Beyond Central Limit Theorem for Higher-Order Inference in Batched Bandits

Yechan Park
Faculty of Economics
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
yechanparkjp@g.ecc.u-tokyo.ac.jp

Ruohan Zhan
Graduate School of Business
Stanford University
655 Knight Way, Stanford, CA 94305, USA
ruohanzhan@gmail.com

Nakahiro Yoshida
Graduate School of Mathematical Sciences
University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan
nakahiro@ms.u-tokyo.ac.jp

Abstract

Adaptive experiments have been gaining traction in a variety of domains, which stimulates a growing literature focusing on post-experimental statistical inference on data collected from such designs. Prior work constructs confidence intervals mainly based on two types of methods: (i) martingale concentration inequalities and (ii) asymptotic approximation to distribution of test statistics; this work contributes to the second kind. The current asymptotic approximation methods however mostly rely on first-order limit theorems, which can have a slow convergence in a data-poor regime. Besides, established results often rely on conditions that noises behave well, which can be problematic when the real-world instances are heavy-tailed or asymmetric. In this paper, we propose a higher-order asymptotic expansion formula for inference on adaptively collected data, which generalizes normal approximation to the distribution of standard test statistics. Our theorem relaxes assumptions on the noise distribution and benefits a higher-order approximation in the distributional distance to accommodate small sample sizes. We complement our results by promising empirical performances in simulations.

1 Introduction

Adaptive experimental designs such as bandit algorithms have received increasing popularity in many applications [13, 1, 7]. Instead of fixing a randomization rule, the experimenter progressively updates the data collection mechanism in response to past observations, so as to reduce experimental costs or optimize sample efficiency to test hypotheses. With the increasing availability of data collected from such designs, we would like to understand the best practice to conduct post-experimental statistical inference. Specifically, we seek to evaluate alternative treatment assignment policies using data collected from batched multi-armed bandit algorithms and construct confidence intervals around the

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

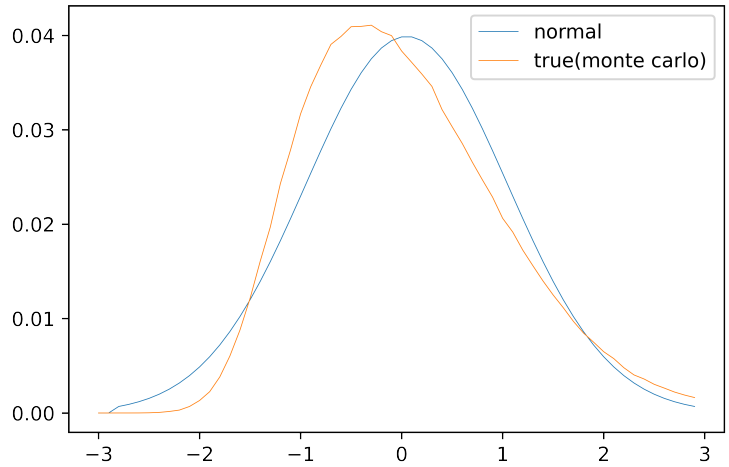

Figure 1: The Monte Carlo distribution of the test statistic based on OLS is far from being normal due to the asymmetric and heavy-tailed noise.

estimations. We would like to do so in a data-efficient and robust way and accommondate regimes where the sample size may be small, and the noise may be heavy-tailed or asymmetric.

There are two types of approaches of inference on adaptive data in the current literature. The first kind is based on martingale concentration inequalities that result in valid confidence intervals for arbitrary stopping times, but such methods are often overly conservative and lose statistical power especially when sample size is small. [11, 12, 6, 14]. Alternatively, asymptotic approximation-based methods construct estimators that have desirable limiting distributions, which are then translated to asymptotically valid confidence intervals [5, 16, 8, 17]. We view our work as complimentary to this line of research. Closest to our setting of batched multi-armed bandits, [16] first construct a studentized statistic per batch and then aggregate them to obtain an estimator with asymptotically normal guarantees, which retains even with nonstationarity.

Despite the growing research efforts, the current asymptotic approximation-based methods for post-experimental inference lack two things:

1. Robustness to distributional assumptions. Prior work has mainly relied on assumptions on sub-gaussian or symmetric noises, which is however problematic since such desirable properties may not be realized in many real-world applications. For example, treatment effect heterogeneity or truncation induces asymmetry in the noises. Thus methods predicated upon such assumptions might not yield correct statistical properties in cases of heavy-tailed or asymmetric noises.

