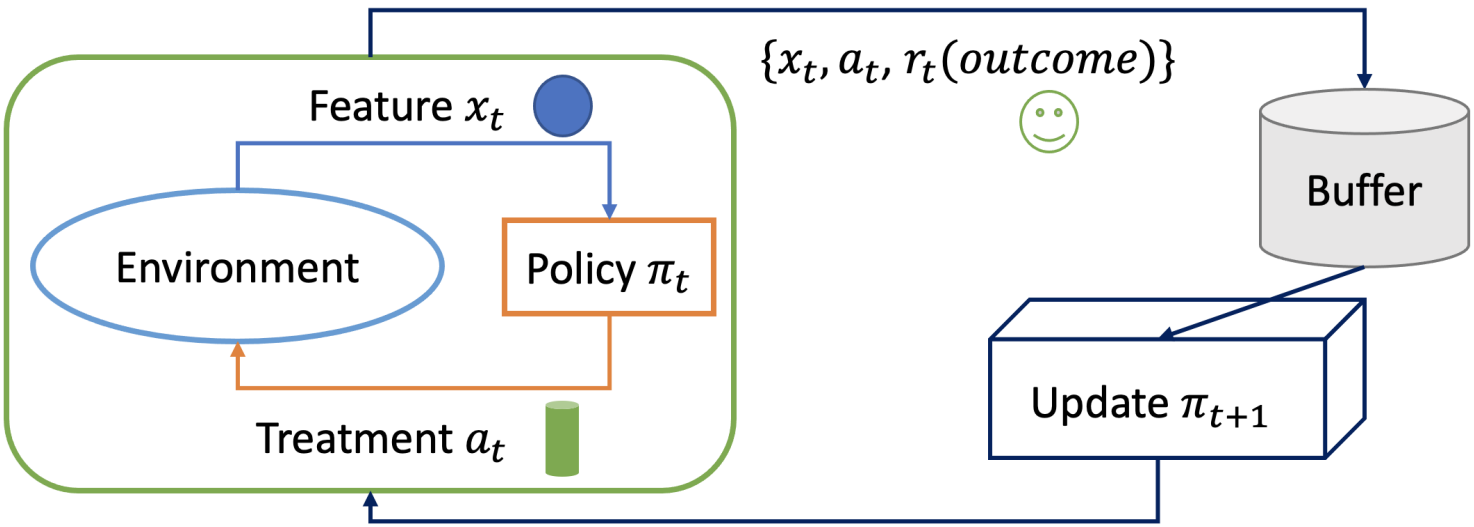
Personalized Sequential Decision Making

Dynamic Pricing (Turvey 2017): Offering customized incentives to increase sales and the level of engagement over time.



Contextual Bandits

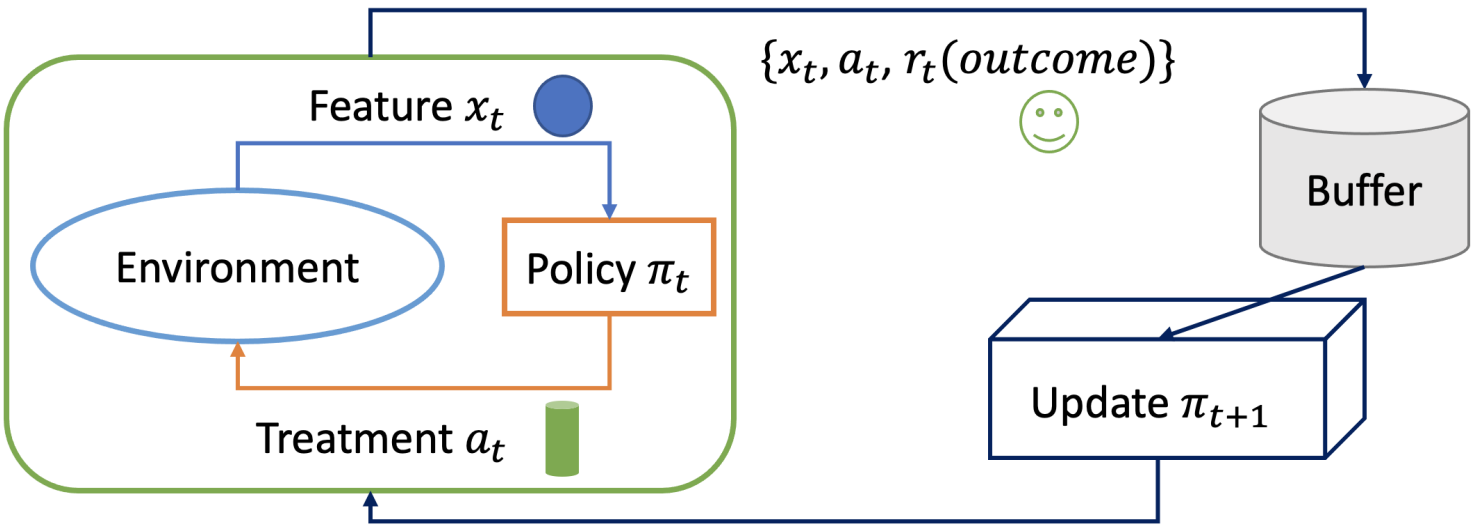
Most bandits works (Lattimore & Szepesvári 2020) focus on the regret analysis, but not evaluate the performance of a bandit policy.



E.g., if the mean outcome of the ongoing rule is much smaller than desired curative effect, the online trial should stop until more effective treatments.

Contextual Bandits

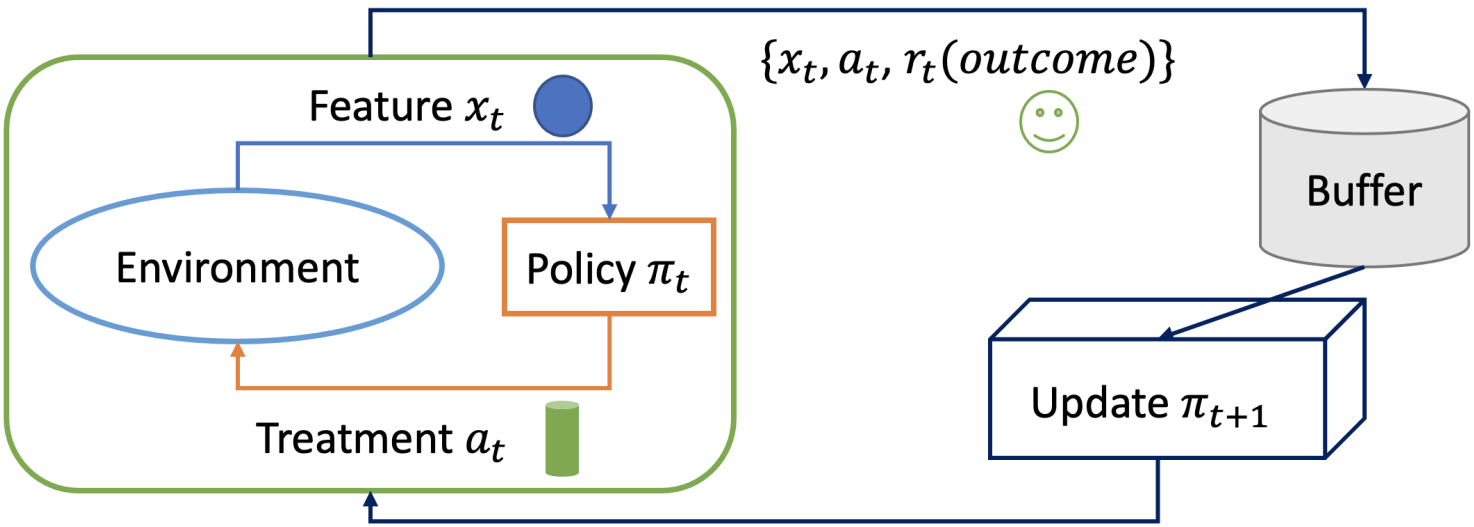
Most bandits works (Lattimore & Szepesvári 2020) focus on the regret analysis, but not evaluate the performance of a bandit policy.



E.g., if the mean outcome of the ongoing rule is much smaller than desired curative effect, the online trial should stop until more effective treatments.

