Supplementary Materials for "Unifying Offline Causal Inference and Online Bandit Learning for Data Driven Decision"

Li Ye
The Chinese University of Hong Kong

Hong Xie
College of Computer Science, Chongqing University

Yishi Lin
Tencent Inc.

John C.S. Lui
The Chinese University of Hong Kong

1 MORE THEORETICAL RESULTS

1.1 General Lower Bound on The Regret

Theorem 8 (General lower bound). *Suppose for any bandit oracle O, $\exists$ a non-decreasing function $h(T)$, s.t. $R(T, \mathcal{A}_{O+\mathcal{E}_\emptyset}) \geq h(T)$ for $\forall T$. Suppose the offline estimator $\mathcal{E}$ returns unbiased outcomes $\{y_j\}_{j=1}^N$ w.r.t. $\{(\boldsymbol{x}_j, a_j)\}_{j=1}^N$. Then for any contextual-independent algorithm $\mathcal{A}_{O+\mathcal{E}}$, we have:*

$$R(T, \mathcal{A}_{O+\mathcal{E}}) \geq h(T) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a = a_j] \right).$$

For any contextual algorithm $\mathcal{A}_{O_c+\mathcal{E}}$, we have

$$R(T, \mathcal{A}_{O_c+\mathcal{E}}) \geq h(T) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a, \boldsymbol{x}_j] - \mathbb{E}[y|a=a_j, \boldsymbol{x}_j] \right).$$

Theorem 8 shows how we can apply the regret *"lower bound"* of online bandit oracles (e.g. [8]) to derive a regret lower bound with logged data . When an algorithm's upper bound meets the lower bound, we get a *nearly optimal* online decision algorithm that uses the logged data. The proof of Theorem 8 is in Section 3.1.

Definition 1 (The value of logged data)**.** *The online learning oracle O has a regret upper bound $g(T)$ after T time slots. Suppose the regret of an algorithm $\mathcal{A}$ that uses logged data is upper bounded by $R(T, \mathcal{A})$. Then, we call $g(T) - R(T, \mathcal{A})$ the "value of logged data" in time T.*

The "value of logged data" quantifies the reduction of regret by using the logged data. The following corollary gives a lower bound on the "value of logged data" for large T.

Corollary 1. *Suppose conditions in Theorem 1 hold. Suppose the offline evaluator returns $\{\tilde{y}_j\}_{j=1}^N$ w.r.t. $\{(\tilde{\boldsymbol{x}}_j, \tilde{a}_j)\}_{j=1}^N$ till time T. If an online bandit oracle satisfies the "no-regret" property, i.e. $\exists$ a regret upper bound $g(T)$, such that $\lim_{T\to\infty} g(T)/T=0$ (and g is concave), then the difference of regret bounds (before and after using offline data) has the following limit for a context-independent algorithm $\mathcal{A}_{O+\mathcal{E}}$:*

$$\lim_{T\to+\infty} g(T) - R(T, \mathcal{A}_{O+\mathcal{E}}) \geq \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a=\tilde{a}_j] \right).$$

For a contextual algorithm $\mathcal{A}_{O_c+\mathcal{E}}$, the limit of such difference

$$\lim_{T\to+\infty} g(T) - R(T, \mathcal{A}_{O_c+\mathcal{E}}) \geq \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a, \tilde{\boldsymbol{x}}_j] - \mathbb{E}[y|a=\tilde{a}_j, \tilde{\boldsymbol{x}}_j] \right).$$

1.2 Problem independent regret upper bound on $\mathcal{A}_{\text{LinUCB+LR}}$

Theorem 9 (Linear regression+LinUCB, problem-independent). *Suppose we have N offline data points. With a probability at least $1-\delta$, the psuedo-regret (here, $V_0 = \boldsymbol{I}_d$ is a $d \times d$ identity matrix)*

$$R(T, \mathcal{A}_{LinUCB+LR}) \leq \sqrt{8(N+T)\beta_T(\delta) \log \frac{\texttt{trace}(V_0)+(N+T)L^2}{\texttt{det}(V_0)}} - \sqrt{8\beta_T(\delta)} \min\{1, ||\boldsymbol{x}||_{\min}\} \frac{2}{L^2} \left(\sqrt{1+NL^2} - 1 \right).$$

Here, $\{\beta_t(\delta)\}_{t=1}^T$ is a non-decreasing sequence where $\beta_t(\delta) \geq 2d(1 + 2\ln(1/\delta))$. In addition, $L=||\boldsymbol{x}||_{\max}$ is the maximum of l_2-norm of the context in any time slot.

The regret upper bound of Theorem 9 consists of two terms. The first term that is from the online bandit oracle is $O(\sqrt{(N+T)\log(N+T)})$. The second term is the reduction of regret by matching logged data which is $-\Omega(\sqrt{N\log(N+T)})$. Comparing with the regret bound $O(\sqrt{T\log(T)})$ for only using the online feedbacks [1], the regret bound changes from $O(\sqrt{T\log(T)})$ to $O(\sqrt{(N+T)\log(N+T)}) - \Omega(\sqrt{N\log(N+T)})$. To illustrate the reduction, we observe that $\sqrt{N+T} - \sqrt{N} = \sqrt{T}\frac{\sqrt{T}}{\sqrt{N+T}+\sqrt{N}} \leq \sqrt{T}$, where "$\sqrt{N+T} - \sqrt{N}$" is for our regret bound with logged data, and "$\sqrt{T}$" is for the previous bound without logged data.

2 MORE EXPERIMENTS AND CODE EXPLAINATION

2.1 Code and experiment settings

Note that we provide the code for reproducibility and one can find the detailed experiment settings in the code. Thus, this section serves as a document of our code.

When we run one experiment, we run the corresponding `python` scripts in the `/experiments` folder. Figure 1 illustrates the Call Graph of one experiment.

Code for the *ϵ-decreasing multi-action forest.* We modify the R package "grf" to implement our multi-action forest. In particular, we implement the `BanditPrediction.cpp` in `grf/core/src` that extends the regression forest (or causal forest) to allow multiple actions under a leaf node. In a typical call for the bandit predictor, the following functions are called in sequence in the file `r-package/grf/R/causal_forest.R`. The order of functions being called is `predict_action`→`causal_predict_action`. Note that although we still use the name `causal_forest` in the names

of our multi-action forest for convenience, our multi-action forest does not call the predictor of "causal forest" but use our own implementation instead.

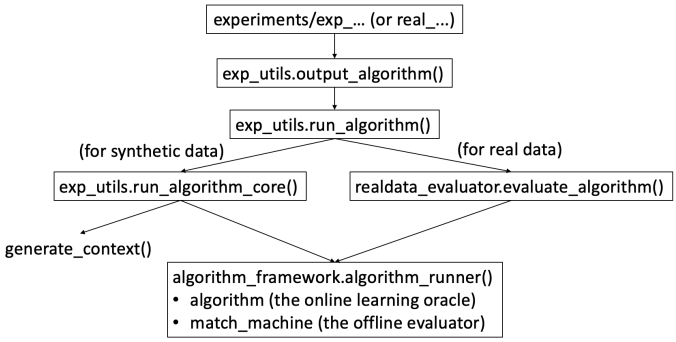

Figure 1: Call Graph of one experiment (in the code)

2.1.1 Settings on the simulation. To do the simulation, we need to simulate an online environment and use it to generate the logged data. To have a unified framework for both the context-independent case and the contextual case, we first have a model to generate the outcome w.r.t. the context and action, and then get the average outcome w.r.t. the actions by summing over all contexts. The simulation code is in `environment.py`.

Note that our method to generate the outcomes for the context-independent case is not restrictive, because the expected reward for each action can be arbitrary. Also, the distribution of reward for each action can be arbitrary by setting different distribution of the contexts.

2.1.2 Settings on WeChat's experiments. Due to confidentiality, we cannot release the data of WeChat. We run the algorithms for five times on each of the eight datasets to get the average.

How we get the propensity scores and weight. In WeChat's datasets, we do not record the propensity score in the observational data. Therefore, we need to estimate the propensity score and the inverse propensity score weighting. For the propensity matching algorithm $\mathcal{A}_{\text{UCB+PSM}}$, we use linear regression to estimate the propensity score. For the inverse propensity score weighting algorithm $\mathcal{A}_{\text{UCB+IPSW}}$, we use WeightIt package[1] to estimate the propensity score (generalized linear model) and get the weight.

Differences-in-differences. We use the differences-in-differences[5] method to pre-process the outcomes in the observational data for practical considerations. Recall that we have two actions (named *"treatment"* and *"control"*). We want to compare the outcomes of these two actions and choose the better one. However, the expected outcome under one action may change over time, e.g. in Chinese new year the users will be more active than in other times. Therefore, we do not want to compare two data items recorded at different time. The differences-in-differences method compares the average change over time in the outcome variable for the treatment group, compared to the average change over time for the control group. Using such time-series data, we can better infer which action has a larger reward.

[1] https://cran.r-project.org/web/packages/WeightIt/WeightIt.pdf

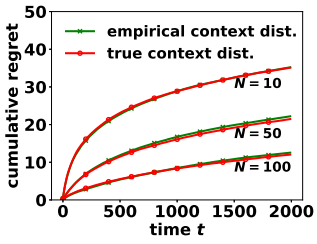

Figure 2: Empirical context distribution ($\mathcal{A}_{\text{UCB+IPSW}}$)

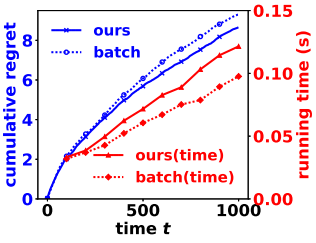

Figure 3: Batch method vs. ours ($\mathcal{A}_{\text{UCB+IPSW}}$)

2.2 Practical Considerations

Exp8: Relaxing knowledge on context distribution. Recall that in our framework Algorithm 1, we propose to use the empirical distribution of the contexts from both offline and online data. In Figure 2, we compare the regret using empirical and true context distribution using synthetic data, where we run the algorithms for 2,000 time to take the average. For various number of logged data $N \in \{10, 50, 100\}$, algorithms that use empirical context distribution or true context distributions have similar regrets. This shows the soundness to use empirical context distribution in our framework. We do not use real data, because for real data we do not know the true context distribution.

Exp9: Comparison to batch method. One variant of our algorithmic framework is to use the logged data all in a batch before the online decisions. In contrast, in our Algorithm 1, we use the logged data before each online decision round t. On synthetic data, Figure 3 shows that our method and the batch method have similar cumulative regrets, although our method is slightly better when t is large. The running time for the two methods increase linearly as the number of online rounds t increases. This shows that both methods are scalable w.r.t. t. Our Algorithm 1 has lower regret when t is large, but is slower compared to its batch variant. We also point out that the batch method do not have theoretical regret guarantee. We do the comparison on real data in our supplement [2].

2.3 Experiments on Unobserved Confounders

For real data of Weixin and Yahoo, probably we do not observe all the confounders [42][13]. Our experiments show that in these real datasets, our algorithms still have the lowest regrets.

We first directly analyze the imapct of unobserved confounders on the regret. Then, we notice that the unobserved confounders create bias in the estimated reward which relates to the "quality of the logged data". Therefore, in the second part, we discuss the impact of the quantity and quality of logged data.

The imapct of unobserved confounders. In Figure 19, we randomly choose a number of confounders and hide them as unobserved. We see that the cumulative regret becomes the lowest when there are no unobserved confounders. When there exists unobserved confounders, the regret do not have a clear relationship with the number of unobserved confounders. This is because when there is some missing information, we do not know whether each part of the missing information has positive or negative impacts on the cumulative regrets.

Impact of the quantity and quality of logged data. Here, we explore the situations where the offline evaluator may return biased samples. In the ideal case, in terms of quantity we have a sufficiently large number of data for each action, and in terms of quality the data records all the confounding factors. In reality, these conditions may not hold.

In Figure 18, we investigate the impacts of both the quantity and quality of data, where we focus on the context-indepedent algorithm $\mathcal{A}_{\text{UCB+IPSW}}$. Recall that the expected rewards for the two actions are 0 and 0.5. Now, in the logged data we add a bias to the first action, and its expected reward becomes "0+bias". We observe that when the bias is 0 or 0.3, the "offline+online" variant $\mathcal{A}_{\text{UCB+IPSW}}$ has the lowest regret. This is because with small bias, the logged data is still informative to select the better action. However, when the bias is as large as 0.9, the "only_online" variant (i.e. UCB) achieves the lowest regret, because the offline estimations are misleading. The impact of the number of logged samples depends on the bias. In the case of zero bias (the left figure), if we have a large number of logged samples (e.g. 100), then our $\mathcal{A}_{\text{UCB+IPSW}}$ algorithm and the "only_offline" IPSW algorithm have low regrets because they use logged data. But when logged data has high bias (the right figure), more logged samples result in a higher regret for algorithms $\mathcal{A}_{\text{UCB+IPSW}}$ and IPSW that use the logged data.

2.4 Thompson Sampling

BanditOracle 4 is the Thompson Sampling algorithm where the reward of the actions are assumed to be Gaussian random variables. Figure 13 in the main paper uses BanditOracle 4. When the reward is of binary values (e.g. in the Yahoo's dataset), one can use the BanditOracle 5 which assume the rewards are Bernoulli random variables. For the Bernoulli Thompson sampling, the mean of the reward has a Beta-distributed posterior distribtion.

BanditOracle 4: Thompson Sampling (Gaussian)

Member variables: the average outcome $\bar{y}_a$ of each action $a \in [K]$, and the number of times n_a that action a was played.
Function play$(\boldsymbol{x})$**:**
  $R_a \leftarrow$ a random variable with normal distribution $\mathcal{N}(\bar{y}_a, \beta^2/(n_a+1))$, for $\forall a \in [K]$.
  $r_a \leftarrow$ is a sample from R_a.
  return $\arg\max_{a\in[K]} r_a$
Function update$(\boldsymbol{x}, a, y)$**:**
  $\bar{y}_a \leftarrow (n_a\bar{y}_a + y)/(n_a+1)$, $n_a \leftarrow n_a + 1$

2.5 Propensity Score Matching for More Than Two Actions

In the main paper, we consider the $\mathcal{A}_{\text{UCB+PSM}}$ algorithm only for two actions $K = 2$. Here, we keep other settings as default and change the number of actions. Figure 4-7 show the cumulative regrets for the $\mathcal{A}_{\text{UCB+PSM}}$ algorithm for the number of actions $K = 2$ to $K = 5$.

Note that the "only_online" algorithm UCB is not affected by the offline evalutor. Therefore, the "only_online" curve can serve as the baseline. First, we observe that when $K > 2$, the "only_offline" PSM algorithm has a high regret, which is much higher than the regret for $K = 2$. Second, when $K > 2$, the cumulative regret for the "offline+online" algorithm $\mathcal{A}_{\text{UCB+PSM}}$ can be higher than that of the "only_online" UCB algorithm. In other words, the propensity score matching offline evaluator does not help reduce the regret by using the offline data. This is because it is difficult to find matched samples with similar propensity vector and our stratification strategy introduces further bias on the estimated reward. Moreover, when $K > 2$, the regret for the "only_offline" PSM algorithm does not necessarily depend on the number of actions K. This is because PSM algorithm cannot effectively use the offline data and the decision depends on some other non-informative factors such as how the values are stratified.

Lessons learned. The original original version of propensity score matching algorithm (with stratification) is not suitable for more than two actions.