2. Sample size considerations. The current asymptotic approximation methods for adaptive inference all rely on first-order limit theorems, a.k.a. central limit theorem. This convergence is slow in a data-poor regime, which is typical if we want to evaluate sub-optimal arms using bandit data. Bandit algorithms are often geared towards minimizing cumulative regret, which leads to meager data collection for arms that do not perform well during the first part of the experiment. This issue exacerbates as the number of arms increases.

To provide intuition, Example 1 shows that batched OLS [16] generates a test statistic that is far from being normal with limited sample size and asymmetric noise.

Example 1. *Consider a two-batch multi-armed bandit experiment run under an ϵ-greedy algorithm, where the noise is distributed as standardized $\Gamma(2, 1)$ with mean 0 and variance 1. Figure 1 presents the Monte Carlo distribution of the test statistic based on batched OLS [16], which is shown to be far from being normal due to the asymmetric and heavy-tailed noise.*

How should we tackle these problems? The confidence sequence approach has an appeal of time-uniform valid inference, but past simulation studies suggest that they may be conservative with small

sample size [16, 5]. This paper provides an alternative method: if the first-order approximation is insufficient, why not resort to higher-orders? With given sample size, our method aims to capture higher-order departure of the test statistic from being normally distributed and translate it into a form of an asymptotic expansion density. This asymptotic expansion formula formally guarantees an error bound of $o(n^{-(p-2)/2})$, where p is a user-specified integer. This asymptotic expansion is used to provide higher-order adjusted z-scores for test statistics, which should be instance-dependent with respect to parameters of the experiment such as sample size. This paper provides the first attempt on higher-order approximation in the context of adaptive experiments.

As we will see, the framework of [15] that was used for a very different purpose (Jump-diffusion process in random environment) can be applied batchwise to the context of batch-wise adaptive data collection, justified by the back propogation formula that provides precise estimate of the approximation error.

Besides, we also modify his framework to incorporate random measurable functions, possibly discontinuous, which corresponds to the assignment policy in the context of adaptive experiments.

The rest of the paper is organized as follows. Section 2 formulates the problem and provides preliminaries. The main results are provided in Section 3, where we estimate an arm value based on an asymptotic expansion density that generalizes the commonly-made normal approximation. Section 4 offers preliminary numerical experiments. Finally, Section 5 concludes.

2 Setup

2.1 Batched Bandit Framework

Suppose the agent interacts with n individuals at batch s, and the experiment has $\mathbb{S} = \{1, ..., S\}$ batches in total. At batch s, the i-th individual is identified with an element $j = (s, i_s)$. We write $\mathbb{J}^n = \{(s, i_s;\ i_s \in \mathbb{I}^n_s,\ s \in \mathbb{S}\}$, where $\mathbb{I}^n_s = \{1, ..., n_s\}$. Let $s(j) = s$ for each individual $j \in \mathbb{J}^n_s$. The batch size n is the parameter that drives the asymptotic theory we will develop.

Given a probability space $(\Omega, \mathcal{F}, P)$, for the individual $j \in \mathbb{J}^n_s$ the agent selects an action/treatment according to an arm assignment policy (to be shortly introduced), which is expressed by a $\overline{k}_s$-dimensional vector $A_j = (A_{j,k_s})_{k_s \in \mathcal{K}_s}$, where $\mathcal{K}_s$ is the action space of size k_s in batch s. Each entry A_{j,k_s} takes values in $\{1, 0\}$ and $\sum_{k_s \in \mathcal{K}_s} A_{j,k_s} = 1$.

The agent then observes a reward R_j from the action A_j, which is written as follows,

$$R_j = A_j^\star \beta_{s(j)} + \dot{\epsilon}_j \quad (j \in \mathbb{J}^n_s) \tag{1}$$

where $\beta_s \in \mathbb{R}^{\overline{k}_s}$ represents the underlying effect of the actions, $\dot{\epsilon}_j$ is an exogenous mean-zero noise that is sampled i.i.d. each time, and the star $\star$ denotes the matrix transpose[1].

