Retrospective Inference for Stochastic Contextual Bandits

Retrospective analysis of historical data can be used to evaluate counterfactual interventions or design future experiments. However, inference in these settings can be quite challenging, particularly if the data was collected in an adaptive manner, for example using a contextual bandit algorithm. When treatment assignments at a point in time depend on past data, naive estimators of policy values may be biased or have non-normal distributions, while the mismatch between the treatment assignment policy used in data-collection and the policy considered for evaluation results in high-variance estimates and thus low power. The issue is exacerbated in contextual bandit experiments, where different regions of the feature space may be assigned to different treatment arms.

We propose a generic estimator that enables the evaluation of arbitrary policies in adaptively-collected data. We prove that our estimator is consistent and asymptotically normal, with improved power over existing alternatives. Our estimator takes into account temporal and contextual variations in sample variances across subgroups of contexts, and constructs adaptive weights for stabilizing and reducing variance of the final estimate. We illustrate the performance of our estimator using simulations, and we find that our estimator outperforms existing baselines with robustness to a variety of setups.

Key words: contextual bandits, policy inference, asymptotic normality, power improvement

1. Introduction

Estimating the benefits of targeted treatment assignment policies is a key problem in a variety of domains. For example, in personalized healthcare we may need to evaluate how particular groups of patients will respond to a given treatment regime (Murphy 2003), whereas in targeted advertising one may want to understand how alternative advertisements perform for different consumer segments (Li et al. 2011).

When estimating the value of a policy, the analyst has two choices: to design and run an experiment specifically for this task, or to conduct a retrospective analysis on historical data. The latter kind of analysis is called *off-policy evaluation*, and can be a much cheaper and straightforward alternative to the former, since it simply re-uses data that is already collected to estimate the effect of a counterfactual policy. However, it requires more sophisticated techniques for statistical inference.

In this paper, we propose an off-policy evaluation method that yields consistent and asymptotically normal estimates for a large class of experimental designs. Our main focus will be on the case of *adaptive* data-collection mechanism such as *contextual bandits* algorithms (e.g., Lattimore and Szepesvári 2018, Russo et al. 2017), in which assignment probabilities can change over time and depend on personal characteristics in complex and nonlinear ways.

Our method allows the researcher to overcome two main challenges inherent in off-policy evaluation in adaptive designs. The first challenge is that adaptive data-collection introduces sequential correlation among the observations, since treatment assignments later in the experiment depend on treatment assignments and outcomes earlier in the experiment. Even when there is no dependence on covariates, this time dependence causes naive estimates to be biased (Villar et al. 2015, Nie et al. 2017, Shin et al. 2019, 2020) and fail to be asymptotically normal (Zhang et al. 2020).

The second challenge is a little more subtle. Off-policy evaluation requires that there exists enough *overlap* between the policy being evaluated and the data-collection mechanism. This issue is defined and analyzed formally later, but intuitively it means that for each region of the feature space, there is sufficient probability that the arm assigned in the historical data is the one that would be selected using counterfactual policy of interest. For example, if a proposed ad targeting policy assigns a certain type of ad to customers in a certain demographic, then our estimates of the value of that policy for this customer segment will be less precise if few such people received this type of ad in the historical data. Overlap is required even in non-adaptive experiments (e.g., Imbens 2004), but in contextual bandit experiments the issue can be be extreme because the data-collection policy tends to concentrate on the optimal treatment for each subgroup. Continuing the example, if the proposed ad is not optimal for that demographic, then the number of observations that can be used to estimate a policy that assigns it is very small, and so our estimates will have high variance. Although the optimal policy as determined by the bandit is of interest, in many settings it is also of interest to test hypotheses about the benefits of different arms, to understand whether alternative policies with other attractive features (e.g. simpler policies) would work as well as the one selected by a bandit, and to gain insight about how each demographic responds to alternative treatments in order to guide future design and innovation.

Our method addresses both of the challenges described above. Building on recent work by Hadad et al. (2019), who focus on the non-contextual case, we propose an estimator that is a weighted average of doubly-robust scores, which incorporate generic machine learning models for nonparametric regression adjustments (Chernozhukov et al. 2016). Our weights are carefully designed so that the resulting estimator is consistent, asymptotically normal, and adaptive both over time and also across data-driven subgroups. We evaluate the robustness of our estimator in various simulation setups. Our experimental results demonstrate that our estimator has expected coverage, and has lower RMSE and standard error compared to existing baselines.

The rest of the paper proceeds as follows. We discuss related literature in more detail below, focusing on recent progress regarding off-policy inference using adaptive data. We formalize our problem in Section 2, and propose our estimator in Section 3. We present our main results in Section 4, and show the empirical evidence in Section 5. Finally, we conclude in Section 6. The key steps and the technical tools of our asymptotic analysis are put in the Appendix.

1.1. Related Literature

While the literature on policy evaluation in the case of iid data is vast (see e.g., Athey and Wager 2017, and references therein), policy inference on adaptive data remains a challenging problem. Many estimators that have desirable properties in the iid case will either be biased or have non-normal asymptotic distributions when applied to data collected adaptively.

The main cause of non-normality in adaptive experiments is failure to guarantee a "variance convergence" condition that is required in martingale central limit theorems (e.g., Hall and Heyde

2014, Section 3.4). In the case of data collected by bandit algorithms, this instability arises when the assignment probabilities fail to converge, which can happen when there is no clear optimal policy. Some recent papers have proposed different ways to avoid this issue. One line of research relies on constructing "local" estimates of policy value that are guaranteed to be stable and then assembled into a single asymptotically normal statistic. For example, Zhang et al. (2020) note that, when data is collected in batches, we can compute policy value estimates within each batch. For large batch sizes, the vector of studentized policy values from each batch is asymptotically jointly normal. In a more general setting, Luedtke and Van Der Laan (2016) show that even when data is not collected in batches one can still construct local policy value estimates normalized by their standard deviation, which can then be aggregated. Hadad et al. (2019) develop this insight further. Focusing on the case of multi-armed bandits, they show that by averaging such local estimates using carefully chosen weights, one can decrease the asymptotic variance of the resulting estimator while keeping its desirable asymptotic properties.

In a different approach, Deshpande et al. (2017) show that an appropriately tuned ridge-like penalty can ensure asymptotic normality of linear regression coefficients under strong correlation, and apply their method to linear bandits. Finally, another line of research sidesteps the need for asymptotic normality and applies martingale concentration inequalities to create finite-sample confidence sequences (Darling and Robbins 1967, Jamieson et al. 2014, Howard et al. 2018).

2. Problem Formulation

We now formulate the problem of inference on stochastic contextual bandits, and begin by introducing some notation used throughout the paper.

For sample size T, let $\mathcal{T} = \{t\}_{t=1}^{T}$. At each time step t, a new individual arrives with *context* X_t, which is a random sample from population density P in the sample space $\mathcal{X}$. An agent has a choice of K *arms* with arm set $\mathcal{W} = \{w\}_{w=1}^{K}$. Based on historical data, the agent updates her arm assignment probability function $e_t(x, w)$ for $x \in \mathcal{X}, w \in \mathcal{W}$, and selects an arm W_t with respect to probabilities $(e_t(X_t, w))_{w \in \mathcal{W}}$. Then, a stochastic *reward* outcome $Y_t = Y_t(W_t)$ is observed, which centers at $\mu(X_t, W_t)$. Define $H_t = \{(X_s, W_s, Y_s)\}_{s=1}^t$ to be a collection of samples up to time t, and let $\widehat{\mu}_t(x, w)$ be a regression adjustment model for $\mu(x, w)$ fitted only on past data H_{t-1}. Given a policy π, each context x is assigned to an arm w with probability $\pi(x, w)$, which yields a policy effect $Q(x, \pi) := \mathbb{E}\left[\sum_{w\in\mathcal{W}} \pi(x, w) Y(w) \middle| x\right]$. We are interested in the average policy effect $Q(\pi) := \mathbb{E}\left[Q(x, \pi)\right]$.

Remark 1. Throughout this paper, we assume that e_t are known. This is likely to be satisfied in practice when the data-collection mechanism is designed by the experimenter.

To simplify notation, we will drop the dependency on policy whenever it does not cause confusion, and write e.g. Q for $Q(\pi)$, $Q(x)$ for $Q(x, \pi)$, and etc. Our goal is to design a consistent and asymptotically normal estimator for off-policy evaluation in contextual bandits, and construct applicable confidence intervals with reduced variance.

3. Adaptive Subgroup Weighting

Our estimator will consist of a weighted average of unbiased *scores* of the form

$$\widehat{\Gamma}_t = \sum_{w\in\mathcal{W}} \pi(X_t, w) \left(\widehat{\mu}_t(X_t, w) + \frac{\mathbf{1}\left[W_t = w\right]}{e_t(X_t, w)} (Y_t - \widehat{\mu}_t(X_t, w))\right). \tag{1}$$

Let us unpack this definition and provide some motivation for it. The object $\widehat{\mu}_t(x, w)$ is an estimate of $\mu(x, w)$ based only on data up to time $t-1$. Using this estimate, we could have come up with a naive estimator of the policy value such as

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{w} \pi(X_t, w) \widehat{\mu}_t(X_t, w).$$

However, due to misspecification we cannot guarantee that the summand $\widehat{\mu}_t(x, w)$ is unbiased for $\mu(x, w)$. This is a problem for us because our results rely on martingale central limit theorems that require the construction of zero-mean martingale arrays. That is why we need to correct the misspecification bias in $\widehat{\mu}_t(x, w)$, and incorporate the second term in (1) as a weighted regression residual, where the weight has been chosen for bias correction. We can verify that

$$\mathbb{E}\left[\widehat{\Gamma}_t \middle| H_{t-1}, X_t\right] = Q(X_t).$$

This important property is what will allow us to derive consistency and asymptotic normality results later on.

