Bayesian Inference of Contextual Bandit Policies via Empirical Likelihood

Jiangrong Ouyang   JIANGRONG.OUYANG@UNIMELB.EDU.AU
Mingming Gong   MINGMING.GONG@UNIMELB.EDU.AU
Howard Bondell   HOWARD.BONDELL@UNIMELB.EDU.AU
School of Mathematics and Statistics
University of Melbourne
Victoria, Australia

Abstract

Policy inference plays an essential role in the contextual bandit problem. In this paper, we use empirical likelihood to develop a Bayesian inference method for the joint analysis of multiple contextual bandit policies in finite sample regimes. The proposed inference method is robust to small sample sizes and is able to provide accurate uncertainty measurements for policy value evaluation. In addition, it allows for flexible inferences on policy comparison with full uncertainty quantification. We demonstrate the effectiveness of the proposed inference method using Monte Carlo simulations and its application to an adolescent body mass index data set.

Keywords: empirical likelihood, contextual bandit, off-policy evaluation, Bayesian inference

1. Introduction

The contextual bandit problem is an extension of the standard multi-armed bandit problem, in which the rewards of arms depend on some context information (Auer et al., 2002; Langford and Zhang, 2007). A contextual bandit policy is a function that guides the player to select more profitable arms based on the given context information. The inference of contextual bandit policies is essential, since it can be used to improve the current policy or to experiment with hypothetical policies of interest. Hence, there is no need to actually run the hypothetical policies in the real world, since we are able to reconstruct the information just as what we would have observed if the desired policies had been applied.

To perform a reliable inference, one should first be able to evaluate policies accurately. However, the evaluation of a bandit policy is not straightforward, since only partial feedback is observed, and the available data is not necessarily collected using the policy we want to evaluate. Hence, we need to resort to off-policy evaluation (Sutton and Barto, 2018). Off-policy evaluation is a problem of evaluating a given target policy using the data collected with another policy, called behavior policy.

There are three types of off-policy evaluation methods: direct method, importance sampling method, and doubly robust method. For the direct method, the expected reward given the arm and context is learned using regression, and then it is used to estimate the

target policy directly. The direct method estimator relies heavily on the accuracy of the regression for the reward. If the regression model is misspecified, then the estimator can be biased and inconsistent.

The main idea of the importance sampling method is to correct the weights of data points collected by the behavior policy so that they align with the target policy (Horvitz and Thompson, 1952). For example, the importance sampling (IS) estimator takes the weighted average of the rewards, where the weights are given by the ratios of the probabilities from the target policy to that from the behavior policy. The IS estimator is unbiased and consistent, yet its variance is often large, especially when the target policy differs significantly from the behavior policy. The self-normalized IS (SNIS) estimator is a scaled IS estimator, normalized by the average of the importance weights, which decreases the mean squared error at the cost of being biased (Swaminathan and Joachims, 2015). Furthermore, the SNIS estimator is bounded within the support of rewards, whereas the IS estimator is not.

The doubly robust estimator combines the direct method and importance sampling method, in which the direct method estimator acts as a baseline and is corrected by the IS estimator (Dudik et al., 2011; Jiang and Li, 2016). If the regression model for the reward is specified correctly, then the mean squared error of the doubly robust estimator achieves the semiparametric lower bound asymptotically. Otherwise, it can be worse than that of the IS and SNIS estimators (Kang and Schafer, 2007; Kallus and Uehara, 2019).

While many policy learning algorithms and evaluation methods are well-developed, policy inference remains difficult, since the distribution of the embedded importance weights is unknown. In this paper, we focus on the inference of contextual bandit policies. Empirical likelihood is a nonparametric inference method, which relies only on a set of estimating equations for the data (Owen, 1988, 1990; Qin and Lawless, 1994). Hence, with appropriate estimating equations, inferences on the target policy can be drawn based on the empirical likelihood. It has been shown that the point estimator based on empirical likelihood generally has a lower mean squared error than the IS, SNIS, and DR estimators (Kallus and Uehara, 2019). In addition, hypothesis tests and confidence intervals can be constructed using the empirical likelihood ratio with a chi-squared approximation (Karampatziakis et al., 2020). However, the empirical likelihood ratio in this setting only behaves like the chi-squared distribution when the sample size is large. For small or medium sample sizes, the confidence intervals are poorly calibrated and suffer from mismatched coverage probabilities. In addition, all the aforementioned methods focus on single policy evaluation, so comparison among potentially correlated policies is infeasible.