For batch s, let $\mathbf{A}^n_s = (A_j;\ j \in \mathbb{J}^n_s)$ represent the action set and $\boldsymbol{\epsilon}^n_s = (\epsilon_j;\ j \in \mathbb{J}^n_s)$, where ϵ_j is an r_s-dimensional random vector, e.g., $\epsilon_j = (\dot{\epsilon}_j, \dot{\epsilon}_j^2 - \sigma^2)^\star$. For $j \in \mathbb{J}^n_s$, let W^n_j be a $\mathsf{d}_s \times \mathsf{r}_s$ random matrix measurable with respect to $\sigma[\mathbf{A}^n_{s(j)}]$ Write $\mathbf{W}^n_s = (W^n_j)_{j \in \mathbb{J}^n_s}$. Consider a weighted sum

$$\mathbb{Z}^n_s = \sum_{j \in \mathbb{J}^n_s} W^n_j \epsilon_j. \tag{2}$$

This $\mathbb{Z}^n_s$ captures the most statistical quantities of interest in estimating the value of a single arm or arm difference for batch s. For example, using data collected only from batch s, the scaled error of the OLS estimator has the following form:

$$\mathbb{Z}^n_s = \sum_{j \in \mathbb{J}^n_s} W^n_j \epsilon_j = \text{diag}((N^n_{s,1})^{-1/2} \sum_{j \in \mathbb{J}^n_s} A_{j,1}\epsilon_j, ..., (N^n_{s,\overline{k}_s})^{-1/2} \sum_{j \in \mathbb{J}^n_s} A_{j,\overline{k}_s}\epsilon_j),$$

where $N^n_{s,k}$ denotes the number of action k being selected in batch s.

[1] Extensions for considering different sets of actions and different distributions of ϵ_j for different batchs, or even incorporating contexts is straightforward though we adopted the present setting for notational simplicity.

We next model the bandit algorithm that determines the policy for each batch. Let $\mathcal{L}_s$ be a measurable space for $s \in \mathbb{S}$. We consider a $\sigma[\mathbf{A}_s^n]$-measurable random map $L_s^n : \Omega \to \mathcal{L}_s$ for every $(n, s) \in \mathbb{N} \times \mathbb{S}$. The variables L_{s-1}^n will be used for making a criterion for selection of an action in batch s. We write $\underline{L}_s^n = (L_1^n, ..., L_s^n)$ for $(L_s)_{s \in \mathbb{S}}$ [2]. In the batched bandits, how the agent updates the data collection policy c_s (the distribution of $\mathbf{A}_s^n$) from a measurable set $\mathfrak{C}_s$ is determined by the average effects of actions at batch $s-1$. We choose the policy c_s by the distribution $q\big((\underline{L}_{s-1}^n, \underline{\mathbb{Z}}_{s-1}^n), dc_s\big)$ over the action space.

Example 2 (ϵ-Greedy algorithm)**.** *The ϵ-Greedy algorithm having $\mathfrak{C}_2 = \{c_2^{(1)}, c_2^{(2)}\}$ with*

$$q\big((\underline{l}_1, \underline{z}_1), dc_2\big) = 1_{\{\mathtt{h}_1(\underline{l}_1, \underline{z}_1) \geq 0\}} \delta_{c_2^{(1)}}(dc_2) + 1_{\{\mathtt{h}_1(\underline{l}_1, \underline{z}_1) < 0\}} \delta_{c_2^{(2)}}(dc_2) \tag{3}$$

for some function $\mathtt{h}_1$.

Example 3 (Thomspon sampling)**.** *A Thompson sampling algorithm is realized as*

$$\begin{aligned} q\big((\underline{l}_1, \underline{z}_1), dc_2\big) = 1_{\{\mathtt{h}_1(\underline{l}_1, \underline{z}_1) \leq a_1\}} \delta_{c_2^{(1)}}(dc_2) + 1_{\{\mathtt{h}_1(\underline{l}_1, \underline{z}_1) > a_2\}} \delta_{c_2^{(2)}}(dc_2) \\ + 1_{\{a_1 < \mathtt{h}_1(\underline{l}_1, \underline{z}_1) \leq a_2\}} \delta_{C_2(l_1, z_1)}(dc_2) \end{aligned} \tag{4}$$

with some constants a_1, a_2 expressing the clipping constraint commonly assumed in batched bandits[16] and some $\mathfrak{C}_2$-valued function C_2 of (l_1, z_1).