Using the scores (1), we could then construct an estimator of the form

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\Gamma}_t,$$

but this estimator is not guaranteed to have an asymptotically normal distribution either. The problem is that the weights $\mathbf{1}[W_t = w]/e_t(X_t, w)$ depend on the reciprocal of the assignment probabilities $e_t(X_t, w)$. When these probabilities are small, the scores can become very large and have explosive variance. Moreover, if $e_t(X_t, w)$ fail to converge to anything in probability, then their oscillatory behavior can also prevent the estimator from converging in distribution to a normal distribution. Our solution to this problem is to replace the uniform $1/T$ weights with our later proposed weights that adapt to the asymptotic behavior of $\widehat{\Gamma}_t$. We call such weights *evaluation weights* and denote them by h. We will allow these weights to change over time and to be a function of observed covariates, but with several caveats that we will discuss next.

Recall that these scores can have explosive variance due to the vanishing and potentially unstable assignment probabilities. Thus, to stabilize the variance of our estimate, we choose evaluation weights to *offset* the score variance, that is, h_t is a decreasing function of the variance of $\widehat{\Gamma}_t$. However, to enable the martingale analysis in our proof, we require h_t to be measurable with respect to data up to $t-1$. We thus compute $\mathrm{Var}\left(\widehat{\Gamma}_t | H_{t-1}\right)$, the score variance conditioning on data up to $t-1$, and construct evaluation weights $h_t(\mathcal{X}; H_{t-1})$ to offset it. This $h_t(\mathcal{X}; H_{t-1})$ is changing over time, and we attribute it to the *temporal* variation in the score variance. This weighting scheme is the contextual bandit extension of that proposed for multi-armed bandits by Hadad et al.. We refer to it as **non-contextual weighting**, as $h_t(\mathcal{X}; H_{t-1})$ do not depend on any specific context.

However, we can even do better if h_t can incorporate context information. We notice that $\mathrm{Var}\left(\widehat{\Gamma}_t | H_{t-1}, X_t\right)$ can be quite different when X_t is in the different region of covariate space $\mathcal{X}$, which we refer to as the *contextual* variation in the score variance. But a naive weighting scheme that lets h_t offset $\mathrm{Var}\left(\widehat{\Gamma}_t|H_{t-1}, X_t\right)$ directly wouldn't give us a correct estimate. This is because when h_t depends on the current context X_t and thus takes the form $h_t(X_t; H_{t-1})$, the estimate would converge to $\mathrm{plim}\frac{\sum_{t\in\mathcal{T}} h_t(X_t;H_{t-1})Q(X_t)}{\sum_{t\in\mathcal{T}} h_t(X_t;H_{t-1})}$ instead of target estimand Q. The reason is that the population density would change with the law of covariates $x \in \mathcal{X}$ redistributed by the weights $h_t(X_t; H_{t-1})$.

To ensure the correct convergence but still make use of the covariates, we thus propose offsetting score variance for different subgroups of contexts. These subgroups are determined beforehand by a discretization function (we will address how to find it soon), so that the contextual variation of score variance is less severe within subgroups as opposed to between subgroups. Now for each subgroup, we can apply the similar approach as non-contextual weighting to get a consistent estimate of the policy value for this subgroup. Then summing up the subgroup results, we are able to estimate the policy value of the entire covariate space consistently. We refer to this approach as **subgroup weighting**, which will be proved later to be asymptotically more efficient than non-contextual weighting. An intuitive explanation is that subgroup weighting can always achieve the same result as non-contextual weighting by applying the same evaluation weights to all subgroups.

Yet in practice, we usually do not have knowledge of such discretization; we thus learn it from the data. This scheme is hence referred to as **adaptive subgroup weighting**. Specifically, we split the samples into two sets. A discretization function is fitted on the first dataset and then guides the second dataset on subgroup weighting. Besides, to achieve the comparable efficiency as if the full samples were used to estimate the policy value, we also generate a non-contextual weighting estimator from the first dataset. Combining estimates from both datasets tailors our final adaptive subgroup weighting estimator.

3.1. Non-contextual Weighting

We now propose our non-contextual weighting scheme. As discussed earlier, non-contextual weights $h_t(\mathcal{X}; H_{t-1})$ offset $\mathrm{Var}\left(\widehat{\Gamma}_t|H_{t-1}\right)$ over time, which deals with the temporal variation in sample score. We first compute this conditional variance.

Proposition 1. *The variance of score $\widehat{\Gamma}_t$ conditioning on history H_{t-1} is*

$$\begin{aligned}\mathrm{Var}\left(\widehat{\Gamma}_t|H_{t-1}\right) = \mathbb{E}\left[\sum_w \frac{\pi^2(X_t,w)(Y_t(w)-\widehat{\mu}_t(X_t,w))^2}{e_t(X_t,w)}\middle| H_{t-1}\right] \\ -\mathbb{E}\left[\left(\sum_w \pi(X_t,w)\left(\widehat{\mu}_t(X_t,w)-\mu(X_t,w)\right)\right)^2\middle| H_{t-1}\right] + \mathbb{E}\left[\left(Q(X_t)-Q\right)^2\right].\end{aligned} \tag{2}$$

Moreover, suppose that $Y_t(w)$ have finite second moments, e_t are strictly positive, and that $\widehat{\mu}_t$ are bounded almost surely. Then, there exist positive L, U such that

$$L\cdot\mathbb{E}\left[\sum_{w\in\mathcal{W}}\frac{\pi^2(X_t,w)}{e_t(X_t,w)}\middle| H_{t-1}\right] \le \mathrm{Var}\left(\widehat{\Gamma}_t|H_{t-1}\right) \le U\cdot\mathbb{E}\left[\sum_{w\in\mathcal{W}}\frac{\pi^2(X_t,w)}{e_t(X_t,w)}\middle| H_{t-1}\right]. \tag{3}$$

This proposition highlights the term $\mathbb{E}\left[\sum_{w\in\mathcal{W}}\frac{\pi^2(X_t,w)}{e_t(X_t,w)}\middle| H_{t-1}\right]$, which can be used as a *surrogate* to represent $\mathrm{Var}\left(\widehat{\Gamma}_t|H_{t-1}\right)$. More importantly,given data up to $t-1$, this surrogate can be approximated as follows, which ensures evaluation weights to be computed from H_{t-1}.

$$\mathbb{E}\left[\sum_{w\in\mathcal{W}}\frac{\pi^2(X_t,w)}{e_t(X_t,w)}\middle| H_{t-1}\right] \approx \left(\sum_{s=1}^{t-1}\sum_{w\in\mathcal{W}}\frac{\pi^2(X_s,w)}{e_t(X_s,w)}\right)\Big/(t-1). \tag{4}$$

Remark 2. This approximation is limited since assignment probabilities e_t are correlated with the historical contexts $\{X_s\}_{s=1}^{t-1}$. However, it is the best we can do provided the available data and the constraint that evaluation weights should be computed from H_{t-1}. One can improve this approximation if an external dataset of contexts is accessible.

We then choose evaluation weights using the notation in Hadad et al. (2019): (i) `propscore`, the inverse of this surrogate; and (ii) `stablevar`, the square root of the inverse. How these weights contribute to convergence is discussed in Section 4, and their numerical performances are demonstrated in Section 5. We formalize the *non-contextual weighting estimator* $\widehat{Q}_{T,N}$ and its estimated variance $\widehat{V}_{T,N}$ as follows:

$$\widehat{Q}_{T,N} = \frac{\sum_{t\in\mathcal{T}} h_t(\mathcal{X};H_{t-1})\widehat{\Gamma}_t}{\sum_{t\in\mathcal{T}} h_t(\mathcal{X};H_{t-1})}, \qquad \widehat{V}_{T,N} = \frac{\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t-\widehat{Q}_{T,N}\right)^2}{\left(\sum_{t\in\mathcal{T}} h_t(\mathcal{X};H_{t-1})\right)^2}. \tag{5}$$

3.2. Subgroup Weighting

We now utilize covariates to further reduce estimation variance. Following the previous discussion, groupwise evaluation weights are applied to account for the contextual variation in the score variance. Specifically, we discretize covariate space $\mathcal{X}$ into G subgroups $\mathcal{G} = \{\mathcal{X}_g\}$. Then, we construct groupwise evaluation weights $h_t(\mathcal{X}_g; H_{t-1})$ to offset $\int_{\mathcal{X}_g} \mathrm{Var}\left(\widehat{\Gamma}_t | H_{t-1}, X_t\right) \mathrm{d}P(X_t)$, which is again represented by a surrogate that can be approximated as follows:

$$\int_{\mathcal{X}_g} \sum_{w \in \mathcal{W}} \frac{\pi^2(X_t, w)}{e_t(X_t, w)} \, \mathrm{d}P(X_t) \approx \frac{\sum_{s=1}^{t-1} \mathbf{1}\left[X_s \in \mathcal{X}_g\right] \sum_{w \in \mathcal{W}} \frac{\pi^2(X_s, w)}{e_t(X_s, w)}}{\sum_{s=1}^{t-1} \mathbf{1}\left[X_s \in \mathcal{X}_g\right]}. \tag{6}$$