Given the poor chi-squared calibration in the frequentist setting, we propose using empirical likelihood in the Bayesian paradigm to conduct proper inference, either with an uninformative prior or by accommodating any prior information about the policy value. Due to the fact that empirical likelihood is not an ordinary likelihood, the Bayesian inference with empirical likelihood does not follow automatically. Lazar (2003) explored the validity of the resultant posterior when the data density is replaced by empirical likelihood. Grendár and Judge (2009) further justified the use of empirical likelihood in the Bayesian paradigm. Later, Bayesian empirical likelihood was adapted in several fields: survey sampling (Rao and Wu, 2010), small area estimation (Chaudhuri and Ghosh, 2011), and quantile regression (Yang and He, 2012). Algorithms for more efficient Bayesian computation have also been developed (Mengersen et al., 2013; Chaudhuri et al., 2017; Yu and Bondell, 2024).

In this paper, we employ empirical likelihood to develop a Bayesian inference method for jointly analyzing multiple contextual bandit policies in finite sample regimes. Our proposed inference method is robust with respect to small sample sizes and is able to provide accurate uncertainty measurements for policy value evaluation. Additionally, flexible inferences can be drawn on policy comparison with full uncertainty quantification.

The remainder of this paper is organized as follows. Section 2 provides the background information and defines the notation used throughout this paper. Section 3 describes the proposed Bayesian empirical likelihood method for policy inference. Section 4 presents two experiments demonstrating the proposed method. Section 5 gives the application to an adolescent body mass index data set. Section 6 contains some discussions.

2. Background and Notation

Bandit Problem. The multi-armed bandit problem is a repeated game proposed by Robbins (1952). In each round, the player makes a choice from K different arms and receives the associated random reward. The goal is to develop a strategy to maximize the total rewards. This game is a paradigmatic example of the exploration–exploitation trade-off dilemma. As the player has no knowledge of the reward distribution associated with each arm initially, exploration is needed. Then the player can exploit the information acquired in the past to gain rewards at the moment or keep exploring for better rewards in the future.

The contextual bandit problem is an extension of the standard multi-armed bandit problem. It further assumes that the reward distributions depend on some context information, and the context information is given to the player for decision making. This setting is more practical, since bandit problems in reality often come with context information. Due to the introduction of context information, contextual bandits have been adopted widely in various areas, such as news article recommendation (Li et al., 2010), personalized medicine (Tewari and Murphy, 2017), design of clinical trials (Villar and Rosenberger, 2018), and advertisement optimization (Narita et al., 2019).

Consider the contextual bandit problem with K arms. In each round, a context $\mathbf{x} \in \mathcal{X}$ is generated from some distribution D. Given this context $\mathbf{x}$, the reward vector $\mathbf{r} = (r_1, r_2, \ldots, r_K)^\top$ is drawn from the conditional distribution $D_{\mathbf{x}}$. Based on the context $\mathbf{x}$, the player selects an arm $a \in \{1, 2, \ldots, K\}$, and receives the reward r_a. Note that only r_a is revealed. Without loss of generality, we assume $r_a \in [0, 1]$.

Let $\pi : \mathcal{X} \to \mathcal{P}$ be a contextual bandit policy, where $\mathcal{P}$ is the set of all probability mass functions over the arms. For a given context $\mathbf{x}$, the policy π guides the player to select an arm based on the probability mass function, $\pi_{\mathbf{x}} = \pi(\mathbf{x})$; that is, $\pi_{\mathbf{x}}(a)$ indicates the probability of choosing arm a under the context $\mathbf{x}$. The value of the policy π is defined as the expected reward, i.e.,

$$\mathrm{v}_\pi = \mathbb{E}_{\mathbf{x} \sim D, a \sim \pi_{\mathbf{x}}}[r_a].$$

Off-Policy Evaluation. Consider the off-policy evaluation, where the offline data is collected with some behavior policy p rather than the target policy π. That is, in each round, the arm is chosen according to the probability mass function $p_{\mathbf{x}}$. We assume that $p_{\mathbf{x}}$ is known or can be well estimated.