Contextual Bandits

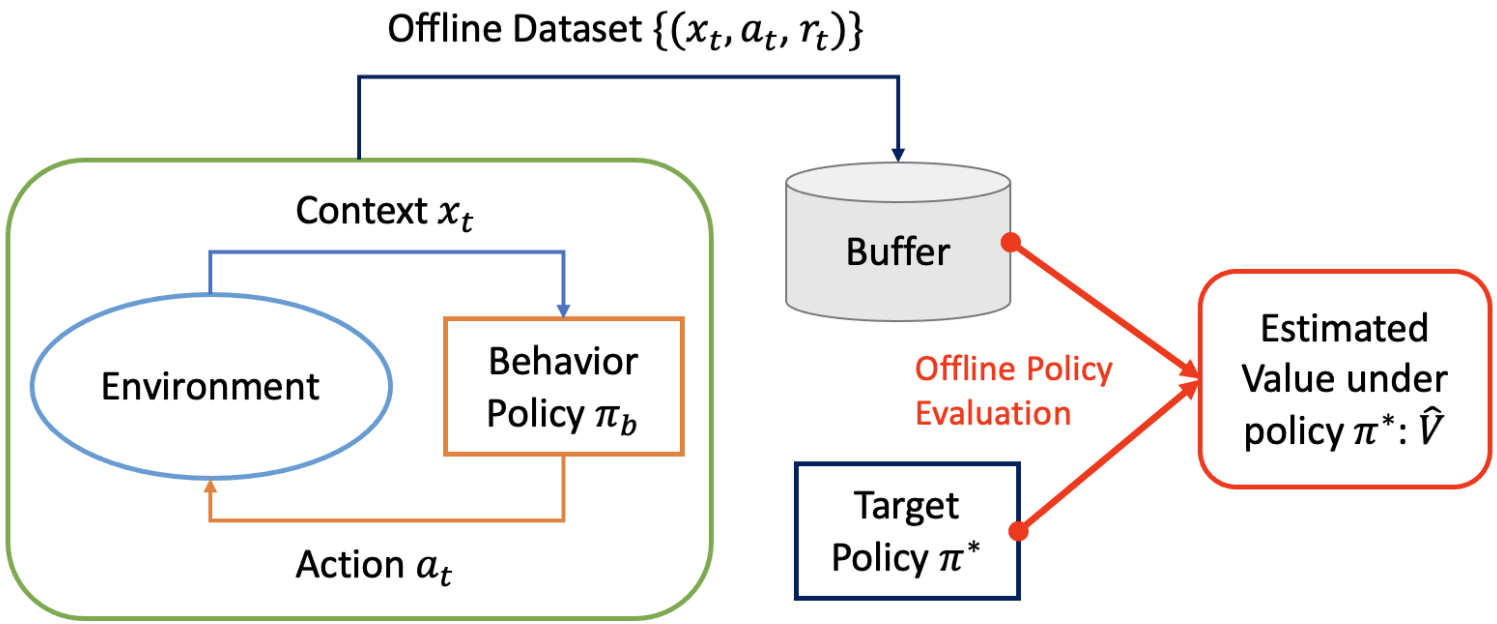
Most bandits works (Lattimore & Szepesvári 2020) focus on the regret analysis, but not evaluate the performance of a bandit policy.



E.g., if the mean outcome of the ongoing rule is much smaller than desired curative effect, the online trial should stop until more effective treatments.

Offline Policy Evaluation (OPE)

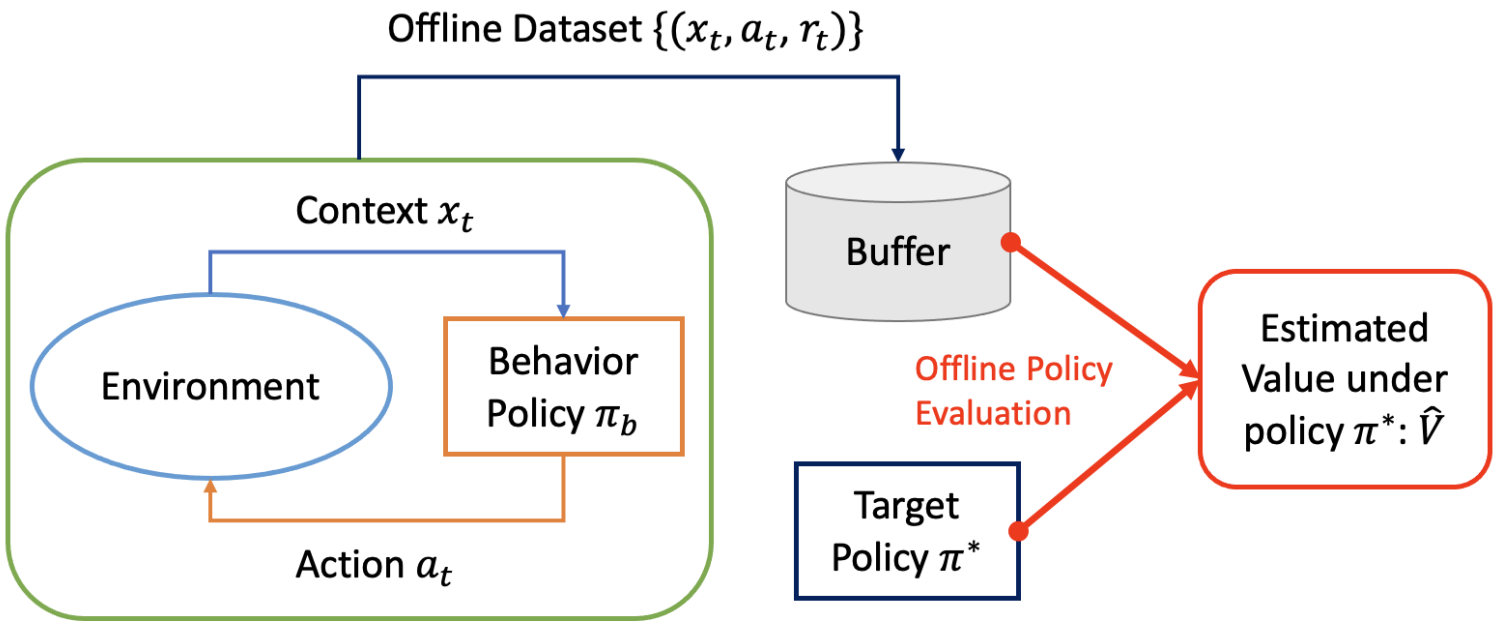
Goal: unbiasedly estimate the expected outcome under a target policy (i.e., value in Dudík et al. (2011)).



Q: Can we evaluate the ongoing policy in real-time? To provide the early-stop of the online experiment and timely feedback from the environment!

Offline Policy Evaluation (OPE)

Goal: unbiasedly estimate the expected outcome under a target policy (i.e., value in Dudík et al. (2011)).



Q: Can we evaluate the ongoing policy in real-time? To provide the early-stop of the online experiment and timely feedback from the environment!

Main Challenges and Related Works

- The data are not identical and independent (i.i.d.) in online learning;
  — OPE (see e.g., Li et al. 2011, Dudík et al. 2011) needed i.i.d. assumption.
- The optimal policy needs to be estimated and updated in real time;
  — Target policy in OPE is fixed and usually known.
- There may not exist a unified best action for all subjects due to heterogeneity;
  — The online sample mean for a fixed arm (see e.g., Nie et al. 2018, Hadad et al. 2019, Zhang et al. 2020).
- The exploration-and-exploitation trade-off (such as upper confidence bound (UCB), Thompson sampling (TS), and ϵ-greedy (EG));
  — Little efforts on quantifying the probability of exploration over time (see e.g., Chu et al. 2011, Zhou 2015, Chambaz et al. 2017, Chen et al. 2020, Bibaut et al. 2021, Zhan et al. 2021).