This allows us to estimate the subgroup policy value $Q(\mathcal{X}_g) = \mathbb{E}\left[\mathbf{1}\left[x \in \mathcal{X}_g\right] Q\right]$ using the *individual subgroup estimator* $\widehat{Q}_T(\mathcal{X}_g)$:

$$\widehat{Q}_T(\mathcal{X}_g) = \frac{\sum_{t \in \mathcal{T}} h_t(\mathcal{X}_g; H_{t-1}) \, \mathbf{1}\left[X_t \in \mathcal{X}_g\right] \widehat{\Gamma}_t}{\sum_{t \in \mathcal{T}} h_t(\mathcal{X}_g; H_{t-1})}. \tag{7}$$

Summing up results from all subgroups yields a *subgroup weighting estimator* $\widehat{Q}_{T,S}$ for the policy value Q and its estimated variance $\widehat{V}_{T,S}$ as follows.

$$\widehat{Q}_{T,S} = \sum_{g \in \mathcal{G}} \widehat{Q}_T(\mathcal{X}_g), \quad \widehat{V}_{T,S} = \sum_{t \in \mathcal{T}} \left(\sum_{g \in \mathcal{G}} \frac{h_t(\mathcal{X}_g; H_{t-1})}{\sum_{s \in \mathcal{T}} h_t(\mathcal{X}_s; H_{t-1})} \left(\mathbf{1}\left[X_t \in \mathcal{X}_g\right] \widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g) \right) \right)^2. \tag{8}$$

We note that non-contextual weighting encompasses subgroup weighting when restricting the number of subgroups to one, while the latter is shown to improve power (Corollary 1). Intuitively, subgroup weights offset the score variance specifically for this subgroup, which should perform no worse than non-contextual weights that remain unchanged across all subgroups.

3.3. Adaptive Subgroup Weighting

Adaptive subgroup weighting is motivated by learning a discretization function for subgroup weighting. Let $\mathcal{T}_1 = \{t\}_{t=1}^{rT}$ and $\mathcal{T}_2 = \{t\}_{t=rT+1}^{T}$ be time steps of the split data. From the first set $\mathcal{D}_1 = \{(X_t, W_t, Y_t)\}_{t \in \mathcal{T}_1}$, we obtain a discretization function $d_T(x)$ and a non-contextual weighting estimator $\widehat{Q}_{T,N}$. We then discretize covariate space into G subgroups, and let $\mathcal{G} = \{g\}_{g=1}^{G}$ be subgroup indices. We thus have a subgroup weighting estimator $\widehat{Q}_{T,S}$ from the second part $\mathcal{D}_2 = \{(X_t, W_t, Y_t)\}_{t \in \mathcal{T}_2}$. Finally, to minimize variance, the *adaptive subgroup weighting estimator* $\widehat{Q}_T$ is a weighted average of both results $\widehat{Q}_{T,N}$ and $\widehat{Q}_{T,S}$, with weights proportional to $1/\widehat{V}_{T,N}$ and $1/\widehat{V}_{T,S}$. The variance of $\widehat{Q}_T$ is hence estimated by $\widehat{V}_T = \widehat{V}_{T,N}\widehat{V}_{T,S}/(\widehat{V}_{T,N}+\widehat{V}_{T,S})$. The procedure is formalized in Algorithm 1.

Algorithm 1 Adaptive Subgroup Weighting

Input: samples $\{(X_t, W_t, Y_t)\}_{t\in\mathcal{T}}$, data-splitting ratio r, number of subgroups G.

Output: adaptive subgroup weighting estimator $\widehat{Q}_T$ and estimated variance $\widehat{V}_T$.

1. Split samples into two sets: $\mathcal{D}_1 = \{(X_t, W_t, Y_t)\}_{t\in\mathcal{T}_1}$ and $\mathcal{D}_2 = \{(X_t, W_t, Y_t)\}_{t\in\mathcal{T}_2}$.
2. Apply non-contextual weighting to dataset $\mathcal{D}_1$, and obtain estimator and estimated variance:
$$\widehat{Q}_{T,N} = \frac{\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})\widehat{\Gamma}_t}{\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})}, \quad \widehat{V}_{T,N} = \frac{\sum_{t\in\mathcal{T}_1} h_t^2(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t - \widehat{Q}_{T,N}\right)^2}{\left(\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})\right)^2}. \tag{9}$$
3. Use $\mathcal{D}_1$ to fit a function $d_T(\cdot)$ that discretizes covariate space $\mathcal{X}$ into G subgroups $\{\mathcal{X}_g\}_{g\in\mathcal{G}}$.
4. Apply subgroup weighting to dataset $\mathcal{D}_2$, and obtain estimator and estimated variance:
$$\widehat{Q}_{T,S} = \sum_{g\in\mathcal{G}} \widehat{Q}_T(\mathcal{X}_g), \quad \widehat{V}_{T,S} = \sum_{t\in\mathcal{T}_2}\left(\sum_{g\in\mathcal{G}} \frac{h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2} h_s(\mathcal{X}_g;H_{t-1})}\right)^2, \tag{10}$$
where $\widehat{Q}_T(\mathcal{X}_g) = \frac{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g;H_{t-1})\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t}{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g;H_{t-1})}$ evaluates subgroup policy value $\mathbb{E}\left[\mathbf{1}\left[x\in\mathcal{X}_g\right]Q(x)\right]$.
5. Aggregate results and get adaptive subgroup weighting estimator and estimated variance:
$$\widehat{Q}_T = \frac{\widehat{V}_{T,S}\widehat{Q}_{T,N} + \widehat{V}_{T,N}\widehat{Q}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}, \quad \widehat{V}_T = \frac{\widehat{V}_{T,N}\widehat{V}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}. \tag{11}$$

We now describe our heuristics in fitting a discretization function d_T on dataset $\mathcal{D}_1$. We want to group contexts with similar conditional variance of scores across $\mathcal{T}_2$. Thus in simulations (Section 5), we fit a decision tree to approximate the final score variance in dataset $\mathcal{D}_1$: $\gamma(x) = \mathbb{E}\left[\sum_{w\in\mathcal{W}} \pi^2(x,w)/e_{rT}(x,w) \middle| H_{rT-1}, x\right]$. In other words, the training covariates and targets for the decision tree are $\{X_s\}_{s\in\mathcal{T}_1}$ and $\{\gamma(X_s)\}_{s\in\mathcal{T}_1}$ respectively.

4. Main Results

In this section, we establish the consistency and asymptotic normality of our estimators, and show that subgroup weighting improves inference power as compared to non-contextual weighting. We start by describing assumptions that are required in our analysis. We then show our main results, and discuss how one can choose evaluation weights to satisfy the assumptions. Finally, we generalize our estimators to evaluate functions that include contrast value between policies.

4.1. Assumptions

Assumption 1 **(Settings)**. *The joint distribution* $(X, Y(w_1), Y(w_2), \ldots, Y(w_K), W_t)$ *satisfies:*

(a) *Probabilistic agent:* $e_t(x,w) > 0, \forall (x,w) \in \mathcal{X} \times \mathcal{W}$*;*

(b) *Unconfoundedness:* $(Y_t(w))_{w \in \mathcal{W}} \perp W_t | X_t$*;*

(c) *Stationarity:* $(X_t, Y_t(w) : w \in \mathcal{W})$ *are independent and identically distributed;*

(d) *Stochastic noise: There exist positive constants* L, U *such that* $\mathbb{E}\left[Y^4(w)\right] \leq U < \infty$ *and* $\mathrm{Var}\left(Y(w)\right) \geq L > 0$ *for any* $(x,w) \in \mathcal{X} \times \mathcal{W}$*.*

The *probabilistic agent* condition requires positive e_t but allows it to vanish, which relaxes the *overlap* assumption in causal inference literature that requires e_t bounded away from zero (Imbens 2004). The *unconfoundedness* condition implies that enough covariates have been measured to capture the dependency between the selected arm and the potential rewards. The *stationarity* condition requires that the population density P that the stochastic environment that generates the rewards is unchanged over time, but the observed rewards are not iid, as which arm to choose at time t relies on the data up to time t. Finally, the *stochastic noise* condition regularizes the tail of the reward distribution, and is satisfied for a variety of density families including sub-Gaussian random variables.

Assumption 2 **(Nuisance Estimators)**. *The regression adjustment model* $\widehat{\mu}_t$ *and the discretization function* d_T *are bounded and converge almost surely to bounded functions* μ_∞ *and* d_∞*.*

We do not require $\widehat{\mu}_t$ to be consistent to restore normality, but a more accurate regression adjustment brings more inference power (see Corollaries 2-4 in the Appendix). Meanwhile, d_T does not need to perfectly group contexts with similar score variances, since subgroup weighting based on any discretization performs at least as well as its non-contextual counterpart asymptotically (see Corollary 1). But a more precise discretization reduces more estimation variance.

We now introduce assumptions on evaluation weights. For clarity, we denote T as a set of time steps and G as a set of subgroup indices invoked in the assumptions. Specifically, replace (T,G) with $(\mathcal{T}_1,\{1\})$ in the matter of dataset $\mathcal{D}_1$, and with $(\mathcal{T}_2,\mathcal{G})$ as regards to dataset $\mathcal{D}_2$.

