Minimax Off-Policy Evaluation for Multi-Armed Bandits

Cong Ma, Banghua Zhu, *Student Member, IEEE*, Jiantao Jiao, *Member, IEEE*, and Martin J. Wainwright, *Senior Member, IEEE*

***Abstract*— We study the problem of off-policy evaluation in the multi-armed bandit model with bounded rewards, and develop minimax rate-optimal procedures under three settings. First, when the behavior policy is known, we show that the Switch estimator, a method that alternates between the plug-in and importance sampling estimators, is minimax rate-optimal for all sample sizes. Second, when the behavior policy is unknown, we analyze performance in terms of the competitive ratio, thereby revealing a fundamental gap between the settings of known and unknown behavior policies. When the behavior policy is unknown, any estimator must have mean-squared error larger— relative to the oracle estimator equipped with the knowledge of the behavior policy— by a multiplicative factor proportional to the support size of the target policy. Moreover, we demonstrate that the plug-in approach achieves this worst-case competitive ratio up to a logarithmic factor. Third, we initiate the study of the partial knowledge setting in which it is assumed that the minimum probability taken by the behavior policy is known. We show that the plug-in estimator is optimal for relatively large values of the minimum probability, but is sub-optimal when the minimum probability is low. In order to remedy this gap, we propose a new estimator based on approximation by Chebyshev polynomials that provably achieves the optimal estimation error. Numerical experiments on both simulated and real data corroborate our theoretical findings.**

***Index Terms*— Off-policy evaluation, multi-armed bandits, minimax optimality, importance sampling.**

I. Introduction

VARIOUS forms of sequential decision-making, including multi-armed bandits [1], contextual bandits [2], [3], and Markov decision processes [4], [5], are characterized in terms of policies that prescribe actions to be taken. A central problem in all of these settings is that of policy evaluation—that is, estimating the performance of a given target policy. As a concrete example, given a new policy for deciding between treatments for cancer patients, one would be interested in assessing the effect on mortality when it is applied to a certain population of patients.

Perhaps the most natural idea is to deploy the target policy in an actual system, thereby collecting a dataset of samples, and use them to construct an estimate of the performance. Such an approach is known as on-policy evaluation, since the policy is evaluated using data that were collected under the same target policy. However, on-policy evaluation may not be feasible; in certain applications, it can be costly, dangerous and/or unethical, such as in clinical trials and autonomous driving. In light of these concerns, a plausible work-around is to evaluate the target policy using historical data collected under a different behavior policy; doing so obviates the need for any further interactions with the real environment. This alternative approach is known as *off-policy evaluation*, or OPE for short. Methods for off-policy evaluation have various applications, among them news recommendation [6], online advertising [7], robotics [8], to name just a few. Although OPE is appealing in not requiring collection of additional data, it also presents statistical challenges, in that the target policy to be evaluated is usually different from the behavioral policy that generates the data.

A. Gaps in Current Statistical Understanding of OPE

Recent years have witnessed considerable progress in the development and analysis of methods for OPE. Nonetheless, there remain a number of salient gaps in our current statistical understanding of off-policy evaluation, and these gaps motivate our work.

1) Non-Asymptotic Analysis of OPE: The classical analysis of OPE relies upon asymptotics in which the size of historical dataset, call it n, increases to infinity with all other aspects of the problem set-up held fixed. Such analysis shows that a simple plug-in estimator, to be described in the sequel, is asymptotically efficient for the OPE problem in certain settings [9]. However, such classical analysis fails to capture the modern practice of OPE, in which the sample size n may be of the same order as other problem parameters, such as the number of actions k. Thus, it is of considerable interest to obtain non-asymptotic guarantees on the performance of different methods, along with explicit dependence on different problem parameters. Li *et al.* [10] and Wang *et al.* [11] went

Manuscript received 15 June 2021; revised 11 January 2022; accepted 8 March 2022. Date of publication 25 March 2022; date of current version 13 July 2022. The work of Cong Ma and Martin J. Wainwright was supported in part by NSF under Grant DMS-2015454 and in part by the Office of Naval Research under Grant DOD-ONR-N00014-18-1-2640. The work of Banghua Zhu and Jiantao Jiao was supported in part by NSF under Grant IIS-1901252 and Grant CCF-1909499. *(Corresponding author: Cong Ma.)*

Cong Ma is with the Department of Statistics, The University of Chicago, Chicago, IL 60637 USA (e-mail: congma2015@gmail.com).

Banghua Zhu is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley (UC Berkeley), Berkeley, CA 94720 USA.

Jiantao Jiao and Martin J. Wainwright are with the Department of Electrical Engineering and Computer Sciences and the Department of Statistics, University of California at Berkeley (UC Berkeley), Berkeley, CA 94720 USA.

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3162335.

Digital Object Identifier 10.1109/TIT.2022.3162335

beyond the asymptotic setting and studied the OPE problem for multi-armed bandits and contextual bandits, respectively, from a non-asymptotic perspective. However, as we discuss in the sequel, their analyses and results are applicable only when the sample size n is sufficiently large. In this large sample regime, a number of estimators, including the plug-in, importance sampling and Switch estimators to be discussed in this paper, are all minimax rate-optimal. Thus, analysis of this type falls short of differentiating between different estimators. In particular, are they all rate-optimal for the full range of sample sizes, or is one estimator better than others?

2) Known vs. Unknown Behavior Policies: In practice, the behavior policy generating the historical data might be known or unknown to the statistician, depending on the application at hand. This difference in available knowledge raises a natural question: is there any fundamental difference between OPE problems with known or unknown behavior policies? This question, though natural, appears to have been less explored in the literature. As we noted above from an asymptotic point of view, the plug-in estimator—which requires no information about the behavior policy—is optimal. In other words, asymptotically speaking, knowing the behavior policy brings no extra benefits to solving the OPE problem. Does this remarkable property continue to hold in the finite sample setting?

3) OPE With Partial Knowledge of the Behavior Policy: The known and unknown cases form two extremes of a continuum: in practice, one often has partial knowledge about the behavior policy. For instance, one might have a rough idea on how well the behavior policy covers/approximates the target policy, as measured in terms of likelihood ratios defined by the two policies. Alternatively, there might be a guarantee on the overall exploration level of the behavior policy, as measured by the minimum probability of observing each state/action under the behavior policy. How does such extra knowledge alter the statistical nature of the OPE problem? Can one develop estimators that fully exploit this information and yield improvements over the case of a fully unknown behavior policy?

B. Contributions and Organizations

In this paper, we focus on the off-policy evaluation problem under the *multi-armed bandit* model with bounded rewards. This setting, while seemingly simple, is rich enough to reveal some non-trivial issues in developing optimal methods for OPE.

More concretely, consider a bandit model with a total of k possible actions to take, also known as arms. Any (possibly randomized) policy π can be thought of as a probability distribution over the action space $[k] := \{1, 2, \ldots, k\}$. Given a target policy π_{t} and a collection of action-reward pairs $\{(A_i, R_i)\}_{i=1}^n$ generated i.i.d. from the behavior policy π_{b} and the reward distributions $\{f(\cdot \mid a)\}_{a\in[k]}$, the goal of OPE is to estimate the value function $V_f(\pi_{\mathsf{t}})$ of the target policy π_{t}, given by

$$V_f(\pi_{\mathsf{t}}) := \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a).$$

Here the quantity $r_f(a) := \mathbb{E}_{R\sim f(\cdot|a)}[R]$ denotes the mean reward of the arm a. Our goal is to provide a sharp non-asymptotic characterization of the statistical limits of the OPE problem in three different settings: (i) when the behavior policy π_{b} is known; (ii) when π_{b} is unknown; and (iii) when we have partial knowledge about π_{b}. Along the way, we also develop computationally efficient procedures that achieve the minimax rates, up to a universal constant, for all sample sizes. The detailed statements of our main results are deferred to Section III, but let us highlight here our contributions that we make in each of the three settings.

1) Known Behavior Policy: First, when the behavior policy π_{b} is known to the statistician, we sharply characterize the minimax risk of estimating the target value function $V_f(\pi_{\mathsf{t}})$ in Theorem 1. Notably, this bound holds for all sample sizes, in contrast to previous statistical analysis of OPE, which are either asymptotic or valid only when the sample size is sufficiently large. In addition, we show in Proposition 1 that the so-called Switch estimator achieves this optimal risk. The family of Switch estimators interpolate between two base estimators: a direct method based on the plug-in principle applied to actions in some set S, and an importance sampling estimate applied to its complement S^c. Our theory identifies a simple convex program that specifies the optimal choice of subset: solving this program specifies a threshold level of the likelihood ratio at which to switch between the two base estimators. We prove that this choice yields a minimax-optimal estimator, one that reduces the variance of the importance sampling estimator alone.

2) Unknown Behavior Policy: Moving onto the case when the behavior policy π_{b} is completely unknown, we first argue that the global minimax risk is no longer a sensible criterion to measure the performance of different estimators. Instead, we propose a different metric, namely the *minimax competitive ratio*, that measures the performance of an estimator against the best achievable via an oracle—in this setting, an oracle with the knowledge of the behavior policy. With this new metric in place, we uncover a fundamental statistical gap between the known and unknown behavior policy cases in Theorem 3. More specifically, when evaluating a target policy π_{t} that can take at most s actions (for some $s \in \{1, 2, \ldots, k\}$), any estimator without the knowledge of the behavior policy must pay multiplicative factor of s (modulo a log factor) compared to the oracle Switch estimator given knowledge of the behavior policy. We further demonstrate that the plug-in estimator alone achieves this optimal worst-case competitive ratio (up to a log factor), illustrating its near-optimality in the unknown π_{b} case (cf. Theorem 2).

3) Partially Known Behavior Policy: In the third part of the paper, we initiate the study of the middle ground between the previous two extreme cases: what if we have some partial knowledge regarding the behavior policy? More concretely, we assume the knowledge of the minimum probability $\min_{a\in[k]} \pi_{\mathsf{b}}(a)$ that is taken by the behavior policy in Section III-C. Under such circumstance, we first show that the plug-in estimator is sub-optimal when the behavior policy is less exploratory—that is, in the regime $\min_{a\in[k]} \pi_{\mathsf{b}}(a) \ll (\log k)/n$. We then propose a new estimator based on

approximation by Chebyshev polynomials and show that it is optimal in estimating a large family of target policies. It is worth pointing out that this optimality is established under a different but closely related Poisson sampling model—instead of the usual multinomial sampling one—with the benefit of simplifying the analysis.

C. Related Work

Off-policy evaluation has been extensively studied in the past decades and by now there has been an immense body of literature on this topic. Here we limit ourselves to discussion of work directly related to the current paper.

1) Various Estimators for OPE: There exist two classical approaches to the OPE problem. The first is a direct method based on the plug-in principle: it estimates the value of the target policy using the reward and/or the transition dynamics estimated from the data. In the multi-armed bandit setting, the direct method uses the data to estimate the mean rewards, and plugs these estimates into the expression for the target value function. The other approach is based on importance sampling [12], also known as inverse propensity scoring (IPS) in the causal inference literature. It reweights the observed rewards according to the likelihood ratios between the target and the behavior policies. Both methods are widely used in practice; we refer interested readers to the recent empirical study [13] for various forms of these estimators. A number of authors [11], [14] have proposed hybrid estimators that involve a combination of these two approaches, a line of work that inspired our analysis of the Switch estimator. In this context, our novel contribution is to specify a particular set for switching between the two estimators, and showing that the resulting Switch estimator is minimax-optimal for any sample size.

2) Statistical Analysis of OPE: Statistical analysis of OPE can be separated into two categories: asymptotic and non-asymptotic. On one hand, the asymptotic properties of the OPE estimators are quite well-understood, with plug-in methods known to be asymptotically efficient [9], and asymptotically minimax optimal in multi-armed bandits [10]. Moving beyond bandits, a Cramér-Rao lower bound was recently provided for tabular Markov decision processes [15], and approaches based on the plug-in principle were shown to approach this limit asymptotically [16], [17].

Relative to such asymptotic analysis, there are fewer non-asymptotic guarantees for OPE; of particular relevance are the two papers [10], [11]. Li *et al.* [10] also studied the OPE problem under the multi-armed bandit model, but under different assumptions on the reward distributions than this paper. They proved a minimax lower bound that holds when the sample size is large enough, but did not give matching upper bounds in this regime. Wang *et al.* [11] extended this line of analysis to the contextual bandit setting with uncountably many contexts. They provided matching upper and lower bounds, but again ones that only hold when the sample size is sufficiently large. Notably, in this large sample regime and under the bounded reward condition of this paper, all three estimators (plug-in, importance sampling and Switch) are minimax optimal up to constant factors. Thus, restricting attention to this particular regime fails to uncover the benefits of the Switch estimator. This paper provides a complete picture of the non-asymptotic behavior of these estimators for the OPE problem, showing that only the Switch estimator is minimax-optimal for all sample sizes.

3) Estimation of Nonsmooth Functionals via Function Approximation: The OPE problem with an unknown behavior policy is intimately connected to the problem of estimating nonsmooth functionals. Portions of our analysis and the development of the Chebyshev estimator exploit this connection. The use of function approximation in functional estimation was pioneered by Ibragimov *et al.* [18], and was later generalized to nonsmooth functionals by Lepski *et al.* [19] and Cai and Low [20]. The underlying techniques have been used to devise optimal estimators for a variety of nonsmooth functionals, including Shannon entropy [21]–[23], KL divergence [24], support size [25]–[27], among others. Our development of the Chebyshev estimator is largely inspired by the paper [27] on estimating the support size, which can be viewed as a special case of OPE.

4) Notation: For the reader's convenience, let us summarize the notation used throughout the remainder of the paper. We reserve boldfaced symbols for vectors. For instance, the symbol $\mathbf{0}$ denotes the all-zeros vector, whose dimension can be inferred from the context. For a positive integer k, we refer to $[k]$ as the set $\{1, 2, \ldots, k\}$. For a finite set S, we use $|S|$ to denote its cardinality. We denote by $\mathbb{1}\{\mathcal{E}\}$ the indicator of the event $\mathcal{E}$. For any distribution μ on $\mathbb{R}$, we denote by $\mathrm{supp}(\mu)$ its support. For any distribution π on $[k]$ and any subset $S \subseteq [k]$, we define $\pi(S) := \sum_{a \in S} \pi(a)$. We follow the convention that $0/0 = 0$.

II. Background and Problem Formulation

In this section, we introduce the multi-armed bandit model with stochastic rewards, and then formally define the off-policy evaluation (OPE) problem in this bandit setting. We also introduce two existing estimators—the plug-in and the importance sampling estimators—for the OPE problem.

A. Multi-Armed Bandits and Value Functions

A multi-armed bandit (MAB) model is specified by an action space $\mathcal{A}$ and a collection of reward distributions $f := \{f(\cdot \mid a)\}_{a \in \mathcal{A}}$, where $f(\cdot \mid a)$ is the reward distribution associated with the action or arm a. Throughout the paper, we focus on the MAB model with k possible actions, and we index the action space $\mathcal{A}$ by $[k] = \{1, 2, \ldots, k\}$. In addition, we assume that the collection of reward distributions f belongs to the family of distributions with bounded support—that is,

$$\mathcal{F}(r_{\max}) := \{f \,|\, \mathrm{supp}(f(\cdot \mid a)) \subseteq [0, r_{\max}] \text{ for each } a \in [k]\}. \tag{1}$$

When the maximum reward $r_{\max}$ is understood from the context, we adopt the shorthand $\mathcal{F}$ for this class of distributions.

A (randomized) policy π is simply a distribution over the action space $[k]$, where $\pi(a)$ specifies the probability of

selecting the action a. Correspondingly, we can define the value function $V_f(\pi)$ of the policy π to be

$$V_f(\pi) := \sum_{a\in[k]} \pi(a) r_f(a), \tag{2}$$

where $r_f(a) := \mathbb{E}_f[R \mid A = a]$ denotes the mean reward under f given that action a is taken. Here R denotes a reward random variable distributed according to $f(\cdot \mid a)$.

B. Observation Model and Off-Policy Evaluation

Suppose that we have collected a collection of pairs $\{(A_i, R_i)\}_{i=1}^n$, where the action A_i is randomly drawn from the behavior policy π_{b}, whereas the reward R_i is distributed according to the reward distribution $f(\cdot \mid A_i)$. Given a target policy π_{t}, the goal of *off-policy evaluation* (OPE) is to evaluate the value function of the target policy, given by

$$V_f(\pi_{\mathsf{t}}) = \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a). \tag{3}$$

Note that this problem is non-trivial because the data $\{(A_i, R_i)\}_{i=1}^n$ is collected under the behavior policy π_{b}, which is typically distinct from the target policy π_{t}.

C. Plug-In and Importance Sampling Estimators

A variety of estimators have been designed to estimate the value function $V_f(\pi_{\mathsf{t}})$. Here we introduce two important ones most relevant to our development, namely the plug-in estimator and the importance sampling estimator. We note that in some of the literature, the plug-in estimator is also known as the regression estimator.

1) Plug-In Estimator: Perhaps the simplest method is based on applying the usual plug-in principle. Observe that the only unknown quantities in the definition (3) of the value function are the mean rewards $\{r_f(a)\}$. These unknown quantities can be estimated by their empirical counterparts

$$\widehat{r}(a) := \begin{cases} \frac{1}{n(a)} \sum_{i=1}^n R_i \mathbb{1}\{A_i = a\}, & \text{if } n(a) \geq 1, \text{ and} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $n(a) := \sum_{i=1}^n \mathbb{1}\{A_i = a\}$ denotes the number of times that action a is observed in the data set. Substituting these empirical estimates into the definition of the value function yields the plug-in estimator

$$\widehat{V}_{\mathsf{plug}} := \sum_{a\in[k]} \pi_{\mathsf{t}}(a) \widehat{r}(a), \tag{5}$$

Observe that this estimator is fully agnostic to the behavior policy. Thus, it can also be used when the behavior policy π_{b} is unknown, a setting that we also study in the sequel.

2) Importance Sampling Estimator: An alternative estimator, one which does require knowledge of the behavior policy, is based on the idea of importance sampling. More precisely, let $\rho(a) := \pi_{\mathsf{t}}(a)/\pi_{\mathsf{b}}(a)$ denote the likelihood ratio associated with the action a. The importance sampling (IS) estimator is given by

$$\widehat{V}_{\mathsf{IS}} := \frac{1}{n} \sum_{i=1}^n \rho(A_i) R_i. \tag{6}$$

In words, it weighs the observed reward R_i based on the corresponding likelihood ratio $\rho(A_i)$. As long as $\rho(a) < \infty$ for all $a \in [k]$, the importance sampling estimator $\widehat{V}_{\mathsf{IS}}$ is an unbiased estimate of $V_f(\pi_{\mathsf{t}})$. Note that the IS estimate relies on knowledge of the behavior policy π_{b} via its use of the likelihood ratio.

It is worth mentioning that the plug-in estimator can also be viewed as the importance sampling estimator in which the weights are estimated from the data.

III. Main Results

We now move onto the main results of this paper. We begin in Section III-A with results in the case when the behavior policy π_{b} is known *a priori*. In Section III-B, we provide guarantees when the behavior policy is completely unknown, whereas Section III-C is devoted to the setting where certain partial knowledge about the behavior policy, say the minimum value $\min_{a\in[k]} \pi_{\mathsf{b}}(a)$, is known.

A. Switch Estimator With Known π_{b}

When the behavior policy is known, both the plug-in estimator and the importance sampling estimator are applicable. In fact, they belong to the family of *Switch* estimators, as introduced in past[1] work [11], [28]. For any subset $S \subseteq [k]$, we define the Switch estimator associated with S as

$$\widehat{V}_{\mathsf{switch}}(S) := \sum_{a\in S} \pi_{\mathsf{t}}(a)\widehat{r}(a) + \frac{1}{n}\sum_{i=1}^n \rho(A_i) R_i \mathbb{1}\{A_i \notin S\}, \tag{7}$$

where $\widehat{r}(a)$ is the empirical mean reward defined in equation (4). By making the choices $S = [k]$ or $S = \emptyset$, respectively, the Switch estimator $\widehat{V}_{\mathsf{switch}}(S)$ reduces either to the plug-in estimator (5) or to the IS estimator (6). Choices of S intermediate between these two extremes allow us to interpolate (or switch) between the plug-in estimator and the IS estimator.

The following proposition, whose proof is relatively elementary, provides a unified performance guarantee for the family of Switch estimators.

Proposition 1: For any subset $S \subseteq [k]$, we have

$$\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] \leq 3 r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a\notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}. \tag{8}$$

See Section IV-A for the proof of this claim.

Given the family of Switch estimators $\{\widehat{V}_{\mathsf{switch}}(S)\}_{S\subseteq[k]}$, it is natural to ask: how to choose the subset S among all possible subsets of the action space? The unified upper bounds established in Proposition 1 offer us a reasonable guideline: one should select a subset S to minimize the error bound (8), i.e.,

$$\min_{S\subseteq[k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a\notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}. \tag{9}$$

[1]To be clear, the paper [11] considers a restricted version of this family, in which the subset S is restricted to be of the form $\{a \in [k] \mid \rho(a) \geq \tau\}$ for some threshold $\tau \geq 0$, whereas we define the estimator for any set.