BanditOracle 5: Thompson Sampling (Bernoulli)

Member variables: the number of "1"'s s_a (success) in the feedback for each action $a \in [K]$, and the number of "0"'s f_a (failure) in the feedback for each action $a \in [K]$.
Function play$(\boldsymbol{x})$**:**
  $R_a \leftarrow$ a random variable with beta distribution $Beta(s_a, f_a)$, for $\forall a \in [K]$.
  $r_a \leftarrow$ is a sample from R_a.
  return $\arg\max_{a\in[K]} r_a$
Function update$(\boldsymbol{x}, a, y)$**:**
  if $y = 1$ **then**
    $s_a \leftarrow s_a + 1$
  else
    $f_a \leftarrow f_a + 1$

2.6 Experiment on Other Settings of Synthetic Data

We will extend the default experiment settings in three aspects: (1) the number of actions, (2) the propensity score function $ps(\boldsymbol{x}, a)$, and (3) the outcome function $f(\boldsymbol{x}, a)$.

The number of actions. In Figure 8-11, we increase the number of actions from 3 to 8 for the $\mathcal{A}_{\text{UCB+IPSW}}$ algorithm. First, we observe that for each number of actions, our $\mathcal{A}_{\text{UCB+IPSW}}$ algorithm always has a lower regret compared to its two variants. Second, we observe that as the number of actions increases, the difference between the regret of the "offline+online" algorithm $\mathcal{A}_{\text{UCB+IPSW}}$ and the regret of the "only_online" UCB algorithm becomes smaller. This is because when we have more actions, we need more logged data so that the numbers of logged data are sufficient for each actions.

The propensity score function. In the main paper, we set the propensity score function to $ps(\boldsymbol{x}, a) = \exp(s_a)/(\sum_{a=0}^{K-1}\exp(s_a))$, where $s_a = \exp(\rho \boldsymbol{x}^T \boldsymbol{\theta}_a(\mathbb{E}[y|a] - \mathbb{E}[y|(a+1) \mod K]))$ and $\rho = -1$. The parameter ρ controls the correlation between the action and the outcome given the contexts. Negative ρ indicates the following negative correlation: when $\rho < 0$, if an action has a higher expected

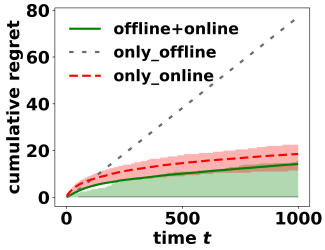

Figure 4: $\mathcal{A}_{\text{UCB+PSM}}$, $K = 2$

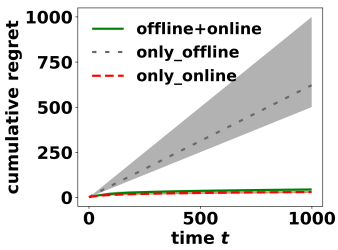

Figure 5: $\mathcal{A}_{\text{UCB+PSM}}$, $K = 3$

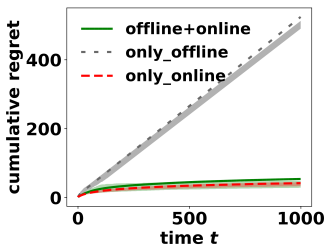

Figure 6: $\mathcal{A}_{\text{UCB+PSM}}$, $K = 4$

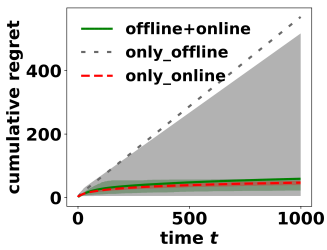

Figure 7: $\mathcal{A}_{\text{UCB+PSM}}$, $K = 5$

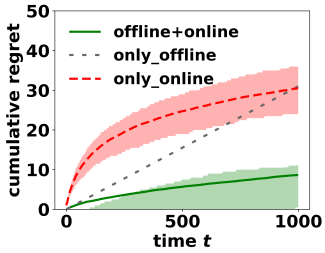

Figure 8: $\mathcal{A}_{\text{UCB+IPSW}}$, $K = 3$

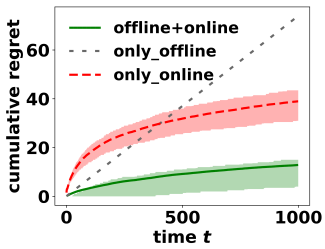

Figure 9: $\mathcal{A}_{\text{UCB+IPSW}}$, $K = 4$

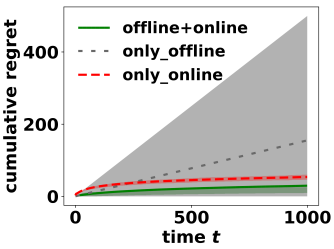

Figure 10: $\mathcal{A}_{\text{UCB+IPSW}}$, $K = 6$

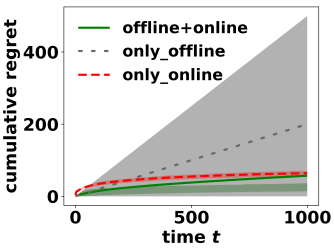

Figure 11: $\mathcal{A}_{\text{UCB+IPSW}}$, $K = 8$

reward, then the samples of this action will be selected with a higher probability if the sample reward is lower. In the following experiment, we explore more settings where $\rho = 0$ or $\rho = 1$. Here, $\rho = 0$ means that each action will have the same propensity score, i.e., each action will be selected with equal probability.

The outcome function. In our main paper, the default outcome function is the linear function $y = f(\boldsymbol{x}, a) = \boldsymbol{x}^T\boldsymbol{\theta}_a + b_a$. Here, we consider two variants of the outcome function. The first is the sigmoid function $y = 1/(1 + \exp(-\boldsymbol{x}^T\boldsymbol{\theta}_a + b_a))$. The second is the binary outcome $y \in \{0, 1\}$ where $y = 1$ with probability $1/(1 + \exp(-\boldsymbol{x}^T\boldsymbol{\theta}_a + b_a))$. We point out that the expected reward for the "sigmoid" and the "binary" settings are the same.

Figure 15, 16 and 17 are the results for the linear outcome function, the sigmoid outcome function and the binary outcome function respectively. We observe that the outcome function significantly affects the performance of the algorithms. For sigmoid outcome function, our "offline+online" algorithm and the "only_offline" algorithm almost have zero regret. It means that the 100 logged samples provide enough information for the decision maker to distinguish the action with the highest expected reward. When the outcome is binary, our "offline+online" algorithm has a lower regret than the "only_online" UCB algorithm. Although the sigmoid function and the binary outcome function correspond to the same expected reward for each action, the regret is higher for the binary outcome because the binary outcome function implies a larger variance of the outcome.

2.7 Linear vs. Forest Model on Yahoo's Data

In Figure 20 and Figure 21, we compare the cumulative reward for $\mathcal{A}_{\text{LinUCB+LR}}$ and $\mathcal{A}_{\text{Fst+MoF}}$ on Yahoo's data. We see that the two algorithms result in similar cumulative regrets. Recall that the user features in the Yahoo's data were learned via a linear model. In other words, our non-paramtric forest model achieves comparable performance with the LinUCB even on the "linear" dataset.

2.8 Comparison to Batch Method on Real Data

The batch version of our algorithmic framework is outlined as Algorithm 2. There are several differences between the batch variant and our original algorithmic framework in Algorithm 1. First, in the online phase (Line 13-16) of the batch variant, we do not use the offline data. Second, in Line 7 of Algorithm 2, the action a is not generated by the online learning oracle, but is a fixed value inside the for-loop. Because not all the actions are generated by the bandit oracle, we cannot directly use the theoretical results of existing bandit algorithms.

In Figure 22, we show that the cumulative regrets for the batch method and our method are almost indisdinguishable on Yahoo's dataset. This further validate our observation in the main paper on the synthetic data.

2.9 Detailed results of WeChat's experiments

Figure 9 shows the results for three variants of algorithms $\mathcal{A}_{\text{UCB+IPSW}}$ on the WeChat dataset. We see our algorithm that uses both the logged data and online feedbacks has the lowest regret. After 200,000 rounds, our algorithm $\mathcal{A}_{\text{UCB+IPSW}}$ reduces the total regret by 45.1% (or 21.7%) compared to the "only_online" UCB algorithm (or the "only_offline" IPSW algorithm).

Due to confidentiality, we cannot release the WeChat data. But we list the experiment results on each of the eight datasets in Figure 24, 25 and 26, so as to show the details of our experiments on WeChat.

Figure 23 lists the cummulative regret for the three variants of algorithm $\mathcal{A}_{\text{PSM+UCB}}$ on the eight datasets from WeChat. Each figure is for each dataset. There are two kinds of typical results. An example of the first kind is the results for the first dataset

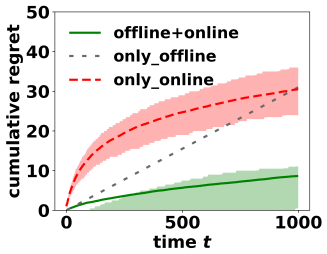

Figure 12: $\mathcal{A}_{\text{UCB+IPSW}}$, $\rho = -1$

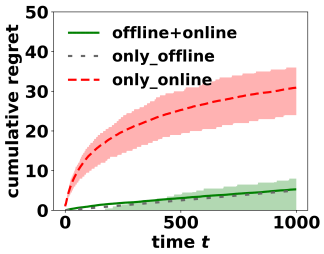

Figure 13: $\mathcal{A}_{\text{UCB+IPSW}}$, $\rho = 0$

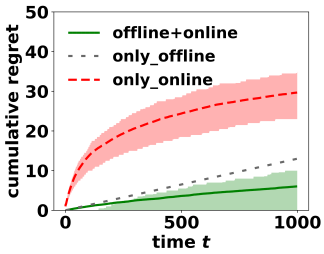

Figure 14: $\mathcal{A}_{\text{UCB+IPSW}}$, $\rho = 1$

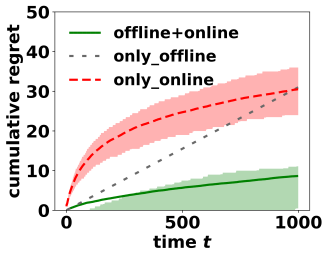

Figure 15: $\mathcal{A}_{\text{UCB+IPSW}}$, linear function

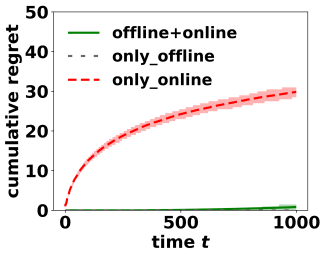

Figure 16: $\mathcal{A}_{\text{UCB+IPSW}}$, sigmoid function

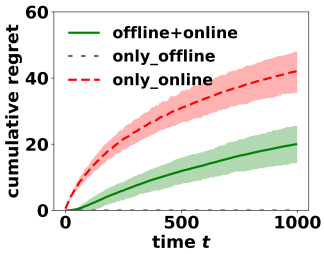

Figure 17: $\mathcal{A}_{\text{UCB+IPSW}}$, binary outcome

Figure 18: The impact of the bias and the number of logged samples on the total regrets for $\mathcal{A}_{\text{UCB+IPSW}}$ (T=500)

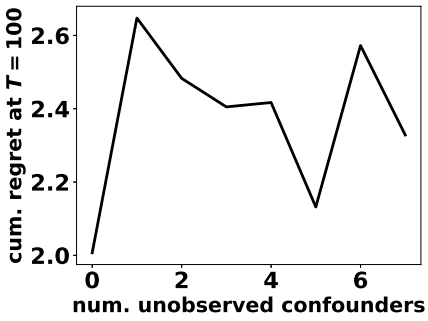

Figure 19: The impact of unobserved confounders for $\mathcal{A}_{\text{UCB+IPSW}}$

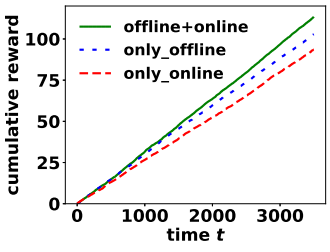

Figure 20: $\mathcal{A}_{\text{LinUCB+LR}}$ on Yahoo's data

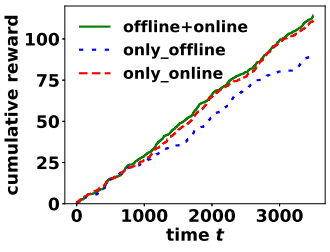

Figure 21: $\mathcal{A}_{\text{Fst+MoF}}$ on Yahoo's data

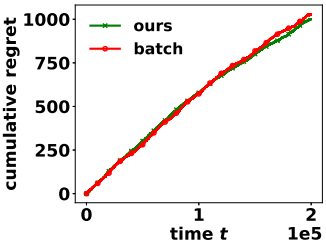

Figure 22: Batch mode vs. our method on Yahoo's data

(row 1, column 1 in Figure 23). For this kind of results, the regret for the "only_offline" variant is low, and "offline+online" variant has lower regret than the "only_online" variant. We observe the first kind of results because the samples from the offline evaluator have small biases so that making decisions only based on logged data is good enough, and initializing the online decision maker

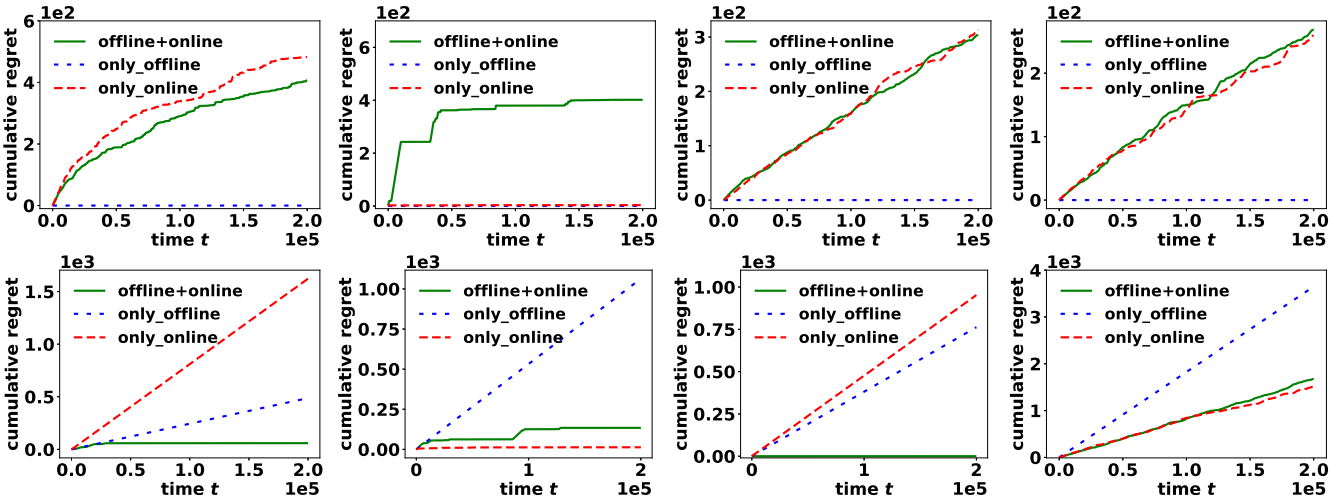

Figure 23: The cumulative regrets on the eight datasets for algorithm $\mathcal{A}_{\text{UCB+PSM}}$

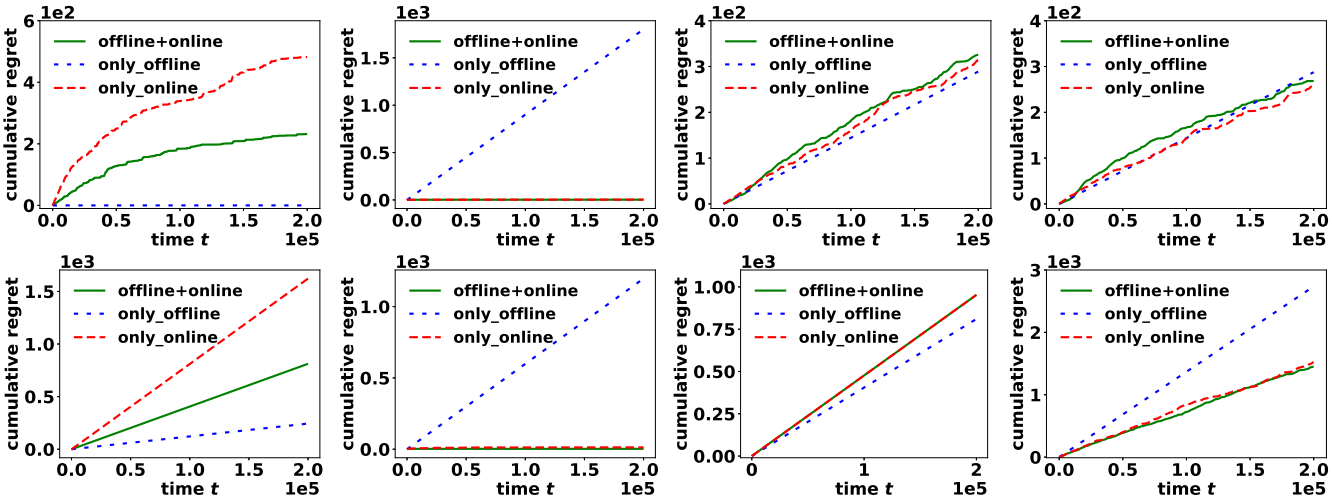

Figure 24: The cumulative regrets on the eight datasets for algorithm $\mathcal{A}_{\text{UCB+IPSW}}$

with the logged data is useful compared to only relying on the online feedbacks. An example of the second kind is the results for the sixth dataset (row 2, column 2). For this kind of results, the regret for the "only_offline" variant is high and it increases linearly in time t. Moreover, the regret for the "only_online" variant is low, which is slightly lower than our "offline+online" variant. We observe the second kind of results because the samples from the offline evalutor have high biases so that making decisions only based on logged data is bad, and the initialization from the logged data needs the corrections from online feedbacks. In the first kind of results, our "offline+online" variant outperforms the "only_online" and in the second kind of results, our "offline+online" variant is slightly worse the "only_online" variants. That is why on average of the eight datasets, our "offline+online" variant outperforms the "only_online". In the second kind of results, our "offline+online" variant can correct the wrong initialization, and thus significantly outperforms the "only_offline" variant whose regret is linear in time t. That is why on average of the eight datasets, our "offline+online" variant outperforms the "only_offline" variant.