Assumption 3 **(Weights)**. *For each subgroup $g\in\mathsf{G}$, the evaluation weights $h_t(\mathcal{X}_g;H_{t-1})$ satisfy*

(a) *Lyapunov condition:*

$$\frac{\sum_{t\in\mathsf{T}} h_t^4(\mathcal{X}_g;H_{t-1})\mathbb{E}\left[\sum_w \frac{\pi^4(x,w)}{e_t^3(x,w)}\middle| H_{t-1}, x\in\mathcal{X}_g\right]}{\left(\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t^2(\mathcal{X}_g;H_{t-1})\sum_w\frac{\pi^2(x,w)}{e_t(x,w)}\middle| x\in\mathcal{X}_g\right]\right)^2}\xrightarrow{L_1} 0. \tag{12}$$

(b) *Variance convergence:*

$$\frac{\sum_{t\in\mathsf{T}} h_t^2(\mathcal{X}_g;H_{t-1})\mathbb{E}\left[\sum_w \frac{\pi^2(x,w)(Y_t(w)-\mu_\infty(x,w))^2}{e_t(x,w)}\middle| H_{t-1}, x\in\mathcal{X}_g\right]}{\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t^2(\mathcal{X}_g;H_{t-1})\sum_w\frac{\pi^2(x,w)(Y_t(w)-\mu_\infty(x,w))^2}{e_t(x,w)}\middle| x\in\mathcal{X}_g\right]}\xrightarrow{L_1} 1, \quad \textit{and} \quad \frac{\sum_{t\in\mathsf{T}} h_t^2(\mathcal{X}_g;H_{t-1})}{\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t^2(\mathcal{X}_g;H_{t-1})\right]}\xrightarrow{L_1} 1. \tag{13}$$

(c) *Infinite sampling:*

$$\frac{\sum_{t\in\mathsf{T}} h_t^2(\mathcal{X}_g;H_{t-1})\mathbb{E}\left[\sum_w \frac{\pi^2(x,w)}{e_t(x,w)}\middle| H_{t-1}, x\in\mathcal{X}_g\right]}{\left(\sum_{t\in\mathsf{T}}(h_t(\mathcal{X}_g;H_{t-1}))\right)^2}\xrightarrow{p} 0 \tag{14}$$

(d) *Weight convergence:*

$$\frac{\sum_{t\in\mathsf{T}} h_t(\mathcal{X}_g;H_{t-1})}{\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t(\mathcal{X}_g;H_{t-1})\right]}\xrightarrow{p} 1 \quad \textit{and} \quad \frac{\sum_{t\in\mathsf{T}} h_t(\mathcal{X}_{g_1};H_{t-1})h_t(\mathcal{X}_{g_2};H_{t-1})}{\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t(\mathcal{X}_{g_1};H_{t-1})h_t(\mathcal{X}_{g_2};H_{t-1})\right]}\xrightarrow{L_1} 1. \tag{15}$$

The *Lyapunov* and *infinite sampling* conditions are satisfied with `propscore` and `stablevar` evaluation weights as long as the assignment probabilities have a floor that decays slower than $1/t$, which essentially require sufficient samples from each arm. The *variance convergence* condition is particularly important to ensure stabilized variance of the estimator (see details in the following proof sketch and Remark 4). The second term $\frac{\sum_{t\in\mathsf{T}} h_t^2(\mathcal{X}_g;H_{t-1})}{\sum_{t\in\mathsf{T}}\mathbb{E}\left[h_t^2(\mathcal{X}_g;H_{t-1})\right]}$ can be eliminated when $\sum_w\frac{\pi^2(x,w)}{e_t(x,w)}$ is large, which usually happens due to the mismatch between the evaluation policy and the data-collection mechanism. The final *weight convergence* condition helps us to further stabilize the variance and covariance when we need to aggregate different estimates for subgroup weighting and adaptive subgroup weighting (see details in the following proof sketch).

4.2. Central Limit Theorem

We now present the asymptotic properties of our estimators. The following theorem shows that our estimators are asymptotically unbiased, and can be used to construct reliable confidence intervals as paired with the corresponding estimated variances.

THEOREM 1. *Suppose that Assumptions 1 and 2 are satisfied. Suppose that the non-contextual weights and the subgroup weights both satisfy Assumption 3. Then, the adaptive subgroup weighting estimator $\widehat{Q}_T$ formalized in Algorithm 1 are consistent for the true policy value Q, and together with the estimated variances $\widehat{V}_T$ formalized in Algorithm 1, forms an asymptotically normal studentized statistic: $\frac{\widehat{Q}_T - Q}{\sqrt{\widehat{V}_T}} \xrightarrow{d} \mathcal{N}(0,1)$.*

REMARK 3. The consistency and asymptotic normality of non-contextual weighting estimator $\widehat{Q}_{T,N}$ and subgroup weighting estimator $\widehat{Q}_{T,S}$ are implied by Theorem 1. By choosing the data-splitting ratio to be 1, adaptive subgroup weighting is reduced to non-contextual weighting. By choosing the data-splitting ratio to be 0 and fixing the discretization function, adaptive subgroup weighting is reduced to subgroup weighting.

In the interest of space, we show a road map of our proof here, and put the key steps and technical tools of our analysis in the Appendix. We first show how to prove the convergence of the non-contextual weighting estimator $\widehat{Q}_{T,N}$. The t-statistic of non-contextual weighting is $\frac{\widehat{Q}_{T,N} - Q}{\sqrt{\widehat{V}_{T,N}}}$. After cancelling out the $\sum_{t\in\mathcal{T}} h_t(\mathcal{X}; H_{t-1})$ in both numerator and denominator, the t-statistic is simplified to be

$$\frac{\widehat{Q}_{T,N} - Q}{\sqrt{\widehat{V}_{T,N}}} = \frac{\sum_{t\in\mathcal{T}} h_t(\mathcal{X}; H_{t-1})(\widehat{\Gamma}_t - Q)}{\sqrt{\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X}; H_{t-1})(\widehat{\Gamma}_t - \widehat{Q}_T)^2}}. \tag{16}$$

To show its convergence, we construct a zero-mean martingale array (martingale difference sequence or MDS) as follows:

$$\xi_{T,t} = \frac{\sum_{t\in\mathcal{T}} h_t(\mathcal{X}; H_{t-1})(\widehat{\Gamma}_t - Q)}{\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X}; H_{t-1})(\widehat{\Gamma}_t - Q)^2\right]}}. \tag{17}$$

Under Assumptions 1-3(b), we show this MDS $\xi_{T,t}$ satisfies the conditions in martingale CLT in Hall and Heyde (2014), and thus $\sum_{t\in\mathcal{T}} \xi_{T,t} \xrightarrow{d} \mathcal{N}(0,1)$.

To see the consistency of $\widehat{Q}_{T,N}$, we have

$$\widehat{Q}_{T,N} - Q = \underbrace{\frac{\sum_{t\in\mathcal{T}} h_t(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - Q)}{\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - Q)^2\right]}}}_{\sum_{t\in\mathcal{T}}\xi_{t,T}\xrightarrow{p}\mathcal{N}(0,1)} \underbrace{\frac{\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - Q)^2\right]}}{\sum_{t\in\mathcal{T}} h_t(\mathcal{X};H_{t-1})}}_{\text{vanishes by Assumptions 3(b),3(c)}}. \tag{18}$$

To further obtain the CLT of $\widehat{Q}_{T,N}$, we finally show the convergence of its variance, that is,

$$\frac{\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - \widehat{Q}_T)^2}{\mathbb{E}\left[\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - Q)^2\right]} \xrightarrow{p} 1. \tag{19}$$

The convergence of $\widehat{Q}_T(\mathcal{X}_g)$ can be verified using the exact same procedure.

However, for subgroup weighting estimator $\widehat{Q}_{T,S}$ and adaptive subgroup weighting estimator $\widehat{Q}_T$, we need to aggregate results of different estimates. Thus their t-statistics can not be simplified as that in (16), and particularly we have the term $\sum_{t\in\mathcal{T}} h_t(\mathcal{X}_g;H_{t-1})$ retained. Moreover, there exists correlation between different estimates, which involves the term $\sum_{t\in\mathcal{T}} h_t(\mathcal{X}_{g_1};H_{t-1})h_t(\mathcal{X}_{g_2};H_{t-1})$. Therefore, we additionally require Assumption 3(d) to ensure these two terms converging, which results in the convergence of the variance of our estimators and the covariance between them. Then, following a similar procedure as before, we construct a MDS and use its convergence to invoke the consistency and asymptotic normality of the corresponding estimator.

Yet despite the additional condition required in the asymptotic analysis of subgroup weighting, it can improve power on non-contextual weighting, as stated in the following corollary. This can be directly verified by checking the asymptotic variance form of subgroup weighting if one uses the same evaluation weights across subgroups (see details in Corollaries 3 and 4 in the Appendix).

Corollary 1 **(Power improvement of subgroup weighting)**. *Given Assumptions 1-3, when estimating from the same data, if the subgroup weights are entirely replaced by non-contextual weights, then the subgroup weighting estimator has the same asymptotic variance as that of the non-contextual weighting estimator.*

Remark 4. As we can see, subgroup weighting brings more efficiency by reducing more estimation variance, but it requires additional Assumption 3(d) to ensure the convergence of the variance and covariance. For non-contextual weighting, one only needs to ensure the convergence of

$\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t-\widehat{Q}_T)^2$ to achieve convergent variance. By Assumption 3(a), this is reduced to the convergence of $\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})\mathbb{E}\left[(\widehat{\Gamma}_t-\widehat{Q}_T)^2\middle| H_{t-1}\right]$, which can be generally achieved by convergent $\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})\mathbb{E}\left[\frac{\pi^2(x,w)}{e_t(x,w)}\middle| H_{t-1}\right]$, as reflected in Proposition 1. Then, one can use the "stick-breaking" approach proposed in Hadad et al. (2019) to heuristically approximate that $\sum_{t\in\mathcal{T}} h_t^2(\mathcal{X};H_{t-1})\mathbb{E}\left[\frac{\pi^2(x,w)}{e_t(x,w)}\middle| H_{t-1}\right]=1$. In the setting of contextual bandits, this means recursively constructing $h_t^2(\mathcal{X};H_{t-1})$ such that

$$h_t^2(\mathcal{X};H_{t-1})\sum_{i=1}^{t-1}\frac{\pi^2(X_i,w)}{e_t(X_i,w)}\Big/(t-1)=\left(1-\sum_{t=1}^{t-1}h_s^2(\mathcal{X};H_{t-1})\sum_{j=1}^{s-1}\frac{\pi^2(X_j,w)}{e_s(X_j,w)}\Big/(s-1)\right)\lambda_t(\mathcal{X};H_{t-1}),\qquad(20)$$

where $\lambda_t(\mathcal{X};H_{t-1})$ reflects the ratio of current h_t-adjusted score variance to the remaining variance. The `stablevar` weights can be interpreted by choosing $\lambda_t(\mathcal{X};H_{t-1})=1/(T-t+1)$.