Main Challenges and Related Works

- The data are not identical and independent (i.i.d.) in online learning;
  — OPE (see e.g., Li et al. 2011, Dudík et al. 2011) needed i.i.d. assumption.
- The optimal policy needs to be estimated and updated in real time;
  — Target policy in OPE is fixed and usually known.
- There may not exist a unified best action for all subjects due to heterogeneity;
  — The online sample mean for a fixed arm (see e.g., Nie et al. 2018, Hadad et al. 2019, Zhang et al. 2020).
- The exploration-and-exploitation trade-off (such as upper confidence bound (UCB), Thompson sampling (TS), and ϵ-greedy (EG));
  — Little efforts on quantifying the probability of exploration over time (see e.g., Chu et al. 2011, Zhou 2015, Chambaz et al. 2017, Chen et al. 2020, Bibaut et al. 2021, Zhan et al. 2021).

Main Challenges and Related Works

- The data are not identical and independent (i.i.d.) in online learning;
  — OPE (see e.g., Li et al. 2011, Dudík et al. 2011) needed i.i.d. assumption.
- The optimal policy needs to be estimated and updated in real time;
  — Target policy in OPE is fixed and usually known.
- There may not exist a unified best action for all subjects due to heterogeneity;
  — The online sample mean for a fixed arm (see e.g., Nie et al. 2018, Hadad et al. 2019, Zhang et al. 2020).
- The exploration-and-exploitation trade-off (such as upper confidence bound (UCB), Thompson sampling (TS), and ϵ-greedy (EG));
  — Little efforts on quantifying the probability of exploration over time (see e.g., Chu et al. 2011, Zhou 2015, Chambaz et al. 2017, Chen et al. 2020, Bibaut et al. 2021, Zhan et al. 2021).

Main Challenges and Related Works

- The data are not identical and independent (i.i.d.) in online learning;
  — OPE (see e.g., Li et al. 2011, Dudík et al. 2011) needed i.i.d. assumption.
- The optimal policy needs to be estimated and updated in real time;
  — Target policy in OPE is fixed and usually known.
- There may not exist a unified best action for all subjects due to heterogeneity;
  — The online sample mean for a fixed arm (see e.g., Nie et al. 2018, Hadad et al. 2019, Zhang et al. 2020).
- The exploration-and-exploitation trade-off (such as upper confidence bound (UCB), Thompson sampling (TS), and ϵ-greedy (EG));
  — Little efforts on quantifying the probability of exploration over time (see e.g., Chu et al. 2011, Zhou 2015, Chambaz et al. 2017, Chen et al. 2020, Bibaut et al. 2021, Zhan et al. 2021).

Policy Evaluation in Online Learning

Goal: unbiasedly estimate the value of the optimal policy that the bandit policy is approaching and infer the corresponding estimate in online learning.

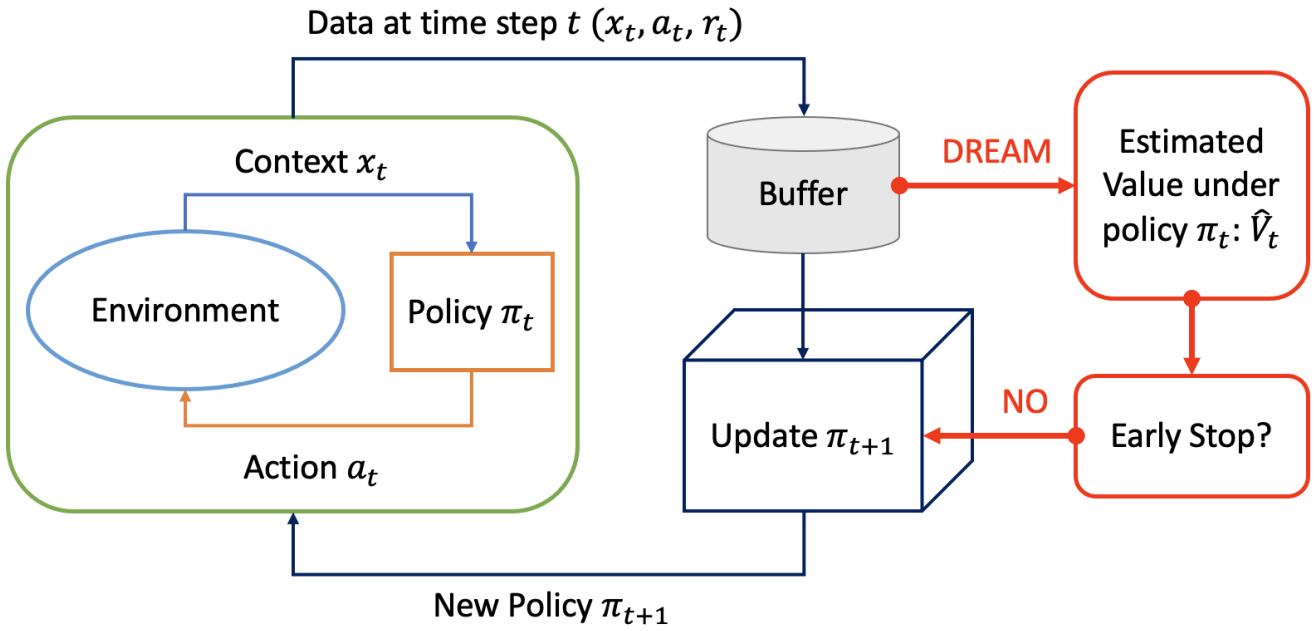


Figure: The architecture of doubly robust interval estimation (DREAM) method.

Framework (Contextual Bandits)

- Let $\boldsymbol{X} \in \mathcal{X}$ be d dimensional context (included 1 for the intercept), $A \in \mathcal{A} = \{0, 1\}$ as the action, and $R \in \mathcal{R}$ as the reward/outcome.
- At time t with a context $\boldsymbol{x}_t$ drawn from $P_{\mathcal{X}}$, choose an action a_t by a policy $\pi : \mathcal{X} \to \mathcal{A}$ given history $\mathcal{H}_{t-1} = \{\boldsymbol{x}_i, a_i, r_i\}_{1 \le i \le t-1}$, and receive reward r_t.
- Suppose $R \equiv \mu(\boldsymbol{x}, a) + e$, where the conditional mean outcome function is

$$\mu(\boldsymbol{x}, a) \equiv \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a),$$

  and e is independent σ-subguassian and at time t, e_t independent of $\mathcal{H}_{t-1}$.
- **Value**: $V(\pi) \equiv \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mathbb{E}\{R|\boldsymbol{X}, A = \pi(\boldsymbol{X})\}\right] = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi(\boldsymbol{X})\}\right]$.
- **Optimal policy**:

$$\pi^*(\boldsymbol{x}) \equiv \arg\max_{a \in \mathcal{A}} \mu(\boldsymbol{x}, a) = \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\}, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

- Goal: estimate **optimal value** $V^* \equiv V(\pi^*) = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi^*(\boldsymbol{X})\}\right]$.

Framework (Contextual Bandits)

- Let $\boldsymbol{X} \in \mathcal{X}$ be d dimensional context (included 1 for the intercept), $A \in \mathcal{A} = \{0, 1\}$ as the action, and $R \in \mathcal{R}$ as the reward/outcome.
- At time t with a context $\boldsymbol{x}_t$ drawn from $P_{\mathcal{X}}$, choose an action a_t by a policy $\pi : \mathcal{X} \to \mathcal{A}$ given history $\mathcal{H}_{t-1} = \{\boldsymbol{x}_i, a_i, r_i\}_{1 \le i \le t-1}$, and receive reward r_t.
- Suppose $R \equiv \mu(\boldsymbol{x}, a) + e$, where the conditional mean outcome function is

$$\mu(\boldsymbol{x}, a) \equiv \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a),$$

  and e is independent σ-subguassian and at time t, e_t independent of $\mathcal{H}_{t-1}$.
- **Value**: $V(\pi) \equiv \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mathbb{E}\{R|\boldsymbol{X}, A = \pi(\boldsymbol{X})\}\right] = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi(\boldsymbol{X})\}\right]$.
- **Optimal policy**:

$$\pi^*(\boldsymbol{x}) \equiv \arg\max_{a \in \mathcal{A}} \mu(\boldsymbol{x}, a) = \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\}, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

- Goal: estimate **optimal value** $V^* \equiv V(\pi^*) = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi^*(\boldsymbol{X})\}\right]$.