At first glance, the minimization problem (9) is combinatorial in nature, which indicates the possible computational hardness in solving it. Fortunately, it turns out that such an "ambitious" goal can instead be achieved via solving a tractable convex program. To make this claim precise, let us consider the following convex program

$$\min_{v \in \mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\}, \quad (10)$$

where $v = (v(1), v(2), \ldots, v(k))^\top$ is a vector of decision variables. The convex program (10) can be viewed as a convex relaxation of the combinotorial problem (9). Let $v^\star$ be a minimizer of this optimization problem (10), whose existence is guaranteed by the coerciveness of the objective function. Correspondingly, we define

$$S^\star := \{a \mid v^\star(a) \neq 0\} \quad (11)$$

to be the support of $v^\star$. It turns out that the choice $S = S^\star$ solves the best subset selection problem (9) up to a constant factor. We summarize in the following:

Proposition 2: There exists a universal constant $c > 0$ such that

$$\min_{S \subseteq [k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\} \geq c \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\}. \quad (12)$$

See Section IV-B for the proof of the optimality of $S^\star$.

Thus, we conclude that the among the family of Switch estimators, the optimal estimator is given by

$$\widehat{V}_{\mathsf{switch}}(S^\star) := \sum_{a \in S^\star} \pi_{\mathsf{t}}(a) \widehat{r}(a) + \frac{1}{n} \sum_{i=1}^n \rho(A_i) R_i \mathbb{1}\{A_i \notin S^\star\}. \quad (13a)$$

In view of Proposition 1, it enjoys the following performance guarantee

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S^\star) - V_f(\pi_{\mathsf{t}}))^2] \leq 3 r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\}. \quad (13b)$$

From now on, we shall refer to $\widehat{V}_{\mathsf{switch}}(S^\star)$ as the Switch estimator.

1) Is the Switch Estimator Optimal?: The above discussion establishes the optimality of the Switch estimator $\widehat{V}_{\mathsf{switch}}(S^\star)$ among the family of estimators (7) parameterized by a choice of subset S. However, does the Switch estimator continue to be optimal in a larger context? This question can be assessed by determining whether it achieves, say up to a constant factor, the *minimax risk* given by

$$\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}) := \inf_{\widehat{V}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2], \quad (14)$$

Here the infimum ranges over all measurable functions $\widehat{V}$ of the data $\{(A_i, R_i)\}_{i=1}^n$, whereas the supremum is taken over all reward distributions f belonging to our family $\mathcal{F}$ of bounded mean distributions. The following theorem provides a lower bound on this minimax risk:

Theorem 1: There exists a universal positive constant c such that for all pairs $(\pi_{\mathsf{b}}, \pi_{\mathsf{t}})$, we have

$$\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}) \geq c\, r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\}.$$

See Section IV-C for the proof of this lower bound.

By combining Theorem 1 and the upper bound (13b) on the mean-squared error of the Switch estimator $\widehat{V}_{\mathsf{switch}}(S^\star)$, we obtain a finite-sample characterization of the minimax risk up to universal constants—namely

$$\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}) \asymp r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\}. \quad (15)$$

Consequently, we see that the Switch estimator $\widehat{V}_{\mathsf{switch}}(S^\star)$ is optimal among all estimators in a minimax sense.

In order to gain intuition for this optimality result, it is helpful to consider some special cases.

a) Degenerate case of on-policy evaluation: First, consider the degenerate setting $\pi_{\mathsf{t}} = \pi_{\mathsf{b}}$, so that our OPE problem actually reduces to a standard on-policy evaluation problem. In this case, the IS estimator reduces to the standard Monte Carlo estimate

$$\widehat{V}_{\mathsf{IS}} = \frac{1}{n} \sum_{i=1}^n \rho(A_i) R_i = \frac{1}{n} \sum_{i=1}^n R_i.$$

A straightforward calculation shows that it has mean-squared error $r_{\max}^2/n$, which we claim is order-optimal. To reach this conclusion from our expression (15) for the minimax risk, it suffices to check that $v^\star = 0$ is a minimizer of the optimization problem (10). This fact can be certified by showing that the all-zeros vector 0 obeys the first-order optimality condition associated with the convex program (10). More precisely, for all actions $a \in [k]$, we have

$$\sqrt{\frac{1}{8n}} \frac{\pi_{\mathsf{t}}(a)/\pi_{\mathsf{b}}(a)}{\sqrt{\sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a)]^2}{\pi_{\mathsf{b}}(a)}}} = \sqrt{\frac{1}{8n}} \leq \frac{1}{2}. \quad (16)$$

b) Large-sample regime: Returning to the general off-policy case ($\pi_{\mathsf{t}} \neq \pi_{\mathsf{b}}$), suppose that the sample size n satisfies a lower bound of the form

$$n \geq c \frac{\max_{a \in [k]} \rho^2(a)}{\sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a)} \quad (17)$$

for a sufficiently large constant c. In this case, the all-zeros vector 0 is again optimal for the convex program (10) since the first-order optimality condition (16) is met as long as c is large enough. As a consequence, we conclude that the Switch estimator reduces to the IS estimator in the *large-sample regime* defined by the lower bound (17). In this regime, the IS estimator achieves mean-squared error $r_{\max}^2 \cdot \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a)/n$. Under the bounded reward condition, this result recovers the rate provided by Li *et al.* [10] in the large sample regime (17) up to a constant factor; see Theorem 1 in their paper.

It is worthwhile elaborating further on the connections with the paper of Li *et al.* [10]: they studied classes of reward

distributions that are parameterized by bounds on their means (as implied by our bounded rewards) and variances. In this sense, their analysis is finer-grained than our study of bounded rewards only. However, when their results are specialized to bounded reward distributions (1), their minimax risk result (cf. equation (2) in their paper) applies only in the large sample regime defined by the lower bound (17). As we have discussed, when this lower bound holds, the IS estimator itself is order optimal, so analysis restricted to this regime fails to reveal the tradeoff between the two terms in the minimax rate (15), and in particular, the potential sub-optimality of the IS estimator (as reflected by the presence of the additional term $\pi_{\mathsf{t}}^2(S^\star)$ in the minimax bound).

2) A Closer Look at the Switch Estimator: In this subsection, we take a closer look at some properties of the Switch estimator, and in particular its connection to truncation of the likelihood ratio.

a) Link to likelihood truncation: We begin by investigating the nature of the best subset $S^\star$, as defined in equation (11). Let us assume without loss of generality that the actions are ordered according to the likelihood ratios—viz.

$$\rho(1) \leq \rho(2) \leq \cdots \leq \rho(k). \tag{18}$$

Under this condition, unraveling the proof of Proposition 2 shows that the optimal subset $S^\star$ takes the form

$$S^\star = \{s, s+1, \ldots, k\} \quad \text{for some integer } s \in [k]. \tag{19}$$

Here it should be understood that the choice $s = k$ corresponds to $S^\star = \emptyset$. Thus, the significance of the optimization problem (10) is that it specifies the optimal threshold at which to truncate the likelihood ratio. Although we cannot provide closed form solutions to the optimal threshold, it is interesting to see that the optimal subset $S^\star$ automatically singles out the set of actions with large likelihood ratios, which agrees with the intuition that the IS estimator has large variance for such actions.

As noted previously, Wang *et al.* [11] studied the subfamily of Switch estimators obtained by varying the truncation thresholds of the likelihood ratios. Similar to Li *et al.* [10], they studied the large sample regime in which the IS estimator without any truncation is already minimax optimal up to constant factors. This fails to explain the benefits of truncating large likelihood ratios and the associated Switch estimator. In contrast, the key optimization problem (10) informs us of the optimal subset $S^\star$ and hence an optimal truncation threshold, which allows the Switch estimator $\widehat{V}_{\mathsf{switch}}(S^\star)$ to optimally estimate the target value function for all sample sizes. This result is especially relevant for smaller sample sizes in which the problem is challenging, and the IS estimator can exhibit rather poor behavior.

b) Role of the plug-in component: The Switch estimator (13a) is based on applying the plug-in principle to the actions in $S^\star$ with large likelihood ratios. However, doing so is not actually necessary to achieve the optimal rate of convergence (15). In fact, if we simply estimate the mean reward by zero for any action in $S^\star$, then we obtain the estimate

$$\widehat{V} := \frac{1}{n} \sum_{i=1}^{n} \rho(A_i) R_i \mathbb{1}\{A_i \notin S^\star\}, \tag{20}$$

which is also minimax-optimal up to a constant factor. The intuition is that for actions on the support $S^\star$, the likelihood ratios are so large that the off-policy data is essentially useless, and can be ignored. It suffices to use the zero estimate, yielding a squared bias of the order $\pi_{\mathsf{t}}^2(S^\star)$. On the other hand, for actions in the complement $(S^\star)^c$, the likelihood ratios are comparatively small, so that the off-policy data should be exploited.

We note that truncated IS estimators of the type (20) have been explored in empirical work on counterfactual reasoning [29] and reinforcement learning [30]; our work appears to be the first to establish their optimality for general likelihood ratios. Also noteworthy is the paper by Ionides [31], who analyzed the rate at which the truncation level should decay, assuming that the likelihood ratios decay at a polynomial rate. Our theory, while focused on finite action spaces, instead works for any configuration of the likelihood ratios, and in addition provides a precise truncation level instead of only a rate.

3) Numerical Experiments: In this section, we report the results of some simple numerical experiments on simulated data that serve to illustrate the possible differences between the three methods: Switch, plug-in and IS estimators. We performed experiments with the uniform target policy (i.e., $\pi_{\mathsf{t}}(a) = 1/k$ for all actions $a \in [k]$), and for each action a, we defined the reward distribution $f(\cdot \mid a)$ to be an equi-probable Bernoulli distribution over $\{0, 1\}$, so that $r_{\max} = 1$.

For each choice of k, we constructed a behavior policy of the following form

$$\pi_{\mathsf{b}}(1) = \pi_{\mathsf{b}}(2) = \cdots = \pi_{\mathsf{b}}(\sqrt{k}) = \frac{1}{k^2}, \quad \text{and}$$
$$\pi_{\mathsf{b}}(\sqrt{k}+1) = \pi_{\mathsf{b}}(\sqrt{k}+2) = \cdots = \pi_{\mathsf{b}}(k) = \frac{1 - \frac{1}{k^{3/2}}}{k - \sqrt{k}}.$$

In words, we set the first $\sqrt{k}$ actions with a low probability $\frac{1}{k^2}$, whereas for the remaining $k - \sqrt{k}$ actions, the behavior probabilities are relatively large, which is close to $\frac{1}{k}$. As we will see momentarily, this choice allows us to demonstrate interesting differences between the three estimators.

As is standard in high-dimensional statistics [32], we study a sequence of such problems indexed by the pair (n, k); in order to obtain an interesting slice of this two-dimensional space, we set $n = 1.5k$. For such a sequence of problems, we can explicitly compute that the mean-squared errors of the three estimators scale as follows:

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \asymp 1, \tag{21a}$$
$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{IS}} - V_f(\pi_{\mathsf{t}}))^2] \asymp n^{-1/2}, \quad \text{and} \tag{21b}$$
$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S^\star) - V_f(\pi_{\mathsf{t}}))^2] \asymp n^{-1}. \tag{21c}$$

The purpose of our numerical experiments is to illustrate this theoretically predicted scaling.

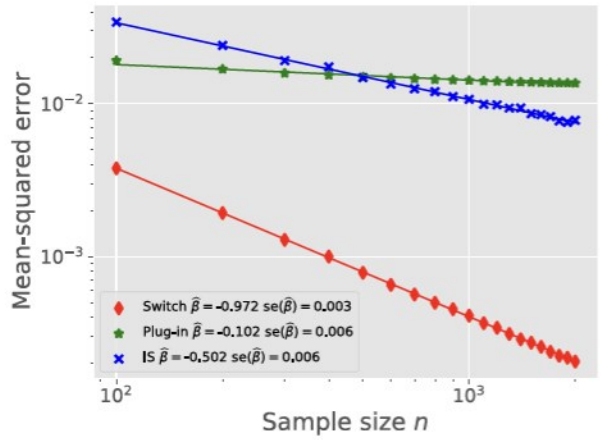

Fig. 1. Log-log plot of the estimation errors vs. the sample size. All results reported are averaged over 10^5 random trials. The legends also contain the estimated slopes and their standard errors obtained by performing linear regressions on the logarithms of the error and the sample size.

Figure 1 shows the mean-squared errors of these three estimators versus the sample size n, plotted on a log-log scale. The results are averaged over 10^5 random trials. As can be seen from Figure 1, the Switch estimator performs better than the two competitors uniformly across different sample sizes. Note that our theory (21) predicts that the mean-squared errors should scale as $n^{-\beta^\star}$, where $\beta^* \in \{0, 1/2, 1\}$ for the plug-in, IS, and SWITCH estimators respectively. In order to assess these theoretical predictions, we performed a linear regression of the log MSE on $\log(n)$, thereby obtaining an estimated exponent $\widehat{\beta}$ for each estimator. These estimates and their standard errors are shown in the legend of Figure 1. Clearly, the estimated slopes are quite close to the theoretical predictions (21).

B. OPE When π_b Is Unknown: Competitive Ratio

Our analysis thus far has taken the behavior policy π_b to be known. This condition, while realistic in some settings, is unrealistic in others. Accordingly, we now turn to the version of the OPE problem in which the only knowledge provided are the action-reward pairs $\{(A_i, R_i)\}_{i=1}^n$. Note that the importance sampling estimator $\widehat{V}_{\mathsf{IS}}$ and Switch estimators $\widehat{V}_{\mathsf{switch}}(S)$ are no longer applicable, since they require knowledge of the behavior policy. Consequently, we are led to the natural question: what is an optimal estimator when π_b is unknown? Before answering this question, one needs to first settle upon a suitable notion of optimality.

1) Optimality via the Minimax Competitive Ratio: The first important observation is that when the behavior policy is unknown, the global minimax risk is no longer a suitable metric for assessing optimality. Indeed, for any target policy π_t, one can construct a "nasty" behavior policy π_b such that for any estimator $\widehat{V}$, we have a lower bound of the form

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_b\otimes f}[(\widehat{V} - V_f(\pi_t))^2] \geq c\, r_{\max}^2,$$

for some universal constant $c > 0$. For this reason, if we measure optimality according to the global minimax risk, then the trivial "always return zero" estimator $\widehat{V} \equiv 0$ is optimal, and hence the global minimax risk is not a sensible criterion in this setting.

This pathology arises from the fact that the adversary has too much power: it is allowed to choose an arbitrarily bad behavior policy while suffering no consequences for doing so. In order to mitigate this deficiency, it is natural to consider the notion of a *competitive ratio*, as is standard in the literature on online learning [33]. An analysis in terms of the competitive ratio measures the performance of an estimator against the best achievable by some oracle—in this case, an oracle equipped with the knowledge of π_b.

For a given target policy π_t and behavior policy π_b, recall the definition (14) of the minimax risk $\mathcal{R}_n^\star(\pi_t; \pi_b)$; it corresponds to smallest mean-squared error that can be guaranteed, uniformly over a class of reward distributions $\mathcal{F}$, by any method equipped with the oracle knowledge of π_b. Given an estimator $\widehat{V}$ and a reward distribution f, we can measure its performance relative to this oracle lower bound via the *competitive ratio*

$$\mathcal{C}\big(\widehat{V}; \pi_t, \pi_b, f\big) := \frac{\mathbb{E}_{\pi_b\otimes f}[(\widehat{V} - V_f(\pi_t))^2]}{\mathcal{R}_n^\star(\pi_t; \pi_b)}. \tag{22}$$

An estimator $\widehat{V}$ with a small competitive ratio—that is, close to 1—is guaranteed to perform almost as well as the oracle that knows the behavior policy π_b. On the other hand, a large competitive ratio indicates poor performance relative to the oracle.

As one concrete example, the "always return zero" estimator $\widehat{V} \equiv 0$ is far from ideal when considered in terms of the competitive ratio (22). Indeed, suppose that $\pi_b = \pi_t$ and $r_f(a) = r_{\max}/2$; we then have

$$\begin{aligned}\sup_{\pi_b, f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_b\otimes f}[(\widehat{V} - V_f(\pi_t))^2]}{\mathcal{R}_n^\star(\pi_t; \pi_b)} &\geq \frac{\mathbb{E}_{\pi_t\otimes f}[(\widehat{V} - V_f(\pi_t))^2]}{\mathcal{R}_n^\star(\pi_t; \pi_b)} \\ &\overset{(i)}{\asymp} \frac{\mathbb{E}_{\pi_t\otimes f}[(V_f(\pi))^2]}{r_{\max}^2/n} \overset{(ii)}{\asymp} n.\end{aligned} \tag{23}$$

Here step (i) follows from the fact that $\widehat{V} \equiv 0$ by definition, along with the scaling $\mathcal{R}_n^\star(\pi_t; \pi_t) \asymp \frac{r_{\max}^2}{n}$ established in Section III-A.1. Step (ii) follows from the assumption that $r_f(a) = r_{\max}/2$, which implies that $\mathbb{E}_{\pi_t\otimes f}[(V_f(\pi))^2] = r_{\max}^2/4$. Thus, we see that the "always return zero" estimator $\widehat{V} \equiv 0$ performs extremely badly relative to the oracle, and its competitive ratio further degrades as the sample size n increases.

2) Competitive Ratio of the Plug-in Estimator: As we have emphasized earlier, the plug-in approach is applicable even if the behavior policy π_b is unknown. The following theorem provides a guarantee on its behavior in terms of the competitive ratio:

Theorem 2: There exists a universal constant $c > 0$ such that for any target policy π_t, the plug-in estimator $\widehat{V}_{\mathsf{plug}}$ satisfies the bound

$$\sup_{\pi_b, f\in\mathcal{F}} \mathcal{C}\big(\widehat{V}; \pi_t, \pi_b, f\big) \leq c\,|\operatorname{supp}(\pi_t)|.$$

See Section IV-D for the proof of this theorem.

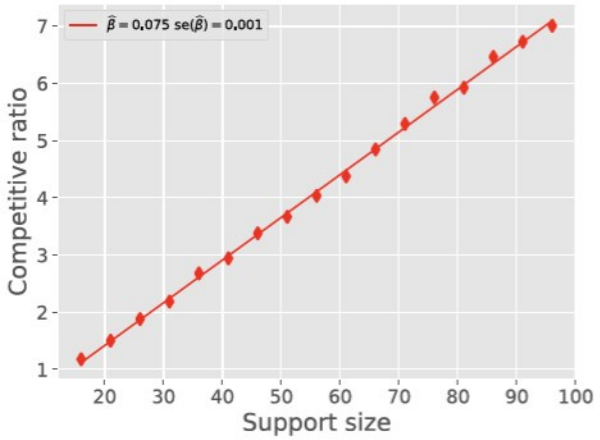

Fig. 2. Illustration of the competitive ratio the plug-in estimator vs. the support size of the target policy. Throughout the experiments, we set $k = 100, n = 2k$. The behavior policy obeys $\pi_{\mathsf{b}}(a) = (n \log k)^{-1}$ for $a \in [k-1]$, and $\pi_{\mathsf{b}}(k) = 1 - (k-1)/(n \log k)$. For each support size s, we take the target policy π_{t} to be the uniform distribution over $[s]$. Since $\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})$ is not known precisely, we use the mean-squared error of the Switch estimator as a surrogate, which is correct up to a constant. The results reported are averaged over 10^4 Monte Carlo trials.

Several remarks are in order. Note that the upper bound on the competitive ratio is at most ck, achieved for a target distribution that places mass on all k actions. Comparing this worst-case guarantee with that of the "always return zero" estimator (23) shows that plug-in estimator is strictly better as soon as the sample size n exceeds a multiple of k. Note that this is a relatively mild condition on the sample size. In addition, Theorem 2 guarantees worst-case competitive ratio of the plug-in estimator scales linearly with the support size of π_{t}. This showcases the automatic adaptivity of the plug-in estimator to the target policy under consideration. See Figure 2 for a numerical illustration of this phenomenon.

We note that Li *et al.* [10] established a similar guarantee (see Theorem 3 in their paper [10]). One importance difference is that their guarantee only holds in the large sample regime (cf. the restriction (17)), whereas ours covers the full spectrum of the sample size. Moreover, their upper bound is proportional to k for any target policy, and so does not reveal the adaptivity of the plug-in estimator to the support size.

3) Is the Plug-In Estimator Optimal?: A natural follow-up question is to investigate the optimality of the plug-in approach—in the sense of the worst-case competitive ratio—in the unknown π_{b} case. It turns out that, the plug-in estimator is close to optimal, as demonstrated by the following theorem.

Theorem 3: Suppose that the sample size is lower bounded as $n \geq c\frac{k}{\log k}$ for a positive constant c. Then for each $s \in \{1, 2, \ldots, k\}$, there exists a target policy π_{t} supported on s actions and

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \mathcal{C}\big(\widehat{V}; \pi_{\mathsf{t}}, \pi_{\mathsf{b}}, f\big) \geq c' \max\left\{\frac{s}{\log k}, 1\right\}, \tag{24}$$

where $c' > 0$ is a universal constant.

See Section IV-E for the proof of Theorem 3.