Remark 5. We note that the asymptotic variance of $\widehat{Q}_{T,S}$ is smaller than the sum of those of $\{\widehat{Q}_T(\mathcal{X}_g)\}_{g\in\mathcal{G}}$. As discussed before, when the variances of $\{\widehat{V}_T(\mathcal{X}_g)\}_{g\in\mathcal{G}}$ and $\widehat{V}_{T,S}$ converge, we can use the expected values of them as their asymptotic variances (See Corollaries 2-4 in the Appendix for details). Thus we have,

$$\mathbb{E}\left[\widehat{V}_{T,S}\right]=\underbrace{\sum_{t\in\mathcal{T}_2}\sum_{g\in\mathcal{G}}\mathbb{E}\left[\left(\frac{h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-Q(\mathcal{X}_g))}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_t(\mathcal{X}_s;H_{t-1})\right]}\right)^2\right]}_{\text{sum of }\mathbb{E}\left[\widehat{V}_T(\mathcal{X}_g)\right]}-\sum_{g_1\neq g_2\in\mathcal{G}}\mathbb{E}\left[\frac{h_t(\mathcal{X}_{g_1};H_{t-1})h_t(\mathcal{X}_{g_2};H_{t-1})Q(\mathcal{X}_{g_1})Q(\mathcal{X}_{g_2})}{\left(\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_{g_1};H_{t-1})\right]\right)\left(\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_{g_2};H_{t-1})\right]\right)}\right].$$

Intuitively, different individual subgroup estimators $\widehat{Q}_T(\mathcal{X}_g)$ are correlated with each other, since the realized context X_t can only fall into one subgroup; this helps reduce the total variance in the aggregated estimator $\widehat{Q}_{T,S}$.

4.3. Generalization to Function Evaluation

Finally, we emphasize that our results above are not restricted to policy evaluation, but can be applied to evaluating more general functions. Consider a function space where eligible elements $f(x,w)$ are (i) bounded: $\|f(x,w)\|_\infty<\infty$; and (ii) non-constant: $\mathbb{E}\left[\sum_{w\in\mathcal{W}} f^2(x,w)\right]>0$. Suppose that we are interested in estimating the function value $Q(f)=\mathbb{E}\left[\sum_w f(x,w)Y(w)\right]$. Then, under regularity conditions, all estimators referenced above have the same asymptotic properties. One direct application is estimating the contrast between policies, where $f(x,w)=\pi_1(x,w)-\pi_2(x,w)$ for different policies π_1 and π_2. We demonstrate this deployment in the experiments below.

5. Simulation Studies

In this section, we evaluate our estimators in simulations and address the following questions:

(i) Does our weighting-based compensation for variations in $\mathrm{Var}\left(\widehat{\Gamma}_t | H_{t-1}, X_t\right)$ improve performance? Specifically, non-contextual weighting offsets score variance over time to account for the temporal variation, and adaptive subgroup weighting further addresses the contextual variation by applying groupwise evaluation weights to data-driven subgroups.

(ii) Compared to non-contextual weighting over the entire dataset, adaptive subgroup weighting makes two changes: (a) subgroup weighting on dataset $\mathcal{D}_2$; and (b) aggregation of estimates using datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Which change drives performance improvement?

(iii) Is adaptive subgroup weighting robust to different simulation setups? In which scenarios should this method be favored?

(iv) How does the data splitting ratio affect the performance of adaptive subgroup weighting?

Data Generating Process. We consider a tree data-generating process. Covariates (x_1, x_2) are generated from a two-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, I_2)$. The space $\mathcal{X}$ is split into four regions: $\{x_1 < 0.5, x_2 < 0.5\}$, $\{x_1 < 0.5, x_2 > 0.5\}$, $\{x_1 > 0.5, x_2 < 0.5\}$, and $\{x_1 > 0.5, x_2 > 0.5\}$. Four arms generate rewards non-contextually within each region, and have region-wise expected rewards $[1, 0, 0, 0]$, $[0.99, 1, 0, 0]$, $[0, 0, 1, 0]$, and $[0, 0, 0, 1]$ with additive iid $\mathcal{N}(0, 1)$ noise.

Agents. We consider Thompson sampling agent (Thompson 1933). A floor is imposed on assignment probabilities and is polynomially decayed with batch indices.

Policies and Contrast. We consider two policies and the contrast value between them: the *optimal contextual policy*, which assigns each context to its optimal arm; and the *best-arm policy*, which always pulls the first arm that has the largest expected reward.

Weighting Schemes. We consider three weighting schemes applied to scores with ridge regression adjustments: non-contextual weighting (NW), adaptive subgroup weighting (ASW), and aggregated non-contextual weighting (ANW), where ANW applies non-contextual weighting to both datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ and then aggregates the estimates. We introduce ANW to investigate the

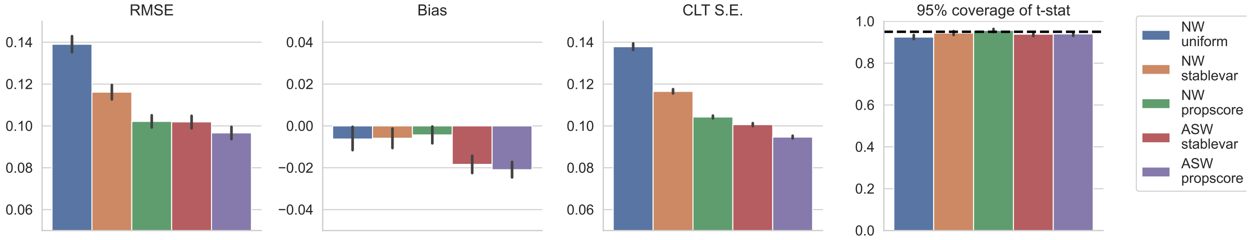

Figure 1 Estimates of contrast between optimal contextual policy and best arm policy. Adaptive subgroup weighting (ASW) has smaller RMSE and more inference power (less standard error for expected coverage) than non-contextual weighting (NW), though it introduces slightly more bias.

sources of variance reduction from ASW. Weights are chosen in the following forms: `uniform` weights which share the same value among all scores, `propscore` and `stablevar` evaluation weights and their corresponding subgroup variants.

Results. Performances are compared based on 3000 replications with sample size $T = 4000$.

Firstly, We show that compensation for variations in score variances does improve performance. We apply weights `NW-uniform`, `NW-propscore`, `NW-stablevar`, `ASW-propscore` and `ASW-stablevar` to estimate contrast between the optimal contextual policy and the best-arm policy. Figure 1 illustrates their performances on RMSE, bias, standard error and coverage. We see that by accounting for temporal variation, non-contextual weighting improves power and RMSE as compared to uniform weighting; adaptive subgroup weighting further refines the results by moreover considering the contextual variation. We notice the relatively large bias in adaptive subgroup weighting. This is because only partial data is used to correct misspecification bias in $\widehat{\mu}_t$, and thus more samples are required to reach consistency.

Secondly, to understand the sources of performance improvement from adaptive subgroup weighting, we compare three weighting schemes: NW, ASW, and ANW with `propscore` evaluation weights. Figure 2 shows RMSE when evaluating the optimal contextual policy and the best-arm policy. It demonstrates that aggregation of estimates yields a slight improvement, but the main effect comes from subgroup weighting.

Thirdly, we conclude from Figure 2 that adaptive subgroup weighting outperforms non-contextual weighting across various evaluation policies, floor-decay rates, and data-splitting ratios.