In Figure 24, we have similar observations as Figure 23. In Figure 25 and Figure 26, we show the detailed cumulative rewards (on the eight datasets) for three variants of contextual algorithms $\mathcal{A}_{\text{LinUCB+LR}}$ and $\mathcal{A}_{\text{Fst+MoF}}$ respectively.

3 PROOFS

In our main paper, we have Theorem 1, 3, 7. We give proofs of these three theorems in Section 3.1, 3.2, 3.3.

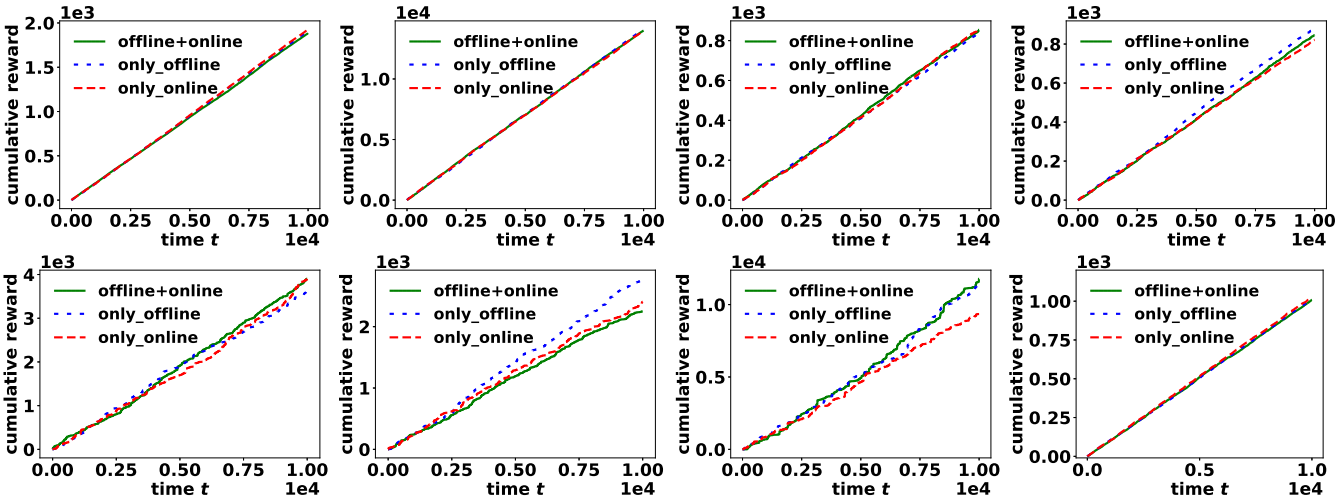

Figure 25: The cumulative rewards on the eight datasets for algorithm $\mathcal{A}_{\text{LinUCB}}$

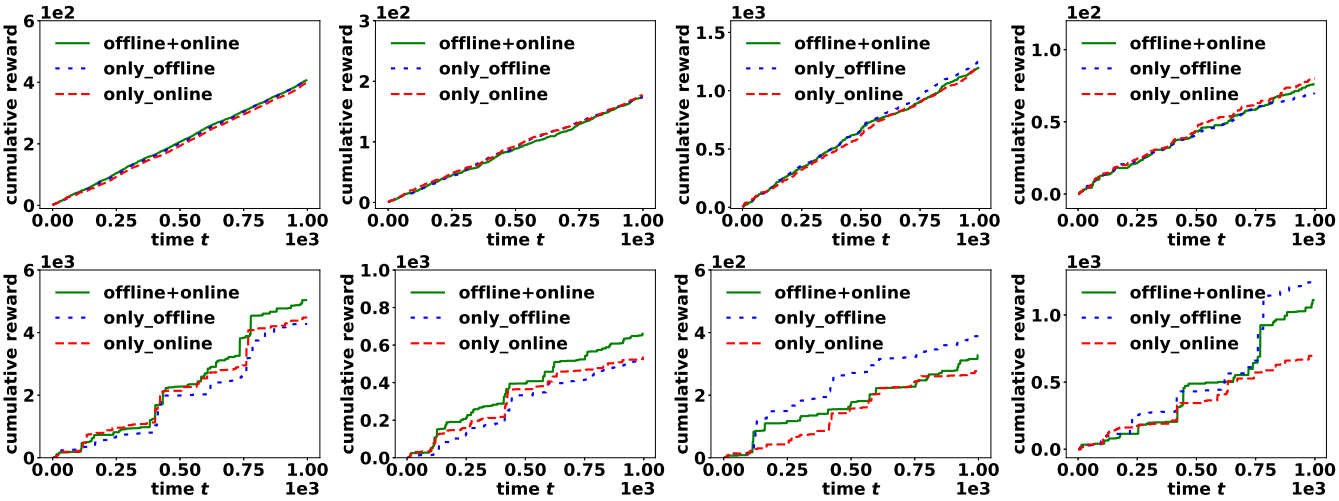

Figure 26: The cumulative rewards on the eight datasets for algorithm $\mathcal{A}_{\text{Fst+MoF}}$

3.1 General Regret Upper and Lower Bounds (Theorem 1 and Theorem 8)

Now, we prove the general upper bound of our framework.[2]

Proof of Theorem 1. The proof follows the idea described in Section 3.2. Online learning oracle is called for $N+T$ times, including N times with synthetic feedbacks and T times with real feedbacks. Denote the total pseudo-regret in these $N+T$ time slots as $R(\mathcal{A}_{O+\mathcal{E}_\emptyset}, N+T)$. Because the condition (2) ensures that our offline evaluator returns unbiased i.i.d. samples in different time slots, the online bandit oracle cannot distinguish these offline samples from online samples. (This is because the regret bound only depends on the expected rewards of each arm and the offline evaluator $\mathcal{E}$ is unbiased.) Then according to the regret bound of the online learning oracle, we have

$$R(\mathcal{A}_{O+\mathcal{E}_\emptyset}, N+T) \le g(N+T). \tag{1}$$

Moreover, we could decompose the total *expected* regret of the online learning oracle as

$$R(\mathcal{A}_{O+\mathcal{E}_\emptyset}, N+T) = \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a=\tilde{a}_j] \right) + R(\mathcal{A}_{O+\mathcal{E}}, T) \tag{2}$$

On the right hand side of (2), the first term $\sum_{j=1}^{N} (\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a=\tilde{a}_j])$ is the cummulative regret of the bandit oracle in the offline phase, and the second term $R(\mathcal{A}_{O+\mathcal{E}}, T)$ is the cumulative

[2] We have a technical condition that regret bounds of the online bandit oracle $g(T)$ only depends on expected rewards of each arm (e.g. the regret bound of UCB [4] only depends on the expected reward).

Algorithm 2: Algorithmic Framework - Batch Variant

Initialize the *OfflineEvaluator* with logged data $\mathcal{L}$
Initialize the *BanditOracle*
//The offline phase
for $a \in [K]$ **do**
  while *True* **do**
    $\boldsymbol{x} \leftarrow$ *context_generator()* //from CDF $F_{\boldsymbol{X}}(\cdot)$
    $y \leftarrow$ *OfflineEvaluator*.**get_outcome**$(\boldsymbol{x}, a)$
    if $y \neq$ NULL **then**
      BanditOracle.**update**$(\boldsymbol{x}, a, y)$
    else //offline evaluator cannot synthesize a feedback
      break
//The online phase
for $t = 1$ *to* T **do**
  $a_t \leftarrow$ *BanditOracle*.**play**$(\boldsymbol{x}_t)$ //online play
  $y_t \leftarrow$ the outcome from the online environment
  BanditOracle.**update**$(\boldsymbol{x}_t, a_t, y_t)$

regret in the online phase. Combining (1) and (2), we get

$$R(\mathcal{A}_{O+\mathcal{E}}, T) \leq g(N+T) - \sum_{j=1}^{N} (\mathbb{E}[y|a^*] - \mathbb{E}[y|\tilde{a}_j]),$$

which concludes our proof for the context-independent case. For the contextual case, the proof is similar and we only need to replace $\mathbb{E}[y|a]$ with $\mathbb{E}[y|a, \boldsymbol{x}]$. □

Proof of Corollary 1. Based on Theorem 1, we only need to show $\lim_{T\to+\infty} g(N+T) - g(T) = 0$. Before we start our proof, we want to point out that regret bounds of many bandit algorithms have "no-regret" property. For example, the regret bound $g(T)$ for UCB is proportional to $\log(T)$, the regret bound $g(T)$ for EXP3 is proportional to $\sqrt{T}$. These functions w.r.t. T are sub-linear and concave. These functions are concave because as the oracle receives more online feedbacks, it makes better decisions and thus has less regret per time slot. For the concave function, $\frac{g(N+T)-g(T)}{N}$ is decreasing in T. We claim that $\lim_{T\to+\infty} \frac{g(N+T)-g(T)}{N} = 0$. Otherwise, there will be a $l > 0$, such that $\frac{g(N+T)-g(T)}{N} \geq l$, for $T \geq T_0$ where T_0 is a constant. It means that gradient of $g(T)$ is larger than l when T is large. Then, $\lim_{T\to+\infty} g(T)/T \geq l$ which contradicts to the "no-regret" property.

Then $N \times \lim_{T\to+\infty} \frac{g(N+T)-g(T)}{N} = N \times 0 = 0$. Now we have

$$\begin{aligned}
&\lim_{T\to+\infty} g(T) - R(T, \mathcal{A}_{O+\mathcal{E}}) \\
&= \lim_{T\to+\infty} (g(T) - g(N+T)) + \lim_{T\to+\infty} (g(N+T) - R(T, \mathcal{A}_{O+\mathcal{E}})) \\
&\geq 0 + \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a = \tilde{a}_j] \right),
\end{aligned}$$

which completes our proof for the context-independent case. For the contextual case, the proof is similar and we only need to replace $\mathbb{E}[y|a]$ with $\mathbb{E}[y|a, \boldsymbol{x}]$. □

Theorem 8 (General lower bound). *Suppose for any bandit oracle O, $\exists$ a non-decreasing function $h(T)$, s.t. $R(T, \mathcal{A}_{O+\mathcal{E}_\emptyset}) \geq h(T)$ for $\forall T$. Suppose the offline estimator $\mathcal{E}$ returns unbiased outcomes $\{y_j\}_{j=1}^N$ w.r.t. $\{(\boldsymbol{x}_j, a_j)\}_{j=1}^N$. Then for any contextual-independent algorithm $\mathcal{A}_{O+\mathcal{E}}$, we have:*

$$R(T, \mathcal{A}_{O+\mathcal{E}}) \geq h(T) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a = a_j] \right).$$

For any contextual algorithm $\mathcal{A}_{O_c+\mathcal{E}}$, we have

$$R(T, \mathcal{A}_{O_c+\mathcal{E}}) \geq h(T) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a, \boldsymbol{x}_j] - \mathbb{E}[y|a=a_j, \boldsymbol{x}_j] \right).$$

Proof of Theorem 8. After decomposing the total regret to the offline phase and online phase, we have for any bandit oracle O

$$\begin{aligned}
R(T, \mathcal{A}_{O+\mathcal{E}}) &= R(T+N, \mathcal{A}_{O+\mathcal{E}_\emptyset}) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a=a_j] \right) \\
&\geq h(T+N) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a=a_j] \right). \qquad (3)
\end{aligned}$$

Next, for a non-decreasing function $h(\cdot)$ we have

$$h(T+N) \geq h(T). \qquad (4)$$

Combining (3) and (4), we have

$$R(T, \mathcal{A}_{O+\mathcal{E}}) \geq h(T) - \sum_{j=1}^{N} \left(\max_{a\in[K]} \mathbb{E}[y|a] - \mathbb{E}[y|a = a_j] \right),$$

which concludes our proof for the unbiased estimators. For the contextual case, the proof is similar and we only need to replace $\mathbb{E}[y|a]$ with $\mathbb{E}[y|a, \boldsymbol{x}]$. □

3.2 Regret Bounds for Context-Independent Algorithms $\mathcal{A}_{\text{UCB+EM}}$ and $\mathcal{A}_{\text{UCB+PSM}}$ (Theorem 2 and Theorem 3)

Proof of Theorem 2. The proof consists of three steps. The first step is to decompose the regret as "the total regret" - "the virtual regret". In the second step, we give a bound to the virtual regret. In the third step, we bound the total regret.

The idea of the proof is similar to the proof of the general upper bounds Theorem 1. According to Assumption 2 (ignorability), the exact-matching offline evaluator returns unbiased outcomes. Since all the decisions are made by the online learning oracle, we can apply the regret bound of the UCB algorithm, and minus the regrets of *virtual* plays for the samples returned by the exact matching evaluator.

Step 1: As usual, to analyze a UCB-like algorithm, we count the number of times we draw each arm.

Definition 2. *λ_a is defined as the expected number of rounds that the a_{th} arm is pulled by the online learning oracle.*

We say an *"offline evaluator returns the a_{th} arm"* if $\mathcal{I}(\boldsymbol{x}, a) \neq \emptyset$ in Line 5 of OfflineEvaluator 1 ($\mathcal{E}_{\text{EM}}$), and meanwhile, the context-action pair $(\boldsymbol{x}, a)$ is *matched* by the offline evaluator. Otherwise, if $\mathcal{I}(\boldsymbol{x}, a) = \emptyset$ in Line 5 of OfflineEvaluator 1, we say $(\boldsymbol{x}, a)$ is *unmatched.*

Definition 3. *Let M_a be the number of times that the offline evaluator returns the a_{th} arm.*

Recall that $\Delta_a = \mathbb{E}[y|a^*] - \mathbb{E}[y|a]$. Then, the expected regret

$$R(\mathcal{A}_{\text{UCB+EM}}, T) = \sum_{a\in[K]} \mathbb{E}[(\lambda_a - M_a)]\Delta_a. \tag{5}$$

Now, we count the number of times M_a that an action a is matched by the exact matching offline evaluator. Denote $M(\boldsymbol{x}^c, a)$ as the number of times the pair $(\boldsymbol{x}^c, a)$ is matched by the offline evaluator, hence $\sum_{c\in[C]} M(\boldsymbol{x}^c, a) = M_a$. We note that M_a is the number of "virtual plays".