As shown in the proof of Theorem 3, for each given integer $s \in [k]$, the lower bound (24) is met by taking the target policy that chooses actions uniformly from the set $\{1, 2, \ldots, s\}$. This lower bound shows that when evaluating a policy with support size s, the gap—between performance when knowing the behavior policy π_{b} relative to not knowing it—scales as s up to a logarithmic factor; thus, these two settings are very different in terms of their statistical difficulty. In addition, comparing the lower bound in Theorem 3 with the upper bound provided in Theorem 2, one can see that the plug-in estimator $\widehat{V}_{\mathsf{plug}}$ is optimal up to a logarithmic factor, measured by the worst-case competitive ratio.

C. OPE With Lower Bounds on the Minimum Exploration Probability

The preceding subsections consider two extreme cases in which the behavior policy is either known or completely unknown. This leaves us with an interesting middle ground: what if we have some partial knowledge regarding the behavior policy? How can such information be properly exploited by estimators?

In this section, we initiate the investigation of these questions by focusing on a particular type of partial knowledge—namely, the *minimum exploration probability* $\min_{a \in [k]} \pi_{\mathsf{b}}(a)$. More precisely, for a given scalar $\nu \geq 0$, consider the collection of distributions

$$\Pi(\nu) := \big\{\pi \mid \min_{a \in [k]} \pi(a) \geq \nu\big\}.$$

Given that any randomized policy π must sum to one, i.e., $\sum_{a \in [k]} \pi(a) = 1$, this family is non-empty only when $\nu \in [0, 1/k]$. Our goal in this section is to characterize the difficulty of the OPE problem when it is known that $\pi_{\mathsf{b}} \in \Pi(\nu)$ for some choice of ν. We first analyze the plug-in estimator, which does *not* require knowledge of ν. We then derive a minimax lower bound, which shows that the plug-in estimator is sub-optimal for certain choices of ν. In the end, we design an alternative estimator, based on approximation by Chebyshev polynomials, that has optimality guarantees for a large family of target policies, albeit under a different but closely related Poisson sampling model.

1) Performance of the Plug-in Estimator: We begin with establishing a performance guarantee for the plug-in estimator.

Theorem 4: There exist universal constants $c, c' > 0$ such that for any $\pi_{\mathsf{b}} \in \Pi(\nu)$, one has

$$\begin{aligned}&\sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \\ &\quad \leq c\left\{r_{\max}^2 \cdot \exp(-2n\nu) + \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})\right\}. \end{aligned} \tag{25a}$$

In addition, if $\nu \geq \frac{\log k}{n}$, then we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \leq c' \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}). \tag{25b}$$

See Section IV-F for the proof of these two claims.

Two interesting observations are worth making. First, if the behavior policy is *sufficiently exploratory*, in the sense that it belongs to the family $\Pi(\nu)$ for some $\nu \geq \frac{\log k}{n}$, then the plug-in estimator $\widehat{V}_{\mathsf{plug}}$ achieves the optimal estimation error $\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})$ up to a constant factor. In other words, the side

condition $\min_{a\in[k]} \pi_{\mathsf{b}}(a) \geq \frac{\log k}{n}$ is sufficient for the plug-in approach to perform optimally.

On the other hand, when the behavior policy is less exploratory—meaning that $\nu < \frac{\log k}{n}$—its mean-squared error involves the additional term $r_{\max}^2 \cdot \exp(-2n\nu)$. As shown in the proof of the upper bound (25a), this extra price stems from bias of the plug-in estimator: if we fail to observe rewards for some action a, then the plug-in estimator has no avenue for estimating the mean reward $r_f(a)$; any estimate that it makes incurs a bias of the order $\pi_{\mathsf{t}}^2(a) \cdot r_{\max}^2$. When $\pi_{\mathsf{b}}(a) = \nu$, such an event takes place with probability on the order of $\exp(-n\nu)$.

2) Is the Plug-in Estimator Optimal Under Partial Knowledge?: Is the extra price $r_{\max}^2 \cdot \exp(-2n\nu)$ necessary for all estimators? In order to answer this question, we need to characterize the constrained minimax risk

$$\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu) := \inf_{\widehat{V}} \sup_{(\pi_{\mathsf{b}}, f)\in\Pi(\nu)\times\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2], \tag{26}$$

where the supremum is taken over all possible behavior policies in $\Pi(\nu)$, and reward distributions in $\mathcal{F}$. In view of our guarantee (25b) for the plug-in approach, it can be seen that when $\nu \geq \frac{\log k}{n}$, then

$$\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu) \asymp \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^{\star}(\pi; \pi_{\mathsf{b}}),$$

Consequently, the plug-in estimator is optimal when the behavior policy is sufficiently exploratory.

As a result, in the remainder of this section, we concentrate on the regime $\nu < \frac{\log k}{n}$. We begin by stating a minimax lower bound in this regime:

Theorem 5: Consider the case $\nu < \frac{\log k}{n}$. If ν further satisfies $\nu \leq \frac{1}{2k}$ and $\nu \geq c\frac{1}{n\log k}$ for some sufficiently large constant $c > 0$, then there exists another universal positive constant c' such that

$$\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu) \geq c' \left\{ \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^{\star}(\pi; \pi_{\mathsf{b}}) + r_{\max}^2 \cdot \exp(-200\sqrt{n\nu\log k}) \right\}. \tag{27}$$

See Section IV-G for the proof of this theorem.

Note that if $\nu \lesssim \frac{1}{n\log k}$, the worst-case risk is lower bounded as $\Omega(r_{\max}^2)$. Combining the lower bound in Theorem 5 with the upper bound shown in Theorem 4, we conclude that the plug-in approach is minimax optimal up to constants once $\nu \gtrsim \log k/n$. However, observe that there remains a gap between the upper and lower bounds when the behavior policy is known to be less exploratory—that is, in the regime $\frac{1}{n\log k} \lesssim \nu \ll \frac{\log k}{n}$.

3) Optimal Estimators via Chebyshev Polynomials in the Poisson Model: In this section, we devote ourselves to the design of optimal estimators when the behavior policy π_{b} is less exploratory, meaning that $\pi_{\mathsf{b}} \in \Pi(\nu)$ for some $\frac{1}{n\log k} \lesssim \nu \ll \frac{\log k}{n}$.

a) The Poisson model: In order to bring the key issues to the fore, we analyze the estimator under the Poissonized sampling model that is standard in the functional estimation literature. Recall that in the multinomial observation model, the action counts $\{n(a), a \in [k]\}$ follow a multinomial distribution with parameters n and π_{b}. In the alternative Poisson model, the total number of samples is assumed to be random, distributed according to a Poisson distribution with parameter n. As a result, the action counts obey $n(a) \overset{\text{ind.}}{\sim} \mathsf{Poi}(n\pi_{\mathsf{b}}(a))$ for $a \in [k]$. Correspondingly, we can define the minimax risk under the Poisson model as

$$\mathcal{R}_{\mathsf{P}}^{\star}(\pi_{\mathsf{t}}, n, \nu) := \inf_{\widehat{V}} \sup_{(\pi_{\mathsf{b}}, f)\in\Pi(\nu)\times\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2], \tag{28}$$

where the expectation is taken under the Poisson model.

Although the two sampling models differ, the difference is not actually essential in terms of characterizing minimax risks. In particular, the corresponding risks $\mathcal{R}_{\mathsf{P}}^{\star}(\pi_{\mathsf{t}}, n, \nu)$ and $\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu)$ are closely related, as demonstrated by the following lemma.

Lemma 1: For any $\beta \in (0, 1)$, we have

$$\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu) \leq \frac{\mathcal{R}_{\mathsf{P}}^{\star}(\pi_{\mathsf{t}}, (1-\beta)n, \nu)}{1 - \exp(-n\beta^2/2)}.$$

See Appendix A for the proof of this bound.

Setting $\beta = 1/2$ in the above lemma reveals that $\mathcal{R}_{\mathsf{M}}^{\star}(\pi_{\mathsf{t}}, n, \nu) \lesssim \mathcal{R}_{\mathsf{P}}^{\star}(\pi_{\mathsf{t}}, \frac{1}{2}n, \nu)$. Consequently, in order to obtain an upper bound on the risk under multinomial sampling, it suffices to control the risk under the Poisson model.

b) The Chebyshev estimator: Now we turn to the construction of the optimal estimator under the Poisson model. It turns out that Chebyshev polynomials play a central role in such a construction. Recall that the Chebyshev polynomial with degree L is given by

$$Q_L(x) := \cos(L \arccos x) \tag{29}$$

for $x \in [-1, 1]$. Correspondingly, for any pair of scalars such that $r > \ell > 0$, we can define a shifted and scaled polynomial

$$P_L(x) = -\frac{Q_L\left(\frac{2x-r-\ell}{r-\ell}\right)}{Q_L\left(\frac{-r-\ell}{r-\ell}\right)} =: \sum_{d=0}^{L} a_d x^d,$$

where a_d denotes the coefficient of x^d. Using the coefficients of this polynomial as a building block, we then define a function, with domain the set of nonnegative integers, given by

$$g_L(j) = \begin{cases} a_j \frac{j!}{n^j} + 1, & \text{for } j = 0, 1, \ldots, L, \text{ and} \\ 1 & \text{if } j > L. \end{cases}$$

In terms of these quantities, the *Chebyshev estimator* takes the form

$$\widehat{V}_{\mathsf{C}} := \sum_{a\in[k]} \pi_{\mathsf{t}}(a)\widehat{r}(a) g_L(n(a)), \tag{30}$$

where $\widehat{r}(a)$ is the empirical mean reward defined in equation (4). In words, when the action count $n(a)$ is larger than the

degree L, one uses the usual sample mean reward $\widehat{r}(a)$. On the other hand, when the action count $n(a)$ is below this threshold, the Chebyshev estimator rescales the empirical mean reward by the value $g_L(n(a))$. The goal of this rescaling is to reduce the bias of the plug-in estimate.

A little calculation helps to provide intuition regarding this bias-reduction effect. Under the Poisson sampling model, the biases of the Chebyshev estimator and the plug-in estimator are given by

$$\mathbb{E}[\widehat{V}_{\mathsf{C}}] - V_f(\pi_{\mathsf{t}}) = \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a) e^{-n\pi_{\mathsf{b}}(a)} P_L(\pi_{\mathsf{b}}(a)), \text{ and}$$

$$\mathbb{E}[\widehat{V}_{\mathsf{plug}}] - V_f(\pi_{\mathsf{t}}) = \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a) e^{-n\pi_{\mathsf{b}}(a)},$$

respectively. For the plug-in estimator, if we allow the behavior policy π_{b} to range over the family $\Pi(\nu)$, then the bias can be as large as $r_{\max} \cdot \exp(-n\nu)$.

By construction, the Chebyshev polynomial P_L is the unique degree-L polynomial such that $P_L(0) = -1$, and that is closest in sup norm to the all-zeros function on the interval $[\ell, r]$; see Exercise 2.13.14 in the book [34]. By suitable choices of the triple (ℓ, r, L), we can shape the additional modulation factor $P_L(\pi_{\mathsf{b}}(a))$ so as to reduce the bias of the plug-in estimator. The following theorem makes this intuition precise:

Theorem 6: Suppose that the target policy satisfies the bound

$$\sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) \le \frac{1}{k^\gamma} \quad \text{for some scalar } \gamma > 0, \tag{31}$$

and that we implement the Chebyshev estimator with

$$\ell := \nu, \quad r := c_1 \log k/n, \quad \text{and} \quad L = c_0 \log k,$$

for some sufficiently large constant $c_1 > 0$ and some constant $c_0 \le \gamma/7$. Then under the Poisson sampling model, there exists a pair of positive constants (c, c') such that

$$\sup_{(\pi_{\mathsf{b}}, f)\in\Pi(\nu)\times\mathcal{F}} \mathbb{E}\left[\left(\widehat{V}_{\mathsf{C}} - V_f(\pi_{\mathsf{t}})\right)^2\right] \le c\left\{r_{\max}^2 \exp(-c'\sqrt{n\nu \log k}) + \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})\right\}. \tag{32}$$

See Section IV-H for the proof of this upper bound.

Several comments are in order. First, when $\nu < \log k/n$, the worst-case risk (32) of the Chebyshev estimator (30) matches the lower bound (27) derived in Theorem 5, which showcases the optimality of the Chebyshev estimator when the partial knowledge ν is available.

Second, the restriction (31) on the evaluation policy is worth emphasizing. In words, the constraint (31) requires the target policy to be somewhat "de-localized"—that is, there is no action a that has an extremely large probability mass. As an example, the uniform policy on $[k]$ satisfies such a constraint.

Third, it should be noted that the Chebyshev estimator requires knowledge of the minimum exploration probability ν. This property makes it less practically applicable a priori, and how to design an estimator that can adapt to the nested family of behavior policies $\{\Pi(\nu)\}_{0<\nu\le 1/k}$ is an interesting question for future work.

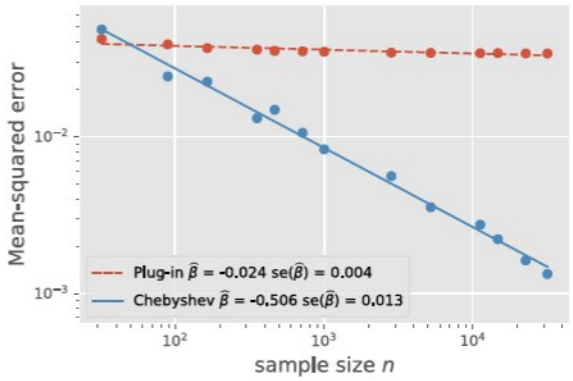

Fig. 3. Log-log plot of the estimation errors vs. the sample size. All results reported are averaged over 10^4 random trials. The legends also contain the estimated slopes obtained by performing a linear regression of the log error on the log sample size.

4) Numerical Experiments: We conclude this section with experiments on both simulated data and real data to assess the performance of the Chebyshev estimator relative to other choices.

a) Simulated data: We begin with some experiments on simulated data. As in our previous simulations, we fix the target policy to be uniform over $[k]$, and for each action $a \in [k]$, we choose the reward distribution $f(\cdot \mid a)$ to be an equi-probable Bernoulli distribution over $\{0, 1\}$, so that $r_{\max} = 1$. For each k, we define the behavior policy

$$\pi_{\mathsf{b}}(1) = 1 - \frac{k-1}{k^{1.5}}, \quad \text{and}$$
$$\pi_{\mathsf{b}}(2) = \pi_{\mathsf{b}}(3) = \cdots = \pi_{\mathsf{b}}(k) = \frac{1}{k^{1.5}}.$$

Again, we consider a particular scaling of the pair (n, k) that highlights interesting differences. In particular, when the sample size n scales as $n = k^{1.5}$, then our theory predicts that the plug-in and Chebyshev estimators should have mean-squared error scaling as

$$\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \asymp 1, \quad \text{and}$$
$$\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{\mathsf{C}} - V_f(\pi_{\mathsf{t}}))^2] \asymp n^{-\delta} \quad \text{for some } \delta > 0,$$

respectively. Figure 3 plots the mean-squared errors of the two estimators vs. the sample size n on a log-log scale. The results are averaged over 10^4 random trials. It is clear from Figure 3 that the Chebyshev estimator performs better than the one based on the plug-in principle. Based on the estimated slopes (as shown in the legend), the mean-squared error of the Chebyshev estimator decays as $n^{-1/2}$, while consistent with our theory, that of the plug-in estimator nearly plateaus.

To further evaluate the performance of the Chebyshev estimator, we conduct the same experiment as before under diverse target policies. More specifically, fixing some $\alpha \ge 0$, we set

$$\pi_{\mathsf{t}}(i) \propto 1/i^\alpha \quad \text{for all } i \ge 1.$$

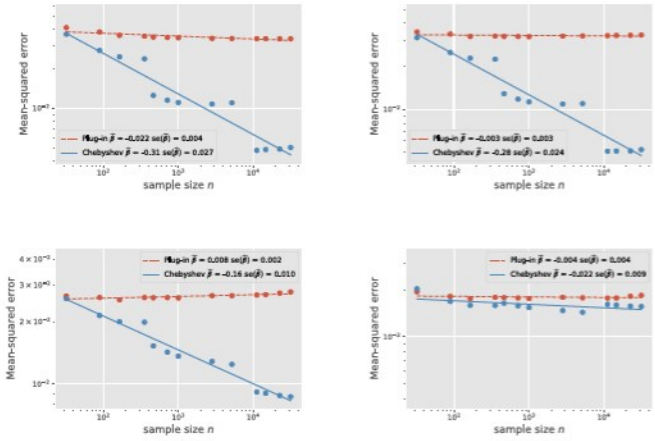

Fig. 4. Performance comparison between the plug-in and the Chebyshev estimators under non-uniform target policies. The parameter α takes the value of $0.1, 0.5, 1, 1.5$ from left to right, up to bottom.

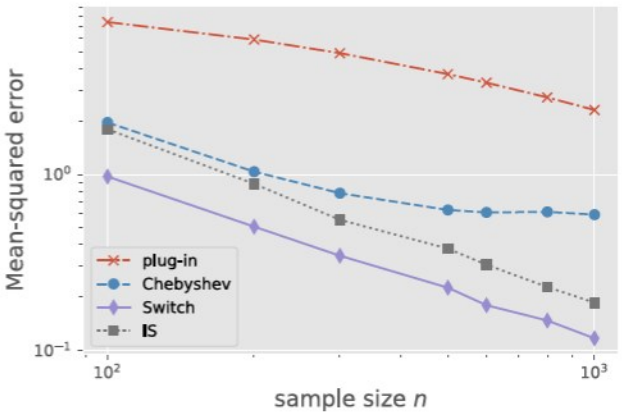

Fig. 5. Mean-squared errors of four different estimators vs. the sample size n on the MovieLens 25M data set. The results are averaged over 10^4 trials. See the text for further details on the experimental set-up.

Clearly, when $\alpha = 0$, the target policy π_{t} becomes uniform, while a larger α corresponds to more non-uniform test policies. Figure 4 reports the results for both the Chebyshev estimator and the plug-in estimator for four different values of α: 0.1, 0.5, 1, 1.5. It can be seen that the Chebyshev estimator performs uniformly better than the plug-in estimator. However, as the target policy becomes peaked (i.e., α increases) and the assumption in equation (31) is violated, the performance improvement is less significant. This partially suggests the necessity of the assumption (31).

b) Real data: We now turn to some experiments with the MovieLens 25M data set [35]. In order to form a bandit problem, we extracted a random subset of 500 movies that each have at least 10 ratings. This subset of movies defines an action space $[k]$ with $k = 500$. For each movie, we average its rating over all samples in order to define the mean reward $r_f(a)$ associated with the movie a. This is the ground truth that defines our problem instance. Setting the target policy to be uniform, our goal is to estimate the mean rating of these 500 movies—that is, the quantity $\sum_{a\in[k]} r_f(a)/k$.

In order to evaluate our methods, we need to generate an off-policy dataset. In order to do so, we uniformly subsample n ratings from the set of all ratings on our subset of 500 movies. This procedure implicitly defines a behavior policy that is very different from the uniform target policy, because the number of ratings for each movie vary drastically. Given such an off-policy dataset, we evaluate the mean-squared errors of four different estimators—the plug-in estimator, the IS estimator, the Switch estimator as well as the Chebyshev estimator. We repeat this procedure for a total of 10^4 trials for a range of sample sizes n.

Figure 5 plots the mean-squared error (averaged over the trials) versus the sample size n for the four estimators. To be clear, the Switch estimator and the IS estimator have the luxury of knowing the behavior policy whereas the Chebyshev estimator is given minimum exploration probability. The plug-in estimator requires no side information. Given the oracle knowledge of the behavior policy, the Switch estimator always outperforms other estimators, including the IS estimator with the same knowledge. In addition, the Chebyshev estimator outperforms plug-in estimator, especially in the small sample regime. These qualitative changes are consistent with our theoretical predictions.

IV. Proofs

We now turn to the proofs of the main results presented in Section III. We begin in Section IV-A with the proof of Proposition 1, followed by the proof of Proposition 2 in Section IV-B. Sections IV-C through IV-H are devoted to the proofs of Theorems 1 through 6.

A. Proof of Proposition 1

We begin with the standard bias-variance decomposition

$$\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] = \left(\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[\widehat{V}_{\mathsf{switch}}(S)] - V_f(\pi_{\mathsf{t}})\right)^2 + \mathsf{Var}\left(\widehat{V}_{\mathsf{switch}}(S)\right). \tag{33}$$

Our proof involves establishing the following two bounds

$$\left(\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[\widehat{V}_{\mathsf{switch}}(S)] - V_f(\pi_{\mathsf{t}})\right)^2 \le r_{\max}^2 \pi_{\mathsf{t}}^2(S), \quad \text{and} \tag{34a}$$

$$\mathsf{Var}\left(\widehat{V}_{\mathsf{switch}}(S)\right) \le 2r_{\max}^2 \left\{\pi_{\mathsf{t}}^2(S) + \frac{\sum_{a\notin S}\pi_{\mathsf{b}}(a)\rho^2(a)}{n}\right\}. \tag{34b}$$

The claim of the proposition follows by substituting these bounds into the bias-variance decomposition (33).