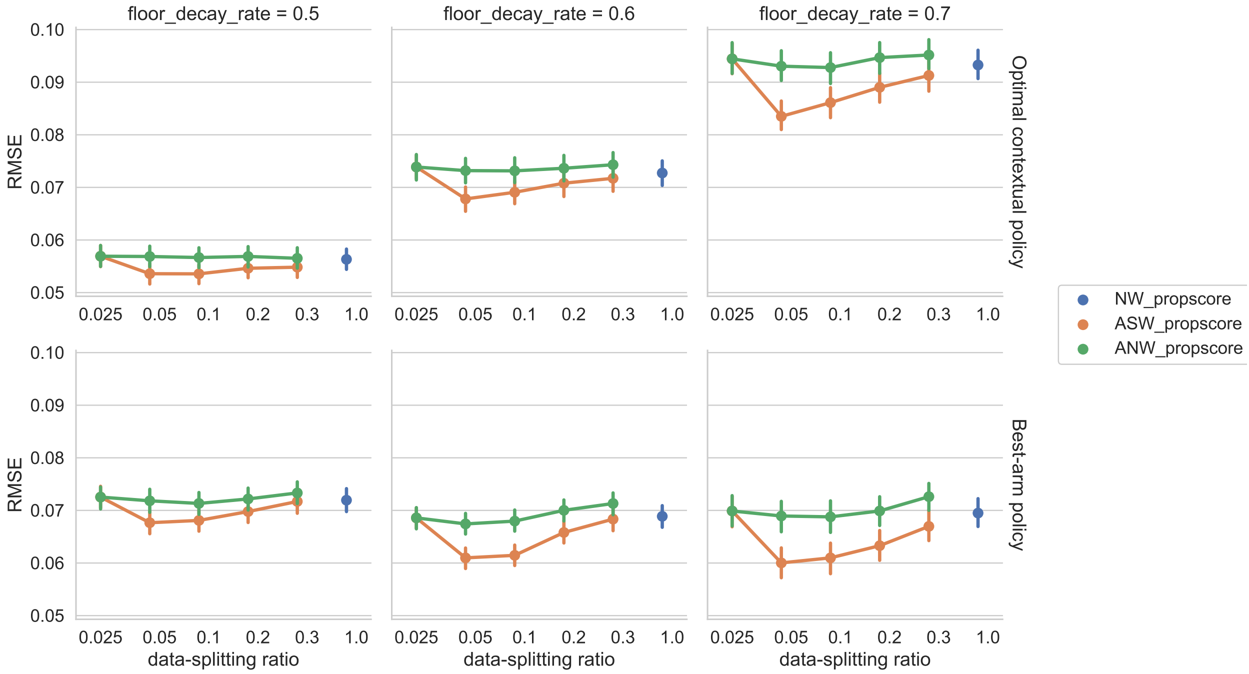

Figure 2 RMSE of evaluating optimal contextual policy and best-arm policy across different floor-decay rates and data-splitting ratios. When the data-splitting ratio equals 1, adaptive subgroup weighting (ASW)/ aggregated non-contextual weighting (ANW) reduces to non-contextual weighting (NW). The performance improvement of ASW is mostly driven by subgroup weighting, and ASW is more advantageous when applied to more adaptive data, e.g., a faster floor decay.

It is also more advantageous when used on more adaptive data. By varying the floor-decay rates of assignment probabilities, we alter the degree of adaptivity. More improvement is observed in cases of a faster floor decay, as compared to a slower floor decay. This phenomenon is consistent with our analysis. More adaptivity brings more variation of score variance in covariate space, which motivates our design of adaptive subgroup weighting.

Finally, the data-splitting ratio should be small to fully utilize subgroup weighting. The ideal ratio should correspond to the earliest time when the agent identifies the difference of score variance in covariate space. In our simulations, this usually happens after collecting the first two batches.

6. Conclusion

Datasets collected during adaptive experiments can provide a cheap and efficient way of estimating the effect of new policies without having to run a dedicated experiment. Such analyses can also be crucial for providing empirical evidence to motivate further experimentation, for example to justify future clinical trials of medical treatments (Vogelstein et al. 2020).

However, off-policy evaluation on adaptive data can be challenging. Adaptivity brings dependency among samples, and aggravates the overlap between the data-collection and evaluation policies. Many estimators hence suffer from huge variance and non-normal limit distributions.

In this paper, we address both issues simultaneously, and propose adaptive subgroup weighting, a generic estimator to evaluate policies in a large class of adaptive experiment designs that includes stochastic contextual bandits. We show that our estimator is consistent and asymptotically normal, and exhibits reduced variance when compared to existing alternatives. Empirically, our estimator has anticipated coverage and outperforms baselines with lower RMSE and standard error.

A number of important research directions remain open. Adaptive subgroup weighting gains more inference power by offsetting sample variances both over time and across data-driven subgroups. However, in our current analysis, the convergence of its variance additionally requires the convergence of evaluation weights. Alternative approaches such as non-contextual weighting can ensure convergent variance with carefully designed evaluation weights, but lose inference power as compared to adaptive subgroup weighting (see details in Corollary 1 and Remark 4). It remains unclear whether there really exists a tradeoff between efficiency and robustness, or it is due to a limitation of our proof technique. We conjecture that the weight convergence is essential for convergent variance of an aggregated estimator, but one can explore weighting schemes that simultaneously stabilize variance and have convergent weights.

Finally, our approach can serve as a starting point to study other problems, such as optimal policy learning on adaptively collected data. Several efficient estimators have been proposed for optimal policy estimation with fast regret decay (Athey and Wager 2017, Luedtke and Chambaz 2017). The key thing in policy learning is to prove a universal bound on the policy value for a set of policies. An important future problem is the extension of our central limit theorem to a uniform one, which would imply analogous results for policy learning from adaptively randomized data.

Appendix

In this appendix, we present the key steps in our analysis and the technical tools used in our proof. The following martingale central limit theorem is derived from Corollary 3.1 in Hall and Heyde (2014)[1], and will be invoked throughout the proof.

PROPOSITION 2 **(Martingale CLT)**. *Let $\{\xi_{T,t}, \mathcal{F}_{T,t}\}_{t\in\mathcal{T}}$ be a martingale difference sequence with finite fourth moment. Suppose that $\sum_{t\in\mathcal{T}} \mathbb{E}\left[\xi_{T,t}^2 | H_{t-1}\right] \xrightarrow{p} 1$(**conditional variance convergence**) and that $\sum_{t\in\mathcal{T}} \mathbb{E}\left[\xi_{T,t}^4 \middle| H_{t-1}\right] \xrightarrow{p} 0$ (**conditional 4^{th} moment decay**). Then, $\sum_{t\in\mathcal{T}} \xi_{T,t} \xrightarrow{d} \mathcal{N}(0,1)$.*

For each estimator formalized in Algorithm 1, we construct an auxiliary MDS. Give the assumptions in Section 4.1, we successively show the convergence of our estimators by first proving the convergence of corresponding MDS. Asymptotic results will be given in the order of individual subgroup estimator $\widehat{Q}_T(\mathcal{X}_g)$, non-contextual weighting estimator $\widehat{Q}_{T,N}$, subgroup weighting estimator $\widehat{Q}_{T,S}$, and adaptive subgroup weighting estimator $\widehat{Q}_T$.

A: Convergence of Individual Subgroup Estimator $\widehat{Q}_T(\mathcal{X}_g)$

The individual subgroup estimator $\widehat{Q}_T(\mathcal{X}_g)$ is applied to dataset $\mathcal{D}_2$ and estimates subgroup policy value $Q(\mathcal{X}_g) = \mathbb{E}\left[\mathbf{1}\left[x \in \mathcal{X}_g\right] Q\right]$. Its definition and estimated variance are as follows:

$$\widehat{Q}_T(\mathcal{X}_g) = \frac{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g; H_{t-1})\,\mathbf{1}\left[X_t \in \mathcal{X}_g\right]\widehat{\Gamma}_t}{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g; H_{t-1})}, \quad \widehat{V}_T(\mathcal{X}_g) = \frac{\sum_{t\in\mathcal{T}_2} h_t^2(\mathcal{X}_g; H_{t-1})\left(\mathbf{1}\left[X_t \in \mathcal{X}_g\right]\widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g)\right)^2}{\left(\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g; H_{t-1})\right)^2}. \tag{21}$$

Let its auxiliary MDS $\{\xi_{T,t}(\mathcal{X}_g)\}_{t\in\mathcal{T}_2}$ be

$$\xi_{T,t}(\mathcal{X}_g) = \frac{h_t(\mathcal{X}_g; H_{t-1})(\mathbf{1}\left[X_t \in \mathcal{X}_g\right]\widehat{\Gamma}_t - Q(\mathcal{X}_g))}{\sqrt{\sum_{t\in\mathcal{T}_2} \mathbb{E}\left[h_t^2(\mathcal{X}_g; H_{t-1})(\mathbf{1}\left[X_t \in \mathcal{X}_g\right]\widehat{\Gamma}_t - Q(\mathcal{X}_g))^2\right]}}. \tag{22}$$

We justify this definition by the fact that subgroup g is always determined given H_{t-1} for $t > rT$; thus, $\mathbb{E}\left[\xi_{T,t}(\mathcal{X}_g) \middle| H_{t-1}\right] = 0$. The conditional variance convergence of $\xi_{T,t}(\mathcal{X}_g)$ is a result of Proposition 1 and Assumption 3(b). For its conditional fourth moment, we have the following lemma.

[1] The original conditional Lindeberg condition in Hall and Heyde (2014) can be derived by those in Proposition 2.

LEMMA 1 **(Behavior of MDS 4^{th} conditional moments)**. *Given Assumptions 1 and 2, there exists a constant $C_4 > 0$ such that the 4^{th} conditional moment of MDS $\xi_{T,t}(\mathcal{X}_g)$ almost surely satisfies*

$$\mathbb{E}\left[\xi^4_{T,t}(\mathcal{X}_g)\middle| H_{t-1}\right] \leq C_4 \cdot \frac{h^4_t(\mathcal{X}_g;H_{t-1})\mathbb{E}\left[\mathbf{1}\left[X_t \in \mathcal{X}_g\right]\sum_w \frac{\pi^4(X_t,w)}{e^3_t(X_t,w)}\middle| H_{t-1}\right]}{\left(\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[\mathbf{1}\left[X_s \in \mathcal{X}_g\right]h^2_s(\mathcal{X}_g;H_{t-1})\sum_w \frac{\pi^2(X_t,w)}{e_s(X_s,w)}\right]\right)^2}. \tag{23}$$

Together with Assumption 3(a), Lemma 1 immediately implies the decay of conditional fourth moment of $\xi_{T,t}(\mathcal{X}_g)$. We thus have its convergence.