Framework (Contextual Bandits)

- Let $\boldsymbol{X} \in \mathcal{X}$ be d dimensional context (included 1 for the intercept), $A \in \mathcal{A} = \{0, 1\}$ as the action, and $R \in \mathcal{R}$ as the reward/outcome.
- At time t with a context $\boldsymbol{x}_t$ drawn from $P_{\mathcal{X}}$, choose an action a_t by a policy $\pi : \mathcal{X} \to \mathcal{A}$ given history $\mathcal{H}_{t-1} = \{\boldsymbol{x}_i, a_i, r_i\}_{1 \le i \le t-1}$, and receive reward r_t.
- Suppose $R \equiv \mu(\boldsymbol{x}, a) + e$, where the conditional mean outcome function is

$$\mu(\boldsymbol{x}, a) \equiv \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a),$$

  and e is independent σ-subguassian and at time t, e_t independent of $\mathcal{H}_{t-1}$.
- **Value**: $V(\pi) \equiv \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mathbb{E}\{R|\boldsymbol{X}, A = \pi(\boldsymbol{X})\} \right] = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi(\boldsymbol{X})\} \right]$.
- **Optimal policy**:

$$\pi^*(\boldsymbol{x}) \equiv \arg\max_{a \in \mathcal{A}} \mu(\boldsymbol{x}, a) = \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\}, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

- Goal: estimate **optimal value** $V^* \equiv V(\pi^*) = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi^*(\boldsymbol{X})\} \right]$.

Framework (Contextual Bandits)

- Let $\boldsymbol{X} \in \mathcal{X}$ be d dimensional context (included 1 for the intercept), $A \in \mathcal{A} = \{0, 1\}$ as the action, and $R \in \mathcal{R}$ as the reward/outcome.
- At time t with a context $\boldsymbol{x}_t$ drawn from $P_{\mathcal{X}}$, choose an action a_t by a policy $\pi : \mathcal{X} \to \mathcal{A}$ given history $\mathcal{H}_{t-1} = \{\boldsymbol{x}_i, a_i, r_i\}_{1 \le i \le t-1}$, and receive reward r_t.
- Suppose $R \equiv \mu(\boldsymbol{x}, a) + e$, where the conditional mean outcome function is

$$\mu(\boldsymbol{x}, a) \equiv \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a),$$

  and e is independent σ-subguassian and at time t, e_t independent of $\mathcal{H}_{t-1}$.
- **Value**: $V(\pi) \equiv \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mathbb{E}\{R|\boldsymbol{X}, A = \pi(\boldsymbol{X})\}\right] = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi(\boldsymbol{X})\}\right]$.
- **Optimal policy**:

$$\pi^*(\boldsymbol{x}) \equiv \arg\max_{a \in \mathcal{A}} \mu(\boldsymbol{x}, a) = \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\}, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

- Goal: estimate **optimal value** $V^* \equiv V(\pi^*) = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi^*(\boldsymbol{X})\}\right]$.

Framework (Contextual Bandits)

- Let $\boldsymbol{X} \in \mathcal{X}$ be d dimensional context (included 1 for the intercept), $A \in \mathcal{A} = \{0, 1\}$ as the action, and $R \in \mathcal{R}$ as the reward/outcome.
- At time t with a context $\boldsymbol{x}_t$ drawn from $P_{\mathcal{X}}$, choose an action a_t by a policy $\pi : \mathcal{X} \to \mathcal{A}$ given history $\mathcal{H}_{t-1} = \{\boldsymbol{x}_i, a_i, r_i\}_{1 \le i \le t-1}$, and receive reward r_t.
- Suppose $R \equiv \mu(\boldsymbol{x}, a) + e$, where the conditional mean outcome function is

$$\mu(\boldsymbol{x}, a) \equiv \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a),$$

  and e is independent σ-subguassian and at time t, e_t independent of $\mathcal{H}_{t-1}$.
- **Value**: $V(\pi) \equiv \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mathbb{E}\{R|\boldsymbol{X}, A = \pi(\boldsymbol{X})\}\right] = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi(\boldsymbol{X})\}\right].$
- **Optimal policy**:

$$\pi^*(\boldsymbol{x}) \equiv \arg\max_{a \in \mathcal{A}} \mu(\boldsymbol{x}, a) = \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\}, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

- Goal: estimate **optimal value** $V^* \equiv V(\pi^*) = \mathbb{E}_{\boldsymbol{X} \sim P_{\mathcal{X}}} \left[\mu\{\boldsymbol{X}, \pi^*(\boldsymbol{X})\}\right]$.

Example: Upper Confidence Bound(UCB) (Li et al. 2010)

- Select action at time t:

$$a_t = \arg\max_{a \in \mathcal{A}} \widehat{\mu}_{t-1}(\boldsymbol{x}_t, a) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, a),$$

- with two actions, i.e.,

$$a_t = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)\right\}. \quad (1)$$

- Estimated optimal policy under the UCB at time step t is

$$\widehat{\pi}_t(\boldsymbol{x}) = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}, 0)\right\}, \quad \forall \boldsymbol{x} \in \mathcal{X}. \quad (2)$$

Example: Upper Confidence Bound(UCB) (Li et al. 2010)

- Select action at time t:

$$a_t = \arg\max_{a \in \mathcal{A}} \widehat{\mu}_{t-1}(\boldsymbol{x}_t, a) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, a),$$

- with two actions, i.e.,

$$a_t = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)\right\}. \quad (1)$$

- Estimated optimal policy under the UCB at time step t is

$$\widehat{\pi}_t(\boldsymbol{x}) = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}, 0)\right\}, \quad \forall \boldsymbol{x} \in \mathcal{X}. \quad (2)$$

Example: Upper Confidence Bound(UCB) (Li et al. 2010)

- Select action at time t:

$$a_t = \arg\max_{a \in \mathcal{A}} \widehat{\mu}_{t-1}(\boldsymbol{x}_t, a) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, a),$$

- with two actions, i.e.,

$$a_t = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) + c_t \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)\right\}. \quad (1)$$

- Estimated optimal policy under the UCB at time step t is

$$\widehat{\pi}_t(\boldsymbol{x}) = \mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}, 0)\right\}, \quad \forall \boldsymbol{x} \in \mathcal{X}. \quad (2)$$

Probability of Exploration and Exploitation

Status of Exploration

$$\mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\},$$

where the action taken by the bandit is different from the estimated optimal action that exploits the historical information.

Probability of Exploration

$$\kappa_t(\boldsymbol{x}_t) \equiv \Pr\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\} = \mathrm{E}[\mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\}].$$

Probability of Exploitation

$$\Pr\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} = 1 - \kappa_t(\boldsymbol{x}_t) = \mathrm{E}[\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}].$$

Doubly Robust Interval Estimation

In view of the regret analysis for the contextual bandits,

$$\left|\sum_{t=1}^{T}(r_t - V^*)\right| = \tilde{\mathcal{O}}(\sqrt{dT}) \quad \longrightarrow \quad \sqrt{T}\Big(\frac{1}{T}\sum_{t=1}^{T} r_t - V^*\Big) = \tilde{\mathcal{O}}(1),$$

since the dimension d is finite, where $\tilde{\mathcal{O}}$ is the asymptotic order up to some logarithm factor. A simple average of the total outcome is not a good estimator!

Doubly Robust Mean Outcome Estimator

$$\widehat{V}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(x_t)\}}{1-\widehat{\kappa}_t(x_t)}\Big[r_t - \widehat{\mu}_{t-1}\{x_t, \widehat{\pi}_t(x_t)\}\Big] + \widehat{\mu}_{t-1}\{x_t, \widehat{\pi}_t(x_t)\}, \tag{3}$$

where T is the current time step or the termination time, with a variance estimator as

$$\widehat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\widehat{\pi}_t(x_t)\widehat{\sigma}_{1,t-1}^2 + \{1-\widehat{\pi}_t(x_t)\}\widehat{\sigma}_{0,t-1}^2}{1-\widehat{\kappa}_t(x_t)} + \Big[\widehat{\mu}_T\{x_t, \widehat{\pi}_T(x_t)\} - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_T\{x_t, \widehat{\pi}_T(x_t)\}\Big]^2\right). \tag{4}$$

where $\widehat{\sigma}_{a,t}^2$ is an estimator for $\sigma_a^2 = \mathrm{E}(e^2|A=a)$, for $a = 0, 1$.