1) Proof of the Bias Bound (34a)*:* Using the shorthand $\mathbb{E}$ for the expectation over $\pi_{\mathsf{b}} \otimes f$, we have

$$\begin{aligned}\mathbb{E}[\widehat{V}_{\mathsf{switch}}(S)] &= \sum_{a\in S}\pi_{\mathsf{t}}(a)\mathbb{E}[\widehat{r}(a)] + \mathbb{E}[\rho(A_i)R_i\mathbb{1}\{A_i\notin S\}] \\ &= \sum_{a\in S}\pi_{\mathsf{t}}(a)\mathbb{E}[\widehat{r}(a)] + \sum_{a\notin S}\pi_{\mathsf{t}}(a)r_f(a).\end{aligned}$$

Recalling the definition (3) of $V_f(\pi_{\mathsf{t}})$, we have

$$\left(\mathbb{E}[\widehat{V}_{\mathsf{switch}}(S)] - V_f(\pi_{\mathsf{t}})\right)^2 = \left(\sum_{a\in S} \pi_{\mathsf{t}}(a)\left\{\mathbb{E}[\widehat{r}(a)] - r_f(a)\right\}\right)^2 \le r_{\max}^2 \pi_{\mathsf{t}}^2(S),$$

where the final inequality follows from the bound $\left|\mathbb{E}[\widehat{r}(a)] - r_f(a)\right| \le r_{\max}$.

2) Proof of the Variance Bound (34b)*:* Using the inequality $\mathsf{Var}(X+Y) \le 2\mathsf{Var}(X) + 2\mathsf{Var}(Y)$, we have

$$\mathsf{Var}\left(\widehat{V}_{\mathsf{switch}}(S)\right) \le 2\mathsf{Var}\left(\sum_{a\in S}\pi_{\mathsf{t}}(a)\widehat{r}(a)\right) + 2\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^n \rho(A_i)R_i\mathbb{1}\{A_i \notin S\}\right).$$

The first term can be bounded as

$$\mathsf{Var}\left(\sum_{a\in S}\pi_{\mathsf{t}}(a)\widehat{r}(a)\right) \le \mathbb{E}\left[\left(\sum_{a\in S}\pi_{\mathsf{t}}(a)\widehat{r}(a)\right)^2\right] \le r_{\max}^2\left(\sum_{a\in S}\pi_{\mathsf{t}}(a)\right)^2 = r_{\max}^2\pi_{\mathsf{t}}^2(S),$$

where the penultimate relation arises from the fact that $|\widehat{r}(a)| \le r_{\max}$. With regards to the variance brought by importance sampling, one has

$$\begin{aligned}
&\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^n \rho(A_i)R_i\mathbb{1}\{A_i\notin S\}\right)\\
&= \frac{1}{n}\mathsf{Var}\left(\rho(A_i)R_i\mathbb{1}\{A_i\notin S\}\right)\\
&\le \frac{1}{n}\mathbb{E}[\rho^2(A_i)R_i^2\mathbb{1}\{A_i\notin S\}]\\
&\le r_{\max}^2\frac{\mathbb{E}[\rho^2(A_i)\mathbb{1}\{A_i\notin S\}]}{n}\\
&= r_{\max}^2\frac{\sum_{a\notin S}\pi_{\mathsf{b}}(a)\rho^2(a)}{n},
\end{aligned}$$

where the last inequality uses the fact that $|R_i| \le r_{\max}$. Combining the two terms yields the claimed variance bound (34b).

B. Proof of Proposition 2

Recall that the subset $S^\star$ corresponds to the support set of the solution $v^\star$ to the convex program (10). Here we state an important connection between the objective value of the problem (10) and this support set:

Lemma 2: We have

$$\min_{v\in\mathbb{R}^k}\left\{\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]}|v(a)|\right\} \asymp \pi_{\mathsf{t}}(S^\star) + \sqrt{\frac{\sum_{a\notin S^\star}\pi_{\mathsf{b}}(a)\rho^2(a)}{n}}, \tag{35}$$

where $\asymp$ denotes equality up to a universal constant. See Appendix B-A for the proof of this lemma.

In light of the equivalence (35) as well as the sandwich bound $x^2+y^2 \le (x+y)^2 \le 2x^2+2y^2$ for any $x, y > 0$, proving the bound (12) reduces to establishing the lower bound

$$\min_{S\subseteq[k]}\left\{\frac{1}{2}\pi_{\mathsf{t}}(S) + \sqrt{\frac{\sum_{a\notin S}\pi_{\mathsf{b}}(a)\rho^2(a)}{8n}}\right\} \ge \min_{v\in\mathbb{R}^k}\left\{\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]}|v(a)|\right\}. \tag{36}$$

Letting $\xi(a) \in \{0,1\}$ denote a binary indicator variable for the event $\{a \in S\}$, the optimization problem on the left hand side is equivalent to

$$p^* = \min_{\boldsymbol{\xi}\in\mathbb{R}^k}\left\{\frac{1}{2}\sum_{a\in[k]}\pi_{\mathsf{t}}(a)\xi(a) + \sqrt{\frac{\sum_{a\in[k]}\pi_{\mathsf{b}}(a)\rho^2(a)(1-\xi(a))}{8n}}\right\}$$
$$\text{s.t. } \xi(a)\in\{0,1\} \text{ for all } a\in[k].$$

By relaxing to the requirement $\xi(a) \in [0,1]$, we obtain a convex lower bound

$$p^* \ge \min_{\boldsymbol{\xi}\in\mathbb{R}^k}\left\{\frac{1}{2}\sum_{a\in[k]}\pi_{\mathsf{t}}(a)\xi(a) + \sqrt{\frac{\sum_{a\in[k]}\pi_{\mathsf{b}}(a)\rho^2(a)(1-\xi(a))}{8n}}\right\}$$
$$\text{s.t. } \xi(a)\in[0,1] \text{ for all } a\in[k].$$

Since the inclusion $\xi(a) \in [0,1]$ guarantees that $1-\xi(a) \ge (1-\xi(a))^2$, we can further relax to obtain the lower bound

$$p^* \ge \min_{\boldsymbol{\xi}\in\mathbb{R}^k}\left\{\frac{1}{2}\sum_{a\in[k]}\pi_{\mathsf{t}}(a)\xi(a) + \sqrt{\frac{\sum_{a\in[k]}\pi_{\mathsf{b}}(a)\rho^2(a)(1-\xi(a))^2}{8n}}\right\}$$
$$\text{s.t. } \xi(a)\in[0,1] \text{ for all } a\in[k].$$

Now we are ready to prove the claimed bound (36). Applying the change of variables $\xi(a) = v(a)/\pi_{\mathsf{t}}(a)$, we can transform the problem above into

$$\min_{\boldsymbol{v}}\left\{\frac{1}{2}\sum_{a\in[k]}v(a) + \sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}}\right\}.$$

subject to the constraints $v(a) \in [0, \pi_{\mathsf{t}}(a)]$ for all $a \in [k]$. In fact, following some simple calculations, one can see that it is equivalent to the following unconstrained problem

$$\min_{v\in\mathbb{R}^k}\left\{\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]}|v(a)|\right\} \tag{37}$$

Note that this is identical to the lower bound we aim at in equation (36), and hence the proof is finished.

C. Proof of Theorem 1

Our proof is based on using Le Cam's method to lower bound the minimax risk. In doing so, a key step is to construct two similar reward distributions $f_1, f_2 \in \mathcal{F}$ such that the absolute distance $|V_{f_1}(\pi) - V_{f_2}(\pi)|$ is large.

For each action a, let $f_2(\cdot \mid a)$ be a Bernoulli distribution on the set $\{0, r_{\max}\}$ with parameter $\frac{1}{2}$, let $f_1(\cdot \mid a)$ be a Bernoulli

distribution over $\{0, r_{\max}\}$ with parameter $\frac{1}{2} + \delta(a)$ for some $\delta(a) \in [0, \frac{1}{2}]$. From Le Cam's inequality (see Theorem 36.8 in the article [36]), we have the lower bound

$$\inf_{\widehat{V}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2] \geq \frac{1}{8}(V_{f_1}(\pi) - V_{f_2}(\pi))^2 e^{-n \mathsf{D}_{\mathrm{KL}}(\pi_{\mathsf{b}} \otimes f_1 \,\|\, \pi_{\mathsf{b}} \otimes f_2)}.$$

With this choice of f_1 and f_2, one can verify that

$$\begin{aligned}\mathsf{D}_{\mathrm{KL}}(\pi_{\mathsf{b}} \otimes f_1 \,\|\, \pi_{\mathsf{b}} \otimes f_2) &= \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \mathsf{D}_{\mathrm{KL}}(f_1 \,\|\, f_2) \\ &\leq 4 \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \delta^2(a),\end{aligned}$$

where the inequality arises from the relation $\mathsf{D}_{\mathrm{KL}}(\mathsf{Bern}(\frac{1}{2} + \delta(a)) \,\|\, \mathsf{Bern}(\frac{1}{2})) \leq 4\delta^2(a)$. In addition, it is easily seen that

$$(V_{f_1}(\pi) - V_{f_2}(\pi))^2 = \tfrac{1}{4} r_{\max}^2 \Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) \delta(a) \Big)^2.$$

Therefore, we can obtain a lower bound on the minimax risk—one that is optimal within this particular family—by solving the optimization problem

$$\begin{aligned}\max_{\boldsymbol{\delta} \in \mathbb{R}^k} \quad & r_{\max}^2 \Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) \delta(a) \Big)^2 \\ \text{subject to } & 4 \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \delta^2(a) \leq \frac{1}{2n} \text{ and} \\ & \delta(a) \in [0, \tfrac{1}{2}] \text{ for all } a \in [k]. \end{aligned} \tag{38}$$

First, we make the observation that the optimization problem (38) is equivalent to the following optimization problem in the sense that they share the same minimizer and the minimum values are in a one-to-one correspondence

$$\begin{aligned}\max_{\boldsymbol{\delta} \in \mathbb{R}^k} \quad & \Big\{ \sum_{a \in [k]} \pi_{\mathsf{t}}(a) \delta(a) \Big\} \\ \text{subject to } & \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \delta^2(a) \leq \frac{1}{8n} \text{ and} \\ & \delta(a) \in [0, \tfrac{1}{2}] \text{ for all } a \in [k]. \end{aligned} \tag{39}$$

Note that this is a convex problem with quadratic constraints. We find it easier to look at its dual formulation, which is supplied in the following lemma.

Lemma 3: The Fenchel dual problem of the optimization program (39) is given by

$$\min_{\boldsymbol{v} \in \mathbb{R}^k} \quad \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\}, \tag{40}$$

which shares the same optimal objective value as that of the problem (39).

See Appendix C for the proof of this result.

Fortunately, the minimum value of the dual program (40) has been characterized in Lemma 2, namely

$$\begin{aligned}&\min_{\boldsymbol{v} \in \mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\} \\ &\asymp \pi_{\mathsf{t}}(S^\star) + \sqrt{\frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n}},\end{aligned}$$

where $\asymp$ denotes equality up to universal constants.

Taking the preceding equivalence relationships collectively and use the elementary relation $(x+y)^2 \asymp x^2 + y^2$, we can arrive at the desired conclusion

$$\begin{aligned}&\inf_{\widehat{V}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2] \\ &\geq c\, r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a \notin S^\star} \pi_{\mathsf{b}}(a) \rho^2(a)}{n} \right\}\end{aligned}$$

with c a universal positive constant.

D. Proof of Theorem 2

The mean-squared error of $\widehat{V}_{\mathsf{plug}}$ can be decomposed as

$$\begin{aligned}&\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \\ &= \Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\ &\quad + \sum_{a \in [k]} \pi_{\mathsf{t}}^2(a) \sigma_f^2(a) \mathbb{E}\left[\tfrac{\mathbb{1}\{n(a) > 0\}}{n(a)} \right] \\ &\quad + \mathsf{Var}\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) \mathbb{1}\{n(a) > 0\} \Big).\end{aligned} \tag{41}$$

See Appendix A.6 of the paper [10] for the calculations underlying this decomposition. Here the first term represents the squared bias of $\widehat{V}_{\mathsf{plug}}$, while the remaining two correspond to the variance of $\widehat{V}_{\mathsf{plug}}$.

Fix any behavior policy π_{b}, and let $S^\star$ be defined as in equation (11). Our proof consists of upper bounding the squared bias and variance in terms of functions of $\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})$. More precisely, we prove the following two bounds

$$\begin{aligned}&\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\ &\quad \leq c\, |\operatorname{supp}(\pi_{\mathsf{t}})|\, \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}), \quad \text{and} \end{aligned} \tag{42a}$$

$$\begin{aligned}&\sum_{a \in [k]} \pi_{\mathsf{t}}^2(a) \sigma_f^2(a) \mathbb{E}\left[\tfrac{\mathbb{1}\{n(a) > 0\}}{n(a)} \right] \\ &\quad + \mathsf{Var}\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) \mathbb{1}\{n(a) > 0\} \Big) \\ &\quad \leq c'\, \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}),\end{aligned} \tag{42b}$$

for some universal constants (c, c'). Since the bounds (42) hold for any π_{b}, the desired conclusion follows by combining the above two bounds. In the sequel, we focus on establishing the bounds (42).

1) Proof of the Bias Bound (42a)*:* Beginning with the squared bias, we have

$$\begin{aligned}
&\Big(\sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2 \\
&\overset{\text{(i)}}{\le} r_{\max}^2 \Big(\sum_{a\in[k]} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2 \\
&\overset{\text{(ii)}}{\le} 2r_{\max}^2 \Big\{\Big(\sum_{a\in S^\star} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2 \\
&\quad + \Big(\sum_{a\in (S^\star)^c} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2\Big\} \\
&\overset{\text{(iii)}}{\le} 2r_{\max}^2 \Big\{\pi_{\mathsf{t}}^2(S^\star) + \Big(\sum_{a\in (S^\star)^c} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2\Big\}.
\end{aligned} \tag{43}$$

Here the first inequality (i) arises from the assumption $|r_f(a)| \le r_{\max}$, the second relation (ii) applies the inequality $(a+b)^2 \le 2(a^2+b^2)$, and the last one (iii) uses the fact $(1-\pi_{\mathsf{b}}(a))^n \le 1$.

Applying the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\Big(\sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2 \\
&\quad = \Big(\sum_{a\in(S^\star)^c\cap \operatorname{supp}(\pi_{\mathsf{t}})} \pi_{\mathsf{t}}(a)(1-\pi_{\mathsf{b}}(a))^n\Big)^2 \\
&\quad \le |(S^\star)^c \cap \operatorname{supp}(\pi_{\mathsf{t}})| \left(\sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}^2(a)(1-\pi_{\mathsf{b}}(a))^{2n}\right) \\
&\quad \le |\operatorname{supp}(\pi_{\mathsf{t}})| \sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}^2(a)(1-\pi_{\mathsf{b}}(a))^n \\
&\quad \le |\operatorname{supp}(\pi_{\mathsf{t}})| \sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}^2(a)\frac{1}{n\pi_{\mathsf{b}}(a)},
\end{aligned}$$

where the last inequality follows from the bound $(1-\pi_{\mathsf{b}}(a))^n \le 1/(n\pi_{\mathsf{b}}(a))$. Combining the preceding two bounds with our previous expression (15) for $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$ yields the claimed bound (42a).

2) Proof of the Variance Bound (42b)*:* We now move onto the two variance terms. Since $\sigma_f^2(a) \le r_{\max}^2$, we can write

$$\begin{aligned}
&\sum_{a\in[k]} \pi_{\mathsf{t}}^2(a)\sigma_f^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(1)>0\}}{n(a)}\right] \\
&\le r_{\max}^2 \left\{\sum_{a\in S^\star} \pi_{\mathsf{t}}^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right] + \sum_{a\notin S^\star} \pi_{\mathsf{t}}^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right]\right\}.
\end{aligned}$$

Since $\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right] \le 1$, the first term on the right-hand side can be upper bounded as

$$\sum_{a\in S^\star} \pi_{\mathsf{t}}^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right] \le \sum_{a\in S^\star} \pi_{\mathsf{t}}^2(a).$$

From Lemma 1 of the paper [10], we have the bound $\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right] \le \frac{5}{n\pi_{\mathsf{b}}(a)}$, which implies the second term is upper bounded as

$$\sum_{a\notin S^\star} \pi_{\mathsf{t}}^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(a)>0\}}{n(a)}\right] \le \frac{5}{n}\sum_{a\notin S^\star} \pi_{\mathsf{t}}^2(a)\frac{1}{\pi_{\mathsf{b}}(a)}.$$

By combining these two inequalities, we conclude that

$$\sum_{a\in[k]} \pi_{\mathsf{t}}^2(a)\sigma_f^2(a)\mathbb{E}\left[\frac{\mathbb{1}\{n(1)>0\}}{n(a)}\right] \le c\,\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}), \tag{44}$$

for a universal constant c.

Turning to the second quantity in the variance, we have

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a)\mathbb{1}\{n(a)>0\}\right) \\
&\le \sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) r_f^2(a)(1-\pi_{\mathsf{b}}(a))^n \\
&\le r_{\max}^2 \left\{\sum_{a\in S^\star} \pi_{\mathsf{t}}^2(a)(1-\pi_{\mathsf{b}}(a))^n + \sum_{a\notin S^\star} \pi_{\mathsf{t}}^2(a)(1-\pi_{\mathsf{b}}(a))^n\right\} \\
&\overset{(i)}{\le} r_{\max}^2 \left\{\sum_{a\in S^\star} \pi_{\mathsf{t}}^2(a) + \sum_{a\notin S^\star} \pi_{\mathsf{t}}^2(a)\frac{1}{n\pi_{\mathsf{b}}(a)}\right\} \\
&\le c\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}),
\end{aligned}$$

where step (i) follows from the elementary inequalities $(1-\pi_{\mathsf{b}}(a))^n \le 1$ and $(1-\pi_{\mathsf{b}}(a))^n \le 1/(n\pi_{\mathsf{b}}(a))$. Combining this bound with our earlier inequality (44) yields the claimed bound (42b).

E. Proof of Theorem 3

We first show that the competitive ratio is lower bounded by 1. From the elementary inequality $\inf\sup \ge \sup\inf$, we see that

$$\begin{aligned}
&\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}},f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})} \\
&\quad \ge \sup_{\pi_{\mathsf{b}}} \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})} \overset{(i)}{=} 1,
\end{aligned}$$

where equality (i) follows from the definition (14) of $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$.

The remainder of our analysis is to prove a lower bound in terms of $s/\log k$. Throughout this analysis, we consider a target distribution π_{t} that is uniform over $\{1,2,\ldots,s\}$, and we consider the set

$$\Pi(\nu) := \{\pi_{\mathsf{b}} \mid \min_{a\in[k]} \pi_{\mathsf{b}}(a) \ge \nu\}, \quad \text{where } \nu := 1/(n\log k).$$

Since $\Pi(\nu)$ is a subset of all behavior policies, we have

$$\begin{aligned}
&\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}},f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})} \\
&\quad \ge \inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \frac{1}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2] \\
&\quad \ge \frac{\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]}{\sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})},
\end{aligned} \tag{45}$$

where in the second line we use the trivial upper bound $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) \le \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$ for any $\pi_{\mathsf{b}} \in \Pi(\nu)$.

In view of the lower bound (45), the proof can be decomposed into two steps: (1) Provide a lower bound on the numerator; and (2) Establish an upper bound on the denominator.

1) Step 1: The required lower bound on the numerator in equation (45) can be obtained by applying Theorem 5. More specifically, doing so with $\nu = 1/(n \log k)$ yields

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2] \gtrsim r_{\max}^2. \tag{46}$$

In light of this lower bound and equation (45), any upper bound on $\sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})$ yields a valid lower bound on the competitive ratio.

2) Step 2: We now turn to the upper bound on the denominator of equation (45). By combining Lemma 2 with the characterization (15) of $\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})$, we find that

$$\begin{aligned}
&\left[\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})/r_{\max}^2\right]^{1/2} \\
&\asymp \inf_{v \in \mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\},
\end{aligned}$$

which further implies

$$\begin{aligned}
&\sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \left[\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})/r_{\max}^2\right]^{1/2} \\
&\asymp \sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \inf_{v \in \mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\} \\
&\leq \inf_{v \in \mathbb{R}^k} \sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\}.
\end{aligned}$$

Here, the last inequality arises from the elementary fact $\sup \inf \leq \inf \sup$. Focusing on the inner maximization problem, we can easily see that

$$\begin{aligned}
&\sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \\
&\leq \sqrt{\frac{\log k}{8} \sum_{a \in [k]} [\pi_{\mathsf{t}}(a) - v(a)]^2} + \frac{1}{2} \sum_{a \in [k]} |v(a)|,
\end{aligned}$$

where we substitute in the definition of ν. Combine the previous two bounds together to reach

$$\begin{aligned}
&\sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \left[\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})/r_{\max}^2\right]^{1/2} \\
&\lesssim \inf_{v \in \mathbb{R}^k} \sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \left\{ \sqrt{\frac{\log k}{8} \sum_{a \in [k]} [\pi_{\mathsf{t}}(a) - v(a)]^2} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\} \\
&\leq \sqrt{\frac{\log k}{8s}},
\end{aligned}$$

where the final inequality follows by setting $v = 0$ and using the definition of π_{t}. Consequently, we have established the upper bound

$$\sup_{\pi_{\mathsf{b}} \in \Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}) \lesssim \frac{\log k}{s} \cdot r_{\max}^2. \tag{47}$$

Combining equation (45) with equations (46), and (47) yields the desired conclusion.