Step 2: (lower bound of M_a) The lower bound of M_a corresponds to the lower bound of *regret of virtual play*. Note that when some context-action pair $(\boldsymbol{x}, a)$ is unmatched, the matching process for action a will stop. We consider the following two cases: (1) the matching process does not stop at T. In this case the expected number $\mathbb{E}[M(\boldsymbol{x}^c, a)] = \lambda_a \mathbb{P}[\boldsymbol{x}^c]$, because the context and action are generated independently for the context-independent decisions. (2) the matching process terminates before T. In this case, we run out of the samples with $(\boldsymbol{x}^{\tilde{c}}, a)$. Suppose the unmatched context-action pair is $(\boldsymbol{x}^{\tilde{c}}, a)$ (there are still samples for some other context $\boldsymbol{x}$), then the expected number of matched sample for some other context $\boldsymbol{x}^c$ is $\mathbb{E}[M(\boldsymbol{x}^c, a)] = N(\boldsymbol{x}^{\tilde{c}}, a)\frac{\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}$. This is because the $M(\boldsymbol{x}^{\tilde{c}}, a) = N(\boldsymbol{x}^{\tilde{c}}, a)$ and $\frac{\mathbb{E}[M(\boldsymbol{x}^c,a)]}{\mathbb{E}[M(\boldsymbol{x}^{\tilde{c}},a)]} = \frac{\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}$. The unmatched context can be any $\boldsymbol{x}^c$ $\forall c \in [C]$. Consider the worst case, then $M(\boldsymbol{x}^c, a) \geq \min_{\tilde{c}\in[C]} N(\boldsymbol{x}^{\tilde{c}}, a)\frac{\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}$. Note that when $\tilde{c} = c$, we have $\frac{N(\boldsymbol{x}^{\tilde{c}},a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]} = N(\boldsymbol{x}^{\tilde{c}}, a)$. Combining the counts of $M(\boldsymbol{x}^c, a)$ in the above two cases, we have

$$\mathbb{E}[M_a] \geq \sum_{c\in[C]} \min\left\{\min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, \lambda_a \mathbb{P}[\boldsymbol{x}^c]\right\}. \tag{6}$$

Combine (5) and (6), and we note λ_a=$\lambda_a \sum_{c\in[C]} \mathbb{P}[\boldsymbol{x}^c]$ (because $\sum_{c\in[C]} \mathbb{P}[\boldsymbol{x}^c] = 1$ by definition), then

$$R(\mathcal{A}_{\text{UCB+EM}}, T) \leq \sum_{a\in[K]} \Delta_a \times \left(\sum_{c\in[C]} \mathbb{E}\left[\max\{\lambda_a \mathbb{P}[\boldsymbol{x}^c] - \min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, 0\}\right]\right).$$

We have the following equality:

$$\begin{aligned}
&\max\{\lambda_a \mathbb{P}[\boldsymbol{x}^c] - \min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, 0\} \\
=&\max\{l_a \mathbb{P}[\boldsymbol{x}^c] + (\lambda_a - l_a)\mathbb{P}[\boldsymbol{x}^c] - \min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, 0\} \\
=&\max\{l_a \mathbb{P}[\boldsymbol{x}^c] - \min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, 0\} + (\lambda_a - l_a)\mathbb{P}[\boldsymbol{x}^c].
\end{aligned}$$

where we define

$$l_a \triangleq \lceil (8\ln(T + \mathbb{E}[\sum_{a\in[K]} M_a]))/\Delta_a^2 \rceil. \tag{7}$$

Then, $l_a \geq \mathbb{E}[\lceil \mathbb{E}[8\ln(T + \sum_{a\in[K]} M_a)]\rceil]$ because $\ln(\cdot)$ is a concave function (according to Jensen's inequality, the right term takes the expectation out). According Assumption 1 and 3 (stable unit in offline and online cases) and "the reward y is bounded in $[0, 1]$", we can apply the results in paper of Auer et al.[4] and $\mathbb{E}[\lambda_a - l_a] \leq 1 + \frac{\pi^2}{3}$ for some sub-optimal action $a \neq a^*$. Therefore, we have

$$R(\mathcal{A}_{\text{UCB+EM}}, T) \leq \sum_{a\in[K]} \left((1 + \frac{\pi^2}{3}) + \sum_{c\in[C]} \max\{l_a \mathbb{P}[\boldsymbol{x}^c] - \min_{\tilde{c}\in[C]} \frac{N(\boldsymbol{x}^{\tilde{c}}, a)\mathbb{P}[\boldsymbol{x}^c]}{\mathbb{P}[\boldsymbol{x}^{\tilde{c}}]}, 0\}\right)\Delta_a. \tag{8}$$

Step 3: (upper bound of M_a) To get an upper bound for l_a, we now give an upper bound for the expected number of samples that are matched, i.e. $\mathbb{E}[\sum_{a\in[K]} M_a]$. Recall that we denote the number of matched samples with context $\boldsymbol{x}^c$ and arm a as $M(\boldsymbol{x}^c, a)$. Then, because it cannot exceed the number of data samples, we have "the trivial bound"

$$\mathbb{E}[M(\boldsymbol{x}^c, a)] \leq N(\boldsymbol{x}^c, a). \tag{9}$$

Also, because the expected number of matched samples cannot exceed the expected number of times the action is selected, we have "the refined bound"

$$\mathbb{E}[M(\boldsymbol{x}^c, a)] \leq \mathbb{E}[\lambda_a]\mathbb{P}[\boldsymbol{x}^c]. \tag{10}$$

Therefore, combining (9) and (10), we have

$$\mathbb{E}[M(\boldsymbol{x}^c, a)] \leq \max\{N(\boldsymbol{x}^c, a), \lambda_a \mathbb{P}[\boldsymbol{x}^c]\}.$$

Then,

$$\begin{aligned}
&\mathbb{E}[\sum_{a\in[K]} M_a] \leq \sum_{c\in[C]} \sum_{a\in[K]} \min\{N(\boldsymbol{x}^c, a), \lambda_a \mathbb{P}[\boldsymbol{x}^c]\} \\
=&-\sum_{c\in[C]} \sum_{a\in[K]} \max\{-N(\boldsymbol{x}^c, a), -\lambda_a \mathbb{P}[\boldsymbol{x}^c]\} \\
=&\sum_{c\in[C]} \sum_{a\in[K]} N(\boldsymbol{x}^c, a) - \\
&\sum_{c\in[C]} \sum_{a\in[K]} \max\left\{N(\boldsymbol{x}^c,a) - N(\boldsymbol{x}^c,a), N(\boldsymbol{x}^c,a) - \mathbb{E}[\lambda_a]\mathbb{P}[\boldsymbol{x}^c]\right\} \\
=&N - \sum_{c\in[C]} \sum_{a\in[K]} \max\{0, N(\boldsymbol{x}^c, a) - \mathbb{E}[\lambda_a]\mathbb{P}[\boldsymbol{x}^c]\} \\
\leq&N - \sum_{c\in[C]} \sum_{a\in[K]} \max\{0, N(\boldsymbol{x}^c, a) - (8\frac{\ln(T+N)}{\Delta_a^2} + 1 + \frac{\pi^2}{3})\mathbb{P}[\boldsymbol{x}^c]\}.
\end{aligned} \tag{11}$$

Recall that N is the number of all logged samples. The last equation is because $\lambda_a \leq 8\frac{\ln(T+N)}{\Delta_a^2} + 1 + \frac{\pi^2}{3}$ according to the paper [4].

Plug-in (7) and (11) to (8), then we have the upper bound claimed by our Theorem. □

Proof of Theorem 3. The proof is similar to the proof of Theorem 2 for $\mathcal{A}_{\text{UCB+EM}}$. The only difference is that for propensity score matching, the only context to be matched is the propensity score.

First, we will show that by matching the propensity score, the expected reward in each round for each arm is not changed.

The expected reward when we choose action a is

$$\mathbb{E}[y|a] = \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}]\mathbb{E}[y|a, \boldsymbol{x}],$$

where $\mathbb{E}[y|a, \boldsymbol{x}]$ is the expected reward when the context is $\boldsymbol{x}$ and the action is a. We then consider the expected reward when we use

the propensity score matching strategy. Let us denote the propensity score of choosing an action $\tilde{a}$ under context $\tilde{\boldsymbol{x}}$ as

$$p(\tilde{\boldsymbol{x}}, \tilde{a}) = \mathbb{P}[a = \tilde{a} | \boldsymbol{x} = \tilde{\boldsymbol{x}}].$$

By the propensity matching procedure, the expected reward of choosing an action $\tilde{a}$ is

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}]\mathbb{E}[y|\boldsymbol{p}{=}\boldsymbol{p}(\boldsymbol{x}), a{=}\tilde{a}] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \left(\sum_{c\in[Q]} \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \mathbb{E}[y|\boldsymbol{p}{=}\boldsymbol{p}_c, a{=}\tilde{a}] \right) \\
&= \sum_{c\in[Q]} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \mathbb{E}[y|\boldsymbol{p}{=}\boldsymbol{p}_c, a{=}\tilde{a}].
\end{aligned}$$

and we have

$$\begin{aligned}
\mathbb{E}[y|\boldsymbol{p}{=}\boldsymbol{p}_c, a{=}\tilde{a}] &= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[y|\boldsymbol{x},\tilde{a}] \times \mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \boldsymbol{p}_c(\tilde{a})}{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \times \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \boldsymbol{p}_c(\tilde{a})} \\
&= \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[y|\boldsymbol{x},\tilde{a}] \times \mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}}}{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \times \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}}}.
\end{aligned}$$

Therefore, we have the expected reward

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}]\mathbb{E}[y|\boldsymbol{p}{=}\boldsymbol{p}(\boldsymbol{x}), a{=}\tilde{a}] \\
&= \sum_{c\in[Q]} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \frac{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[y|\boldsymbol{x},\tilde{a}]\mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}}}{\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}}} \\
&= \sum_{c\in[Q]} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[y|\boldsymbol{x},\tilde{a}]\mathbb{P}[\boldsymbol{x}] \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[y|\boldsymbol{x},\tilde{a}]\mathbb{P}[\boldsymbol{x}] = \mathbb{E}[y|\tilde{a}]. \qquad (12)
\end{aligned}$$

The last but one equation is from our assumption that all the propensity scores are belong to a finite set $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_Q\}$, and thus $\sum_{c\in[Q]} \mathbb{1}_{\{\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{p}_c\}} = 1$ (namely, the propensity score belongs to some value in the set).

Hence, our propensity score matching method unbiasedly estimate the $\mathbb{E}[y|\tilde{a}]$ for any action $\tilde{a}$.

With such unbiasedness property, the remaining is the same as Theorem 2, except that the contexts $\boldsymbol{x}$ is replaced by the propensity score $\boldsymbol{p}$ verbatim. □

3.3 Regret Bound for Contextual Algorithm $\mathcal{A}_{\mathbf{Fst+}\mathcal{E}_\emptyset}$ (Theorem 7)

Proof of Theorem 7. The proof of Theorem 7 consists of four parts. First, Lemma 10 will show that if the exploration rate is $\epsilon_t = t^{-1/2(1-\beta)}$, then in each data item of the dataset up till time T, any action $a \in [K]$ will be played with a probability at least $\varepsilon_T = \frac{1}{K} T^{-1/2(1-\beta)}$, i.e. $\mathbb{P}[A_t = a|X = \boldsymbol{x}] \geq \varepsilon_t$. Second, Lemma 11 will show that when each action was played with probability at least ε_t at time t, then the estimation error at that time will be asymptotically bounded. Third, based on the previous asymptotic results, our Lemma 12 will show that when the number of samples is large, the estimation error by our multi-action forest estimator will be small with high probability. Fourth, we use Lemma 11 and Lemma 12 to conclude that the cummulative regret will be small.

Step 1: Recall that in each time slot t, we have a probability ϵ_t to draw a random action. Step 1 is to show that the ϵ-decreasing strategy will create an overlap condition for the dataset of online feedbacks. Moreover, we show that compared to a constant exploration rate (instead of our ϵ-decreasing exploration), our strategy is not doing over-exploration up to a logarithmic factor.

Lemma 10. *We have the following bound for the sum of power*

$$T^{1-p} \leq \sum_{t=1}^{T} t^{-p} \leq T^{1-p} \log(T)^p, \quad \text{for some } p \in (0,1). \qquad (13)$$

Applying to our case, we let $p = -\epsilon_0 = 1/2(1-\beta)$, and

$$T^{1+\epsilon_0} \leq \sum_{t=1}^{T} t^{\epsilon_0} \leq T^{1+\epsilon_0} \log(T)^{-\epsilon_0}.$$

Moreover, in the dataset collected till time T, for a randomly picked data point (X, Y, A), we have $\mathbb{P}[A = a|X = x] \geq \frac{1}{K} T^{-1/2(1-\beta)}$.

Proof. The left inequality is easy to show. As t^{-p} decreases in t, $T^{-p} \leq t^{-p}$ for any $t \leq T$, and thus $T^{1-p} = \sum_{t=1}^{T} T^{-p} \leq \sum_{t=1}^{T} t^{-p}$. Now, we show the right inequality. According to Cauchy-Schwartz inequality (note that $1/p > 1$),

$$\begin{aligned}
\frac{\sum_{t=1}^{T} t^{-p}}{T} &\leq \left(\frac{\sum_{t=1}^{T} (t^{-p})^{1/p}}{T} \right)^p \\
&= \left(\frac{\sum_{t=1}^{T} t^{-1}}{T} \right)^p \leq \left(\frac{\log(T)}{T} \right)^p.
\end{aligned}$$

Then, we get the inequality $\sum_{t=1}^{T} t^{-p} \leq T^{1-p} \log(T)^p$ that is (13). Then, we note that the expected total number of times to do the random exploration is $\sum_{t=1}^{T} t^{\epsilon_0}$ till time T. Thus, the expected number of times that we do the exploration in a randomly picked time slot is $(\sum_{t=1}^{T} t^{\epsilon_0})/T$. For a randomly picked data item, the probability that an action is played $\mathbb{P}[A = a|X = \boldsymbol{x}]$ is greater than or equal to $\frac{1}{K}$ times the probability that we do exploration in a randomly picked time slot. Therefore, $\mathbb{P}[A = a|X = \boldsymbol{x}] \geq \frac{1}{K}(\sum_{t=1}^{T} t^{\epsilon_0})/T = \frac{1}{K} T^{\epsilon_0}$. □

In Lemma 10, our main purpose is to give a lower bound on the overlap (or "exploration") probability. In particular, the lower bound T^{1-p} corresponds to a fixed rate of exploration $\epsilon_t = T^{-p}$ for $\forall t$. Then, for our ϵ-decreasing strategy we give an upper bound and a lower bound compaing to two fixed-exploration-rate strategies.

Step 2: In Lemma 10, we have shown that our ϵ-decreasing exploration gives a *"dynamic"* overlap condition, i.e. ε_t changes in t. In contrast, the usual overlap condition (e.g. [12]) states a constant overlap probability. Now, we will show that under this dynamic overlap condition, we have the asymptotic convergence and normality properties for our multi-action forest estimator.

We first introduce the notation $\lesssim$. Here, $f(s) \lesssim g(s)$ means that $\lim_{s\to+\infty} \frac{f(s)}{g(s)} \leq 1$.