Doubly Robust Interval Estimation

In view of the regret analysis for the contextual bandits,

$$\left|\sum_{t=1}^{T}(r_t - V^*)\right| = \tilde{\mathcal{O}}(\sqrt{dT}) \quad \longrightarrow \quad \sqrt{T}\Big(\frac{1}{T}\sum_{t=1}^{T} r_t - V^*\Big) = \tilde{\mathcal{O}}(1),$$

since the dimension d is finite, where $\tilde{\mathcal{O}}$ is the asymptotic order up to some logarithm factor. A simple average of the total outcome is not a good estimator!

Doubly Robust Mean Outcome Estimator

$$\widehat{V}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)}\Big[r_t - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big] + \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}, \tag{3}$$

where T is the current time step or the termination time, with a variance estimator as

$$\widehat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\widehat{\pi}_t(\boldsymbol{x}_t)\widehat{\sigma}_{1,t-1}^2 + \{1 - \widehat{\pi}_t(\boldsymbol{x}_t)\}\widehat{\sigma}_{0,t-1}^2}{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)} + \left[\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\} - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\}\right]^2\right). \tag{4}$$

where $\widehat{\sigma}_{a,t}^2$ is an estimator for $\sigma_a^2 = \mathrm{E}(e^2|A = a)$, for $a = 0, 1$.

Bounding the Probability of Exploration

(A1) (Boundness) Exist $L_{\boldsymbol{x}} > 0$ such that $\|\boldsymbol{x}\|_\infty \leq L_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, and $\Sigma = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}} \left(\boldsymbol{x}\boldsymbol{x}^\top \right)$ has minimum eigenvalue $\lambda_{\min}(\Sigma) > \lambda$ for some $\lambda > 0$.

(A2) (Clipping) For any $a \in \mathcal{A}$ and time $t \geq 1$, there exists an sequence positive and non-increasing $\{p_i\}_{i=1}^t$ s.t. $\lambda_{\min}\{t^{-1} \sum_{i=1}^t \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top\} > p_t \lambda_{\min}(\boldsymbol{\Sigma})$.

Theorem 1 (short version for UCB only)

In the online contextual bandit optimization using UCB, assuming (A1) and (A2), let $\Delta_{\boldsymbol{x}_t} \equiv \mu(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 0)$, then with probability approaching 1, we have

$$\kappa_t(\boldsymbol{x}_t) \leq 2 \exp\left\{ -\sqrt{\lambda}/(4L_{\boldsymbol{x}}) \left| \sqrt{p_{t-1} c_t^2} - \sqrt{\lambda (t-1) p_{t-1}^2} \Delta_{\boldsymbol{x}_t} / L_{\boldsymbol{x}} \right| \right\}.$$

If $p_{t-1} c_t^2 < (t-1) p_{t-1}^2 \to \infty$, the upper bound for κ_t decays at $\mathcal{O}\{\exp(-\sqrt{t p_t^2})\}$.

Bounding the Probability of Exploration

(A1) (Boundness) Exist $L_{\boldsymbol{x}} > 0$ such that $\|\boldsymbol{x}\|_\infty \leq L_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, and $\Sigma = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}} \left(\boldsymbol{X}\boldsymbol{X}^\top \right)$ has minimum eigenvalue $\lambda_{\min}(\Sigma) > \lambda$ for some $\lambda > 0$.

(A2) (Clipping) For any $a \in \mathcal{A}$ and time $t \geq 1$, there exists an sequence positive and non-increasing $\{p_i\}_{i=1}^t$ s.t. $\lambda_{\min}\{t^{-1} \sum_{i=1}^t \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top\} > p_t \lambda_{\min}(\boldsymbol{\Sigma})$.

Theorem 1 (short version for UCB only)

In the online contextual bandit optimization using UCB, assuming (A1) and (A2), let $\Delta_{\boldsymbol{x}_t} \equiv \mu(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 0)$, then with probability approaching 1, we have

$$\kappa_t(\boldsymbol{x}_t) \leq 2 \exp\left\{ -\sqrt{\lambda}/(4L_{\boldsymbol{x}}) \left| \sqrt{p_{t-1} c_t^2} - \sqrt{\lambda (t-1) p_{t-1}^2} \Delta_{\boldsymbol{x}_t} / L_{\boldsymbol{x}} \right| \right\}.$$

If $p_{t-1} c_t^2 < (t-1) p_{t-1}^2 \to \infty$, the upper bound for κ_t decays at $\mathcal{O}\{\exp(-\sqrt{t p_t^2})\}$.

Bounding the Probability of Exploration

(A1) (Boundness) Exist $L_{\boldsymbol{x}} > 0$ such that $\|\boldsymbol{x}\|_\infty \leq L_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, and $\Sigma = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}} \left(\boldsymbol{X}\boldsymbol{X}^\top \right)$ has minimum eigenvalue $\lambda_{\min}(\Sigma) > \lambda$ for some $\lambda > 0$.

(A2) (Clipping) For any $a \in \mathcal{A}$ and time $t \geq 1$, there exists an sequence positive and non-increasing $\{p_i\}_{i=1}^t$ s.t. $\lambda_{\min}\{t^{-1} \sum_{i=1}^t \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top\} > p_t \lambda_{\min}(\boldsymbol{\Sigma})$.

Theorem 1 (short version for UCB only)

In the online contextual bandit optimization using UCB, assuming (A1) and (A2), let $\Delta_{\boldsymbol{x}_t} \equiv \mu(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 0)$, then with probability approaching 1, we have

$$\kappa_t(\boldsymbol{x}_t) \leq 2 \exp \left\{ -\sqrt{\lambda}/(4L_{\boldsymbol{x}}) \left| \sqrt{p_{t-1} c_t^2} - \sqrt{\lambda (t-1) p_{t-1}^2} \Delta_{\boldsymbol{x}_t} / L_{\boldsymbol{x}} \right| \right\}.$$

If $p_{t-1} c_t^2 < (t-1) p_{t-1}^2 \to \infty$, the upper bound for κ_t decays at $\mathcal{O}\{\exp(-\sqrt{t p_t^2})\}$.

Asymptotic Normality and Robustness for DREAM

(A3) (Margin Condition) Exist γ and δ such that for $0 < M \leq \delta$,

$$\Pr\{0 < |\mu(\boldsymbol{X}, 1) - \mu(\boldsymbol{X}, 0)| \leq M\} = \mathcal{O}(M^{\gamma}), \quad \forall \boldsymbol{X} \in \mathcal{X}.$$

(A4) (Rate Double Robustness) For $a \in \mathcal{A}$,

$$\mathrm{E}_{\boldsymbol{X} \in \mathcal{X}} |\mu(\boldsymbol{X}, a) - \widehat{\mu}_t(\boldsymbol{X}, a)||\kappa_t(\boldsymbol{X}) - \widehat{\kappa}_t(\boldsymbol{X})| = o_p(t^{-1/2}), \text{ for } a \in \mathcal{A}.$$

Theorem 2

Under conditions in Theorem 1, $tp_t^2 \to \infty$ as $t \to \infty$, and (A3)-(A4), we have

$$\sqrt{T}(\widehat{V}_T - V^*) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{DR}^2\right),$$

$$\widehat{\sigma}_T^2 \xrightarrow{p} \sigma_{DR}^2 = \int_x \frac{\pi^*(\boldsymbol{x})\sigma_1^2 + \{1 - \pi^*(\boldsymbol{x})\}\sigma_0^2}{1 - \kappa_\infty(\boldsymbol{x})} dP_{\mathcal{X}} + \mathrm{Var}\left[\mu\{\boldsymbol{x}, \pi^*(\boldsymbol{x})\}\right] < \infty,$$

with $\kappa_\infty(\cdot, \cdot) \equiv \lim_{t \to \infty} \kappa_t(\cdot, \cdot)$ and $\sigma_a^2 = \mathrm{E}(e_t^2 | A_t = a)$.