F. Proof of Theorem 4

Recall the decomposition (41) of the mean-squared error incurred by the plug-in estimator:

$$\begin{aligned}
&\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \\
&= \Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\
&\quad + \sum_{a \in [k]} \pi_{\mathsf{t}}^2(a) \sigma_f^2(a) \mathbb{E}\left[\tfrac{\mathbb{1}\{n(a) > 0\}}{n(a)}\right] \\
&\quad + \mathsf{Var}\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) \mathbb{1}\{n(a) > 0\} \Big).
\end{aligned} \tag{48a}$$

As shown in equation (42b), the variance components are well-behaved in the sense that

$$\begin{aligned}
&\sum_{a \in [k]} \pi_{\mathsf{t}}^2(a) \sigma_f^2(a) \mathbb{E}\left[\tfrac{\mathbb{1}\{n(a) > 0\}}{n(a)}\right] \\
&\quad + \mathsf{Var}\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) \mathbb{1}\{n(a) > 0\} \Big) \\
&\quad \leq c' \, \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})
\end{aligned} \tag{48b}$$

for some universal constant c'. As a result, we only need to focus on the squared bias term. Corresponding to the statements in Theorem 4, we split the proof into two cases: (1) ν is arbitrary, and (2) $\nu \geq \log k/n$.

1) Case 1: We begin with the case of an arbitrary ν. Since $\min_a \pi_{\mathsf{b}}(a) \geq \nu$, one has $(1 - \pi_{\mathsf{b}}(a))^n \leq \exp(-n\pi_{\mathsf{b}}(a)) \leq \exp(-n\nu)$. This combined with the fact that $|r_f(a)| \leq r_{\max}$ yields

$$\begin{aligned}
&\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\
&\quad \leq r_{\max}^2 \Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) \exp(-n\nu) \Big)^2 \\
&\quad = r_{\max}^2 \exp(-2n\nu),
\end{aligned} \tag{48c}$$

where the last equality arises from the fact that $\sum_{a \in [k]} \pi_{\mathsf{t}}(a) = 1$. Combining the bounds (48a), L(48b) and (48c) yields the claimed bound (25a).

2) Case 2: Now suppose that ν is lower bounded as $\nu \geq \log k/n$. By applying the bias bound (43), we find that

$$\begin{aligned}
&\Big(\sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\
&\leq 2 r_{\max}^2 \Big\{ \pi_{\mathsf{t}}^2(S^\star) + \Big(\sum_{a \in (S^\star)^c} \pi_{\mathsf{t}}(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \Big\}.
\end{aligned} \tag{49a}$$

We then apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
&\Big(\sum_{a \in (S^\star)^c} \pi_{\mathsf{t}}(a) (1 - \pi_{\mathsf{b}}(a))^n \Big)^2 \\
&\quad \leq k \cdot \sum_{a \in (S^\star)^c} \pi_{\mathsf{t}}^2(a) (1 - \pi_{\mathsf{b}}(a))^{2n}
\end{aligned}$$

$$\leq \sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}^2(a)(1-\pi_{\mathsf{b}}(a))^n$$
$$\leq \sum_{a\in(S^\star)^c} \pi_{\mathsf{t}}^2(a)\frac{1}{n\pi_{\mathsf{b}}(a)}. \tag{49b}$$

Here, the middle line follows from $(1-\pi_{\mathsf{b}}(a))^n \leq \exp(-n\pi_{\mathsf{b}}(a)) \leq \exp(-n\nu) \leq 1/k$, and the last inequality holds since $(1-\pi_{\mathsf{b}}(a))^n \leq \frac{1}{n\pi_{\mathsf{b}}(a)}$. Combining the bounds (48a), (48b), (49a), (49b) with the expression (15) of $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$, we arrive at the desired bound (25b).

G. Proof of Theorem 5

Without loss of generality, we may assume that the actions are ordered such that $\pi_{\mathsf{t}}(1) \geq \pi_{\mathsf{t}}(2) \geq \cdots \geq \pi_{\mathsf{t}}(k)$. First, observe that

$$\begin{aligned}\mathcal{R}_{\mathsf{M}}^\star(\pi_{\mathsf{t}},n,\nu) &= \inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}\in\Pi(\nu),f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2] \\ &\overset{\text{(i)}}{\geq} \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2] \\ &\overset{\text{(ii)}}{=} \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}),\end{aligned} \tag{50}$$

where the inequality (i) follows from the fact that $\inf\sup \geq \sup\inf$, whereas the equality (ii) uses the definition (14) of $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$. This yields the first term in the lower bound (27). In particular, if

$$\pi_{\mathsf{t}}^2(1)\exp(-2n\nu) \geq \exp(-200\sqrt{n\nu\log k}), \tag{51}$$

then the bound (50) tells us that

$$\begin{aligned}\mathcal{R}_{\mathsf{M}}^\star(\pi_{\mathsf{t}},n,\nu) &\geq \sup_{\pi_{\mathsf{b}}\in\Pi(\nu)} \mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) \\ &\overset{\text{(i)}}{\geq} \frac{1}{4}r_{\max}^2\pi_{\mathsf{t}}^2(1)(1-\nu)^n \\ &\overset{\text{(ii)}}{\geq} \frac{1}{4}r_{\max}^2\pi_{\mathsf{t}}^2(1)\exp(-2n\nu) \\ &\overset{\text{(iii)}}{\geq} \frac{1}{4}r_{\max}^2\exp(-200\sqrt{n\nu\log k}).\end{aligned}$$

Here, the first relation (i) uses Theorem 1 in the paper [10], the second inequality (ii) uses the elementary relation $(1-\nu)^n \geq e^{-2n\nu}$ for $\nu \in [0,\frac{1}{2}]$, while the last one (iii) arises from the condition (51). In words, when the largest mass $\pi_{\mathsf{t}}(1)$ in $\{\pi_{\mathsf{t}}(a)\}$ is sufficiently large (cf. the condition (51)), the term $\sup_{\pi_{\mathsf{b}}\in\Pi(\nu)}\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})$ dominates $r_{\max}^2\cdot\exp(-200\sqrt{n\nu\log k})$, and hence the desired lower bound (27) follows.

Therefore, in the remaining part of this section, we concentrate on establishing the lower bound for the case when

$$\pi_{\mathsf{t}}^2(1)\exp(-2n\nu) \leq \exp(-200\sqrt{n\nu\log k}). \tag{52}$$

We record an immediate consequence of the relation (52) that will be useful later

$$\sum_{a=1}^{k-1}\pi_{\mathsf{t}}^2(a) \leq \pi_{\mathsf{t}}(1) \leq \exp(n\nu-100\sqrt{n\nu\log k}). \tag{53}$$

In order to prove lower bound in this case, it is convenient to make use of a Poissonized model.

1) The Poissonized Model: Recall that in the multinomial observation model, $(n(1),n(2),\ldots,n(k))$ follows a multinomial distribution with parameters n and π_{b}. Here the dependence among the counts $\{n(a)\}_{a\in[k]}$ complicates the analysis. In order to sidestep this dependence, we consider the Poisson model in which each action a is taken with $n(a)$ times with $n(a) \sim \mathsf{Poi}(n\pi_{\mathsf{b}}(a))$ independently across actions. Then conditional on the count $n(a)=t$, a total of t rewards are observed independently for each action a. Note that in the Poissonized model, $\{n(a)\}$ are mutually independent by design, which greatly facilitates the analysis. Another difference that is worth pointing out is that in the original multinomial model, $\sum_{a\in[k]}\pi_{\mathsf{b}}(a)$ must sum to 1, while in the Poisson case, this restriction does not necessarily hold. To account for this fact, we define the following ε-relaxed probability simplex for a parameter $0\leq\varepsilon\leq 1$:

$$\Theta(\nu,\varepsilon) := \Big\{\pi_{\mathsf{b}} \geq \mathbf{0} \mid \min_{a\in[k]}\pi_{\mathsf{b}}(a)\geq\nu,\ \Big|\sum_{a\in[k]}\pi_{\mathsf{b}}(a)-1\Big| \leq \varepsilon\Big\}. \tag{54}$$

Correspondingly we can define the minimax risk over the relaxed parameter set under the Poisson model as

$$\mathcal{R}_{\mathsf{P}}^\star(\pi_{\mathsf{t}},n,\nu,\varepsilon) := \inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}\in\Theta(\nu,\varepsilon),f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2], \tag{55}$$

where we note that the expectation $\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[\cdot]$ is taken under the Poissonized model. It turns out that the risks under these two models, i.e., the multinomial model and the Poisson model, are closely related, as shown in the following lemma.

Lemma 4: For any $\varepsilon \in (0,1/4)$, the following relation holds:

$$\mathcal{R}_{\mathsf{P}}^\star(\pi_{\mathsf{t}},n,\nu,\varepsilon) \leq \mathcal{R}_{\mathsf{M}}^\star(\pi_{\mathsf{t}},n/2,\nu/(1+\varepsilon)) + e^{-3n/72}\cdot r_{\max}^2.$$

See Appendix D for the proof of this claim.

Lemma 4 shows that it suffices to establish a good lower bound on $\mathcal{R}_{\mathsf{P}}^\star(\pi_{\mathsf{t}},n,\nu,\varepsilon)$, and the remainder of our analysis focuses on this sub-problem.

2) The Bernoulli Reward Model: Recall that f can be any reward distribution supported on $[0,r_{\max}]$. For the purpose of the lower bound, we can restrict our attention to Bernoulli reward models, in which each action is associated with a Bernoulli reward distribution over $\{0,r_{\max}\}$ with the parameter $r(a)/r_{\max}$. Here $0\leq r(a)\leq r_{\max}$ is a parameter associated with the action a. This Bernoulli reward model, in conjunction with the Poisson sampling model yields two observations $\{P_a,N_a\}$ for each action a

$$P_a \sim \mathsf{Poi}(n\pi_{\mathsf{b}}(a)\tfrac{r(a)}{r_{\max}}),\ \text{and}\ N_a \sim \mathsf{Poi}(n\pi_{\mathsf{b}}(a)(1-\tfrac{r(a)}{r_{\max}})).$$

Here, P_a denotes the number of positive rewards (i.e., the rewards with value $r_{\max}$) obtained for arm a, while N_a denotes the number of "negative" rewards—meaning those with value 0. We naturally have

$$\mathcal{R}_{\mathsf{P}}^\star(\pi_{\mathsf{t}},n,\nu,\varepsilon) = \inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}\in\Theta(\nu,\varepsilon),f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]$$

$$\geq \inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}} \in \Theta(\nu,\varepsilon), \{r(a)\}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2], \tag{56}$$

where the set of numbers $\{r(a)\}$ dictates the Bernoulli reward model f.

3) A Useful Reparameterization: Now we introduce a reparameterization of the models introduced above. Let us denote

$$\nu_{\mathsf{p}}(a) := \pi_{\mathsf{b}}(a)\tfrac{r(a)}{r_{\max}}, \quad \text{and} \quad \nu_{\mathsf{n}}(a) := \pi_{\mathsf{b}}(a)(1 - \tfrac{r(a)}{r_{\max}}). \tag{57}$$

Using this notation, we can translate our target $V_f(\pi_{\mathsf{t}})$ into

$$\begin{aligned} T(\boldsymbol{\theta}) := V_f(\pi_{\mathsf{t}}) &= \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r(a) \\ &= r_{\max} \cdot \sum_{a\in[k]} \pi_{\mathsf{t}}(a) \frac{\nu_{\mathsf{p}}(a)}{\nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a)}, \end{aligned}$$

where we denote by $\boldsymbol{\theta} \in \mathbb{R}^{2K}$ the collection of parameters under the new parametrization, i.e.,

$$\boldsymbol{\theta} := \big(\nu_{\mathsf{p}}(1), \nu_{\mathsf{n}}(1), \nu_{\mathsf{p}}(2), \nu_{\mathsf{n}}(2), \cdots, \nu_{\mathsf{p}}(k), \nu_{\mathsf{n}}(k)\big)^{\top}.$$

Note that there is a one-to-one mapping between the original parameterization $\{\pi_{\mathsf{b}}(a), r(a)\}$ and the new one $\boldsymbol{\theta}$. Hence, with an abuse of notation, we shall denote

$$\begin{aligned} \Theta(\nu,\varepsilon) := \Big\{ \boldsymbol{\theta} \geq \mathbf{0} \mid \Big| \sum_{a\in[k]} \nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a) - 1 \Big| \leq \varepsilon, \\ \nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a) \geq \nu, \quad \text{for all } a \in [k] \Big\}. \end{aligned}$$

With this set of notation in place, the lower bound in equation (56) can be equivalently written as

$$\inf_{\widehat{V}} \sup_{\boldsymbol{\theta}\in\Theta(\nu,\varepsilon)} \mathbb{E}_{\boldsymbol{\theta}}[(\widehat{V}(\boldsymbol{Z}) - T(\boldsymbol{\theta}))^2], \tag{58}$$

where $\boldsymbol{Z} := (P_1, N_1, \cdots, P_k, N_k)^{\top} \in [n]^{2k}$ denotes the observations following the Poisson sampling and the Bernoulli reward models.

We intend to invoke Lemma 10 to obtain a good lower bound. It all boils down to constructing two prior distributions Ξ_0, Ξ_1 over the parameter space $\Theta(\nu,\varepsilon)$ such that the functional values $T(\boldsymbol{\theta})$ are well separated under different priors, while at the same time one cannot differentiate those two distributions based on the data alone.

4) A Construction of Two Priors Ξ_0, Ξ_1 Over $\Theta(\nu,\varepsilon)$: The construction of the two priors hinges on the existence of two random variables X, X', introduced in the following lemma.

Lemma 5: There exist two random variables X, X' supported on $[0,1]$ such that

$$\mathbb{E}\left[\frac{X}{X + \frac{n\nu}{8\log k}}\right] - \mathbb{E}\left[\frac{X'}{X' + \frac{n\nu}{8\log k}}\right] \geq \frac{1}{2}\exp\left(-96\sqrt{n\nu\log k}\right) \tag{59a}$$

$$\mathbb{E}[X] = \mathbb{E}[X'] = \frac{n\nu}{8\log k}; \tag{59b}$$

$$\mathbb{E}\left[X^j\right] = \mathbb{E}\left[(X')^j\right], \qquad \text{for all } 1 \leq j \leq \lceil 48\log k \rceil. \tag{59c}$$

See Appendix D for the proof of this claim.

Now we are ready to construct two "helper" priors Γ_0, Γ_1 on $\Theta(\nu,\varepsilon)$. Under both priors, we always set $\nu_{\mathsf{n}}(a) = \nu$ for $1 \leq a \leq k-1$, and $\nu_{\mathsf{n}}(k) = 0$. Under Γ_0, we let $X_1, \cdots, X_{k-1}$ be i.i.d. copies of X, and set $\nu_{\mathsf{p}}(a) = \frac{8\log k}{n} \cdot X_a$ for $1 \leq a \leq k-1$, $\nu_{\mathsf{p}}(k) = 1 - (k-1)\nu - (k-1)\frac{8\log k}{n}\mathbb{E}[X]$. Similarly, under Γ_1, we let $X'_1, \cdots, X'_{k-1}$ be i.i.d. copies of X', and set $\nu_{\mathsf{p}}(a) = \frac{8\log k}{n} \cdot X'_a$ for $1 \leq a \leq k-1$, $\nu_{\mathsf{p}}(k) = 1 - (k-1)\nu - (k-1)\mathbb{E}[X']$. It is straightforward to check that under both priors Γ_0 and Γ_1

$$\begin{aligned} \nu_{\mathsf{p}}(k) &= 1 - (k-1)\nu - (k-1)\frac{8\log k}{n} \cdot \frac{n\nu}{8\log k} \\ &= 1 - 2(k-1)\nu \geq \nu, \end{aligned}$$

as long as $\nu \leq \frac{1}{2k}$. Finally, for $i \in \{0,1\}$, we define the prior Ξ_i to be the push-forward measure of the restriction of Γ_i to the following set:

$$\begin{aligned} E_i := \Theta(\nu, 1/5) \\ \cap \left\{ \boldsymbol{\theta} \mid \big|T(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}\sim\Gamma_i}[T(\boldsymbol{\theta})]\big| \leq \frac{\exp(-96\sqrt{n\nu\log k})}{8} \cdot r_{\max} \right\}. \end{aligned} \tag{60}$$

5) Application of Le Cam's Method: Now we are positioned to invoke Le Cam's method, in the form of Lemma 10, with the choices

$$\begin{aligned} \xi &:= \frac{\mathbb{E}_{\Xi_0}[T(\boldsymbol{\theta})] + \mathbb{E}_{\Xi_1}[T(\boldsymbol{\theta})]}{2}, \\ s &:= \frac{\exp(-96\sqrt{n\nu\log k})}{16} \cdot r_{\max}, \quad \text{and} \\ \Theta &:= \Theta(\nu, 1/5). \end{aligned}$$

From the constructions of the priors Ξ_0 and Ξ_1, we have

$$\begin{aligned} &\mathbb{E}_{\Xi_0}[T(\boldsymbol{\theta})] - \mathbb{E}_{\Xi_1}[T(\boldsymbol{\theta})] \\ &= r_{\max} \cdot \mathbb{E}_{\Xi_0}\left[\sum_{a\in[k]} \pi_{\mathsf{t}}(a)\frac{\nu_{\mathsf{p}}(a)}{\nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a)}\right] \\ &\quad - r_{\max} \cdot \mathbb{E}_{\Xi_1}\left[\sum_{a\in[k]} \pi_{\mathsf{t}}(a)\frac{\nu_{\mathsf{p}}(a)}{\nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a)}\right] \\ &= r_{\max}\left(\sum_{1\leq a\leq k-1} \pi_{\mathsf{t}}(a)\right) \\ &\quad \cdot \left\{ \mathbb{E}\left[\frac{X}{X + \frac{n\nu}{8\log k}}\right] - \mathbb{E}\left[\frac{X'}{X' + \frac{n\nu}{8\log k}}\right] \right\} \\ &\geq \frac{1}{4} r_{\max} \exp\left(-96\sqrt{n\nu\log k}\right) = 4s, \end{aligned}$$

where the last inequality arises from the property of X, X' (cf. the inequality (59a)) as well as the fact that $\sum_{1\leq a\leq k-1} \pi_{\mathsf{t}}(a) \geq 1/2$. Consequently, $T(\boldsymbol{\theta}) \leq \xi - s$ almost surely for $\boldsymbol{\theta} \sim \Xi_1$, and $T(\boldsymbol{\theta}) \geq \xi + s$ almost surely for $\boldsymbol{\theta} \sim \Xi_0$, which immediately implies $\beta_0 = \beta_1 = 0$ in Lemma 10.

It remains to control the total variation distance between Ξ_0 and Ξ_1. To begin with, denoting $\Gamma'_i = \Gamma_i \circ (\boldsymbol{\theta}^{\otimes n})^{-1}$ for $i \in \{0,1\}$, the triangle inequality gives

$$\mathsf{TV}(\Xi_0, \Xi_1) \leq \mathsf{TV}(\Xi_0, \Gamma'_0) + \mathsf{TV}(\Gamma'_1, \Xi_1) + \mathsf{TV}(\Gamma'_0, \Gamma'_1)$$

$$\leq \Xi_0(E_0^c) + \Xi_1(E_1^c) + \mathsf{TV}(\Gamma_0', \Gamma_1').$$

Regarding the last term $\mathsf{TV}(\Gamma_0', \Gamma_1')$, we invoke Lemma 11 to obtain

$$\begin{aligned}&\mathsf{TV}(\Gamma_0', \Gamma_1')\\ &\leq (k-1)\mathsf{TV}(\mathbb{E}_X[\mathsf{Poi}(8\log k \cdot X)] - \mathbb{E}_{X'}[\mathsf{Poi}(8\log k \cdot X')])\\ &\leq \left(\frac{16e\log k}{48\log k}\right)^{48\log k} \leq \frac{1}{10}\end{aligned}$$

as long as the number k of actions is sufficiently large. For the remaining two terms $\Xi_0(E_0^c)+\Xi_1(E_1^c)$, we concentrate on the term $\Xi_0(E_0^c)$ and the same argument and bound apply to the term $\Xi_1(E_1^c)$. By definition, one has

$$\begin{aligned}\Xi_0(E_0^c) \leq &\mathbb{P}_{\Xi_0}\left(\left|\sum_{a\in[k]} \nu_{\mathsf{p}}(a)+\nu_{\mathsf{n}}(a)-1\right| \geq \frac{1}{5}\right)\\ &+ \mathbb{P}_{\Xi_0}\left(|\,T(\boldsymbol{\theta}) - \mathbb{E}_{\Xi_0}[T(\boldsymbol{\theta})]\,| \geq \frac{\exp(-96\sqrt{n\nu\log n}}{8}\cdot r_{\max}\right).\end{aligned} \tag{61}$$

In regard to the first term, one has

$$\sum_{a\in[k]} \nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a) - 1 = \frac{8\log k}{n}\cdot\Big(\sum_{a=1}^{k-1} X_a - \mathbb{E}[X_a]\Big),$$

which together with the Chebyshev's inequality gives

$$\begin{aligned}&\mathbb{P}_{\Xi_0}\left(\left|\sum_{a\in[k]} \nu_{\mathsf{p}}(a) + \nu_{\mathsf{n}}(a) - 1\right| \geq \frac{1}{5}\right)\\ &= \mathbb{P}\left(\left|\frac{8\log k}{n}\cdot\Big(\sum_{a=1}^{k-1} X_a - \mathbb{E}[X_a]\Big)\right| \geq \frac{1}{5}\right)\\ &\leq \frac{k40^2\log^2 k\mathsf{Var}(X)}{n^2}\\ &\leq \frac{k40^2\log^2 k}{n^2} \leq \frac{1}{10}.\end{aligned}$$

Here the penultimate inequality follows from the fact that $X \in [0,1]$ and hence $\mathsf{Var}(X) \leq 1$, and the last relation holds as long as $k \gg \sqrt{k}\log k$.