Finally, we denote $\mathcal{G}_s^n = \sigma\big[L_{s'}^n, \mathbb{Z}_{s'}^n;\ s' \leq s\big]$, and let $\mathcal{L} = \Pi_{s \in \mathbb{S}} \mathcal{L}_s$ and $\mathsf{d} = \sum_{s \in \mathbb{S}} \mathsf{d}_s$. We assume the following condition holds unless stated otherwise:

$$\boldsymbol{\epsilon}_s^n \amalg \left(\mathcal{G}_{s-1}^n \vee \sigma[\mathbf{A}_s^n]\right) \quad (s \in \mathbb{S}). \tag{5}$$

For notational simplicity, we denote the conditional expectation $E[\,\cdot\,|\mathcal{G}_{s-1}^n]$ as $E_{s-1}[\,\cdot\,]$ and the conditional expectation $E\big[\ \cdot\ |\mathcal{G}_{s-1}^n \vee \sigma[\mathbf{A}_s^n]\big]$ as $E_{s-1, \mathbf{A}_s^n}[\,\cdot\,]$. For $\mathbf{w}_s^n = (w_j^n)_{j \in \mathbb{J}_s^n} \in \mathbb{R}^{\mathsf{d}_s \mathsf{r}_s n_s}$, $P_s^n(\mathbf{w}_s^n, dz_s) = P^{\sum_{j \in \mathbb{J}_s^n} w_j^n \epsilon_j}(dz_s)$ represents the distribution of $\sum_{j \in \mathbb{J}_s^n} w_j^n \epsilon_j$ [3] We assume a representation of a regular conditional distribution of $(L_s^n, \mathbf{W}_s^n)$ given $\mathcal{G}_{s-1}^n$ as

$$P^{(L_s^n, \mathbf{W}_s^n)}(dl_s, d\mathbf{w}_s^n | \mathcal{G}_{s-1}^n) = \int_{\mathfrak{C}_s} q\big((\underline{L}_{s-1}^n, \underline{\mathbb{Z}}_{s-1}^n), dc_s\big) \nu_{c_s}^n(dl_s, d\mathbf{w}_s^n) \tag{6}$$

for a probability distribution $\nu_{c_s}^n = \eta_{c_s}^{(L_s^n, \mathbf{W}_s^n)}$ of $(L_s^n, \mathbf{W}_s^n)$ given c_s.

2.2 Off-policy Evaluation with Backwards Induction

When the experiment finishes, we want to conduct hypothesis testing like the value of the arm $\beta_{s(j)}$ for scientific discovery. We can generically define hypothesis testing problems mathematically as follows. Given a measurable function $f : \mathcal{L} \times \mathbb{R}^{\mathsf{m}} \to \mathbb{R}$, where $\mathsf{m} = \sum_{s \in \mathbb{S}} \mathsf{m}_s$, our estimand is the value of the expectation $\overline{\mathbf{E}}_n = E\big[f\big(\underline{L}_S^n, \underline{\mathbb{Y}}_S^n\big)\big]$, where $\mathbb{Y}_s^n = Y_s^n(L_s^n, \mathbb{Z}_s^n)$ is an m_s-dimensional random variable. For standard hypotheses and confidence intervals, f is usually an indicator function for whether the test statistic is smaller than some particular threshold.

If we know the true underlying density $P_s^n(\mathbf{w}_s^n, dz_s)$, we can iteratively calculate the desired expectation from a simple exercise analogous to dynamic programming, with a slight modification with a truncation $1_{\{L_s^n \in \mathcal{A}_s^n\}}$, where $\mathcal{A}_s^n$ is a measurable set of $\mathcal{L}_s$. If the clipping constraint holds, $\mathbf{E}_n$ almost equals to $\mathbf{E}_n := E\left[f\big(\underline{L}_S^n, \underline{\mathbb{Y}}_S^n\big) \prod_{s=1}^S 1_{\{L_s^n \in \mathcal{A}_s^n\}}\right]$ [4]. Theorem1 formally proves this with the aid of the following lemma.