Lemma 11 (Asymptotic bias and variance). *Suppose that we have n i.i.d. training examples $(X_i, Y_i, A_i) \in [0,1]^d \times \mathbb{R} \times [k]$. Suppose the ignorability Assumption 2 holds. Finally, suppose that all potential outcome distributions $(X_i, Y_i(a))$ for $\forall a \in [K]$ satisfy the same regularity assumptions as the pair (X_i, Y_i) did in the statement*

of Theorem 3.1 in [15]. Under this data-generating process, suppose the trained $\mathcal{F}$ (in Line 11) is honest, α-regular with $\alpha \leq 0.2$ in the sense of Definition 1 and 2, and symmetric random-split (in the sense of Definition 3 and 5 in [15]) multi-action forest. Denote $A \triangleq \frac{\pi}{d}\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$ where $\pi \in [0,1]$ is a constant in Definition 3 of [15]. Suppose in the fixed logged data of n samples,

$$\mathbb{P}[A = a|X = \boldsymbol{x}] > \varepsilon_n, \text{ for each } a \in [K], \text{ for any } \boldsymbol{x}. \tag{14}$$

where ε_n is a constant. Then for $s = n^\beta$ where $\beta = 1 - \frac{2A}{2+3A}$

$$|\mathbb{E}[\hat{\mu}_n(\boldsymbol{x}, a)] - \mu(\boldsymbol{x}, a)| \lesssim Md\left(\frac{\varepsilon_n s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}}. \tag{15}$$

In addition, there exists a sequence $\{\sigma_n\}_{n=1}^T$ where $\sigma_n = O(\frac{s}{n})$, $\frac{\mathbb{E}[\hat{\mu}_n(\boldsymbol{x},a)] - \hat{\mu}_n(\boldsymbol{x},a)}{\sigma_n(\boldsymbol{x})} \Rightarrow \mathcal{N}(0,1)$ for $\forall a$, where "$\Rightarrow$" means "converges in distribution". Here, $\hat{\mu}_n(\boldsymbol{x}, a) \triangleq \frac{1}{B}\sum_{b\in[B]} \hat{L}_b(\boldsymbol{x}, a)$ is the prediction by the multi-action forest, with n data samples.

Proof. The proof mirrors the proof of Theorem 4.1 in [15] (or Theorem 11 in its arXiv version[3]). The main steps involve bounding the bias of multi-action forests with an analogue to Theorem 3.2 in [15] (or Theorem 3 in its arXiv version) and their incrementality using an analogue to Theorem 3.3 in [15] (or Theorem 5 in its arXiv version). In general, the same arguments as used with regression forest in [15] goes through, but the constants in the results get worse by a factor ε_n that is the least probability that an action is played in the training data. Given these results, the subsampling-based argument from Section 3.3.2 in [15] can be reproduced almost verbatim, and the final proof of this Theorem is identical to that of Theorem 3.1 in [15] (or Theorem 1 in its arXiv version).

As an ensemble method, the multi-action forest uses a subsample s out of n data points to train a tree. The subsample of data is denoted as $\mathcal{D}_s = (Z_1, \ldots, Z_s) = ((X_{i_1}, Y_{i_1}, A_{i_1}), \ldots, (X_{i_s}, Y_{i_s}, A_{i_s}))$. [15] use the notation X_i while we use the notation $\boldsymbol{x}_i$.

Bias. In this part, we want to show (we copy (15) below) :

$$|\mathbb{E}[\hat{\mu}_n(\boldsymbol{x}, a)] - \mu(\boldsymbol{x}, a)| \lesssim Md\left(\frac{\varepsilon_n s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}}.$$

To establish this claim, we first seek with an analogue to Lemma 2 in the arXiv version of [15], except now s in (31) is replaced by $s_{\min}$, i.e., the minimum of the number of cases (i.e. the minimum number of observations for all the actions $a \in [K]$). Then, $s_{\min}/s \gtrsim \varepsilon_n$, because with probability at least ε_n an action will be taken, so a variant of Equation (32) in [15] where we replace s with $\varepsilon_n s$ still holds for large s. Notice that $\hat{\mu}(\boldsymbol{x}, a)$ is a estimate of $\mathbb{E}[Y(a)|X = \boldsymbol{x}]$ (or $\mu(\boldsymbol{x}, a)$)[4]. Then, we get (15) following the results of Theorem 3.2 in [15] (or Theorem 3 in its arXiv version).

We copy the definition of ν-incrementality (Definition 6 of [15]) here.

Definition 4. *The predictor T is $\nu(s)$-incremental at $\boldsymbol{x}$ if*

$$var[\mathring{T}(\boldsymbol{x}; Z_1, \ldots, Z_s)]/var[\boldsymbol{x}; Z_1, \ldots, Z_s] \gtrsim \nu(s),$$

where $\mathring{T}$ is the Hájek projection

$$\mathring{T} = \mathbb{E}[T] + \sum_{i=1}^{n}(\mathbb{E}[T|Z_i] - \mathbb{E}[T]). \tag{16}$$

In our notation, $f(s) \gtrsim g(s)$ means that $\liminf_{s\to\infty} f(s)/g(s) \geq 1$.

Incrementality. Suppose that the conditions of Lemma 3.2 of [15] (or Lemma 4 in its arXiv version) hold and that T is an honest α-regular multi-action tree in the sense of Definition 1 and 2. Suppose moreover that $\mathbb{E}[Y(a)|X = \boldsymbol{x}]$ and $\mathrm{Var}[Y(a)|X = \boldsymbol{x}]$ for $\forall a \in [K]$ are all Lipschitz continuous at $\boldsymbol{x}$, and that $\mathrm{Var}[Y|X = \boldsymbol{x}] > 0$. Suppose, finally, that the overlap condition (14) holds with $\varepsilon_n > 0$. Then, T is $\nu(s)$-incremental at $(\boldsymbol{x}, a)$ with

$$\nu(s) = \varepsilon_n C_{f,d}/\log(s)^d,$$

where $C_{f,d}$ is the constant from Lemma 3.2 of [15] (or Lemma 4 in its arXiv version).

To prove this claim, we follow the argument of the proof of Lemma 3.2 of [15] (or Lemma 4 in its arXiv version). Like the proof in [15], we focus on the case where $f(x) = 1$, in which case we use $C_{f,d} = 2^{-(d+1)}(d-1)!$. We begin by setting up notation as in the proof of Lemma 3.2 of [15] (or Lemma 4 in its arXiv version). We write the estimation for the action a as $T^a(\boldsymbol{x}; \mathcal{D}) = \sum_{i=1}^s S_i^a Y_i$, where

$$S_i^a = \begin{cases} |\{i : X_i \in L(\boldsymbol{x}; \mathcal{D}_s), A_i = a\}|^{-1} & \text{if } X_i \in L(\boldsymbol{x}; \mathcal{D}_s) \text{ and } A_i = a \\ 0 & else; \end{cases}$$

where $L(\boldsymbol{x}; \mathcal{D}_s)$ denotes the leaf containing $\boldsymbol{x}$ in the tree trained with a subsample of data $\mathcal{D}_s$.

We also define the quantities

$$P_i^a = 1_{\{X_i \text{ is a } k\text{-PNN of } x \text{ among points with action } a\}}.$$

where k-PNN (k-potential nearest neighbor) is defined in Definition 7 in Section 3.3.1 of [15].

Because T^a is a k-PNN predictor, $P_i^a = 0$ implies that $S_i^a = 0$. Moreover, by regularity of tree T^a of the forest $\mathcal{F}$, we know that the number of leaf samples $|\{i : X_i \in L(\boldsymbol{x}; \mathcal{D})\}| \geq k$. Thus, we can verify that

$$\mathbb{E}[S_1^a|Z_1] \leq \frac{1}{k}\mathbb{E}[P_1|Z_1] \tag{17}$$

We are now ready to use the same machinery as the Proof of Lemma 4 in the arXiv version of [15]. Similar to the Proof of Theorem 11 in the arXiv version of [15], the random variable P_1^a now satisfy

$$\mathbb{P}\left[\mathbb{E}[P_1^a|Z_1] \geq \frac{1}{s^2\mathbb{P}[A_1 = a]^2}\right] \lesssim k \times \frac{2^{d+1}\log(s)^d}{(d-1)!}\frac{1}{s\mathbb{P}[A_1 = a]}; \tag{18}$$

by the argument in (17) and ε_n-overlap (14), (18) immediately implies that

$$\mathbb{P}\left[\mathbb{E}[S_1^a|Z_1] \geq \frac{1}{k\varepsilon_n^2 s^2}\right] \lesssim k\frac{2^{d+1}\log(s)^d}{(d-1)!}\frac{1}{\varepsilon_n s}.$$

By construction, we know that (because $\sum_{i=1}^s S_i^a = 1$ by definition)

$$\mathbb{E}[S_1^a|Z_1] = \mathbb{E}[S_1^a] = \frac{1}{s},$$

[3] The paper's arXiv version is available at: https://arxiv.org/pdf/1510.04342.pdf

[4] Here, we actually do not need the ignorability Assumption 2 (a.k.a. unconfoundedness) because the bandit algorithm does online intervention and we can directly get the feedback of $Y(a)$

which by the same argument as [15] implies that

$$\mathbb{E}[\mathbb{E}[S_1^a|Z_1]^2] \gtrsim \frac{(d-1)!}{2^{d+1}\log(s)^d}\frac{\varepsilon_n}{ks}. \tag{19}$$

The second part of the proof follows from a straight-forward adaptation of the proof of Theorem 5 in the arXiv version of [15].

So far, we have proved the tree estimator $T^a(\boldsymbol{x})$ is $\nu(s)$-incremental at $\boldsymbol{x}$ with $\nu(s) = \varepsilon_n C_{f,d}/\log(s)^d$. One can check that the proofs for Lemma 3.5 of [15] (or Lemma 7 in its arXiv version) still goes through verbatim because the proof of Lemma 3.5 in [15] uses the properties of the ensemble of forest, and our multi-action forest uses the same ensemble technique via subsampling.

Now, we are going to show the result in Theorem 3.4 in [15] (or Theorem 8 in its arXiv version), as follows:
claim: (in Theorem 3.4 of [15]) "Suppose, $\mathbb{E}[|Y-\mathbb{E}[Y|X=\boldsymbol{x}]|^{2+\delta}|X=\boldsymbol{x}] \le M$ for some constants $\delta, M > 0$, uniformly over all $x \in [0,1]^d$. Then, there exists a sequence $\sigma_n(\boldsymbol{x},a) \to 0$ such that

$$\frac{\hat{\mu}_n(\boldsymbol{x},a) - \mathbb{E}[\hat{\mu}_n(\boldsymbol{x},a)]}{\sigma_n(\boldsymbol{x},a)} \Rightarrow \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is the standard normal distribution. " Now we prove the above claim following the proof of Theorem 3.4 in [15] (or Theorem 8 in its arXiv version). We focus on the trees w.r.t. the action a. Using the notation from Lemma 7 in the arXiv version of [15], let $\sigma_n(\boldsymbol{x},a)^2 = s^2/nV_1$ be the variance of $\mathring{\hat{\mu}}$ ($\mathring{\hat{\mu}}$ is the Hájek projection of $\hat{\mu}$ defined in (16)) where V_1 is defined in (41) in the arXiv version of [15]. We know that

$$\sigma_n^2 = \frac{s}{n}sV_1 \le \frac{s}{n}\mathrm{Var}[T^a].$$

Here, the variance of the base learner $\mathrm{Var}[T^a]$ is finite by the Assumption in Lemma 3.3 in [15]. So $\sigma_n \to 0$ as desired. Now, by our previous argument on the incremental property, combined with Lemma 3.5 in [15], we have ($\mathring{T}^a$ is the Hájek projection of T^a)

$$\begin{aligned}\frac{1}{\sigma_n^2}\mathbb{E}\left[\left((\hat{\mu}_n(\boldsymbol{x},a)) - \mathring{\hat{\mu}}(\boldsymbol{x},a)\right)^2\right] &\le \left(\frac{s}{n}\right)^2\frac{\mathrm{Var}[T^a]}{\sigma_n^2}\\ &= \frac{s}{n}\mathrm{Var}[T^a]/\mathrm{Var}[\mathring{T}^a]\\ &\lesssim \frac{s}{n}\frac{\log(s)^d}{\varepsilon_n C_{f,d}/4}\\ &\to 0.\end{aligned} \tag{20}$$

Compared to the Proof of Theorem 8 in the arXiv version of [15], the difference is that we add a term ε_n for the incremental property. We have $\frac{s}{n}\frac{\log(s)^d}{\varepsilon_n C_{f,d}/4} \to 0$ by plugging in $s = n^\beta$ and $\varepsilon_n \ge n^{-\frac{1}{2}(1-\beta)}$. Then, following the proof of Theorem 8 in the arXiv version of [15], all we need to check is that $\mathring{\hat{\mu}}$ is asymptotically normal. One way to do so is using the Lyapunov central limit theorem (e.g. [6]). Writing

$$\mathring{\hat{\mu}}(\boldsymbol{x},a) = \frac{s}{n}\sum_{i=1}^n(\mathbb{E}[T^a|Z_i] - \mathbb{E}[T]), \tag{21}$$

it suffices to check the following Lyapunov's condition[5][6]:

$$\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\left[|\mathbb{E}[T^a|Z_i] - \mathbb{E}[T^a]|^{2+\tilde{\delta}}\right] / \left(\sum_{i=1}^n \mathrm{Var}[\mathbb{E}[T^a|Z_i]]\right)^{1+\tilde{\delta}/2} = 0 \tag{22}$$

Using notation in the above discussion about incrementality, we write $T^a = \sum_{i=1}^n S_i^a Y_i$. Thanks to honesty, we can verify that for any index $i > 1$, Y_i is independent of S_i^a conditionally on X_i and Z_1, and so (in the following, we slightly abuse the notation and Y stands for $Y(a)$ for some action a)

$$\begin{aligned}&\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]\\ =&\mathbb{E}[S_1^a(Y_1 - \mathbb{E}[Y_1|X_1])|Z_1] + \left(\mathbb{E}\left[\sum_{i=1}^n S_i^a\mathbb{E}[Y_i|X_i]|Z_1\right] - \mathbb{E}[T^a]\right).\end{aligned}$$

Note that the two right-hand-side terms above are both mean-zero. By Jensen's inequality, we also have that

$$\begin{aligned}&2^{-(1+\tilde{\delta})}\mathbb{E}\left[|\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]|^{2+\tilde{\delta}}\right]\\ \le&\mathbb{E}\left[|\mathbb{E}[S_1(Y_1 - \mathbb{E}[Y_1|X_1])|Z_1]|^{2+\tilde{\delta}}\right]\\ &+\mathbb{E}\left[\left|\mathbb{E}\left[\sum_{i=1}^n S_i^a\mathbb{E}[Y_i|X_i]|Z_1\right] - \mathbb{E}[T^a]\right|^{2+\tilde{\delta}}\right].\end{aligned} \tag{23}$$