Moving on to the second term in the equation (61), we have via the Chebyshev's inequality that

$$\begin{aligned}&\mathbb{P}_{\Xi_0}\left(|\,T(\boldsymbol{\theta}) - \mathbb{E}_{\Xi_0}[T(\boldsymbol{\theta})]\,| \geq \frac{\exp(-96\sqrt{n\nu\log k}}{8}\cdot r_{\max}\right)\\ &\leq \frac{8^2\mathsf{Var}\left[\sum_{1\leq a\leq k-1}\pi_{\mathsf{t}}(a)\cdot\frac{X_a}{X_a+\frac{n\nu}{8\log k}}\right]}{\exp(-96\sqrt{n\nu\log k})}\\ &\leq \frac{8^2\sum_{1\leq a\leq k-1}\pi_{\mathsf{t}}^2(a)}{\exp(-96\sqrt{n\nu\log k})}.\end{aligned}$$

Recall that we are working under the assumption (52) in which the restriction (53) holds. This observation leads to

$$\begin{aligned}&\mathbb{P}_{\Xi_0}\left(|T(\boldsymbol{\theta}) - \mathbb{E}_{\Xi_0}[T(\boldsymbol{\theta})]| \geq \frac{\exp(-96\sqrt{n\nu\log k}}{8}\cdot r_{\max}\right)\\ &\leq 8^2\pi_{\mathsf{t}}(1)\exp(96\sqrt{n\nu\log k})\\ &\leq 8^2\exp(n\nu - 4\sqrt{n\nu\log k})\\ &\leq 8^2\exp(-3\sqrt{n\nu\log k}) \leq \frac{1}{10}.\end{aligned}$$

Here the last line holds under the assumption that $\nu \leq \frac{\log k}{n}$ and $n\nu\log k \gg 1$.

In all, we have arrived at the conclusion that

$$\mathsf{TV}(\Xi_0, \Xi_1) \leq \frac{3}{10},$$

which allows us to combine Lemma 10 and Lemma 4 to finish the proof.

H. Proof of Theorem 6

Recall that we work under the Poisson sampling model, and hence throughout this section we use the shorthand notation $\mathbb{E}[\cdot]$ to denote expectation under the Poisson model.

Denoting by $\mathsf{Bias}(\widehat{V}_{\mathsf{C}}) := \mathbb{E}[\widehat{V}_{\mathsf{C}}] - V_f(\pi_{\mathsf{t}})$ the bias of the Chebyshev estimator, we have the usual bias-variance decomposition—namely

$$\mathbb{E}\left[\left(\widehat{V}_{\mathsf{C}} - V_f(\pi_{\mathsf{t}})\right)^2\right] = \left(\mathsf{Bias}(\widehat{V}_{\mathsf{C}})\right)^2 + \mathsf{Var}(\widehat{V}_{\mathsf{C}}).$$

As before, we break the analysis into two parts, namely controlling the bias and variance, and aim at proving the following two bounds:

$$\left(\mathsf{Bias}(\widehat{V}_{\mathsf{C}})\right)^2 \leq 16 r_{\max}^2\exp(-2\sqrt{\tfrac{c_0^2}{c_1}n\nu\log k}), \quad \text{and} \tag{62a}$$

$$\mathsf{Var}(\widehat{V}_{\mathsf{C}}) \leq c'\{\mathcal{R}_n(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) + c_0 r_{\max}^2\log k\cdot k^{4c_0-\gamma}\}, \tag{62b}$$

with $c' > 0$ a universal constant. Taking the above two bounds together, we can deduce that

$$\begin{aligned}&\mathbb{E}\left[\left(\widehat{V}_{\mathsf{C}} - V_f(\pi_{\mathsf{t}})\right)^2\right]\\ &\leq 16 r_{\max}^2\exp(-2\sqrt{\tfrac{c_0^2}{c_1}n\nu\log k})\\ &\quad + c'\mathcal{R}_n(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) + c'c_0 r_{\max}^2\log k\cdot k^{4c_0-\gamma}\\ &\leq c\left\{r_{\max}^2\exp(-2\sqrt{\tfrac{c_0^2}{c_1}n\nu\log k}) + \mathcal{R}_n(\pi_{\mathsf{t}};\pi_{\mathsf{b}})\right\},\end{aligned}$$

as long as $c_0 \leq \gamma/7$, and $c > 0$ is an absolute constant. Taking the supremum over π_{b} completes the proof of Theorem 6.

The remaining two sections are devoted to establishing the bounds (62).

1) Proof of the Bias Bound (62a)*:* It is easily seen from the definition (30) of the Chebyshev estimator that

$$\begin{aligned}\mathbb{E}[\widehat{V}_{\mathsf{C}}] &= \sum_{a\in[k]}\pi_{\mathsf{t}}(a)\mathbb{E}\left[\widehat{r}(a)g_L(n(a))\right]\\ &= \sum_{a\in[k]}\pi_{\mathsf{t}}(a)\sum_{j=0}^{\infty}\mathbb{E}\left[\widehat{r}(a)g_L(n(a)) \mid n(a)=j\right]\mathbb{P}\left(n(a)=j\right)\\ &= \sum_{a\in[k]}\pi_{\mathsf{t}}(a)r_f(a)\sum_{j=0}^{\infty}g_L(j)\mathbb{P}\left(n(a)=j\right),\end{aligned}$$

where the last relation hinges on the fact that $g_L(0) = 0$. Consequently, the bias of $\widehat{V}_{\mathsf{C}}$ is given by

$$\mathsf{Bias}(\widehat{V}_{\mathsf{C}}) = \sum_{a\in[k]}\pi_{\mathsf{t}}(a)r_f(a)\left\{\sum_{j=0}^{\infty}g_L(j)\mathbb{P}\left(n(a)=j\right) - 1\right\}$$

$$= \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a) \left\{ \sum_{j=0}^{\infty} (g_L(j)-1)\, \mathbb{P}\left(n(a)=j\right) \right\}$$
$$= \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a) \left\{ e^{-n\pi_{\mathsf{b}}(a)} \sum_{j=0}^{L} a_j (\pi_{\mathsf{b}}(a))^j \right\}$$
$$= \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r_f(a) e^{-n\pi_{\mathsf{b}}(a)} P_L(\pi_{\mathsf{b}}(a)). \tag{63}$$

Here the middle line uses the fact that $\sum_{j=0}^{\infty} \mathbb{P}\left(n(a)=j\right) = 1$, and the last one follows from the definitions of $g_L(j)$ and $P_L(\pi_{\mathsf{b}}(a))$.

In light of equation (63), the key in bounding the bias is to control $e^{-n\pi_{\mathsf{b}}(a)} P_L(\pi_{\mathsf{b}}(a))$, which is supplied in the following lemma.

Lemma 6: For any $\pi_{\mathsf{b}}(a) \geq \nu$, one has

$$\left| e^{-n\pi_{\mathsf{b}}(a)} P_L(\pi_{\mathsf{b}}(a)) \right| \leq 4 \exp\left(-L\sqrt{\ell/r}\right).$$

See Appendix E-A for the proof of this claim.

In all, this leads us to conclude that

$$|\mathsf{Bias}(\widehat{V}_{\mathsf{C}})| \leq 4 r_{\max} \sum_{a\in[k]} \pi_{\mathsf{t}}(a) \exp\left(-L\sqrt{\ell/r}\right)$$
$$\leq 4 r_{\max} \exp\left(-\sqrt{\tfrac{c_0^2}{c_1} n\nu \log k}\right),$$

where we have used the definitions $\ell = \nu$, $r = c_1 \log k/n$ as well as the relation $|r_f(a)| \leq r_{\max}$. This establishes the bias upper bound (62a).

2) Proof of the Variance Bound (62b)*:* Now we move on to the variance of the Chebyshev estimator $\widehat{V}_{\mathsf{C}}$. Thanks to the independence brought by the Poisson model, we have

$$\mathsf{Var}(\widehat{V}_{\mathsf{C}}) = \mathsf{Var}\Big(\sum_{a\in[k]} \pi_{\mathsf{t}}(a) \widehat{r}(a) g_L(n(a)) \Big)$$
$$= \sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) \mathsf{Var}\left(\widehat{r}(a) g_L(n(a))\right). \tag{64}$$

Applying the law of total variance yields the decomposition

$$\mathsf{Var}\left(\widehat{r}(a) g_L(n(a))\right)$$
$$= \underbrace{\mathsf{Var}\left(\mathbb{E}\left[\widehat{r}(a) g_L(n(a)) \mid n(a)\right]\right)}_{=:\alpha_1}$$
$$+ \underbrace{\mathbb{E}\left[\mathsf{Var}\left(\widehat{r}(a) g_L(n(a)) \mid n(a)\right)\right]}_{=:\alpha_2}. \tag{65}$$

Suppose for the moment that the two terms α_1 and α_2 obey (whose proof are deferred to Appendix E-B)

$$\alpha_1 \leq r_{\max}^2 \left\{ 2\, e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2} c_0 \log k \cdot k^{4c_0} \right\}, \quad \text{and} \tag{66a}$$
$$\alpha_2 \leq r_{\max}^2 \left\{ 2e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2} c_0 \log k \cdot k^{4c_0} + 2\min\left\{1, \frac{5}{n\pi_{\mathsf{b}}(a)}\right\} \right\}. \tag{66b}$$

Then combing the preceding bounds together yields

$$\mathsf{Var}(\widehat{V}_{\mathsf{C}}) \leq r_{\max}^2 \sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) \Big\{ 2\, e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2} c_0 \log k \cdot k^{4c_0}$$
$$+ 2\min\left\{1, \frac{5}{n\pi_{\mathsf{b}}(a)}\right\} \Big\}$$
$$\leq r_{\max}^2 \sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) \left\{ \frac{1}{2} c_0 \log k \cdot k^{4c_0} + 4\min\left\{1, \frac{5}{n\pi_{\mathsf{b}}(a)}\right\} \right\},$$

where the last inequality follows from the elementary bound $e^{-n\pi_{\mathsf{b}}(a)} \leq \min\left\{1, \frac{1}{n\pi_{\mathsf{b}}(a)}\right\}$. Repeating the analysis of the plug-in estimator, we find that

$$r_{\max}^2 \sum_{a\in[k]} \pi_{\mathsf{t}}^2(a) \min\left\{1, \frac{5}{n\pi_{\mathsf{b}}(a)}\right\} \leq c\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}),$$

for some constant $c > 0$. This bound combined with the assumption $\sum_a \pi_{\mathsf{t}}^2(a) \leq k^{-\gamma}$ implies another positive constant $c' > 0$ such that

$$\mathsf{Var}(\widehat{V}_{\mathsf{C}}) \leq c' \{ \mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}}) + c_0 r_{\max}^2 \log k \cdot k^{4c_0-\gamma} \},$$

which finishes the proof of the variance upper bound (62b).

V. Discussion

In this paper, we have studied the off-policy evaluation problem for multi-armed bandits with bounded rewards in three different settings. First, when the behavior policy is known, we showed that the Switch estimator, which interpolates between the plug-in and importance sampling estimator, is minimax optimal. Second, when the behavior policy is unknown, we analyzed performance in terms of a competitive ratio, and showed that the plug-in estimator is near-optimal. Third, we took some initial steps into the intermediate regime, when partial knowledge of the behavior policy is given in the form of the minimum probability over all actions. We showed that the plug-in approach, while optimal in some regimes, can be sub-optimal, and we developed an estimator based on Chebyshev polynomials that is provably optimal for a large family of target distributions.

This paper focused purely on multi-armed bandits, and extending non-asymptotic analysis of this type to contextual bandits and Markov decision processes is certainly of interest. In addition to such extensions, our study leaves a few interesting technical questions to answer. Let us single out three of them to conclude.

A. Extension to Other Reward Distributions

Our focus throughout the paper has been on the family (1) of reward distributions with bounded support. In practice, one might encounter distributions with possibly unbounded support but controlled moments (e.g., sub-Gaussian and sub-exponential distributions), or bounds on variance or other moments. In these more general settings, it is not *a priori* clear that a linear[2] procedure, such as the Switch estimator, need be optimal. However, we believe that the underlying idea of truncating the likelihood ratio should be useful in general. From a technical perspective, the set of bounded reward distributions is convex, which allows us demonstrate that Bernoulli rewards are the hardest instances within this family;

[2]To be clear, the Switch estimator is linear with respect to the observed rewards.

see the proof of the lower bound in Theorem 1. If we move beyond bounded rewards, the set of reward distributions can be non-convex in general, which introduces new challenges.

B. Known and Unknown π_{b} Cases

Our current characterization of the gap between these two cases relies on the support size of the target policy, which allows us to demonstrate the near-optimality of the plug-in estimator in the unknown π_{b} case. However, the support size is a discontinuous function of the target distribution, which makes it sensitive to small perturbations. Is it possible to characterize the gap using a smooth function of the target distribution? Examining the proof of the lower bound for the competitive ratio, the quantity $1/(\sum_{a\in[k]}\pi_{\mathsf{t}}^2(a))$ appears to be a plausible "soft" alternative to the support size. It remains to be seen whether the plug-in estimator satisfies an upper bound in terms of this alternative quantity.

C. Adaptivity to the Minimum Exploration Probability

The Chebyshev estimator proposed in this paper requires the knowledge of the minimum exploration probability $\min_{a\in[k]}\pi_{\mathsf{b}}(a)$. In practice, this minimum probability may not be known. A natural question, then, is whether it is possible devise an estimator that adapts to this minimum probability—that is, exhibits the same optimal behavior without knowing the minimum probability in advance. If not, what is the price for adaptivity?

APPENDIX A
PROOF OF LEMMA 1

We note that the proof of this result follows that of Lemma 1 in the paper [27]; we include the details here for completeness. The minimax risk $\mathcal{R}_{\mathsf{P}}(\pi_{\mathsf{t}},(1-\beta)n,\nu)$ can be rewritten as

$$\mathcal{R}_{\mathsf{P}}(\pi_{\mathsf{t}},(1-\beta)n,\nu) = \inf_{\{\widehat{V}_m\}}\sup_{\pi_{\mathsf{b}}\in\Pi(\nu),f\in\mathcal{F}}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{n'}-V_f(\pi_{\mathsf{t}}))^2],$$

where $\{\widehat{V}_m\}_{m\geq 0}$ denotes a family of estimators corresponding to the sample size m, and $n'\sim\mathsf{Poi}((1-\beta)n)$. Using the Bayes risk as a lower bound of the minimax risk, we have

$$\mathcal{R}_{\mathsf{P}}(\pi_{\mathsf{t}},(1-\beta)n,\nu)\geq\sup_{\Omega}\inf_{\{\widehat{V}_m\}}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{n'}-V_f(\pi_{\mathsf{t}}))^2],$$

where Ω is a prior on the parameter space $\Pi(\nu)\times\mathcal{F}$. Note that for any sequence of estimators $\{\widehat{V}_m\}$,

$$\begin{aligned}&\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{n'}-V_f(\pi_{\mathsf{t}}))^2]\\&=\sum_{m\geq 0}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_m-V_f(\pi_{\mathsf{t}}))^2\mid n'=m]\mathbb{P}[n'=m]\\&\geq\sum_{m=0}^{n}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_m-V_f(\pi_{\mathsf{t}}))^2]\mathbb{P}[n'=m].\end{aligned}$$

Taking the infimum on both sides yields

$$\begin{aligned}&\inf_{\{\widehat{V}_m\}}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{n'}-V_f(\pi_{\mathsf{t}}))^2]\\&\geq\sum_{m=0}^{n}\inf_{\widehat{V}_m}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_m-V_f(\pi_{\mathsf{t}}))^2]\mathbb{P}[n'=m]\end{aligned}$$

Observe that for any fixed prior Ω, the mapping $m\mapsto\inf_{\widehat{V}_n}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_m-V_f(\pi_{\mathsf{t}}))^2]$ is decreasing in m, and hence

$$\begin{aligned}&\inf_{\{\widehat{V}_m\}}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_{n'}-V_f(\pi_{\mathsf{t}}))^2]\\&\geq\sum_{m=0}^{n}\inf_{\widehat{V}_n}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_n-V_f(\pi_{\mathsf{t}}))^2]\mathbb{P}[n'=m]\\&=\inf_{\widehat{V}_n}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_n-V_f(\pi_{\mathsf{t}}))^2]\mathbb{P}(n'\leq n)\\&\geq\inf_{\widehat{V}_n}\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_n-V_f(\pi_{\mathsf{t}}))^2](1-\exp(-n\beta^2/2)).\end{aligned}$$

Taking the supremum over all possible priors on both sides and invoking the minimax theorem (cf. Theorem 46.5 in the book [37]) conclude the proof.

APPENDIX B
AUXILIARY RESULTS UNDERLYING PROPOSITION 2

In this section, we prove various auxiliary results that underlie the proof of Proposition 2, including Lemma 2, used in the proof itself, as well as Lemma 7, which is used to prove Lemma 2.

A. Proof of Lemma 2

To begin with, we make a few simple observations regarding the optimization problem (37) and the desired equivalence (35).

- First, for any action $\pi_{\mathsf{t}}(a)=0$, one must have $v^\star(a)=0$. At the same time, $\pi_{\mathsf{t}}(a)=0$ implies $\rho(a)=0$. Therefore on both sides of the equation (35), the contributions from actions a with $\pi_{\mathsf{t}}(a)=0$ are zero. Consequently, without loss of generality, we assume that $\pi_{\mathsf{t}}(a)>0$ for all $a\in[k]$.
- Second, if $\pi_{\mathsf{b}}(a)=0$ for some $a\in[k]$, then one must have $v^\star(a)=\pi_{\mathsf{t}}(a)>0$, which further implies $a\in S^\star$. As a result, the action a contributes $\pi_{\mathsf{t}}(a)$ to both sides of the equation (35). Consequently, we assume without loss of generality that $\pi_{\mathsf{b}}(a)>0$ for all $a\in[k]$.
- Last but not least, it is straightforward to check that $\mathbf{0}\leq v^\star\leq\pi_{\mathsf{t}}$.

In what follows, we separate the proof into three cases: (1) $v^\star=\pi_{\mathsf{t}}$, (2) $v^\star=\mathbf{0}$, and (3) $\mathbf{0}\neq v^\star\neq\pi_{\mathsf{t}}$. The desired equivalence (35) is easy to obtain for the first two cases, while it requires more effort for the last one.

Let us start with the easy cases.

1) Case 1: If $v^\star=\pi_{\mathsf{t}}$, then $S^\star=[k]$, and

$$\begin{aligned}&\min_{\boldsymbol{v}\in\mathbb{R}^k}\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}}+\frac{1}{2}\sum_{a\in[k]}|v(a)|\\&=\frac{1}{2}\sum_{a\in[k]}|\pi_{\mathsf{t}}(a)|=\frac{1}{2}.\end{aligned}$$

At the same time, the right hand side of (35) reads

$$\pi_{\mathsf{t}}(S^\star)+\sqrt{\frac{\sum_{a\notin S^\star}\pi_{\mathsf{b}}(a)\rho^2(a)}{n}}=\pi_{\mathsf{t}}(S^\star)=1.$$

This establishes the claim for the case when $v^\star=\pi_{\mathsf{t}}$.

2) Case 2: If $v^\star = 0$, then $S^\star = \emptyset$, and hence

$$\min_{v\in\mathbb{R}^k} \sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}}+\frac{1}{2}\sum_{a\in[k]}|v(a)| = \sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)]^2}{\pi_{\mathsf{b}}(a)}}.$$

On the other hand, $S^\star = \emptyset$ implies

$$\pi_{\mathsf{t}}(S^\star)+\sqrt{\frac{\sum_{a\notin S^\star}\pi_{\mathsf{b}}(a)\rho^2(a)}{n}} = \sqrt{\frac{1}{n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)]^2}{\pi_{\mathsf{b}}(a)}},$$

which matches desired the equivalence (35).