Lemma 1. *For $s \in \mathbb{S}$,*

[2] This operation by the underline $\underline{\cdot}_s$ will apply to other vectors

[3] In other words, $P_s^n(\mathbf{W}_s^n, dz_s)$ is a regular conditional probability of $\mathbb{Z}_s^n$

[4] The choice of $\mathcal{A}_s^n$ is important so that it will determine the accuracy of approximation $\mathbf{E}_n$ to $\overline{\mathbf{E}}_n$, but the clipping condition that is placed suffices for this condition. Intuitively, when there is too little sample size, there is no possibility of asymptotics.

$$E_{s-1}\left[f_s^n\big(\underline{L}_s^n,\,\underline{\mathbb{Z}}_s^n\big)\prod_{s'=1}^{s}1_{\{L_{s'}^n\in\mathcal{A}_{s'}^n\}}\right]=f_{s-1}^n(\underline{L}_{s-1}^n,\,\underline{\mathbb{Z}}_{s-1}^n)\prod_{s'=1}^{s-1}1_{\{L_{s'}^n\in\mathcal{A}_{s'}^n\}}\quad a.s. \tag{7}$$

Here the product $\prod_{s'=1}^{0}$ *reads* 1.

Theorem 1. *Suppose that* $f\big(\underline{L}_S^n,\underline{\mathbb{Z}}_S^n\big)$ *is integrable. Then*

(a) $\mathbf{E}_n \;=\; f_0^n$, *where* $f_0^n := \int f_1^n\big(l_1,\,z_1\big)P_1^n(\mathbf{w}_1^n,dz_1)1_{\{l_1\in\mathcal{A}_1^n\}}\nu_{c_1}^n(dl_1,d\mathbf{w}_1^n)$,

$f_S^n(\underline{l}_S,\underline{z}_S)=f\big(\underline{l}_S,(Y_s^n(l_s,z_s))_{s\le S}\big)$,

$f_{s-1}^n(\underline{l}_{s-1},\underline{z}_{s-1})$
$=\int_{\mathfrak{C}_s}\int_{\mathcal{L}_s\times\mathbb{R}^{\mathsf{d}_s}}f_s^n\big(\underline{l}_{s-1},l_t,\underline{z}_{s-1},z_s\big)P_s^n(\mathbf{w}_s^n,dz_s)1_{\{l_s\in\mathcal{A}_s^n\}}\nu_{c_s}^n(dl_s,d\mathbf{w}_s^n)$
$\times\, q\big((\underline{l}_{s-1},\underline{z}_{s-1}),dc_s\big)$

for $s\in\mathbb{S}$, $s\ge 2$.

(b) $\overline{\mathbf{E}}_n \;=\; E\big[f\big(\underline{L}_S^n,\underline{\mathbb{Y}}_S^n\big)\big] \;=\; \mathbf{E}_n+\rho_n$ *with* $|\rho_n|\le E\Big[\big|f\big(\underline{L}_S^n,\underline{\mathbb{Z}}_S^n\big)\big|\sum_{s\in\mathbb{S}}1_{\{L_s^n\notin\mathcal{A}_s^n\}}\Big]$.

Theorem 1 on its own does not have much practical relevance, since the true distribution $P_s^n(\mathbf{w}_s^n,dz_s)$ is unknown. Therefore past research on off policy evaluation have aimed to approximate it by some random signed measure $\Psi_{s,p,\mathbf{w}_s^n}^n$ on $\mathbb{R}^d$depending on $\mathbf{w}_s^n$. All past attempts on asymptotic approximation claimed that the gaussian density was sufficient. We claimed that can be problematic especially when the sample size is small and the noise does not behave well.

3 Main Result: Asymptotic Expansion for Batched Bandits

To assess the accuracy of the approximation, we need to estimate the difference between f_0^n and $\widehat{f}_0^n$, where $\widehat{f}_0^n$ is the approximation to the f_0^n by replacing $P_s^n(\mathbf{w}_s^n,dz_s)$ with $\Psi_{s,p,\mathbf{w}_s^n}^n$ for each batch. More precisely, we define $\widehat{f}_s^n(\underline{l}_s,\underline{z}_s)$ by $\widehat{f}_S^n=f_S^n$ and

$$\begin{aligned}\widehat{f}_{s-1}^n(\underline{l}_{s-1},\underline{z}_{s-1}) \;=\; &\int_{\mathfrak{C}_s}\int_{\mathcal{L}_s\times\mathbb{R}^{\mathsf{d}_s}}\widehat{f}_s^n\big(\underline{l}_{s-1},l_t,\underline{z}_{s-1},z_s\big)\Psi_{s,\mathsf{p},\mathbf{w}_s^n}^n(dz_s)1_{\{l_s\in\mathcal{A}_s^n\}}\nu_{c_s}^n(dl_s,d\mathbf{w}_s^n)\\ &\times q\big((\underline{l}_{s-1},\underline{z}_{s-1}),dc_s\big)\end{aligned} \tag{8}$$

for $s\in\mathbb{S}$, if the integral (8) exists.