Asymptotic Normality and Robustness for DREAM

(A3) (Margin Condition) Exist γ and δ such that for $0 < M \leq \delta$,

$$\Pr\{0 < |\mu(\boldsymbol{X}, 1) - \mu(\boldsymbol{X}, 0)| \leq M\} = \mathcal{O}(M^{\gamma}), \quad \forall \boldsymbol{X} \in \mathcal{X}.$$

(A4) (Rate Double Robustness) For $a \in \mathcal{A}$,

$$\mathrm{E}_{\boldsymbol{X} \in \mathcal{X}} |\mu(\boldsymbol{X}, a) - \widehat{\mu}_t(\boldsymbol{X}, a)||\kappa_t(\boldsymbol{X}) - \widehat{\kappa}_t(\boldsymbol{X})| = o_p(t^{-1/2}), \text{ for } a \in \mathcal{A}.$$

Theorem 2

Under conditions in Theorem 1, $tp_t^2 \to \infty$ as $t \to \infty$, and (A3)-(A4), we have

$$\sqrt{T}(\widehat{V}_T - V^*) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{DR}^2\right),$$

$$\widehat{\sigma}_T^2 \xrightarrow{p} \sigma_{DR}^2 = \int_{\boldsymbol{x}} \frac{\pi^*(\boldsymbol{x})\sigma_1^2 + \{1 - \pi^*(\boldsymbol{x})\}\sigma_0^2}{1 - \kappa_\infty(\boldsymbol{x})} dP_{\mathcal{X}} + \mathrm{Var}\left[\mu\{\boldsymbol{x}, \pi^*(\boldsymbol{x})\}\right] < \infty,$$

with $\kappa_\infty(\cdot, \cdot) \equiv \lim_{t \to \infty} \kappa_t(\cdot, \cdot)$ and $\sigma_a^2 = \mathrm{E}(e_t^2 | A_t = a)$.

Algorithm of DREAM

Algorithm 1 DREAM under Contextual Bandits with Clipping

Input: termination time T, and the clipping rate $p_t > \mathcal{O}(t^{-1/2})$;
for Time $t = 1, 2, \cdots T$ **do**
[1] Sample d-dimensional context $\boldsymbol{x}_t \in \mathcal{X}$;
[2] Update $\widehat{\mu}_{t-1}(\cdot, \cdot)$;
[3] Update $\widehat{\pi}_t(\cdot)$ and a_t using the contextual bandit algorithms (such as Equations (1) and (2));
[4] Use the history $\{\mathbb{I}\{\widehat{\pi}_i(\boldsymbol{x}_i) = a_i\}, \boldsymbol{x}_i\}_{1 \le i \le t}$ to estimate $\widehat{\kappa}_t(\cdot)$;
if $\lambda_{\min}(t^{-1}\sum_{i=1}^{t} \mathbb{I}\{a_i = 1 - \widehat{\pi}_t(\boldsymbol{x})\}\boldsymbol{x}_i\boldsymbol{x}_i^\top) < p_t \lambda_{\min}(t^{-1}\sum_{i=1}^{t} \boldsymbol{x}_i\boldsymbol{x}_i^\top)$ **then**
[5] Set $a_t = 1 - \widehat{\pi}_t(\boldsymbol{x}_t)$;
end if
[6] Estimate the value and its variance under the optimal policy using Equations (3) and (4).
end for

Simulation Setting

- Context $\boldsymbol{X} = [X_1, X_2]^\top$, with $X_1, X_2 \overset{i.i.d.}{\sim}$ Uniform$(0, 2\pi)$.
- Outcome given $\boldsymbol{x}$ and a is generated from $R \sim \mathcal{N}\{\mu(\boldsymbol{x}, a), \sigma_a^2\}$, where

$$\begin{aligned}\mu(\boldsymbol{x}, a) &= 2 - a + (5a - 1)\cos(x_1) + (1.5 - 3a)\cos(x_2) \\ &= [1, \cos(x_1), \cos(x_2)][2 - a, 5a - 1, 1.5 - 3a]^\top,\end{aligned}$$

  with equal variances as $\sigma_1 = \sigma_0 = 0.5$.
- Apply three bandit algorithms: UCB, TS, and EG.
- Total decision time $T = 200$ with an initial random exploration as $T_0 = 30$.

Models in Comprison

1. Models of μ and κ_t are both correctly specified; 2. Model of μ is misspecified;
3. Model of κ_t is misspecified; 4. The averaged reward as the value estimator.

Simulation Results

In contextual bandits (see e.g., Chu et al. 2011), the cumulative regret is

$$\left|\sum_{t=1}^{T}(r_t - V^*)\right| = \tilde{\mathcal{O}}(\sqrt{dT}), \quad \longrightarrow \quad \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} r_t - V^*\right) = \tilde{\mathcal{O}}(1).$$

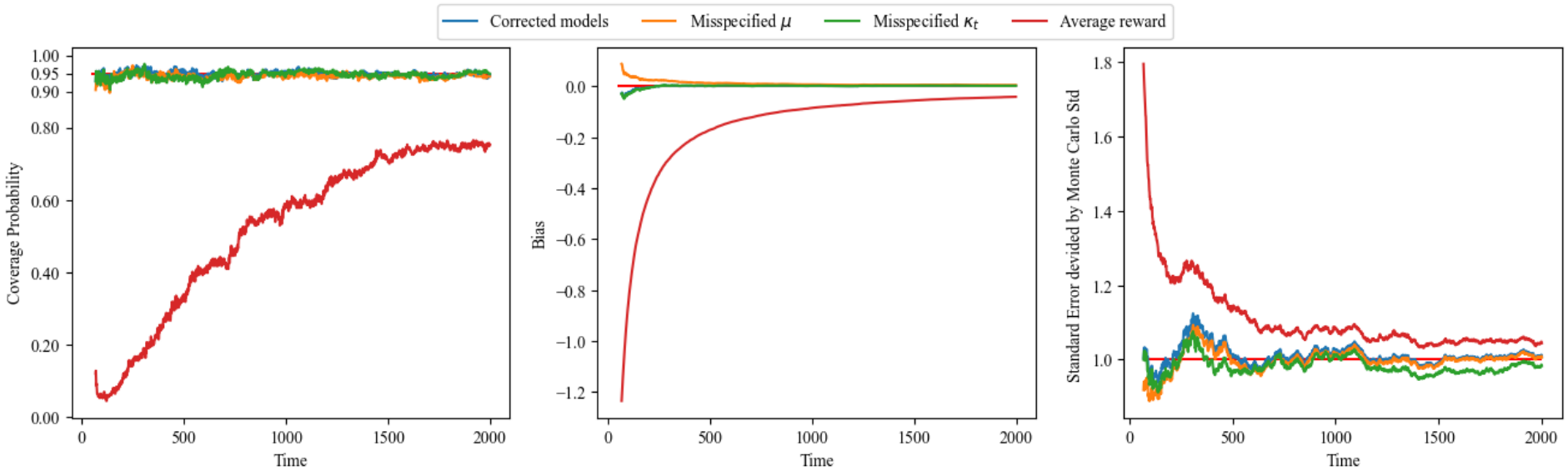


Figure: Results with UCB.

Simulation Results (cont.)

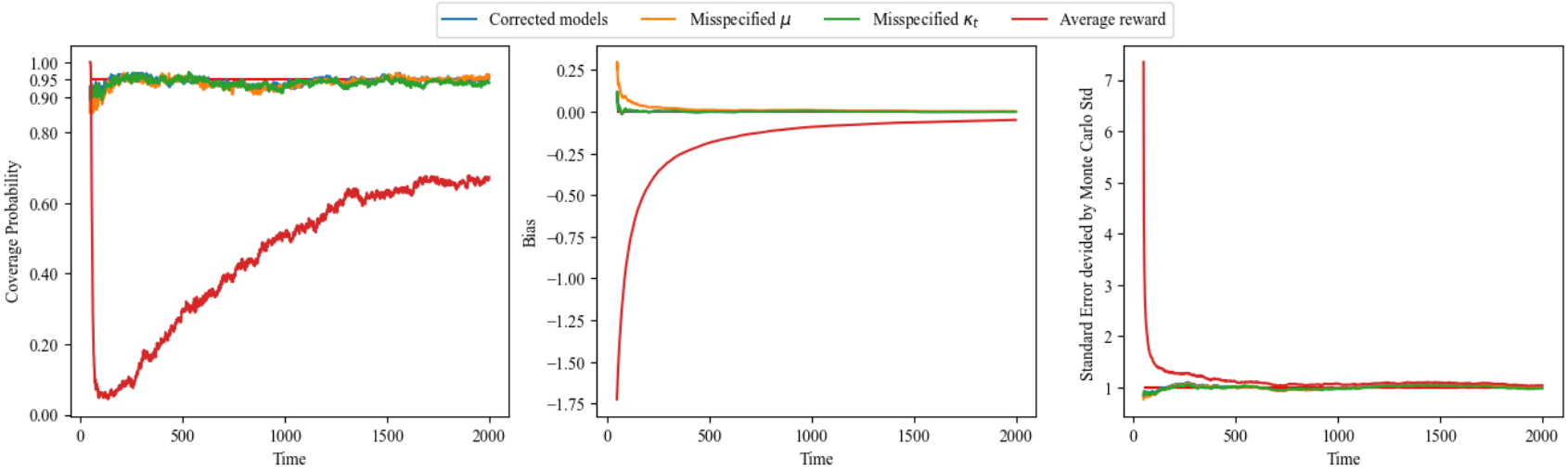


Figure: Results with TS.