3) Case 3: In the end, we focus on the more challenging case when $0 \neq v^\star \neq \pi_{\mathsf{t}}$. In view of the optimality condition of the optimization problem (37), we know that

$$\left[\rho(a)-\frac{v^\star(a)}{\pi_{\mathsf{b}}(a)}\right]^2 = 2n\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)}, \qquad \text{for } a\in S^\star; \tag{67a}$$

$$(\rho(a))^2 \le 2n\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)}, \qquad \text{for } a\notin S^\star. \tag{67b}$$

To simplify the notation hereafter, we denote

$$T_1 := \sum_{a\in S^\star}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)}, \quad \text{and}$$
$$T_2 := \sum_{a\notin S^\star}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)} = \sum_{a\notin S^\star}\pi_{\mathsf{b}}(a)\rho^2(a).$$

A few immediate consequences of the optimality condition is summarized in the following claim, whose proof is deferred to the end of this section.

Lemma 7: Suppose that $0 \neq v^\star \neq \pi_{\mathsf{t}}$ is the minimizer of the optimization problem (37). Then the following conclusions hold:

1) There exists some quantity $\varepsilon \in (0,1)$ such that $\pi_{\mathsf{b}}(S^\star) = (1-\varepsilon)/(2n)$.
2) We have the relation $T_1 = \frac{1-\varepsilon}{\varepsilon}T_2$.
3) The convex program (37) has optimal value $\frac{1}{2}\pi_{\mathsf{t}}(S^\star)+\sqrt{\frac{T_2}{8n}}\,\varepsilon$.

See Appendix B-B for the proof of this claim.

We now use Lemma 7 to establish the equivalence (35) for the third case $0 \neq v^\star \neq \pi_{\mathsf{t}}$. Part (c) of Lemma 7 guarantees that then the desired equivalence (35) holds for any $\varepsilon \in [1/2, 1)$. Therefore, the remainder of our analysis is devoted to the case when $\varepsilon \in (0, 1/2)$, meaning that $\frac{1}{4n} < \pi_{\mathsf{b}}(S^\star) < \frac{1}{2n}$.

Without loss of generality, we assume that the actions are ordered according to their likelihood ratios—that is, $\rho(1) \le \rho(2) \le \cdots \le \rho(k)$. In view of the optimality condition (67) and the restriction $0 < \frac{1}{4n} \le \pi_{\mathsf{b}}(S^\star) \le \frac{1}{2n} < 1$, the subset $(S^\star)^c$ must be of the form $\{1, 2, \ldots, t\}$ for some $t \in [k-1]$, and hence $S^\star = \{t+1, \ldots, k\}$. In words, the support set $S^\star$ contains the actions with larger likelihood ratios $\rho(a)$; see the optimality condition (67). By applying the first optimality condition (67a), we find that

$$\rho^2(t+1) \ge \left[\rho(t+1)-\frac{u^\star(t+1)}{\pi_{\mathsf{b}}(t+1)}\right]^2 = 2n(T_1+T_2) = \frac{2nT_2}{\varepsilon},$$

where the first relation follows from observation that $0 \le u^\star(t+1) \le \pi_{\mathsf{t}}(t+1)$ and the last equality arises from Lemma 7(b). In addition, note that

$$\frac{\pi_{\mathsf{t}}(S^\star)}{\pi_{\mathsf{b}}(S^\star)} = \frac{\sum_{a=t+1}^k \pi_{\mathsf{t}}(a)}{\sum_{a=t+1}^k \pi_{\mathsf{b}}(a)} \ge \frac{\pi_{\mathsf{t}}(t+1)}{\pi_{\mathsf{b}}(t+1)} = \rho(t+1).$$

Combining the previous two bounds yields

$$\frac{\pi_{\mathsf{t}}(S^\star)}{\pi_{\mathsf{b}}(S^\star)} \ge \rho(t+1) \ge \sqrt{\frac{2nT_2}{\varepsilon}}.$$

This inequality, together with the assumption that $\pi_{\mathsf{b}}(S^\star) \ge 1/(4n)$, guarantees that

$$\pi_{\mathsf{t}}(S^\star) \ge \sqrt{\frac{2n}{\varepsilon}T_2}\pi_{\mathsf{b}}(S^\star) \ge \sqrt{\frac{T_2}{8n\varepsilon}} \ge \sqrt{\frac{T_2}{4n}} \ge \sqrt{\frac{T_2}{4n}}\varepsilon.$$

Here the assumption that $\varepsilon \in (0, 1/2)$ is repeatedly used. This together with Lemma 7(c) leads to the conclusion that

$$\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)}}+\frac{1}{2}\sum_{a\in[k]}|v^\star(a)| = \frac{1}{2}\pi_{\mathsf{t}}(S^\star)+\sqrt{\frac{T_2}{8n}}\varepsilon \asymp \pi_{\mathsf{t}}(S^\star) \asymp \pi_{\mathsf{t}}(S^\star)+\sqrt{\frac{T_2}{n}}.$$

As a result, in the case when $\varepsilon \in (0, 1/2)$, the target equivalence (35) follows.

B. Proof of Lemma 7

Summing the first optimality condition (67a) over actions in $S^\star$ yields

$$T_1 = \sum_{a\in S^\star}\pi_{\mathsf{b}}(a)\left[\rho(a)-\frac{v^\star(a)}{\pi_{\mathsf{b}}(a)}\right]^2 = 2n\pi_{\mathsf{b}}(S^\star)\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)} = 2n\pi_{\mathsf{b}}(S^\star)(T_1+T_2), \tag{68}$$

which implies $(S^\star)^c \neq \emptyset$. To see this, assume for the moment that $(S^\star)^c = \emptyset$ and hence $S^\star = [k]$, $T_2 = 0$. The relation above then reduces to

$$T_1 = 2nT_1,$$

which requires $T_1 = 0$ and hence $v^\star = \pi_{\mathsf{t}}$. This contradicts the assumption that $v^\star \neq \pi_{\mathsf{t}}$. Since $(S^\star)^c$ is non-empty and $\rho(a) > 0$, one must have $T_2 > 0$. In addition, since $v^\star \neq 0$, $S^\star$ is nonempty ($\pi_{\mathsf{b}}(S^\star) > 0$), which together with the identity (68) reveals that

$$T_1 > 2n\pi_{\mathsf{b}}(S^\star)T_1.$$

This readily gives the first claim that

$$\pi_{\mathsf{b}}(S^\star) = (1-\varepsilon)\tfrac{1}{2n}$$

for some $\varepsilon \in (0,1)$. With this representation in place, we can also deduce from the equation (68) that

$$T_1 = \frac{\pi_{\mathsf{b}}(S^\star)}{\frac{1}{2n} - \pi_{\mathsf{b}}(S^\star)} T_2 = \frac{1-\varepsilon}{\varepsilon} T_2,$$

which is the second claim. Regarding the last claim, applying the first optimality condition (67a) ensures that for $a \in S^\star$, we have

$$v^\star(a) = \pi_{\mathsf{t}}(a) - \pi_{\mathsf{b}}(a)\sqrt{2n(T_1+T_2)}.$$

As a result, the minimum value obeys

$$\begin{aligned}
&\sqrt{\frac{1}{8n}\sum_{a\in[k]} \frac{[\pi_{\mathsf{t}}(a)-v^\star(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]} |v^\star(a)| \\
&= \sqrt{\frac{1}{8n}}\sqrt{T_1+T_2} + \frac{1}{2}\sum_{a\in S^\star} v^\star(a) \\
&= \sqrt{\frac{1}{8n}}\sqrt{T_1+T_2} + \frac{1}{2}\left(\pi_{\mathsf{t}}(S^\star) - \pi_{\mathsf{b}}(S^\star)\sqrt{2n(T_1+T_2)}\right) \\
&= \frac{1}{2}\pi_{\mathsf{t}}(S^\star) + \sqrt{\frac{1}{8n}}\sqrt{T_1+T_2}\,(1-2n\pi_{\mathsf{b}}(S^\star)).
\end{aligned}$$

Use the first two claims $T_1+T_2 = T_2/\varepsilon$ and $1-2n\pi_{\mathsf{b}}(S^\star) = \varepsilon$ to finish the proof.

Appendix C
Proof of Lemma 3

In this appendix, we derive the dual formulation of the primal problem (39). First, note that we may assume without loss of generality that $\pi_{\mathsf{b}}(a) > 0$. Indeed, if $\pi_{\mathsf{b}}(a) = 0$ for some action a, then the optimal primal variable $\delta(a)$ in the primal problem (39) should be set to $1/2$, while the optimal dual variable $v(a)$ should be $\pi_{\mathsf{t}}(a)$ in the dual formulation (40). Both contribute $\frac{1}{2}\pi_{\mathsf{t}}(a)$ to the objective values. For a scalar $\lambda \geq 0$ and vector $v \geq 0$, the Lagrangian of the primal problem (39) is given by

$$\begin{aligned}
\mathcal{L}(\delta,\lambda,v) &:= -\sum_{a\in[k]} \pi_{\mathsf{t}}(a)\delta(a) \\
&\quad + \lambda\Big(\sum_{a\in[k]} \pi_{\mathsf{b}}(a)\delta^2(a) - \frac{1}{8n}\Big) + \sum_{a\in[k]} v(a)(\delta(a)-\frac{1}{2}) \\
&= -\frac{\lambda}{8n} + \sum_{a\in[k]}\left\{\lambda\pi_{\mathsf{b}}(a)\delta^2(a) + [v(a)-\pi_{\mathsf{t}}(a)]\delta(a) - \frac{1}{2}v(a)\right\}.
\end{aligned}$$

We now compute the dual function $g(\lambda, v) = \inf_{\boldsymbol{\delta}\in\mathbb{R}^k} \mathcal{L}(\delta,\lambda,v)$, and find that

$$g(\lambda,v) = \begin{cases} -\frac{\lambda}{8n} - \frac{1}{4\lambda}\sum_{a\in[k]} \frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)} - \frac{1}{2}, \sum_{a\in[k]} v(a), & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } \lambda > 0, & \\ -\frac{1}{2} \qquad\qquad \text{if } \lambda = 0 \text{ and } v = \pi_{\mathsf{t}}, \text{ and} & \\ -\infty, \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

The value in the first case follows by choosing the optimal $\delta^\star(a) = \frac{\pi_{\mathsf{t}}(a)-v(a)}{2\lambda\pi_{\mathsf{b}}(a)}$.

Since the primal problem (39) satisfies Slater's condition, strong duality holds and hence

$$\begin{aligned}
-\sum_{a\in[k]} \pi_{\mathsf{t}}(a)\delta^\star(a) &= \max_{\lambda\geq 0, v\geq \mathbf{0}} g(\lambda,v) \\
&= \max_{v\geq 0}\left\{-\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}} - \frac{1}{2}\sum_{a\in[k]} v(a)\right\} \\
&= -\min_{v\geq 0}\left\{\sqrt{\frac{1}{8n}\sum_{a\in[k]}\frac{[\pi_{\mathsf{t}}(a)-v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]} v(a)\right\}.
\end{aligned}$$

The constrained optimization problem on the right hand side is equivalent to the unconstrained one (40), which completes the proof.

Appendix D
Proof of Auxiliary Lemmas for Theorem 3

In this section, we collect the proofs of various auxiliary lemmas used in the proof of Theorem 3.

A. Proof of Lemma 4

For each positive integer ℓ, let $\widehat{V}_\ell$ be the optimal estimator under the multinomial sampling model based on ℓ samples, one that achieves the minimax risk $\mathcal{R}^\star_{\mathsf{M}}(\pi_{\mathsf{t}}, \ell, \nu/(1+\varepsilon))$. Now we define a near-optimal estimator in the Poisson sampling model. Let T be the total number of rewards observed in the Poissonized model; by construction, the random variable T follows a $\mathsf{Poi}(n\sum_{a=1}^k \pi_{\mathsf{b}}(a))$ distribution.

Now consider the estimator $\widehat{V}_T$—that is, the minimax optimal estimator based on T samples. This choice yields an upper bound on the Poissonized risk, namely

$$\begin{aligned}
&\mathcal{R}^\star_{\mathsf{P}}(\pi_{\mathsf{t}}, n, \nu, \varepsilon) \\
&\quad\leq \sup_{\pi_{\mathsf{b}}\in\Theta(\nu,\varepsilon), f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}\Big[(\widehat{V}_T - \sum_{a\in[k]}\pi_{\mathsf{t}}(a) r_f(a))^2\Big].
\end{aligned}$$

Further note that for any behavior policy $\pi_{\mathsf{b}} \in \Theta(\nu,\varepsilon)$, the distribution $\tilde{\pi}_{\mathsf{b}} = \pi_{\mathsf{b}}/\|\pi_{\mathsf{b}}\|_1$ satisfies the lower bound $\pi_{\mathsf{b}} \geq \nu/(1+\varepsilon)$. In addition, conditional on the realization of T, the Poisson model is equivalent to the original multinomial model with a discrete distribution $\tilde{\pi}_{\mathsf{b}}$. Combining these two facts together guarantees that, for any $\pi_{\mathsf{b}} \in \Theta(\nu,\varepsilon)$, we have the decomposition

$$\begin{aligned}
&\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}\Big[(\widehat{V}_T - \sum_{a\in[k]}\pi_{\mathsf{t}}(a) r_f(a))^2\Big] \\
&= \sum_{\ell=0}^\infty \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}\Big[(\widehat{V}_T - \sum_{a=1}^k \pi_{\mathsf{t}}(a) r_f(a))^2 \mid T=\ell\Big]\cdot\mathbb{P}(T=\ell) \\
&= \sum_{\ell=0}^\infty \mathbb{E}_{\tilde{\pi}_{\mathsf{b}}\otimes f}\Big[(\widehat{V}_\ell - \sum_{a\in[k]}\pi_{\mathsf{t}}(a) r_f(a))^2 \mid T=\ell\Big]\cdot\mathbb{P}(T=\ell) \\
&\leq \sum_{\ell=0}^\infty \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, \ell, \tfrac{\nu}{1+\varepsilon}\right)\cdot\mathbb{P}(T=\ell),
\end{aligned}$$

where the last line uses the fact that $\widehat{V}_\ell$ is minimax optimal. Since the function $\ell \mapsto \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, \ell, \frac{\nu}{1+\varepsilon}\right)$ is non-increasing, we can write

$$\begin{aligned}
&\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}_T - \sum_{a\in[k]} \pi_{\mathsf{t}}(a) r(a))^2] \\
&\le \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, 0, \tfrac{\nu}{1+\varepsilon}\right)\mathbb{P}(T < \tfrac{n}{2}) + \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, \tfrac{n}{2}, \tfrac{\nu}{1+\varepsilon}\right) \\
&\overset{(i)}{\le} r^2_{\max}\cdot \mathbb{P}(T < \tfrac{n}{2}) + \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, \tfrac{n}{2}, \tfrac{\nu}{1+\varepsilon}\right) \\
&\overset{(ii)}{\le} r^2_{\max}\cdot e^{-3n/72} + \mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, \tfrac{n}{2}, \tfrac{\nu}{1+\varepsilon}\right),
\end{aligned}$$

where step (i) follows from the inequality $\mathcal{R}^\star_{\mathsf{M}}\left(\pi_{\mathsf{t}}, 0, \frac{\nu}{1+\varepsilon}\right) \le r^2_{\max}$, whereas step (ii) follows from a standard tail bound for Poisson random variables; see e.g., Lemma 5 of the paper [38].

B. Proof of Lemma 5

Lemmas 8 and 9 guarantee the existence of random variables U and U' supported on the interval $[\frac{n\nu}{4\log k}, 1]$ such that

$$\begin{aligned}
\mathbb{E}[\tfrac{1}{U}] - \mathbb{E}[\tfrac{1}{U'}] &\ge \frac{4\log k}{n\nu}\cdot\Big(1 - 2\sqrt{\frac{n\nu}{4\log k}}\Big)^{\lceil 48\log k\rceil} \\
&\ge \frac{4\log k}{n\nu}\cdot \exp(-96\sqrt{n\nu\log n}) \\
\mathbb{E}[U^j] &= \mathbb{E}[(U')^j] \quad \text{for all } j = 0, 1, \ldots, \lceil 48\log k\rceil - 1.
\end{aligned}$$

Here we used the fact that $(1-x)^n \ge e^{-2nx}$ for $x \in [0, \frac{1}{2}]$.

Define the shifted random variables $V := U - \frac{n\nu}{8\log k}$ and $V' := U' - \frac{n\nu}{8\log k}$. With these definitions, it is straightforward to check that V and V' take values in the interval $[\frac{n\nu}{8\log k}, 1]$ and they satisfy the bounds

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{V + \frac{n\nu}{8\log k}}\right] - \mathbb{E}\left[\frac{1}{V' + \frac{n\nu}{8\log k}}\right] \\
&\quad\ge \frac{4\log k}{n\nu}\cdot\exp(-96\sqrt{n\nu\log k}), \quad \text{and} \qquad (69)\\
&\mathbb{E}[V^j] = \mathbb{E}[(V')^j], \text{for any } j = 0, 1, \ldots, \lceil 48\log k\rceil - 1.
\end{aligned}$$

Finally, we construct the desired random variables X, X' via changes of variables on V, V'. More specifically, the probability density functions of X, X' are given by

$$\begin{aligned}
P_X(\mathrm{d}x) &:= \left(1 - \mathbb{E}\left[\frac{n\nu/(8\log k)}{V}\right]\right)\delta_0(\mathrm{d}x) \\
&\quad + \frac{n\nu/(8\log k)}{x} P_V(\mathrm{d}x), \quad \text{and} \\
P_{X'}(\mathrm{d}x) &:= \left(1 - \mathbb{E}\left[\frac{n\nu/(8\log k)}{V'}\right]\right)\delta_0(\mathrm{d}x) \\
&\quad + \frac{n\nu/(8\log k)}{x} P_{V'}(\mathrm{d}x).
\end{aligned}$$

This is a valid construction since $V, V' \ge n\nu/(8\log k)$. In addition, we see that $X, X' \in [0,1]$, and for any $j = 1, \ldots, \lceil 48\log k\rceil$,

$$\begin{aligned}
\mathbb{E}[X^j] &= \int_{\frac{n\nu}{8\log k}}^1 x^{j-1}\cdot\frac{n\nu}{8\log k}\cdot P_V(\mathrm{d}x) \\
&= \frac{n\nu\mathbb{E}[V^{j-1}]}{8\log k} = \frac{n\nu\mathbb{E}[(V')^{j-1}]}{8\log k} = \mathbb{E}[(X')^j],
\end{aligned}$$

where we have used the fact that $\mathbb{E}[V^j] = \mathbb{E}[(V')^j]$ for all $j = 0, 1, \ldots, \lceil 48\log k\rceil - 1$. The above formula also tell us that $\mathbb{E}[X] = \mathbb{E}[X'] = \frac{n\nu}{8\log k}\cdot\mathbb{P}(V > 0) = \frac{n\nu}{8\log k}$.

Furthermore, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{X}{X + \frac{n\nu}{8\log k}}\right] &= \int_{\frac{n\nu}{8\log k}}^1 \frac{1}{x + \frac{n\nu}{8\log k}}\frac{n\nu}{8\log k}\cdot P_V(\mathrm{d}x) \\
&= \frac{n\nu}{8\log k}\mathbb{E}\left[\frac{1}{V + \frac{n\nu}{8\log k}}\right],
\end{aligned}$$

Together with equation (69), this relation implies that

$$\mathbb{E}\left[\frac{X}{X+\frac{n\nu}{8\log k}}\right] - \mathbb{E}\left[\frac{X'}{X'+\frac{n\nu}{8\log k}}\right] \ge \frac{1}{2}\exp(-96\sqrt{n\nu\log k}),$$

which concludes the proof.

Appendix E
Auxiliary Results Underlying Theorem 6

In this section, we collect the proofs of some useful results for establishing Theorem 6.

A. Proof of Lemma 6

In order to simplify notation, let us introduce the shorthand

$$\delta := \frac{\ell}{r} = \frac{\nu}{\frac{c_1\log k}{n}}. \qquad (70)$$

We split the proof into two cases, depending on whether $\pi_{\mathsf{b}}(a) \in [\ell, r]$, or $\pi_{\mathsf{b}}(a) > r$.

1) Case 1: $\pi_{\mathsf{b}}(a) \in [\ell, r]$: When $\pi_{\mathsf{b}}(a) \in [\ell, r]$, we have $|Q_L(\frac{2\pi_{\mathsf{b}}(a) - r - \ell}{r-\ell})| \le 1$, and hence

$$|P_L(\pi_{\mathsf{b}}(a))| \le \frac{1}{\left|Q_L\left(\frac{-r-\ell}{r-\ell}\right)\right|} = \Big|\frac{1}{Q_L\left(-\frac{1+\delta}{1-\delta}\right)}\Big|,$$

which implies that $\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \le \frac{1}{|Q_L\left(-\frac{1+\delta}{1-\delta}\right)|}$.