We shall only present the results with assumptions stated verbally due to the page limit and we refer the interested reader to [10] for directions for the full version. The conditions we assume include (i) conditional i.i.d. noise ϵ that hold in most sequential experiments, (ii) regularity conditions such as measurability or existence of moments, and (iii) clipping conditions that are usually implicitly or explicitly assumed in the past literature [15]. The only condition that uniquely arises in the higher-order case is the well-known Cramér condition as follows, for any batch $s\in\mathbb{S}$, there exists a constant B_0 such that

$$\sup_{u\in\mathbb{R}^{d_s}:|u|\ge B_0}|E_{\mathcal{C}^s}[\exp(iu\cdot\epsilon_{(s,1)})]|<1,\qquad \mathcal{C}^s=\mathcal{G}_{s-1}^{\infty}\vee\sigma[\mathbf{A}_s^{\infty}],$$

which places regularity on the characteristic function of the noise distribution and validates the Fourier inversion to get the asymptotic expansion density. This condition generally holds except for lattice distributions that we leave for future work.

Theorem 2 (Informal)**.** *Suppose the conditions stated above hold. There exists a constant* M^* *such that* $\mathbf{E}_n=f_0^n$ *admits the following estimates:*

$$|f_0^n-\widehat{f}_0^n| \;\le\; M^*\Big(n^{-(\mathsf{p}-2+\delta_1)/2}\sum_{s\in\mathbb{S}}\prod_{i=2}^{s}U_i^n\Big), \tag{9}$$

for all $f \in \widehat{\mathcal{D}}(\mathbf{M}, \gamma)$, a class of measurable functions with at most polynomial growth, where

$$U_{s+1}^n = \int (1+|z_s|^\gamma)|\Psi_{s,\mathbf{p},\mathbf{w}_s^n}^n|(dz_s)1_{\{l_s \in \mathcal{A}_s^n\}}\nu_{c_s}^n(dl_s, d\mathbf{w}_s^n)q\big((\underline{l}_{s-1}, \underline{z}_{s-1}), dc_s\big) \quad (10)$$

for $s = 1, ..., S-1$.

Remark 1. *Theorem 2 suggests using $\widehat{f}_0^n$ to approximate $\mathbf{E}_n$. Usually U_{s+1}^n is uniformly bounded in n or grows as slowly as to be controlled by the factor $n^{-\delta_1/2}$ in the error bound of (9).*

We also present an explicit form of the asymptotic expansion density that can be used to calculate the higher-order corrected Z-score for BOLS estimator [16] up to order p (which is a user-specified integer) when the variance is known. We present the case when variance is estimated in [10].

Proposition 1. *The function $d\Psi_{n_s,p,\mathcal{C}^s}/dz(z)$ ($z = (z_1, ..., z_{\overline{k}_s})$) is*

$$\prod_{k \in \mathcal{K}_s} \phi(z_k; \lambda_{n_s,2,\mathcal{C}^s,k}) \times \prod_{k \in \mathcal{K}_s} \Bigg\{1 + \frac{1}{6} n_s^{-1/2} \lambda_{n_s,3,\mathcal{C}^s,k} h_3(z_k; \lambda_{n_s,2,\mathcal{C}^s,k})$$
$$+ n_s^{-1}\left(\frac{1}{24}\lambda_{n,k,4} h_4(z_k; \lambda_{n,k,2}) + \frac{1}{72}\lambda_{n_s,k,3}^2 h_6(z_k; \lambda_{n_s,k,2})\right) + \cdots\Bigg\}, \quad (11)$$

with the summation taken up to order $n_s^{-(p-2)/2}$, where $\lambda_{n_s,r,\mathcal{C}^s,k} = n_s^{(r-2)/2} \cdot (N_{s,k}^n/n_s)^{-(r-2)/2}\, \kappa_{r,\mathcal{C}^s}(\epsilon_{(s,1)})$ with the r-th order $\mathcal{C}^s$conditional cumulant $\kappa_{r,\mathcal{C}^s}(\epsilon_{(s,1)})$ of $\epsilon_{(s,1)}$, and $h_r(z;\Sigma) = (-1)^r \phi(z;0,\Sigma)^{-1}\partial_z^r \phi(z;0,\Sigma)$ is the r-th order Hermite polynomial.