LEMMA 2 **(Convergence of $\sum_{t\in\mathcal{T}_2}\xi_{T,t}(\mathcal{X}_g)$)**. *Given Assumptions 1,2,3(a),3(b) with* (T,G) *initiated by* $(\mathcal{T}_2,\mathcal{G})$*, the MDS* $\xi_{T,t}(\mathcal{X}_g)$ *defined in* (22) *converges:* $\sum_{t\in\mathcal{T}_2}\xi_{T,t}(\mathcal{X}_g)\xrightarrow{d}\mathcal{N}(0,1)$.

To derive the asymptotic properties of $\widehat{Q}_T(\mathcal{X}_g)$, it is helpful to decompose the error as follows:

$$\widehat{Q}_T(\mathcal{X}_g) - Q(\mathcal{X}_g) = \underbrace{\frac{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - Q(\mathcal{X}_g))}{\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_2} h^2_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - Q(\mathcal{X}_g))^2\right]}}}_{\sum_{t\in\mathcal{T}_2}\xi_{T,t}(\mathcal{X}_g)} \underbrace{\frac{\sqrt{\mathbb{E}\left[\sum_{t\in\mathcal{T}_2} h^2_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - Q(\mathcal{X}_g))^2\right]}}{\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g;H_{t-1})}}_{\text{decays by Assumption 3(c)}}. \tag{24}$$

Furthermore, Assumption 3(b) ensures the convergence of unnormalized variance of $\widehat{q}^g_T$. We thus have the consistency and asymptotic normality of $\widehat{Q}_T(\mathcal{X}_g)$. Combined with Assumption 3(d), we can characterize the asymptotic variance of $\widehat{Q}_T(\mathcal{X}_g)$ as follows.

COROLLARY 2 **(Asymptotic variance of $\widehat{Q}_T(\mathcal{X}_g)$)**. *Given Assumptions 1-3 with* (T,G) *initiated by* $(\mathcal{T}_2,\mathcal{G})$*, for each subgroup g, any of the following variances can be used to restore the asymptotic normality of* $\widehat{Q}_T(\mathcal{X}_g)$*: (i)* $\widehat{V}_T(\mathcal{X}_g) = \frac{\sum_{t\in\mathcal{T}_2}\left(h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g)\right)\right)^2}{\left(\sum_{t\in\mathcal{T}_2} h_t(\mathcal{X}_g;H_{t-1})\right)^2}$*; (ii)* $\widehat{v}_{T,g} = \frac{\sum_{t\in\mathcal{T}_2} h^2_t(\mathcal{X}_g;H_{t-1})\mathbb{E}\left[\left(\left(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g)\right)\right)^2\middle| H_{t-1}\right]}{\left(\sum_{t\in\mathcal{T}_2}\mathbb{E}[h_t(\mathcal{X}_g;H_{t-1})]\right)^2}$*; (iii)* $V_{T,g} = \frac{\mathbb{E}\left[\sum_{t\in\mathcal{T}_2}\left(h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t - \widehat{Q}_T(\mathcal{X}_g)\right)\right)^2\right]}{\left(\sum_{t\in\mathcal{T}_2}\mathbb{E}[h_t(\mathcal{X}_g;H_{t-1})]\right)^2}$.

B: Convergence of Non-contextual Weighting Estimator $\widehat{Q}_{T,N}$

The non-contextual weighted estimator $\widehat{Q}_{T,N}$ is applied to dataset $\mathcal{D}_1$ and estimates policy value Q. Its definition and estimated variance are as follows:

$$\widehat{Q}_{T,N} = \frac{\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})\widehat{\Gamma}_t}{\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})}, \quad \widehat{V}_{T,N} = \frac{\sum_{t\in\mathcal{T}_1} h^2_t(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t - \widehat{Q}_{T,N}\right)^2}{\left(\sum_{t\in\mathcal{T}_1} h_t(\mathcal{X};H_{t-1})\right)^2}. \tag{25}$$

Let its auxiliary MDS $\{\xi_{T,t}\}_{t\in\mathcal{T}_1}$ be

$$\xi_{T,t} = \frac{h_t(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t - Q)}{\sqrt{\sum_{t\in\mathcal{T}_1}\mathbb{E}\left[h^2_t(\mathcal{X};H_{t-1})(\mathcal{X}_g;H_{t-1})(\widehat{\Gamma}_t - Q)^2\right]}}. \tag{26}$$

As a spectial case of subgroup weighting by restricting the number of subgroups to one, non-contextual weighting and its MDS $\xi_{T,t}$ immediately inherits the convergence, as stated in Remark 3. Similar to Corollary 2, we characterize the asymptotic variance of $\widehat{Q}_{T,N}$ as follows.

Corollary 3 **(Asymptotic variance of $\widehat{Q}_{T,N}$).** *Under Assumptions 1-3 with* (T,G) *initiated by* $(\mathcal{T}_1,\{1\})$*, any of the following variances can be used to restore the asymptotic normality of* $\widehat{Q}_T^N$*: (i)* $\widehat{V}_{T,N}=\frac{\sum_{t\in\mathcal{T}_1}h_t^2(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t-\widehat{Q}_{T,N}\right)^2}{\left(\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X};H_{t-1})\right)^2}$*; (ii)* $\widehat{v}_{T,N}=\frac{\sum_{t\in\mathcal{T}_1}h_t^2(\mathcal{X};H_{t-1})\mathbb{E}\left[\left(\widehat{\Gamma}_t-Q\right)^2\middle|H_{t-1}\right]}{\left(\sum_{t\in\mathcal{T}_1}\mathbb{E}[h_t(\mathcal{X};H_{t-1})]\right)^2}$*; (iii)* $V_{T,N}=\frac{\mathbb{E}\left[\sum_{t\in\mathcal{T}_1}h_t^2(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t-Q\right)^2\right]}{\left(\sum_{t\in\mathcal{T}_1}\mathbb{E}[h_t(\mathcal{X};H_{t-1})]\right)^2}$.

C: Convergence of Subgroup Weighting Estimator $\widehat{Q}_{T,S}$

The subgroup weighting estimator $\widehat{Q}_{T,S}$ sums up $\widehat{q}_T^g$ to evaluate policy value Q. Its definition and estimated variance are as follows:

$$\widehat{Q}_{T,S}=\sum_{g\in\mathcal{G}}\widehat{Q}_T(\mathcal{X}_g),\quad \widehat{V}_{T,S}=\sum_{t\in\mathcal{T}_2}\left(\frac{\sum_{g\in\mathcal{G}}h_t(\mathcal{X}_g;H_{t-1})\left(\mathbb{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t-\widehat{Q}_T(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}h_s(\mathcal{X}_g;H_{t-1})}\right)^2. \tag{27}$$

Let its auxiliary MDS $\{\zeta_{T,t}\}_{t\in\mathcal{T}_2}$ be

$$\begin{gathered}\zeta_{T,t}=\frac{1}{\sqrt{V_{T,S}}}\sum_{g\in\mathcal{G}}\sqrt{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[\left(\frac{h_s(\mathcal{X}_g;H_{t-1})\left(\mathbb{1}\left[X_s\in\mathcal{X}_g\right]\widehat{\Gamma}_s-Q(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_g;H_{t-1})\right]}\right)^2\right]}\xi_{T,t}^g,\\ \text{where}\quad V_{T,S}=\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})\left(\mathbb{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t-Q(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_g;H_{t-1})\right]}\right)^2\right].\end{gathered} \tag{28}$$

Not that MDS $\zeta_{T,t}$ is a weighted sum of subgroups MDS $\xi_{T,t}(\mathcal{X}_g)$, with weights proportional to the asymptotic standard error of $\widehat{Q}_T(\mathcal{X}_g)$. The normalization term $V_{T,S}$ is the asymptotic variance of $\widehat{Q}_{T,S}$. The conditional variance convergence of $\zeta_{T,t}$ can be deduced from that of each summand $\xi_{T,t}(\mathcal{X}_g)$. To regularize the conditional fourth moment of $\zeta_{T,t}$, we have the following lemma.

Lemma 3. *By the Cauchy-Schwartz inequality, the assignment probabilities satisfy:*

$$1\leq\sum_w\frac{\pi^2(X_t,w)}{e_t(X_t,w)}\leq\sum_w\frac{\pi^3(X_t,w)}{e_t^2(X_t,w)}\leq\sum_w\frac{\pi^4(X_t,w)}{e_t^3(X_t,w)}. \tag{29}$$

Moreover, under Assumptions 1 and 2, for each subgroup g *and* $p=2,3,4$*, there exist positive constants* C_p *such that*

$$\mathbb{E}\left[\left(\widehat{\Gamma}_t-Q(\mathcal{X}_g)\right)^p\middle|H_{t-1},X_t\in\mathcal{X}_g\right]\leq C_p\cdot\mathbb{E}\left[\sum_{w\in\mathcal{W}}\frac{\pi^p(X_t,w)}{e_t^{p-1}(X_t,w)}\middle|H_{t-1},X_t\in\mathcal{X}_g\right]. \tag{30}$$

Expanding $\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\zeta_{T,t}^4\middle|H_{t-1}\right]$ in terms of $\xi_{T,t}(\mathcal{X}_g)$, we deduce the decay of $\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\zeta_{T,t}^4\middle|H_{t-1}\right]$ from those of $\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\xi_{T,t}^4(\mathcal{X}_g)\middle|H_{t-1}\right]$. Collectively, we have the convergence of $\zeta_{T,t}$.