Now, again by honesty (the sample used for estimation will not affect the splitting of decision trees), $\mathbb{E}[S_1^a|Z_1] = \mathbb{E}[S_1^a|X_1]$, and so our uniform $(2+\tilde{\delta})$-moment bounds on the distribution of Y_i conditional on X_i implies that (recall that M is the bounding constant in the Theorem's assumption)

$$\begin{aligned}&\mathbb{E}\left[|\mathbb{E}[S_1^a(Y_1 - \mathbb{E}[Y_1|X_1])|Z_1]|^{2+\tilde{\delta}}\right]\\ =&\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^{2+\tilde{\delta}}\left(|Y_1 - \mathbb{E}[Y_1|X_1]|\right)^{2+\tilde{\delta}}\right]\\ \le&M\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^{2+\tilde{\delta}}\right] \le M\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^2\right],\end{aligned} \tag{24}$$

because $S_1^a \le 1$. Meanwhile, because $\mathbb{E}[Y|X=\boldsymbol{x}]$ is Lipschitz, we can define $u \triangleq \sup\{|\mathbb{E}[Y|X=\boldsymbol{x}]| : \boldsymbol{x} \in [0,1]^d\}$, and see that

$$\begin{aligned}&\mathbb{E}[|\mathbb{E}\left[\sum_{i=1}^n S_i^a\mathbb{E}[Y_i|X_i]|Z_1\right] - \mathbb{E}[T^a]|^{2+\tilde{\delta}}]\\ &\le (2u)^{\tilde{\delta}}\mathrm{Var}\left[\mathbb{E}\left[\sum_{i=1}^n S_i^a\mathbb{E}[Y_i|X_i]|Z_1\right]\right]\\ &\le 2^{1+\tilde{\delta}}u^{2+\tilde{\delta}}\left(\mathbb{E}\left[\mathbb{E}[S_1^a|Z_1]^2\right] + \mathrm{Var}[(n-1)\mathbb{E}[S_2^a|Z_1]]\right)\\ &\le (2u)^{2+\tilde{\delta}}\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^2\right].\end{aligned} \tag{25}$$

Thus, the condition (22) that we need to check simplifies to

$$\lim_{n\to\infty} n\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^2\right] / (n\mathrm{Var}[\mathbb{E}[T^a|Z_1]])^{1+\tilde{\delta}/2} = 0. \tag{26}$$

[5] From now on, the proof are the same as the proof of Theorem 8 in the arXiv version of [15] except that we replace S_i by S_i^a and we replace T by T^a because we have multiple actions

[6] Here, we use the notation $\tilde{\delta}$ instead of the δ in usual Lyapunov condition

Finally, as argued in the proofs of Theorem 5 and Corollary 6 in the arXiv version of [15],

$$\mathrm{Var}[\mathbb{E}[T^a|Z_1]] = \Omega\left(\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^2\right]\mathrm{Var}[Y|X=\boldsymbol{x}]\right).$$

Because the denominator in (26) $\mathrm{Var}[Y|X=\boldsymbol{x}] > 0$ by assumption, we can use (19) in our previous argument on the incrementality. Note that the numerator in (26) satisfies

$$\left(n\mathbb{E}\left[\mathbb{E}[S_1^a|X_1]^2\right]\right)^{-\tilde{\delta}/2} \lesssim \left(\frac{C_{f,d}}{2k}\frac{\varepsilon_n n}{s\log(s)^d}\right)^{-\tilde{\delta}/2},$$

which goes to 0 when we plug in the values $s = O(n^\beta)$ and $\varepsilon_n = n^{-1/2(1-\beta)}$. Compared to the formula in the proof of the arXiv version of [15], we add a factor of ε_n because of the overlap condition for a multi-action tree. □

Step 3: In this step, Lemma 12 shows that for large sample size, the estimation by our estimator is close to the true value with a high probability.

Lemma 12. *For each $\omega' > 0$, there exists a $N_2 > 0$, such that for any $n > N_2$, we have for any $\delta > 0$*

$$\mathbb{P}[|\hat{\mu}_n(\boldsymbol{x},a) - \mathbb{E}[\hat{\mu}_n(\boldsymbol{x},a)]| \le \sigma_n(\boldsymbol{x},a)\delta] \ge 1 - e^{-\delta^2/2} - \omega_n'. \quad (27)$$

Here, $\omega_n' = e^{-\delta^2/2}(4\delta\tilde{\varepsilon} + 2\tilde{\varepsilon}^2) + \frac{C\psi\log n}{\sqrt{n}} + \left(\frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}\right)^{1-2\omega/3}$ which is a function of n, where $\tilde{\varepsilon} \triangleq \left(\frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}\right)^{\omega/3}$. Recall that ω is the small constant in the theorem's statement.

Proof. By Lemma 11, we know that $\frac{\hat{\mu}_n(\boldsymbol{x},a)-\mathbb{E}[\hat{\mu}_n(\boldsymbol{x},a)]}{\sigma_n(\boldsymbol{x},a)} \Rightarrow \mathcal{N}(0,1)$, where $\sigma(\boldsymbol{x},a) \le \frac{s}{n}\mathrm{Var}(T^a)$.

We will first show a property for a normal distributed random variable $X \sim \mathcal{N}(0,1)$, and then discuss the convergence rate towards the normal distribution. For every $\delta > 0$,

$$\mathbb{P}[|X| > \delta] = 2\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-(x+\delta)^2/2}dx,$$

and, for every $x > 0$,

$$e^{-(x+\delta)^2} \le e^{-t^2/2}e^{-x^2/2},$$

hence

$$\begin{aligned}\mathbb{P}[|X| > \delta] &\le 2e^{-\delta^2/2}\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\\ &= 2e^{-\delta^2/2}\mathbb{P}[X>0] = e^{-\delta^2/2}. \end{aligned}\quad (28)$$

Now, we will further show the convergence rate towards the normal distribution. First of all, we will show the convergence of $\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]$ ($\mathring{\hat{\mu}}_n$ is the Hájek projection). We now will show that $\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]$ has finite second absolute moment and finite third absolute moment. For the second absolute moment (variance), we have the following claim: if $\mathbb{E}[|X|^{2+\delta}] \le M$ is bounded for some $\delta > 0$, then $\mathbb{E}[|X|^2] \le M+1$ is also bounded. To prove this claim, we only need to discuss the cases when $|X| \le 1$ or $|X| > 1$. In fact, $\mathbb{E}[|X|^2] = \int_{|X|\le 1}|X|^2 f(X)dX + \int_{|X|>1}|X|^2 f(X)dX \le 1 + \int_{|X|>1}|X|^{2+\delta}f(X)dX \le 1+M$ where $f(\cdot)$ is the probability density function.

For the convergence rate, we have the following lemma:

Lemma 13 ([16]). *Letting $X_1, X_2, \ldots, X_n, \ldots$ be the sequence of independent random variables, $\mathbb{E}[X_i] = \mu_i$, $\mathbb{E}[(X_i-\mu_i)^2] = \sigma_i^2$, $\mathbb{E}[|X_i-\mu_i|^3] = \beta_i$. Let $F(x)$ be the CDF of $\sum_{i=1}^n(X_i-\mu_i)/(\sum_{i=1}^n\sigma_i^2)^{1/2}$, and $\Phi(x)$ be the CDF of the standard normal distribution. Then*

$$\sup_x|F(x)-\Phi(x)| < C\psi\log n/\sqrt{n},$$

where C is a constant, and ψ is a function of the σ_i's and β_i's [7].

In our case, we let X_i to be $\mathbb{E}[T^a|Z_i]$. Next, we consider the $(2+\delta)$ absolute moment $\mathbb{E}[|\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]|^{2+\tilde{\delta}}]$. From the Inequality (23) (24) and (25), we know

$$\begin{aligned}&2^{-(1+\tilde{\delta})}\mathbb{E}\left[\left|\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]\right|^{2+\tilde{\delta}}\right]\\ \le& M\mathbb{E}[\mathbb{E}[S_1^a|X_1]^2] + (2u)^{2+\tilde{\delta}}\mathbb{E}[\mathbb{E}[S_1^a|X_1]^2].\end{aligned}$$

In addition, $\mathbb{E}[S_1^a|X_1] \le 1$ because $S_1^a \le 1$. Then,

$$2^{-(1+\tilde{\delta})}\mathbb{E}\left[\left|\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]\right|^{2+\tilde{\delta}}\right] \le M + (2u)^{2+\tilde{\delta}}.$$

Now, we have $\mathbb{E}\left[|\mathbb{E}[T^a|Z_1] - \mathbb{E}[T^a]|^{2+\tilde{\delta}}\right] \le \left(M + (2u)^{2+\tilde{\delta}}\right)\times 2^{(1+\tilde{\delta})}$. When $\tilde{\delta} = 0$, we have the second absolute moment $\mathbb{E}[|\mathbb{E}[T^a|Z_i] - \mathbb{E}[T]|^2]$ is upper bounded by $4(M+4u^2)$. Similarly, when $\tilde{\delta} = 1$, we have the third absolute moment $\mathbb{E}[|\mathbb{E}[T^a|Z_i] - \mathbb{E}[T]|^2]$ is upper bounded by $8(M+8u^3)$.

We notice that

$$\frac{s^2}{n^2}\sum_{i=1}^n\mathrm{Var}[\mathbb{E}[T^a|Z_i]] = \mathrm{Var}[\mathring{\hat{\mu}}(\boldsymbol{x},a)] = \sigma_n^2.$$

Now, based on the definition of $\mathring{\hat{\mu}}_n(\boldsymbol{x},a)$ in (21), we have

$$\begin{aligned}&\frac{\sum_{i=1}^n(X_i-\mu_i)}{(\sum_{i=1}^n\mathrm{Var}[\mathbb{E}[T^a|Z_i])^{1/2}} = \frac{\sum_{i=1}^n(\mathbb{E}[T^a|Z_i] - \mathbb{E}[T^a]}{(\sum_{i=1}^n\mathrm{Var}[\mathbb{E}[T^a|Z_i] - \mathbb{E}[T^a]])^{1/2}}\\ =&\frac{\frac{n}{s}\mathring{\hat{\mu}}_n(\boldsymbol{x},a)}{(\frac{n^2}{s^2}\mathrm{Var}[\mathring{\hat{\mu}}_n(\boldsymbol{x},a)])^{1/2}} = \frac{\mathring{\hat{\mu}}_n(\boldsymbol{x},a)}{\sigma_n(\boldsymbol{x},a)}.\end{aligned}$$

Thus, $F(x)$ is the CDF of the random variable $\frac{\mathring{\hat{\mu}}_n(\boldsymbol{x},a)}{\sigma_n(\boldsymbol{x},a)}$. According to Lemma 13, we have

$$\sup_x|F(x)-\Phi(x)| < C\psi\log n/\sqrt{n}.$$

Combined the property of normal CDF (28), we have $\mathbb{P}[|\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]| \le \sigma_n\delta] \ge 1 - e^{-\delta^2/2} - \frac{C\psi\log n}{\sqrt{n}}$. Here, $\mathring{\hat{\mu}}_n, \hat{\mu}_n, \sigma_n$ are short for $\mathring{\hat{\mu}}_n(\boldsymbol{x},a)$, $\hat{\mu}_n(\boldsymbol{x},a)$, $\sigma_n(\boldsymbol{x},a)$ respectively.

We now bound the large deviation probability for $\hat{\mu}_n$

$$\begin{aligned}&\mathbb{P}[|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]| \le \sigma_n\delta]\\ \ge&\mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n| + |\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]| + |\mathbb{E}[\hat{\mu}_n] - \mathbb{E}[\mathring{\hat{\mu}}_n]]| \le \sigma_n\delta]\\ =&\mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n| + |\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]| \le \sigma_n\delta]\\ \ge&\mathbb{P}[|\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]| \le \sigma_n\delta - \tilde{\varepsilon}\sigma_n] - \mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n| > \tilde{\varepsilon}\sigma_n] \quad (29)\\ &\text{(the last inequality is because}\\ &\mathbb{P}[|A|+|B| \le \delta] \ge \mathbb{P}[|A| \le \delta - \tilde{\varepsilon}] - \mathbb{P}[|B| > \tilde{\varepsilon}])\end{aligned}$$

[7] When $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, the $\log(n)$ term can be removed according to the Berry–Esseen theorem.

Before further development, we first show that the approximation argument in (20) can be turned into the bound in (30) when $s \geq 4kde^{2d}$ where we recall k is the constant for a regular tree. The source for the approximation is from the proof of Lemma 4 in the arXiv version of [15]. In particular, we modify the approximation in Equation (36) in the arXiv version of [15]. From Corollary 3.2 of [7], we know for the upper incomplete gamma function $\Gamma(d,c)$ we have $\Gamma(d,c) \leq c^{d-1}e^{-c} \times \left(1 + \frac{1}{\frac{c}{d-1}-1}\right)$ where d, c are real values. In the proof of Lemma 4 in the arXiv version of [15], $c = -\log(1 - \exp[-2k\frac{\log(s)}{s-2k+1}])$. One can verify that when $s \geq 4kde^{2d}$, $\frac{c}{d-1} > 2$, and thus $\Gamma(d,c) \leq 2c^{d-1}e^{-c}$.

Moreover, we have $1 - \exp\left[-2k\frac{\log(s)}{s-2k+1}\right] \leq 4k\frac{\log(s)}{s}$ when $s \geq 4k$.

Therefore, when $s \geq \max\{4k, 4kde^{2d}\} = 4kde^{2d}$, the approximation inequality (36) of the arXiv version of [15] is changed to $\mathbb{P}_{x=0}\left[\mathbb{E}[P_1|Z_1] \geq \frac{1}{s^2}\right] \leq \frac{8k}{(d-1)!}\frac{\log(s)^d}{s}$. Note that the upper bound becomes 4 times larger when we change "$\lesssim$" to "$\leq$". Thus, we can finally change the argument in Lemma 4 of arXiv version of [15] as $s\text{Var}[\mathbb{E}[S_1|Z_1]] \geq \frac{4}{k}C_{f,d}/\log(s)^d$. In Theorem 5, we will change the bound to $\frac{\text{Var}[\mathring{T}(x;Z)]}{\text{Var}[T(x;Z)]} \geq \frac{\nu(s)}{4}$. Next, we can change our (20) to

$$\frac{1}{\sigma_n^2}\mathbb{E}\left[\left(\hat{\mu}_n(\boldsymbol{x},a) - \mathring{\hat{\mu}}_n(\boldsymbol{x},a)\right)^2\right] \leq \frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}. \tag{30}$$

For the second term of (29), we have that

$$\begin{aligned}&\mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n| > \sigma_n\tilde{\varepsilon}] = \mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n|^2 > \sigma_n^2\tilde{\varepsilon}^2]\\ \leq&\frac{\mathbb{E}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n|^2]}{\sigma^2\tilde{\varepsilon}^2} \leq 16\frac{s\sigma_n^2}{n}\frac{\log(s)^d}{\varepsilon_n C_{f,d}}/(\sigma_n^2\tilde{\varepsilon}^2) = 16\frac{s}{n}\frac{\log(s)^d}{\varepsilon_n C_{f,d}}/(\tilde{\varepsilon}^2)\end{aligned} \tag{31}$$

Here, the last but one inequality is according to (30) when $s \geq 4kde^{2d}$.

Recall that we let $\tilde{\varepsilon}$ be $\left(\frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}\right)^{\omega/3}$ which $\to 0$ as $n \to \infty$. Recall that $\omega > 0$ is a small constant in our theorem's statement. There exists a N_3, such that when $n > N_3$, we have $\tilde{\varepsilon} < 1$ and $4\tilde{\varepsilon} + 2\tilde{\varepsilon}^2 < 1$.

Then, $\mathbb{P}[|\mathring{\hat{\mu}}_n - \hat{\mu}_n| > \sigma_n\tilde{\varepsilon}] \leq \left(\frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}\right)^{1-2\omega/3}$. Now, we let $N_2 = (4kde^{2d})^{1/\beta}$, so that when $n > N_2$ we have $s > 4kde^{2d}$. So far, we have bound for the second term of the RHS of (29).

For the first term of the RHS of (29), we have

$$\begin{aligned}\mathbb{P}[|\mathring{\hat{\mu}}_n - \mathbb{E}[\mathring{\hat{\mu}}_n]| \leq \sigma_n(\delta - \tilde{\varepsilon})] &\geq 1 - e^{-(\delta-\tilde{\varepsilon})^2/2} - \frac{C\psi\log n}{\sqrt{n}}\\ &\geq 1 - e^{-\delta^2/2}(1 + 4\delta\tilde{\varepsilon} + 2\tilde{\varepsilon}^2) - \frac{C\psi\log n}{\sqrt{n}},\end{aligned} \tag{32}$$

where the last inequality is because $e^x \leq 1 + 2x$ for $x \in [0,1]$.