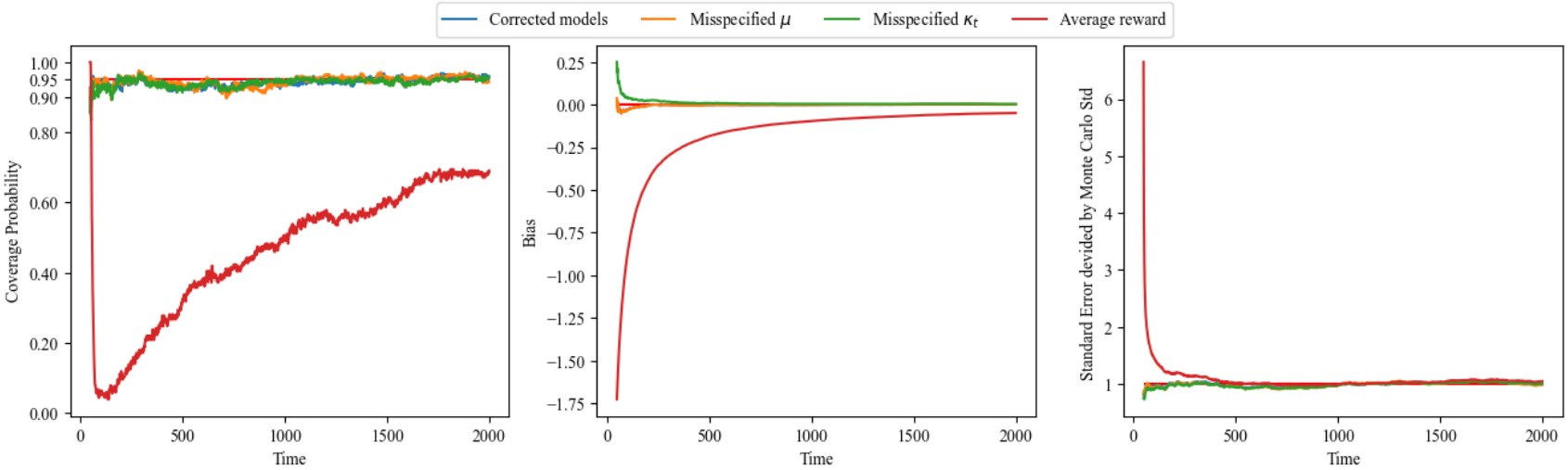


Figure: Results with EG.

Simulation Results: Probability of Exploration

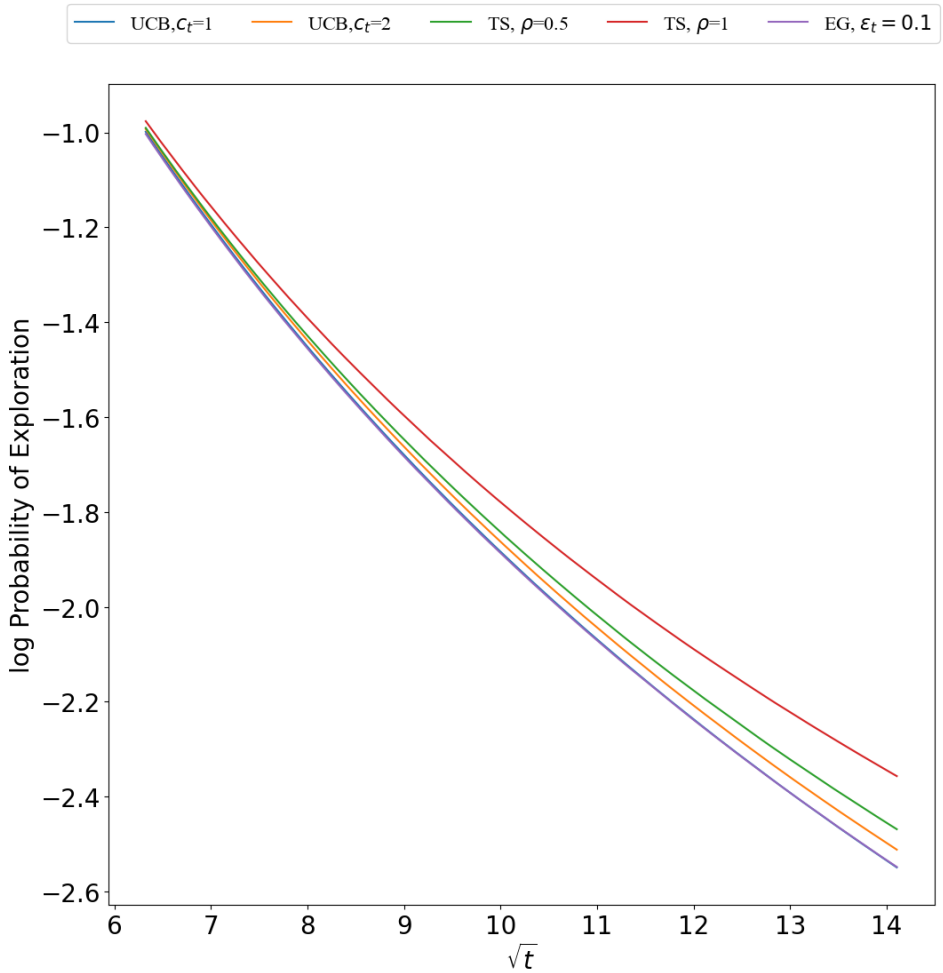


Figure: The logarithm of the probability of exploration for different bandits algorithms under contextual bandits.

Real Data Analysis: OpenML Database (Bischl et al. 2017)

Turn two-class classification tasks into two-armed contextual bandit problems (see e.g., Su et al. 2019):
Draw the context-label pair $\{\boldsymbol{x}_t, Y_t\}$ uniformly at random. Given $\boldsymbol{x}_t$, the bandit selects $a_t \in \{0, 1\}$. The reward is generated from $\mathcal{N}\{\mathbb{I}(a_t = Y_t), 0.5^2\}$.