Remark 2. *Note that all elements of this density are known and calculable for batched bandits. Even if we do not know the distribution and hence the moments of $\epsilon_{(s,1)}$, we derived an asymptotic expansion that is valid up to the first order, by plugging certain estimators in the moments.*

4 Simulations

Due to space restrictions, we only provide the numerical results for the setting described in Example 1. As Figure 2 shows, asymptotic expansion has clear gains over the normal distribution in approximating the true underlying density. While the second order asymptotic expansion seems to have better approximation in most regions of the support, the difference may not be that large. When increasing the sample size for each batch from 30 to 50 as in Figure 3, second order expansion seems to perform better. While a more extensive numerical experiments is necessary to fully assess the marginal benefit, asymptotic expansion methods seems to be better in general than normal approximation based methods.

5 Discussions

In this paper, we have introduced, developed, and justified the first theory for higher-order statistical inference specialized for data collected by batched bandit algorithms. Interesting future work includes

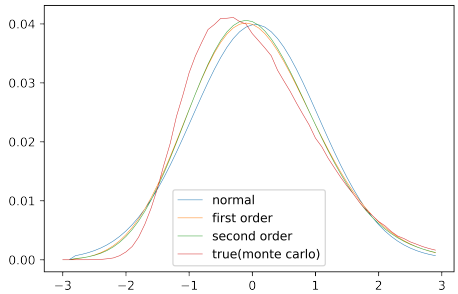

Figure 2: Higher-order asymptotic expansion provides better finite-sample approximation of the test statistics as compared to normal distribution.

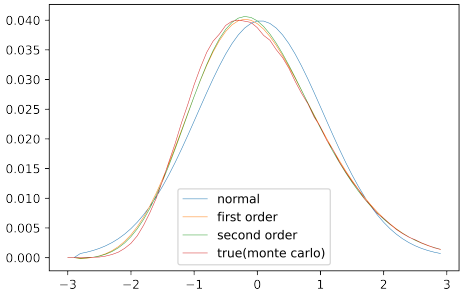

Figure 3: Same setting as Figure 2 but increasing the sample size for each batch from 30 to 50. Higher-order asymptotic expansion provides good performance even for relatively large sample size.

extending our higher-order theory to non-batched case and doing power analysis based on higher-order theory. Having better type-I error control and powered experiments thanks to higher-order inference has a potential to extend the limits on the experimental design, enabling more aggressive algorithms or reducing the sample sizes necessary.

Acknowledgements.

This work was in part supported by Japan Science and Technology Agency CREST JPMJCR2115; Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research No. 17H01702 (Scientific Research); and by a Cooperative Research Program of the Institute of Statistical Mathematics. The authors thank Professor Kengo Kamatani for his valuable comments on numerical computations.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[2] W. Brannath, G. Gutjahr, and P. Bauer. Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107(498):824–832, 2012.

[3] Y. Deshpande, L. Mackey, V. Syrgkanis, and M. Taddy. Accurate inference for adaptive linear models. *PMLR*, 2018.

[4] M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[5] V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. Confidence intervals for policy evaluation in adaptive experiments. *PNAS*, 2021.

[6] S. R. Howard, A. Ramdas, J. Mcauliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic, confidence sequences. *Annals of Statistics*, 2021.

[7] M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. 2020.

[8] G. Lewis and V. Syrgkanis. Double/debiased machine learning for dynamic treatment effects via g-estimation. *arXiv preprint arXiv:2002.07285*, 2020.

[9] A. R. Luedtke and M. J. v. d. Laan. Parametric-rate inference for one-sided differentiable parameters. *Journal of the American Statistical Association*, 113(522):780–788, 2018.

[10] Y. Park and N. Yoshida. Asymptotic expansion for batched bandits. *to be submitted to arXiv*, 2022.

[11] A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

[12] H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[13] M. J. Van Der Laan and S. D. Lendle. Online targeted learning. 2014.

[14] I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

[15] N. Yoshida. Partial mixing and Edgeworth expansion. *Probability Theory and Related Fields*, 129(4):559–624, 2004.

[16] K. W. Zhang, L. Janson, and S. A. Murphy. Inference for batched bandits. *arXiv preprint arXiv:2002.03217*, 2020.

[17] K. W. Zhang, L. Janson, and S. A. Murphy. Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*, 2022.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] In the main section, we assume variance known for the asymptotic expansion densities, so our method should calculate the formula exactly.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]