Combining inequations (32) and (31), we have

$$\begin{aligned}&\mathbb{P}[|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]| \leq \sigma_n\delta]\\ \leq&1 - e^{-\delta^2/2}(1 + 4\delta\tilde{\varepsilon} + 2\tilde{\varepsilon}^2) - \frac{C\psi\log n}{\sqrt{n}} - \left(\frac{s}{n}\frac{16\log(s)^d}{\varepsilon_n C_{f,d}}\right)^{1-2\omega/3}.\end{aligned}$$

Note that $\sigma_n(\boldsymbol{x},a) > 0$, then we get (27) in the statement of Lemma 12. □

Step 4: With the results in Lemma 11 and Lemma 12, we now can prove Theorem 7 that gives an upper bound of the online regret of our ϵ-decreasing multi-action forest algorithm.

Now, let's go back to the proof of Theorem 7. First of all, we decompose the error into two parts

$$|\hat{\mu}(\boldsymbol{x},a) - \mu(\boldsymbol{x},a)| \leq |\hat{\mu}(\boldsymbol{x},a) - \mathbb{E}[\hat{\mu}(\boldsymbol{x},a)]| + |\mathbb{E}[\hat{\mu}(\boldsymbol{x},a)] - \mu(\boldsymbol{x},a)|. \tag{33}$$

From (15) and the definition of "$\lesssim$", we know that there exists an integer $N_1 > 0$ and a constant $C_1 > 0$, such that for any $n \geq N_1$ (and $s = n^\beta$ is a function of n), we have

$$|\mathbb{E}[\hat{\mu}_n(\boldsymbol{x},a)] - \mu_n(\boldsymbol{x},a)| \leq C_1 2Md\left(\frac{\varepsilon_n s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}}. \tag{34}$$

Now we combine (34) and (27). When $n > \max\{N_1, N_2\}$, with probability at least $1 - \frac{1}{2}e^{-\delta^2/2} - \omega_n'$, we have the following error bound

$$|\hat{\mu}_n(\boldsymbol{x},a) - \mu_n(\boldsymbol{x},a)| \leq \sigma_n(\boldsymbol{x},a)\delta + 2C_1 Md\left(\frac{\varepsilon_n s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}}. \tag{35}$$

Now, we turn the error bound (35) into the regret bound. We note that at the beginning of time slot $t+1$, our online learning oracle collects t data points of feedbacks, where we can shuffle the data to be i.i.d. samples satisfying Lemma 11. Then, when $t > N \triangleq \max\{N_1, N_2, N_3\}$, with a probability at least $1 - e^{-\delta^2/2} - \omega_t'$, the regret in round $t+1$ for the online oracle (defined as r_{t+1})

$$\begin{aligned}&r_{t+1} = \mu_t(\boldsymbol{x},a^*) - \mu_t(\boldsymbol{x},a)\\ =&\left[\mu_t(\boldsymbol{x},a^*) - \hat{\mu}_t(\boldsymbol{x},a^*)\right] - \left[\mu_t(\boldsymbol{x},a) - \hat{\mu}_t(\boldsymbol{x},a)\right] + \left[\hat{\mu}_t(\boldsymbol{x},a^*) - \hat{\mu}_t(\boldsymbol{x},a)\right]\\ \leq&\left[\mu_t(\boldsymbol{x},a^*) - \hat{\mu}_t(\boldsymbol{x},a^*)\right] - \left[\mu_t(\boldsymbol{x},a) - \hat{\mu}_t(\boldsymbol{x},a)\right]\\ \leq&|\mu_t(\boldsymbol{x},a^*) - \hat{\mu}_t(\boldsymbol{x},a^*)| + |\mu_t(\boldsymbol{x},a) - \hat{\mu}_t(\boldsymbol{x},a)|\\ \leq&2\sigma_t(\boldsymbol{x},a)\delta + 4C_1 Md\left(\frac{\varepsilon_t s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}} \text{.(recall that } s = t^\beta)\end{aligned}$$

We let $\delta_0 = e^{-\delta^2/2}$, then $\delta = \sqrt{2\log(1/\delta_0)}$. Recall that $\text{Var}[T^a(x)]$ is bounded by V [8] then with probability at least $1 - \delta_0 - \omega_t'$, for $t > N$ we have

$$\begin{aligned}r_{t+1} \leq&2\sqrt{t^{\beta-1}V}\sqrt{2\log(\frac{1}{\delta_0})}\\ &+ 4C_1 Md\left(\frac{\varepsilon_t s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{d}} + \varepsilon_t\Delta_{\max},\end{aligned} \tag{36}$$

[8] It is stated in Lemma 3.3 in [15]. Here, we use the proof of page 38 in the arXiv version of [15] to justify a bound on $\text{Var}[T]$. In our regularity tree, each split has at least k leafs. Thus,

$$k\text{Var}[T(x;Z)] \leq |\{i : X_i \in L(x;Z)\}| \cdot \text{Var}[T(x;Z)] \to_p \text{Var}[Y|X = x].$$

In addition, because of the regularity condition on the moment, $\text{Var}[Y|X=x] = \mathbb{E}[|Y - \mathbb{E}[Y|X=x]|^2|X=x] \leq (M+1)$. Therefore, the variance $\text{Var}[T(x;Z)]$ is bounded.

where $\Delta_{\max}$ denotes the maximum regret for choosing a sub-optimal action as defined in [1][9]. Recall that we denote $A=\frac{\log((1-\alpha)^{-1})\pi}{\log(\alpha^{-1})d}$. Now we denote $\epsilon_0 = -\frac{A}{2+3A}$, and $\varepsilon_t = t^{\epsilon_0}$. One can check that $\beta = 1 - \frac{2A}{2+3A} = \frac{1-A\epsilon_0}{1+A}$.

Here, we notice $(\varepsilon_t s)^{-\frac{1}{2}A} = t^{-\frac{1}{2}A(\beta+\epsilon_0)}$. One can check that by the above parameters setting, each terms in (35) have the same exponent w.r.t. t, i.e.

$$\frac{1}{2}(\beta-1) = -\frac{1}{2}A(\beta+\epsilon_0) = \epsilon_0 = -\frac{A}{2+3A} \tag{37}$$

Then (36) can be rewritten as (with probability at least $1-\delta_0-\omega_t'$)

$$r_{t+1} \le \left(2\sqrt{V}\sqrt{2\log(\frac{1}{\delta_0})}+4C_1Md(2k-1)^{\frac{1}{2}A}+\Delta_{\max}\right)t^{\beta-1}. \tag{38}$$

Consider the probability $\delta_0+\omega_t'$, from (38) we have

$$r_{t+1} \le \left(2\sqrt{V}\sqrt{2\log(\frac{1}{\delta_0})}+4C_1Md(2k-1)^{\frac{1}{2}A}+\Delta_{\max}\right)t^{\beta-1} + (\delta_0+\omega_t')\Delta_{\max}.$$

Let $C_3 \triangleq \left(2\sqrt{V}\sqrt{2\log(\frac{1}{\delta_0})} + 4C_1Md(2k-1)^{\frac{1}{2}A} + \Delta_{\max}\right)$ be a constant. Then, we further denote $p \triangleq \frac{2+3A}{A} > 1$ (where $p = \frac{2}{1-\beta}$) and by Hölder's inequality, when $T > N$ we have

$$\begin{aligned}
R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset}) &= \sum_{t=1}^{N} r_t + \sum_{t=N+1}^{T} r_t \\
&\le \sum_{t=1}^{N} r_t + \left((T-N)\delta_0 + \sum_{t=N+1}^{T}\omega_t'\right)\Delta_{\max} + T^{1-1/p}C_3\left(\sum_{t=1}^{T}(\frac{r_t}{C_3})^p\right)^{1/p} \\
&= \sum_{t=1}^{N} r_t + \left((T-N)\delta_0 + \sum_{t=N+1}^{T}\omega_t'\right)\Delta_{\max} + C_3T^{1-\frac{1}{p}}(\sum_{t=1}^{T}\frac{1}{t})^{\frac{1}{p}} \\
&\le \sum_{t=1}^{N} r_t + \left(\left((T-N)\delta_0 + \sum_{t=N+1}^{T}\omega_t'\right) + N\right)\Delta_{\max} + C_3T^{1-\frac{1}{p}}(\log T)^{\frac{1}{p}},
\end{aligned}$$

where the last inequality holds because $\sum_{t=1}^{T}\frac{1}{t} \le \log T$.

Now, we let $\delta_0 = T^{-\frac{A}{2+3A}}$.

Here,

$$\sum_{t=N+1}^{T}\omega_t' = \sum_{t=N+1}^{T}\left(\delta_0(4\sqrt{2\log(\frac{1}{\delta_0})}\tilde{\varepsilon}+2\tilde{\varepsilon}^2)+\frac{C\psi\log(t)}{\sqrt{t}}+\left(\frac{s}{t}\frac{16\log(s)^d}{\varepsilon_t C_{f,d}}\right)^{1-\frac{2\omega}{3}}\right).$$

[9]For $\Delta_{\max}$ to exist, we have a mild assumption that the average rewards are bounded for each actions.

Recall that when $n > N_3$, $\tilde{\varepsilon} < 1$, and in our parameter setting $\frac{s}{t\varepsilon_t} = t^{-1/2(1-\beta)} = t^{-1/p}$. Hence,

$$\begin{aligned}
\sum_{t=N+1}^{T}\omega_t' \le& \sum_{t=N+1}^{T}\left(\delta_0(4\sqrt{2\log(1/\delta_0)}+2)\right. \\
&\left.+\frac{C\psi\log(t)}{\sqrt{t}}+\left(t^{-\frac{1}{p}}\frac{16\log(s)^d}{C_{f,d}}\right)^{1-2\omega/3}\right) \\
\le&(T-N)(T^{-\frac{1}{p}}(4\sqrt{2\frac{1}{p}\log(T)}+2)) + (T-N)\frac{C\psi\log(T)}{\sqrt{T}} \\
&+(T-N)\left(T^{-\frac{1}{p}}\frac{16\log(T)^d}{C_{f,d}}\right)^{1-2\omega/3} \\
\le&T^{1-\frac{1}{p}}(4\sqrt{2\frac{1}{p}\log(T)}+2)+\sqrt{T}C\psi\log(T) \\
&+T^{1-\frac{1}{p}+\frac{1}{2}\frac{2\omega}{3}}\left(\frac{16\log(T)^d}{C_{f,d}}\right)^{1-2\omega/3}.
\end{aligned}$$

We notice that $1-\frac{1}{p} > \frac{1}{2}$, so the exponent $T^{1-\frac{1}{p}+\frac{1}{3}\omega}$ dominates, and we use another $T^{\frac{1}{6}\omega}$ to hide the $\log(T)$ terms. Then we have $R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset}) = O(T^{1-\frac{1}{p}+\frac{1}{2}\omega})$. Note that $1/p = \frac{1}{2}(1-\beta)$, then

$$\lim_{T\to+\infty}\frac{R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset})}{T^{1-\frac{1}{2}(1-\beta)+\frac{\omega}{2}}} = \lim_{T\to+\infty}\frac{R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset})}{T^{\frac{1+\beta+\omega}{2}}} = 0 \quad \text{for any small } \omega > 0.$$

Thus, using the big-O notation, $\lim_{T\to+\infty}\frac{R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset})}{T} = O(T^{-\frac{A}{2+3A}+\frac{\omega}{2}})$ for any small ω.

Finally, one can verify $1-\frac{A}{2+3A} = \frac{1+\beta}{2}$ which is less than 1. Then, we reach our claim in the theorem that $\lim_{T\to+\infty}\frac{R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset})}{T^{(1+\beta+\omega)/2}} = 0$, and $\lim_{T\to+\infty}\frac{R(T,\mathcal{A}_{\text{Fst+}\mathcal{E}_\emptyset})}{T} = 0$ for any ω that is smaller than $\frac{1-\beta}{2}$. Namely, we have shown that the asymptotic regret is sub-linear w.r.t. T. □

3.4 Regret Bound for Contextual Independent Algorithm $\mathcal{A}_{\text{UCB+IPSW}}$ (Theorem 4)

Proof of Theorem 4. The proof follows the same idea as previous ones. We will first show that the estimation relying on the offline data is unbiased. Second, we use a weighted Chernoff bound to show the effective number of logged samples (a.k.a. Effective Sample Size) in terms of the confidence bound.

Many previous works have shown the inverse propensity weighting method provides an unbiased estimator[14]. In fact, for $\tilde{a} \in [K]$

$$\begin{aligned}
\mathbb{E}[\bar{y}_{\tilde{a}}] &= \frac{\mathbb{E}[\sum_{i\in[-I]}\mathbb{E}[y|\boldsymbol{x}_i,\tilde{a})]\mathbb{E}[\mathbb{1}_{\{a_i=\tilde{a}\}}]/p(\boldsymbol{x}_i,\tilde{a})]}{\sum_{i\in[-I]}\mathbb{E}[\mathbb{1}_{\{a_i=\tilde{a}\}}]/p(\boldsymbol{x}_i,\tilde{a})]} \\
&= \frac{\mathbb{E}[\sum_{i\in[-I]}\mathbb{E}[y|\boldsymbol{x}_i,\tilde{a}]]}{I} \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}}\mathbb{P}[\boldsymbol{x}]\mathbb{E}[y|\boldsymbol{x},\tilde{a}] = \mathbb{E}[\bar{y}_{\tilde{a}}].
\end{aligned}$$

The second equation holds because the probability that we observe the action $\tilde{a}$ is $\mathbb{E}[\mathbb{1}_{\{a_i=\tilde{a}\}}]$ which is the propensity score $p(\boldsymbol{x}_i,\tilde{a})$.

The last equation is because the expectation for data item i is taken over the contexts $\boldsymbol{x}$.

According to Chernoff-Hoeffding bound [11], we have the following Lemma.

Lemma 14. *If $X_1, X_2, \ldots, X_n$ are independent random variables and $A_i \le X_i \le B_i (i = 1, 2, \ldots, n)$, we have the following bounds for the sum $X = \sum_{i=1}^n X_i$:*

$$\mathbb{P}[X \le \mathbb{E}[X] - \delta] \le e^{-\frac{2\delta^2}{\sum_{i=1}^n (B_i - A_i)^2}}.$$

$$\mathbb{P}[X \ge \mathbb{E}[X] + \delta] \le e^{-\frac{2\delta^2}{\sum_{i=1}^n (B_i - A_i)^2}}.$$

In our case to estimate the outcome for an action a, we have $X_i = y_i \frac{\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}{\sum_{i\in[-I]} \mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}$, and $X = \sum_{i\in[-I]} X_i = \bar{y}_a$. Hence the constants $A_i = 0$, $B_i = \frac{\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}{\sum_{i\in[-I]} \mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}[|\bar{y}_a - \mathbb{E}[y|a]| \ge \delta] \\
\le& 2e^{-\frac{2\delta^2}{\sum_{i\in[-I]}\left(\frac{\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}{\sum_{i\in[-I]}\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}\right)^2}} \\
=& 2e^{-\frac{2\delta^2}{\frac{\sum_{i\in[-I]}\left(\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)\right)^2}{\left(\sum_{i\in[-I]}\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)\right)^2}}} \\
=& 2e^{-2\delta^2 \frac{\left(\sum_{i\in[-I]}\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)\right)^2}{\sum_{i\in[-I]}\left(\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)\right)^2}}
\end{aligned}$$

We compare it with the Chernoff-Hoeffding bound used in the UCB algorithm[4]. When we have n_a online samples of arm a,

$$\mathbb{P}[|\bar{y}_a - \mathbb{E}[y|a]| \ge \delta] \le 2e^{-2n_a\delta^2}.$$

By this comparison, we let $n = \widehat{N}_a$ and we will get the same bound.

Now, we show that by using these $\lfloor \widehat{N}_a \rfloor$ samples from logged data, the online bandit UCB oracle will always have a tighter bound than that for $\lfloor \widehat{N}_a \rfloor$ i.i.d. samples from the online environment.