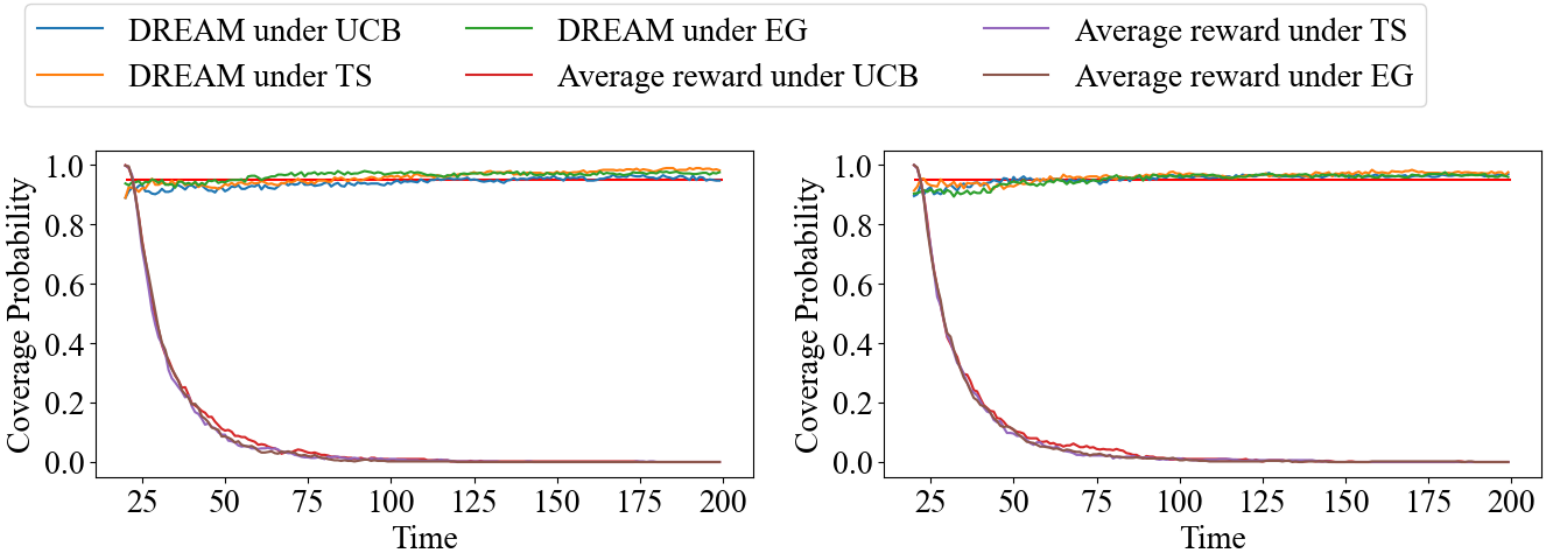


Figure: Left: results for SEA50. Right: results for SEA50000.

Thank You!

The arXiv link.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A. & van der Laan, M. (2021), 'Post-contextual-bandit inference', *Advances in Neural Information Processing Systems* **34**.

Bischl, B., Casalicchio, G., Feurer, M., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N. & Vanschoren, J. (2017), 'Openml benchmarking suites', *arXiv preprint arXiv:1708.03731* .

Chambaz, A., Zheng, W. & van der Laan, M. J. (2017), 'Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward', *Annals of statistics* **45**(6), 2537.

Chen, H., Lu, W. & Song, R. (2020), 'Statistical inference for online decision making: In a contextual bandit setting', *Journal of the American Statistical Association* pp. 1–16.

Chu, W., Li, L., Reyzin, L. & Schapire, R. (2011), Contextual bandits with linear payoff functions, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', JMLR Workshop and Conference Proceedings, pp. 208–214.

Dudík, M., Langford, J. & Li, L. (2011), 'Doubly robust policy evaluation and learning', *arXiv preprint arXiv:1103.4601* .

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S. & Athey, S. (2019), 'Confidence intervals for policy evaluation in adaptive experiments', *arXiv preprint arXiv:1911.02768* .

Lattimore, T. & Szepesvári, C. (2020), *Bandit algorithms*, Cambridge University Press.

Li, L., Chu, W., Langford, J. & Schapire, R. E. (2010), A contextual-bandit approach to personalized news article recommendation, *in* 'Proceedings of the 19th international conference on World wide web', pp. 661–670.

Li, L., Chu, W., Langford, J. & Wang, X. (2011), Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms, *in* 'Proceedings of the fourth ACM international conference on Web search and data mining', pp. 297–306.

Lu, Y., Xu, Z. & Tewari, A. (2021), 'Bandit algorithms for precision medicine', *arXiv preprint arXiv:2108.04782* .

Nie, X., Tian, X., Taylor, J. & Zou, J. (2018), Why adaptively collected data have negative bias and how to correct for it, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 1261–1269.

Su, Y., Dimakopoulou, M., Krishnamurthy, A. & Dudík, M. (2019), 'Doubly robust off-policy evaluation with shrinkage', *arXiv preprint arXiv:1907.09623* .

Turvey, R. (2017), *Optimal Pricing and Investment in Electricity Supply: An Esay in Applied Welfare Economics*, Routledge.

Zhan, R., Hadad, V., Hirshberg, D. A. & Athey, S. (2021), Off-policy evaluation via adaptive weighting with data from contextual bandits, *in* 'Proceedings of

the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining', pp. 2125–2135.

Zhang, K. W., Janson, L. & Murphy, S. A. (2020), 'Inference for batched bandits', *arXiv preprint arXiv:2002.03217* .

Zhou, L. (2015), 'A survey on contextual multi-armed bandits', *arXiv preprint arXiv:1508.03326* .

Bounding the Probability of Exploration

Theorem 1

In the online contextual bandit optimization using UCB, TS, or EG, assuming (A1) and (A2), then with probability approaching 1,
(i) under UCB, let $\Delta_{\boldsymbol{x}_t} \equiv \mu(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 0)$, we have

$$\kappa_t(\boldsymbol{x}_t) \leq 2\exp\left\{-\sqrt{\lambda}/(4L_{\boldsymbol{x}})\left|\sqrt{p_{t-1}c_t^2} - \sqrt{\lambda(t-1)p_{t-1}^2}\Delta_{\boldsymbol{x}_t}/L_{\boldsymbol{x}}\right|\right\};$$

(ii) under TS, we have

$$\kappa_t(\boldsymbol{x}_t) \leq \exp\left\{-\frac{\Delta_{\boldsymbol{x}_t}^2(t-1)p_{t-1}\lambda}{2\rho^2 L_{\boldsymbol{x}}^2}\right\} + 2\exp\left\{-\lambda\Delta_{\boldsymbol{x}_t}\sqrt{(t-1)p_{t-1}^2}/(4\sigma L_{\boldsymbol{x}})\right\};$$

(iii) under EG, we have $\kappa_t(\boldsymbol{x}_t) = \epsilon_t/2$.

Tail Bound of Online Estimator

Consider $\mu(\boldsymbol{x}, a) = \boldsymbol{x}^\top \beta(a)$, where $\beta(\cdot)$ is a smooth function, which can be estimated via a ridge regression based on $\mathcal{H}_{t-1}$ as $\widehat{\beta}_{t-1}(a)$.

Theorem 3

Under conditions in Theorem 1, with κ_t non-increasing, for any $h > 0$, the probability of the online ridge estimator bounded within its true as

$$\begin{aligned} & \Pr\left(\left\|\widehat{\beta}_t(a) - \beta(a)\right\|_1 \leq h\right) \\ \geq \quad & 1 - C_1 \exp(-C_2 t p_t^2) - C_3 \exp(-C_4 t p_t) + C_5 \exp(-C_6 t p_t^2 - C_7 t p_t), \end{aligned}$$

for some constant $\{C_j\}_{1 \leq j \leq 7}$.

This tail bound is asymptotically equivalent to $1 - \exp(-t p_t^2)$.

Asymptotic Normality of Online Estimator

Theorem 4

Under the conditions in Theorem 1 with $tp_t^2 \to \infty$ as $t \to \infty$, we have

$$\sqrt{t}\{\widehat{\mu}_t(\boldsymbol{x}, a) - \mu(\boldsymbol{x}, a)\} \xrightarrow{D} \mathcal{N}\{0, \boldsymbol{x}^\top \sigma^2_{\beta(a)} \boldsymbol{x}\}, \quad \forall \boldsymbol{x} \in \mathcal{X}, \forall a \in \mathcal{A}.$$

where the asymptotic variance is given by

$$\sigma^2_{\beta(a)} = \sigma_a^2 \left[\int \kappa_\infty(\boldsymbol{x}, a) \mathbb{I}\{\boldsymbol{x}^\top \beta(a) < \boldsymbol{x}^\top \beta(1-a)\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\boldsymbol{x}} \right.$$
$$\left. + \int \{1 - \kappa_\infty(\boldsymbol{x}, a)\} \mathbb{I}\left\{\boldsymbol{x}^\top \beta(a) \geq \boldsymbol{x}^\top \beta(1-a)\right\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\boldsymbol{x}} \right]^{-1},$$

with $\kappa_\infty(\cdot, \cdot) \equiv \lim_{t \to \infty} \kappa_t(\cdot, \cdot)$ and $\sigma_a^2 = \mathrm{E}(e_t^2 | a_t = a)$ denoting the conditional variance of e_t given $a_t = a$, for $a = 0, 1$.