2) Case 2: $\pi_{\mathsf{b}}(a) > r$: When $\pi_{\mathsf{b}}(a) > r$, we know that

$$\begin{aligned}
&\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \le \max_{x\in(r,1]} e^{-nx}|P_L(x)| \\
&\le \max_{1<y\le\frac{2-r-\ell}{r-\ell}} \exp\left(-nry(1-\delta)/2\right) Q_L(y)\frac{\exp(-nr(1+\delta)/2)}{\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right|}.
\end{aligned}$$

Lemma 4 from Wu et al. [27] guarantees that if $\beta = O(L)$, then

$$\sup_{y\ge1}\left\{e^{-\beta y}Q_L(y)\right\} = \frac{1}{2}\left(\frac{\alpha+\sqrt{\alpha^2+1}}{e^{\sqrt{1+1/\alpha^2}}}(1+o_L(1))\right)^L, \quad \text{as } L\to\infty,$$

where $\alpha := L/\beta$. We apply this identity with the choices

$$\beta = nr(1-\delta)/2 \le c_1\log k, \qquad \alpha = 2\lambda := 2\frac{c_0}{c_1(1-\delta)},$$

thereby obtaining the inequality

$$\begin{aligned}
&\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \\
&\le \frac{1}{2}\left(\frac{2\lambda+\sqrt{4\lambda^2+1}}{e^{\sqrt{1+1/(4\lambda^2)}}}(1+o_k(1))\right)^L \frac{\exp(-nr(1+\delta)/2)}{\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right|}
\end{aligned}$$

$$= \frac{1}{2}\left(\frac{2\lambda+\sqrt{4\lambda^2+1}}{e^{\sqrt{1+1/(4\lambda^2)}+1/(2\lambda)}}\left(1+o_k(1)+o_\delta(1)\right)\right)^L \frac{1}{\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right|}.$$

By a suitable choice of the universal constants (c_0, c_1)—in particular, by taking $c_0 \ll c_1$— we can make λ as small as we please. This freedom allows us to guarantee that for all $\pi_{\mathsf{b}}(a) \geq r$, we have

$$\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \leq \frac{(1+o_k(1)+o_\delta(1))}{\left|Q_L\left(\frac{-r-\ell}{r-\ell}\right)\right|} \leq \frac{2}{\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right|},$$

as long as k and c_1 are both sufficiently large.

a) Putting pieces together: By combining the previous two cases together, we find that for any $\pi_{\mathsf{b}}(a) \geq \ell$, we have

$$\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \leq \frac{2}{\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right|} \leq 4\left(1-\frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^L.$$

In this argument, the final inequality exploits a basic fact about Chebyshev polynomials, namely that

$$\left|Q_L\left(-\frac{1+\delta}{1-\delta}\right)\right| \geq \frac{1}{2}\left(1-\frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^{-L}.$$

To conclude, we make note of the elementary inequalities $(1-x)^L \leq \exp(-xL)$ for $x \in (0,1)$, and $\frac{2\sqrt{\delta}}{1+\sqrt{\delta}} \geq \sqrt{\delta}$. Substituting these bounds yields the claimed result—viz.

$$\left|e^{-n\pi_{\mathsf{b}}(a)}P_L(\pi_{\mathsf{b}}(a))\right| \leq 4\exp\left(-\frac{2L\sqrt{\delta}}{1+\sqrt{\delta}}\right) \leq 4\exp\left(-L\sqrt{\delta}\right).$$

B. Proof of the Bounds (66)

We prove each of the two bounds (66a) and (66b) in turn.

b) Proof of the inequality (66a)*:* When it comes to α_1, given that $g_L(0) = 0$, one has

$$\mathbb{E}\left[\widehat{r}(a)g_L(n(a)) \mid n(a)\right] = r_f(a)g_L(n(a)),$$

and hence

$$\begin{aligned}\alpha_1 &= \mathsf{Var}\left(r_f(a)g_L(n(a))\right) \leq r_{\max}^2 \mathsf{Var}\left(g_L(n(a))\right)\\ &= r_{\max}^2 \mathsf{Var}\left(g_L(n(a))-1\right).\end{aligned}$$

Here, the inequality arises from the fact that $|r_f(a)| \leq r_{\max}$, and the last identity uses the translation invariance of the variance. Splitting $g_L(n(a))-1$ into $(g_L(n(a))-1)\mathbb{1}\{n(a) \leq L\}$ and $(g_L(n(a))-1)\mathbb{1}\{n(a) > L\}$ and using the elementary inequality $\mathsf{Var}(X+Y) \leq 2\mathsf{Var}(X)+2\mathsf{Var}(Y)$, we can obtain

$$\begin{aligned}\tfrac{\alpha_1}{r_{\max}^2} &\leq 2\mathsf{Var}\left(\{g_L(n(a))-1\}\mathbb{1}\{n(a) \leq L\}\right)\\ &\quad + 2\mathsf{Var}\left(\{g_L(n(a))-1\}\mathbb{1}\{n(a) > L\}\right)\\ &= 2\mathsf{Var}\left(\{g_L(n(a))-1\}\mathbb{1}\{n(a) \leq L\}\right)\\ &\leq 2\mathbb{E}\left(\{g_L(n(a))-1\}^2\mathbb{1}\{n(a) \leq L\}\right).\end{aligned}$$

Here, the equality is due to the fact that $g_L(n(a)) = 1$ for $n(a) > L$. Substitute in the definition of $g_L(n(a))$ to see that

$$\begin{aligned}\frac{\alpha_1}{r_{\max}^2} &\leq 2\sum_{j=0}^{L} e^{-n\pi_{\mathsf{b}}(a)}\frac{[n\pi_{\mathsf{b}}(a)]^j}{j!}(a_j j!/n^j)^2\\ &= 2\,e^{-n\pi_{\mathsf{b}}(a)}\sum_{j=0}^{L} a_j^2 j![\frac{\pi_{\mathsf{b}}(a)}{n}]^j. \qquad (71)\end{aligned}$$

It has been shown in equation (47) in Section 6.1 in the paper [27] that for $j = 1, 2, \ldots L$,

$$|a_j| \leq \frac{1}{2}(\frac{4}{r})^j \exp\left((L+j)h(\frac{2j}{L+j})\right) \leq \frac{1}{2}(\frac{4}{r})^j k^{2c_0}, \quad (72)$$

where $h(\lambda) := -\lambda\log\lambda - (1-\lambda)\log(1-\lambda)$ denotes the binary entropy function. The last inequality holds true since $j \leq L = c_0 \log k$ and $h(\lambda) \leq 1$. Combine the previous two bounds (71) and (72) together to see

$$\begin{aligned}\frac{\alpha_1}{r_{\max}^2} &\leq 2\,e^{-n\pi_{\mathsf{b}}(a)}\left(a_0^2 + \sum_{j=1}^{L} a_j^2 j![\frac{\pi_{\mathsf{b}}(a)}{n}]^j\right)\\ &\leq 2\,e^{-n\pi_{\mathsf{b}}(a)}\left(1+\frac{1}{4}\sum_{j=1}^{L}(\frac{16\pi_{\mathsf{b}}(a)}{nr^2})^j k^{4c_0} j!\right)\\ &\leq 2\,e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2}e^{-n\pi_{\mathsf{b}}(a)}\sum_{j=1}^{L}(\frac{16L\pi_{\mathsf{b}}(a)}{nr^2})^j k^{4c_0},\end{aligned}$$

where the last line follows from the elementary inequality $j! \leq j^j$ and the fact that $j \leq L$. To further upper bound α_1, we consider two separate cases. First, when $16L\pi_{\mathsf{b}}(a) \leq nr^2$, one clearly has

$$e^{-n\pi_{\mathsf{b}}(a)}\sum_{j=1}^{L}(\frac{16L\pi_{\mathsf{b}}(a)}{nr^2})^j k^{4c_0} \leq Lk^{4c_0}.$$

On the other hand, when $16L\pi_{\mathsf{b}}(a) \geq nr^2$, we have

$$\begin{aligned}&e^{-n\pi_{\mathsf{b}}(a)}\sum_{j=1}^{L}(\frac{16L\pi_{\mathsf{b}}(a)}{nr^2})^j k^{4c_0}\\ &\leq Lk^{4c_0}e^{-n\pi_{\mathsf{b}}(a)}\left(\frac{16L\pi_{\mathsf{b}}(a)}{nr^2}\right)^L\\ &= Lk^{4c_0}\exp\left(-n\pi_{\mathsf{b}}(a) + L\log\left(\frac{16c_0 n\pi_{\mathsf{b}}(a)}{c_1^2\log k}\right)\right)\\ &\leq Lk^{4c_0}\end{aligned}$$

as long as $c_1^2 \geq 32c_0^2$. In sum, using the definition $L = c_0\log k$, we arrive at the conclusion that

$$\alpha_1 \leq r_{\max}^2\left\{2\,e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2}c_0\log k \cdot k^{4c_0}\right\}.$$

c) Proof of the inequality (66b)*:* Our next step is to upper bound the second term α_2. We begin by observing that

$$\mathsf{Var}\left(\widehat{r}(a)g_L(n(a)) \mid n(a)\right) = g_L^2(n(a))\frac{\sigma_f^2(a)}{n(a)}\mathbb{1}\{n(a) > 0\},$$

which further implies

$$\begin{aligned}\alpha_2 &\leq r_{\max}^2\mathbb{E}\left[\frac{g_L^2(n(a))}{n(a)}\mathbb{1}\{n(a) > 0\}\right]\\ &= r_{\max}^2\mathbb{E}\left[\frac{g_L^2(n(a))}{n(a)}\mathbb{1}\{0 < n(a) \leq L\}\right]\\ &\quad + r_{\max}^2\mathbb{E}\left[\frac{1}{n(a)}\mathbb{1}\{n(a) > L\}\right].\end{aligned}$$

Here the last identity uses the fact that $g_L(n(a)) = 1$ for $n(a) > L$. Using the inequality $(x+y)^2 \le 2x^2+2y^2$, we can decompose the first term into

$$\begin{aligned}&\mathbb{E}\left[\frac{g_L^2(n(a))}{n(a)}\mathbb{1}\{0 < n(a) \le L\}\right]\\ &\le 2\mathbb{E}\left[\frac{[g_L(n(a))-1]^2}{n(a)}\mathbb{1}\{0 < n(a) \le L\}\right]\\ &\quad + 2\mathbb{E}\left[\frac{1}{n(a)}\mathbb{1}\{0 < n(a) \le L\}\right],\end{aligned}$$

which results in

$$\begin{aligned}\frac{\alpha_2}{r_{\max}^2} &\le 2\mathbb{E}\left[\frac{[g_L(n(a))-1]^2}{n(a)}\mathbb{1}\{0 < n(a) \le L\}\right]\\ &\quad + 2\mathbb{E}\left[\frac{1}{n(a)}\mathbb{1}\{n(a) > 0\}\right].\end{aligned}$$

Note however the first term has been controlled in the analysis of α_1:

$$\begin{aligned}&2\mathbb{E}\left[\frac{[g_L(n(a))-1]^2}{n(a)}\mathbb{1}\{0 < n(a) \le L\}\right]\\ &\le 2\mathbb{E}\left[[g_L(n(a))-1]^2\mathbb{1}\{0 < n(a) \le L\}\right]\\ &\le 2\mathbb{E}\left[[g_L(n(a))-1]^2\mathbb{1}\{n(a) \le L\}\right]\\ &\le 2\, e^{-n\pi_{\mathsf{b}}(a)} + \frac{1}{2}c_0 \log k \cdot k^{4c_0}.\end{aligned}$$

Regarding the second term, Lemma 1 of the paper [10] tells us that[3]

$$\mathbb{E}\left[\frac{1}{n(a)}\mathbb{1}\{n(a) > 0\}\right] \le \min\left\{1, \frac{5}{n\pi_{\mathsf{b}}(a)}\right\}. \tag{73}$$

Combining the preceding bounds yields the stated conclusion (66b).

Appendix F
Some Auxiliary Results

This section gathers some known auxiliary results that are used in our analysis.

A. Best Polynomial Approximation

Given an interval $I := [\ell, r]$ with $\ell > 0$, a positive integer $L > 0$ and a continuous function ϕ on I, let

$$E_L(\phi; I) := \inf_{\{a_i\}} \sup_{x \in I} \left| \sum_{i=0}^{L} a_i x^i - \phi(x) \right|$$

denote the best uniform approximation error of ϕ on I by degree-L polynomials.

In particular, for the function $\phi(x) = 1/x$, the following lemma, proved in Section 2.11.1 of the book [34], provides a precise characterization of $E_L(1/x; [\ell, r])$.

[3]Though Lemma 1 of the paper [10] deals with the case when $n(a)$ is a binomial random variable, the same proof works for the case with Poisson random variables since the multiplicative Chernoff bound used therein also holds for the Poisson case.

Lemma 8: Fix any $r > \ell > 0$ and any positive integer L. Denoting $\delta := \ell/r$, we have

$$2E_L\left(\tfrac{1}{x}, [\ell, r]\right) = \frac{(1+\sqrt{\delta})^2}{\ell}\left(1 - \frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^{L+1}.$$

In fact, the problem of best polynomial approximation is closely related to the problem of moment matching, as shown in the following lemma (cf. Appendix E of the paper [22]).

Lemma 9: The following identity holds:

$$\begin{aligned}2E_L(\phi; I) = \max \quad & \mathbb{E}_{X\sim\mu_1}[\phi(X)] - \mathbb{E}_{X\sim\mu_0}[\phi(X)]\\ \text{subject to} \quad & \mathbb{E}_{X\sim\mu_1}[X^l] = \mathbb{E}_{X\sim\mu_0}[X^l],\\ & l = 0, 1, \ldots, L,\end{aligned}$$

where the maximum is taken over pairs of distributions μ_0, μ_1 supported on the interval I.

B. Minimax Lower Bound via Le Cam's Method

Here we state a version of Le Cam's method for lower bounds based on mixture distributions. Consider a class of distributions $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$, and a target function $T(\theta)$ of the parameter θ. Let Z be a random vector drawn according to some distribution $\mathbb{P}_\theta$, and $\widehat{T}(Z)$ be an arbitrary estimator of the target $T(\theta)$ based on the data Z.

Let Ξ_0, Ξ_1 be two priors on the parameter space Θ. Correspondingly, let F_i denote the marginal distribution of the observation Z under the prior Ξ_i, for $i = 0, 1$. We then have:

Lemma 10: Suppose that there exist some quantities $\xi \in \mathbb{R}, s > 0, 0 \le \beta_0, \beta_1 < 1$ such that

$$\begin{aligned}\Xi_0(\theta : T(\theta) \le \xi - s) &\ge 1 - \beta_0;\\ \Xi_1(\theta : T(\theta) \ge \xi + s) &\ge 1 - \beta_1.\end{aligned}$$

If $\mathsf{TV}(F_1, F_0) \le \nu < 1$, then

$$\inf_{\widehat{T}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left(|\widehat{T}(Z) - T(\theta)| \ge s\right) \ge \frac{1-\nu-\beta_0-\beta_1}{2}.$$

C. Divergence Between Mixtures of Poisson Distributions

Given a nonnegative random variable X, denote by $\mathbb{E}[\mathsf{Poi}(X)]$ the Poisson mixture with respect to the variable X. We then have the following bound, proved as Lemma 3 in the paper [22], on the TV distance between two such Poisson mixtures.

Lemma 11: Let X, X' be random variables supported on $[0, b]$ such that $\mathbb{E}[X^j] = \mathbb{E}[(X')^j]$ for $j = 1, 2, \ldots, L$ for some $L > 2\,eb$. Then the TV distance is bounded as

$$\mathsf{TV}\left(\mathbb{E}[\mathsf{Poi}(X)], \mathbb{E}[\mathsf{Poi}(X')]\right) \le \left(\frac{2eb}{L}\right)^L. \tag{74}$$

References

[1] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[3] A. Tewari and S. A. Murphy, "From ads to interventions: Contextual bandits in mobile health," in *Mobile Health*. Cham, Switzerland: Springer, 2017, pp. 495–517.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[5] C. Szepesvári, "Algorithms for reinforcement learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 4, no. 1, pp. 1–103, 2010.
[6] L. Li, W. Chu, J. Langford, and X. Wang, "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 297–306.
[7] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, "Personalized ad recommendation systems for life-time value optimization with guarantees," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 1806–1812.
[8] A. Irpan, K. Rao, K. Bousmalis, C. Harris, J. Ibarz, and S. Levine, "Off-policy evaluation via off-policy classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5437–5448.
[9] K. Hirano, G. W. Imbens, and G. Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, vol. 71, no. 4, pp. 1161–1189, 2003.
[10] L. Li, R. Munos, and C. Szepesvári, "Toward minimax off-policy value estimation," in *Proc. 8th Int. Conf. Artif. Intell. Statist.*, vol. 38, 2015, pp. 608–616.
[11] Y.-X. Wang, A. Agarwal, and M. Dudık, "Optimal and adaptive off-policy evaluation in contextual bandits," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3589–3597.
[12] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 663–685, 1952.
[13] T. Le Paine *et al.*, "Hyperparameter selection for offline reinforcement learning," 2020, *arXiv:2007.09055*.
[14] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2139–2148.
[15] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 652–661.
[16] M. Yin and Y.-X. Wang, "Asymptotically efficient off-policy evaluation for tabular reinforcement learning," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3948–3958.
[17] Y. Duan, Z. Jia, and M. Wang, "Minimax-optimal off-policy evaluation with linear function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2701–2709.
[18] I. A. Ibragimov, A. S. Nemirovskii, and R. Khas'Minskii, "Some problems on nonparametric estimation in Gaussian white noise," *Theory Probab. Appl.*, vol. 31, no. 3, pp. 391–406, 1987.
[19] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probab. Theory Rel. Fields*, vol. 113, no. 2, pp. 221–253, 1999.
[20] T. T. Cai and M. G. Low, "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional," *Ann. Statist.*, vol. 39, no. 2, pp. 1012–1041, 2011.
[21] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy," in *Proc. Electron. Colloq. Comput. Complex.*, vol. 17, Nov. 2010, p. 179.
[22] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, Jun. 2016.
[23] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.
[24] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of divergences between discrete distributions," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 814–823, Nov. 2020.
[25] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, 2011, pp. 685–694.
[26] G. Valiant and P. Valiant, "Estimating the unseen: Improved estimators for entropy and other properties," *J. ACM*, vol. 64, no. 6, pp. 1–41, 2017.
[27] Y. Wu and P. Yang, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *Ann. Statist.*, vol. 47, no. 2, pp. 857–883, 2019.
[28] A. A. Tsiatis and M. Davidian, "Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Stat. Sci. Rev. J. Inst. Math. Statist.*, vol. 22, no. 4, p. 569, 2007.
[29] L. Bottou *et al.*, "Counterfactual reasoning and learning systems: The example of computational advertising," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3207–3260, Jan. 2013.
[30] P. Wawrzynski and A. Pacut, "Truncated importance sampling for reinforcement learning with experience replay," in *Proc. Int. Multiconf. Comput. Sci. Inf. Technol.*, 2007, pp. 305–315.
[31] E. L. Ionides, "Truncated importance sampling," *J. Comput. Graph. Statist.*, vol. 17, no. 2, pp. 295–311, 2008.
[32] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[33] A. Fiat and G. J. Woeginger, *Online Algorithms: The State of the Art*, vol. 1442. Heidelberg, Germany: Springer, 1998.
[34] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*. Amsterdam, The Netherlands: Elsevier, 2014.
[35] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, Jan. 2015.
[36] J. Lafferty, H. Liu, and L. Wasserman. (2008). *Minimax Theory*. [Online]. Available: http://www.stat.cmu.edu/larry/=sml/Minimax.pdf
[37] H. Strasser, *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, vol. 7. Berlin, Germany: Walter de Gruyter, 2011.
[38] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6343–6354, Nov. 2015.

Cong Ma received the B.Eng. degree from Tsinghua University in 2015 and the Ph.D. degree from Princeton University in 2020. He was a Post-Doctoral Researcher with the Department of Electrical Engineering and Computer Sciences, UC Berkeley. He is currently an Assistant Professor with the Department of Statistics, The University of Chicago. His research interests include mathematics of data science, machine learning, high-dimensional statistics, convex, and nonconvex optimization, as well as their applications to neuroscience. He has received the School of Engineering and Applied Science Award for Excellence from Princeton University in 2019, the AI Labs Fellowship from Hudson River Trading in 2019, and the Student Paper Award from the International Chinese Statistical Association in 2017.

Banghua Zhu (Student Member, IEEE) received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley. His research interests include machine learning, statistics, and information theory.

Jiantao Jiao (Member, IEEE) received the B.Eng. degree (Hons.) in electronic engineering from Tsinghua University, Beijing, China, in 2012, and the M.S. and Ph.D. degrees from Stanford University in 2014 and 2018, respectively. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley. His research interests are in statistical machine learning, mathematical data science, optimization, applied probability, information theory, and their applications in science and engineering.

Martin J. Wainwright (Senior Member, IEEE) received the bachelor's degree in mathematics from the University of Waterloo, Canada, and the Ph.D. degree in electrical engineering and computer sciences (EECS) from the Massachusetts Institute of Technology (MIT). He was a Section Lecturer with the International Congress of Mathematicians in 2014 and the Blackwell Lecturer with the Institute of Mathematical Statistics in 2016. He is currently a Chancellor's Professor with the University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of EECS. His research interests include high-dimensional statistics, information theory, statistical machine learning, and optimization theory. He has been awarded the Alfred P. Sloan Foundation Fellowship in 2005 and the Best Paper Awards from the IEEE Signal Processing Society in 2008 and the IEEE Communications Society in 2010. He was a recipient of the Joint Paper Prize from the IEEE Information Theory and Communication Societies in 2012, the Medallion Lectureship from the Institute of Mathematical Statistics in 2013, and the COPSS Presidents' Award from the Joint Statistical Societies in 2014.