In the online phase, let the number of times to play the action a to be T_a. For the offline samples, let $X_i = y_i \frac{\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)}{\sum_{i\in[-I]}\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)} \frac{\widehat{N}_a}{\widehat{N}_a+T_a}$. For the online samples, let $X^t = y_t \frac{1}{\widehat{N}_a+T_a}$. Let us consider the sequence $\{X_1, \ldots, X_I, X^1, \ldots, X^{T_a}\}$. Now, $X = \sum_{i\in[-I]} X_i + \sum_{t\in[T_a]} X^t$. Then, we have $\mathbb{E}[X] = \mathbb{E}[y|a]$, and $0 \le X_i \le \frac{\widehat{N}_a}{\widehat{N}_a+T_a} B_i (\forall i \in [-I])$, $0 \le X^t \le \frac{1}{\widehat{N}_a+T_a}$. In addition, we have

$$\begin{aligned}
&\left(\frac{\widehat{N}_a}{\widehat{N}_a+T_a}\right)^2 \frac{\sum_{i\in[-I]}\left(\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)\right)^2}{\sum_{i\in[-I]}\mathbb{1}_{\{a_i=a\}}/p(\boldsymbol{x}_i,a_i)} + \sum_{t\in[T_a]}\left(\frac{1}{\widehat{N}_a+T_a}\right)^2 \\
=& \left(\frac{\widehat{N}_a}{\widehat{N}_a+T_a}\right)^2 \left(\frac{1}{\widehat{N}_a}\right) + \frac{T_a}{(\widehat{N}_a+T_a)^2} = \frac{1}{\widehat{N}_a+T_a}.
\end{aligned}$$

Therefore,

$$\mathbb{P}[\bar{y}_a \le \mathbb{E}[y|a] - \delta] \le e^{-2\delta^2(\widehat{N}_a+T_a)},$$

$$\mathbb{P}[\bar{y}_a \ge \mathbb{E}[y|a] + \delta] \le e^{-2\delta^2(\widehat{N}_a+T_a)}.$$

In other words, when we have T_a online samples of an action a, the confidence interval is as if we have $T_a + \widehat{N}_a$ total samples for the bandit oracle. Then, the regret bound reduces to the case where we have $\widehat{N}_a$ offline samples for arm a that do not have contexts. □

3.5 Regret Bound for Contextual Algorithm $\mathcal{A}_{\text{LinUCB+LR}}$ (problem dependent Theorem 9 and problem independent Theorem 6)

Proof of Theorem 9. The proof follows the analytical framework of the paper[1]. Especially, this Theorem corresponds to the Theorem 3 in the paper[1]. The proofs in papers[3][9] have similar ideas.

In particular, we consider that the offline samples have features $\boldsymbol{x}_{-1}, \boldsymbol{x}_{-2}, \ldots, \boldsymbol{x}_{-N}$, and the online samples have features $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^T$. To have a unified index system, we let $\boldsymbol{x}_{N+t} \triangleq \boldsymbol{x}^t$ for $t \ge 1$.

Because we choose the "optimal" action in the online phase, we have the pseudo-regret in time slot t is

$$r_t \le 2\sqrt{\beta_{t-1}(\delta)} \min\{||\boldsymbol{x}_{N+t}||_{V_{N+t-1}^{-1}}, 1\}.$$

Then, we have (recall that in this paper, we set V_0 as a $d \times d$ identity matrix $\boldsymbol{I}_d$)

$$\begin{aligned}
&\sqrt{8\beta_n(\delta)} \sum_{n=1}^N \min\{1, ||\boldsymbol{x}_n||_{V_{n-1}^{-1}}\} + \sum_{t=1}^T r_t \\
\le& \sqrt{8(N+T)\beta_n(\delta) \log \frac{\texttt{trace}(V_0) + (N+T)L^2}{\texttt{det}V_0}}.
\end{aligned}$$

Here, we observe that

$$\begin{aligned}
\sum_{t=1}^T r_t \le& \sqrt{8(N+T)\beta_n(\delta) \log \frac{\texttt{trace}(V_0) + (N+T)L^2}{\texttt{det}V_0}} \\
&- \sqrt{8\beta_n(\delta)} \sum_{n=1}^N \min\{1, ||\boldsymbol{x}_n||_{V_{n-1}^{-1}}\}.
\end{aligned}$$

Now, we give a lower bound of the last term

$$\sqrt{8\beta_n(\delta)} \sum_{n=1}^N \min\{1, ||\boldsymbol{x}_n||_{V_{n-1}^{-1}}\}.$$

Here, $||\boldsymbol{x}||_A = \sqrt{\boldsymbol{x}^T A \boldsymbol{x}} \ge \sqrt{\lambda_{\min}(A)}||\boldsymbol{x}||_2$. We have the following claim that $\lambda_{\min}(V_n^{-1}) \ge \frac{1}{1+(n-1)L^2}$. This is because $\lambda_{\min}(V_n^{-1}) = 1/\lambda_{\max}(V_n)$. In fact, for the symmetric matrices, we have

$$\lambda_{\max}(A + B) \le \lambda_{\max}(A) + \lambda_{\max}(B).$$

We have $\lambda_{\max}(I) = 1$, and $\lambda_{\max}(\boldsymbol{x}\boldsymbol{x}^T) = ||\boldsymbol{x}||_2^2$. Therefore,

$$\lambda_{\max}(V_{n-1}) \le 1 + ||\boldsymbol{x}_1||_2^2 + \ldots + ||\boldsymbol{x}_{n-1}||_2^2 \le 1 + (n-1)||\boldsymbol{x}||_{\max}^2,$$

where we consider $||\boldsymbol{x}_i||_2^2 \le ||\boldsymbol{x}||_{\max}^2$ for $i \in [n]$. Also, we consider $||\boldsymbol{x}_i||_2^2 \ge ||\boldsymbol{x}||_{\min}^2$ for $i \in [n]$.

Let $L = ||\boldsymbol{x}||_{\max}$. Then,

$$\begin{aligned}
&\sum_{n=1}^{N} \min\{1, ||\boldsymbol{x}_n||_{V_{n-1}^{-1}}\} \\
\geq &\sum_{n=1}^{N} \min\{1, ||\boldsymbol{x}||_{\min}\sqrt{\frac{1}{1+(n-1)L^2}}\} \\
\geq &\min\{1, ||\boldsymbol{x}||_{\min}\} \sum_{n=1}^{N} \sqrt{\frac{1}{1+(n-1)L^2}} \\
\geq &\min\{1, ||\boldsymbol{x}||_{\min}\} \sum_{n=1}^{N} \frac{2}{L^2}\left(\sqrt{1+nL^2} - \sqrt{1+(n-1)L^2}\right) \\
= &\min\{1, ||\boldsymbol{x}||_{\min}\} \frac{2}{L^2}\left(\sqrt{1+NL^2} - 1\right).
\end{aligned}$$

Hence, we have the final bound of regret

$$\begin{aligned}
\sum_{t=1}^{T} r_t \leq &\sqrt{8(N+T)\beta_n(\delta)\log\frac{\texttt{trace}(V_0)+(N+T)L^2}{\texttt{det}V_0}} \\
&- \sqrt{8\beta_n(\delta)}\min\{1, ||\boldsymbol{x}||_{\min}\}\frac{2}{L^2}\left(\sqrt{1+NL^2}-1\right).
\end{aligned}$$

□

Compared with the previous regret bound without offline data, the regret bound changes from $O(\sqrt{T})$ to $O(\sqrt{N+T}) - \Omega(\sqrt{N})$. From the view of regret-bound, using offline data does not bring us a large amount of regret-reduction.

We now show a better bound for the problem-dependent case. This corresponds to section 5.2 of the paper[1]. Let Δ_t be the "gap" at step t as defined in the paper of Dani et al.[10]. Intuitively, Δ_t is the difference between the rewards of the best and the "second best" action in the decision set D_t. We consider the samllest gap $\bar{\Delta}_n = \min_{1\leq t\leq n}\Delta_t$.

Proof of Theorem 6. We will first show a high-probability bound, i.e. with probability at least $1-\delta$, the cummulative regret has the bound

$$R(T, \mathcal{A}_{\text{LinUCB+LR}}) \leq \frac{4\beta_{N+T}(\delta)}{\Delta_{\min}} d\log(1+\kappa)$$

when the parameters $\{\beta_t\}_{t=1}^{T}$ ensure the confidence bound in each time slot.

Recall that the contexts of samples returned by the offline evaluator are $\boldsymbol{x}_{-1}, \boldsymbol{x}_{-2}, \ldots, \boldsymbol{x}_{-N}$. We denote $r_t \triangleq \max_{a\in[K]}\mathbb{E}[y_t|\boldsymbol{x}_t, a] - \mathbb{E}[y_t|\boldsymbol{x}_t, a_t]$ as the pseudo-regret in time slot t. Recall that $\beta_t(\delta)$ is the parameter β_t in the t^{th} time slot, and the δ is to emphasize that it is a function of δ. From the proof for the problem-independent bound in paper[1], we know $\sum_{t=1}^{T} r_t \leq \frac{4\beta_{N+T}(\delta)}{\Delta_{\min}}\log\frac{\texttt{det}V_T}{\texttt{det}V_N}$. The following is to bound $\log\frac{\texttt{det}V_{N+T}}{\texttt{det}V_N}$. We have the following lemma.

Lemma 15. *Let $\kappa = \frac{TL^2}{\lambda_{\min}(V_N)}$, then $(1+\kappa)V_N \succcurlyeq V_{T+N}$.*

Proof of Lemma 15. We first consider the case where all the data samples are returned before the first online phase start. Denote the V matrix in the online time slot t after using the logged data as V_{N+t}. Note that $V_{T+N} = V_N + \sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'$. Thus the above lemma is equivalent to $\sum_{t=1}^{T}\boldsymbol{x}_{N+t}\boldsymbol{x}_{N+t}' \preccurlyeq \kappa V_N$. Here, we use $\boldsymbol{x}'$ to denote the transpose of $\boldsymbol{x}$ (to avoid using "$\boldsymbol{x}^T$" with the confusing T). The positive semi-definiteness means that for any $\boldsymbol{x}$ where $||\boldsymbol{x}||_2$=1, we want to have

$$\boldsymbol{x}'\left(\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)\boldsymbol{x} \leq \kappa\boldsymbol{x}'V_N\boldsymbol{x}. \tag{39}$$

In fact $\boldsymbol{x}'\left(\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)\boldsymbol{x} \leq TL^2$, because L is the maximum 2-norm of $\boldsymbol{x}_t$. In addition, $\boldsymbol{x}'V_N\boldsymbol{x} \geq \lambda_{\min}(V_N)$. Hence, we always have (39) for $\forall\boldsymbol{x}$. Hence we proved the above lemma. □

We have $\texttt{det}A \leq \texttt{det}B$ if $A \preccurlyeq B$. Hence,

$$\texttt{det}V_{T+N} \leq \texttt{det}(1+\kappa)V_N = (1+\kappa)^d\texttt{det}V_N.$$

Then, $\log\frac{\texttt{det}V_{N+T}}{\texttt{det}V_N} \leq d\log(1+\kappa)$, which leads to our Theorem.

Now, we set $\beta_t(\delta) = 2d(1+2\ln(1/\delta))$, and the parameter is in the confidence ball with probability at least $1-\delta$. Moreover, we set $\delta = 1/T$. Then, the regret in each time slot can be divided into two parts: (1) the δ probability part (summing up to at most 1, because the outcome is bounded); and (2) the $1-\delta$ probability part (summing up to at most $\frac{8d(1+2\ln(T))}{\Delta_{\min}}d\log(1+\kappa)$). Therefore, the expected cumulative reward has an upper bound $\frac{8d(1+2\ln(T))}{\Delta_{\min}}d\log(1+\kappa)+1$.

Now, plugging in the definition of κ, we have proved

$$R(T, \mathcal{A}_{\text{LinUCB+LR}}) \leq \frac{8d^2(1+2\ln(T))}{\Delta_{\min}}\log\left(1+\frac{TL^2}{\lambda_{\min}(V_N)}\right)+1.$$

□

3.6 Relaxations of The Assumptions on The Logged Data (Theorem 5)

Proof of Theorem 5. Let us consider the number of times that a sub-optimal action is played, using the UCB online bandit oracle. Let us denote the expected reward (or outcome) $\mathbb{E}[y|a]$ for an action a as μ_a. In the t_{th} online round, we make the wrong decision to play an action a only if $(\mu_{a^*} - \mu_a) + \left(\frac{\delta_{a^*}N_a}{N_a+t} - \frac{\delta_a N_a}{N_a+t}\right) < I_a - I_{a^*}$, where I_a is half of the width of the confidence interval $\beta\sqrt{\frac{2\ln(n)}{n_a}}$ for action a, where n_a is the number of times that the online bandit oracle plays action a and $n = \sum_{a\in[K]} n_a$. Now, we only need to consider the case where $\delta_a - \delta_{a^*} \geq 0$. Otherwise, the offline data lets us to have less probability to select the sub-optimal actions, and thus leads to a lower regret.

According to Chernoff bound, when we have

$$(N_a+t)[\Delta_a + \frac{N_a}{N_a+t}(\delta_{a^*} - \delta_a)]^2 \geq 8\ln(N_a+T), \tag{40}$$

the violation probability will be very low. In fact, under (40)

$$\mathbb{P}\left[(\mu_{a^*} - \mu_a) + \left(\frac{\delta_{a^*}N_a}{N_a+t} - \frac{\delta_a N_a}{N_a+t}\right) < I_a - I_{a^*}\right] \leq t^{-4}.$$

Then we can let l_a to be a number such that when $t > l_a$, the inequality (40) is satisfied.

In fact, when $l_a = \lceil 16\frac{\ln(N_a+T)}{\Delta_a^2} + [N_a(\frac{2(\delta_a-\delta_{a^*})}{\Delta_a} - 1)] - N_a\rceil$, (40) is satisfied. Therefore, the expected number of times that we play

an action a is less than

$$l_a + \sum_{t=1}^{T} t^{-4} \leq \left(16\frac{\ln(N_a+T)}{\Delta_a^2} - 2N_a(1-\frac{\max\{0,\delta_a-\delta_{a^*}\}}{\Delta_a}) + (1+\frac{\pi^2}{3})\right).$$

When we sum up over all actions $a \neq a^*$, we get $R(T,\mathcal{A}) \leq \sum_{a\neq a^*} \Delta_a \left(16\frac{\ln(N_a+T)}{\Delta_a^2} - 2N_a(1-\frac{\max\{0,\delta_a-\delta_{a^*}\}}{\Delta_a}) + (1+\frac{\pi^2}{3})\right)$. □

REFERENCES

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*. 2312–2320.

[2] Anonymous. 2020. Supplementary material, Code and Data for "Unifying Offline Causal Inference and Online Bandit Learning for Data Driven Decision". https://1drv.ms/u/s!AuhX-fJM-sJvgy2klyvR0tDNKiVe?e=mgbpYH

[3] Peter Auer. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3, Nov (2002), 397–422.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[5] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119, 1 (2004), 249–275.

[6] Patrick Billingsley. 2008. *Probability and Measure*. John Wiley and Sons.

[7] Jonathan M Borwein, O-Yeat Chan, et al. 2009. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications* 12 (2009), 115–121.

[8] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. 2013. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*. 122–134.

[9] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 208–214.

[10] Varsha Dani, Thomas P Hayes, and Sham M Kakade. 2008. Stochastic linear optimization under bandit feedback. In *COLT*.

[11] Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*. Springer, 409–426.

[12] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[13] Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science* 14, 6 (2013), 570–580.

[14] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*. 814–823.

[15] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[16] Samuel Zahl. 1966. Bounds for the central limit theorem error. *SIAM J. Appl. Math.* 14, 6 (1966), 1225–1245.