Lemma 4 **(Convergence of $\sum_{t\in\mathcal{T}_2}\zeta_{T,t}$).** *Given Assumptions 1-3 with* (T,G) *initiated by* $(\mathcal{T}_2,\mathcal{G})$*, the MDS* $\xi_{T,t}(\mathcal{X}_g)$ *defined in* (22) *converges:* $\sum_{t\in\mathcal{T}_2}\zeta_{T,t}\xrightarrow{d}\mathcal{N}(0,1)$.

Convergence of $\sum_{t\in\mathcal{T}_2}\zeta_{T,t}$ is equivalent to the following convergence:

$$\sum_{t\in\mathcal{T}_2}\zeta_{T,t}=\frac{\sum_{t\in\mathcal{T}_2}\sum_{g\in\mathcal{G}}\frac{h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-Q(\mathcal{X}_g))}{\mathbb{E}\left[\sum_{s\in\mathcal{T}_2}h_s(\mathcal{X}_g;H_{t-1})\right]}}{\sqrt{\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-Q(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_g;H_{t-1})\right]}\right)^2\right]}}\xrightarrow{d}\mathcal{N}(0,1),$$

which appears very close to our targeted statistic

$$\frac{\widehat{Q}_{T,S}-Q}{\sqrt{\widehat{V}_{T,S}(\pi)}}=\frac{\sum_g\frac{\sum_{t\in\mathcal{T}_2}h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-\widehat{Q}_T(\mathcal{X}_g))}{\sum_{t\in\mathcal{T}_2}h_t(\mathcal{X}_g;H_{t-1})}}{\sqrt{\sum_{t\in\mathcal{T}_2}\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-\widehat{Q}_T(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}h_s(\mathcal{X}_g;H_{t-1})}\right)^2}}.$$

Indeed, after some calculation, we obtain the CLT of $\widehat{Q}_{T,S}$ stated in Remark 3. As before, we show alternatives of asymptotic variance for $\widehat{Q}_{T,S}$ in Corollary 4.

Corollary 4 **(Asymptotic variance of $\widehat{Q}_{T,S}$).** *Given Assumptions 1-3 with* (T,G) *initiated by* $(\mathcal{T}_2,\mathcal{G})$*, any of the following variances can be used to restore the asymptotic normality of* $\widehat{Q}_{T,S}$*: (i)* $\widehat{V}_{T,S}=\sum_{t\in\mathcal{T}_2}\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\,[X_s\in\mathcal{X}_g]\widehat{\Gamma}_t-\widehat{Q}_T(\mathcal{X}_g)\right)}{\sum_{s\in\mathcal{T}_2}h_s(\mathcal{X}_g;H_{t-1})}\right)^2$*; (ii)* $\widehat{v}_{T,S}=\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-Q(\mathcal{X}_g))}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_g;H_{t-1})\right]}\right)^2\middle|H_{t-1}\right]$*; (iii)* $V_{T,S}=\sum_{t\in\mathcal{T}_2}\mathbb{E}\left[\left(\sum g\in\mathcal{G}\frac{h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\,[X_t\in\mathcal{X}_g]\widehat{\Gamma}_t-Q(\mathcal{X}_g))}{\sum_{s\in\mathcal{T}_2}\mathbb{E}\left[h_s(\mathcal{X}_g;H_{t-1})\right]}\right)^2\right]$.

D: Convergence of Adaptive Subgroup Weighting Estimator $\widehat{Q}_T$

The adaptive subgroup weighting estimator $\widehat{Q}_T$ uses the entire dataset and estimates policy value Q. It combines non-contextual estimation from dataset $\mathcal{D}_1$ and subgroup estimation from dataset $\mathcal{D}_2$. Its definition and estimated variance are as follows:

$$\widehat{Q}_T=\frac{\widehat{V}_{T,S}\widehat{Q}_{T,N}+\widehat{V}_{T,N}\widehat{Q}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}},\quad\widehat{V}_T=\frac{\widehat{V}_{T,N}\widehat{V}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}.\tag{31}$$

Let its auxiliary MDS $\{\iota_{T,t}\}_{t\in\mathcal{T}}$ be

$$\iota_{T,t}=\begin{cases}\sqrt{V_{T,S}/\left(V_{T,N}+V_{T,S}\right)}\xi_{T,t} & t\in\mathcal{T}_1;\\[2ex]\sqrt{V_{T,N}/\left(V_{T,N}+V_{T,S}\right)}\zeta_{T,t} & t\in\mathcal{T}_2\,.\end{cases}\tag{32}$$

where $V_{T,N}$ and $V_{T,S}$ are the expected variance of estimators $\widehat{Q}_{T,N}$ and $\widehat{Q}_{T,S}$ respectively. The conditional variance convergence and conditional fourth moment decay of $\iota_{T,t}$ immediately follow from those of MDS $\xi_{T,t}$ and $\zeta_{T,t}$, thus invoking the martingale CLT.

Lemma 5 **(Convergence of $\sum_{t\in\mathcal{T}}\iota_{T,t}$).** *Suppose that Assumptions 1 and 2 are satisfied, and that the non-contextual evaluation weights and the subgroup evaluation weights both satisfy Assumption 3. Then, the MDS $\iota^{g}_{T,t}$ defined in* (32) *converges: $\sum_{t\in\mathcal{T}}\iota_{T,t}\xrightarrow{d}\mathcal{N}(0,1)$.*

Convergence of $\sum_{t\in\mathcal{T}}\iota_{T,t}$ is equivalent to

$$\left(\frac{V_{T,S}}{V_{T,N}+V_{T,S}}\frac{\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X};H_{t-1})(\widehat{\Gamma}_t-Q)}{\mathbb{E}\left[\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X};H_{t-1})\right]}+\frac{V_{T,N}}{V_{T,N}+V_{T,S}}\sum_g\frac{\sum_{t\in\mathcal{T}_2}h_t(\mathcal{X}_g;H_{t-1})(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t-Q(\mathcal{X}_g))}{\mathbb{E}\left[\sum_{t\in\mathcal{T}_2}h_t(\mathcal{X}_g;H_{t-1})\right]}\right)\Big/\sqrt{\frac{V_{T,N}V_{T,S}}{V_{T,N}+V_{T,S}}}\xrightarrow{d}\mathcal{N}(0,1),$$

while our targeted statistic is

$$\frac{\widehat{Q}_T-Q}{sqrt\widehat{V}_T}=\left(\frac{\widehat{V}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}\frac{\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X};H_{t-1})\left(\widehat{\Gamma}_t-Q\right)}{\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X};H_{t-1})}+\frac{\widehat{V}_{T,N}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}\sum_g\frac{\sum_{t\in\mathcal{T}_1}h_t(\mathcal{X}_g;H_{t-1})\left(\mathbf{1}\left[X_t\in\mathcal{X}_g\right]\widehat{\Gamma}_t-Q(\mathcal{X}_g)\right)}{\sum_{t\in\mathcal{T}_2}h_t(\mathcal{X}_g;H_{t-1})}\right)\Bigg/\sqrt{\frac{\widehat{V}_{T,N}\widehat{V}_{T,S}}{\widehat{V}_{T,N}+\widehat{V}_{T,S}}}.$$

These two can be bridged by the variance convergences, that is, $\frac{\widehat{V}^N_T}{V^N_T}\xrightarrow{p}1$ and $\frac{\widehat{V}^S_T}{V_{T,S}}\xrightarrow{p}1$, which are ensured by Corollaries 3 and 4. We thus have the CLT of adaptive subgroup weighting.

References

Athey S, Wager S (2017) Efficient policy learning. *arXiv preprint arXiv:1702.02896* .

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2016) Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060* .

Darling D, Robbins H (1967) Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America* 58(1):66.

Deshpande Y, Mackey L, Syrgkanis V, Taddy M (2017) Accurate inference for adaptive linear models. *arXiv preprint arXiv:1712.06695* .

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

Hall P, Heyde CC (2014) *Martingale limit theory and its application* (Academic press).

Howard SR, Ramdas A, McAuliffe J, Sekhon J (2018) Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240* .

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics* 86(1):4–29.

Jamieson K, Malloy M, Nowak R, Bubeck S (2014) lil'ucb: An optimal exploration algorithm for multi-armed bandits. *Conference on Learning Theory*, 423–439.

Lattimore T, Szepesvári C (2018) Bandit algorithms. *preprint* 28.

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

Luedtke A, Chambaz A (2017) Faster rates for policy learning. *arXiv preprint arXiv:1704.06431* .

Luedtke AR, Van Der Laan MJ (2016) Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics* 44(2):713.

Murphy SA (2003) Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2):331–355.

Nie X, Tian X, Taylor J, Zou J (2017) Why adaptively collected data have negative bias and how to correct for it. *arXiv preprint arXiv:1708.01977* .

Russo D, Van Roy B, Kazerouni A, Osband I, Wen Z (2017) A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038* .

Shin J, Ramdas A, Rinaldo A (2019) Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, 7100–7109.

Shin J, Ramdas A, Rinaldo A (2020) On conditional versus marginal bias in multi-armed bandits. *arXiv preprint arXiv:2002.08422* .

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Villar SS, Bowden J, Wason J (2015) Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics* 30(2):199.

Vogelstein JT, Powell M, Koenecke A, Xiong R, Fischer N, Huq S, Khalafallah AM, Papadopoulos N, Kinzler KW, Vogelstein B, et al. (2020) Alpha-1 adrenergic receptor antagonists prevent acute respiratory distress syndrome and death: implications for coronavirus disease 2019 [internet]. *arXiv preprint arXiv:2004.10117* .

Zhang KW, Janson L, Murphy SA (2020) Inference for batched bandits. *arXiv preprint arXiv:2002.03217* .