Since the arms are chosen by the behavior policy p, for the associated rewards r, the target policy π will have different probabilities of receiving them. To leverage these rewards, define the importance weight as the probability ratio,

$$w_a = \frac{\pi_{\mathbf{x}}(a)}{p_{\mathbf{x}}(a)}.$$

It follows that the value of policy π can be calculated by

$$v_\pi = \mathbb{E}_{\mathbf{x} \sim D, a \sim p_{\mathbf{x}}}[w_a r_a].$$

This is the main estimating equation we use to construct the empirical likelihood function. Note that the estimation and inference of policy π depend on the offline data only through the importance weights and the received rewards,

$$\{(w_i, r_i)\}_{i=1}^n,$$

where $w_i = \frac{\pi_{\mathbf{x}_i}(a_i)}{p_{\mathbf{x}_i}(a_i)}$ denotes the observed value of the importance weight in round i in a data set of size n.

Empirical Likelihood. Empirical likelihood is a nonparametric inference method proposed by Owen (1988). It does not require a specification of distributional assumptions for the data, but relies only on a set of estimating equations. Assume that $X_1, \ldots, X_n \in \mathbb{R}$ are independent random variables with a common distribution function F. If the population mean, $\mu = \mu(F)$, is the parameter of interest, then one can use the first moment condition, $\mathbb{E}_F(X - \mu) = 0$, as the estimating equation, and define the log-empirical likelihood for μ as the solution to the following optimization problem,

$$\log L(\mu) = \sup_{\boldsymbol{\mathcal{Q}} \succeq \mathbf{0}} \left\{ \sum_{i=1}^n \log(Q_i) \,\Big|\, \sum_{i=1}^n Q_i = 1, \sum_{i=1}^n Q_i(X_i - \mu) = 0 \right\},$$

where $\boldsymbol{\mathcal{Q}} = (Q_1, Q_2, \ldots, Q_n)^\top$.

More flexible data structures or restrictions on the parameters can be incorporated into the empirical likelihood function by adding extra estimating equations (Qin and Lawless, 1994). Continue with the example above. If we know that the data is Poisson-like, then the second moment condition, $\mathbb{E}_F(X^2 - \mu - \mu^2) = 0$, can be included to ensure the property of $\mathbb{V}\mathrm{ar}_F(X) = \mathbb{E}_F(X)$; that is,

$$\log L(\mu) = \sup_{\boldsymbol{\mathcal{Q}} \succeq \mathbf{0}} \left\{ \sum_{i=1}^n \log(Q_i) \,\Big|\, \sum_{i=1}^n Q_i = 1, \sum_{i=1}^n Q_i(X_i - \mu) = 0, \sum_{i=1}^n Q_i\left(X_i^2 - \mu - \mu^2\right) = 0 \right\},$$

where $\boldsymbol{\mathcal{Q}} = (Q_1, Q_2, \ldots, Q_n)^\top$.

Note that the probabilities are optimized and assigned only to the observed data points. For some parameter values, it is possible that none of the probability allocations satisfies the constraints specified in the optimization problem. In such cases, the log-empirical likelihoods are defined as $-\infty$.

3. Methodology

In this section, we define the empirical likelihood for contextual bandit policies and demonstrate the proposed method for policy inference in the Bayesian paradigm. Additionally, we present a practical method for performing the computations required for the Bayesian inference.

3.1 Joint Empirical Likelihood for Multiple Policies

Let $\mathbf{v} \in [0,1]^\ell$ be the values of a set of ℓ policies, $\pi_1, \pi_2, \ldots, \pi_\ell$. We define the empirical likelihood jointly for these values, $\mathbf{v}$, with the estimating equation,

$$\mathbf{v} = \mathbb{E}_{\mathbf{x} \sim D, a \sim p_{\mathbf{x}}}[\mathbf{w}_a r_a],$$

where

$$\mathbf{w}_a = \left(\frac{\pi_{1,\mathbf{x}}(a)}{p_{\mathbf{x}}(a)}, \frac{\pi_{2,\mathbf{x}}(a)}{p_{\mathbf{x}}(a)}, \ldots, \frac{\pi_{\ell,\mathbf{x}}(a)}{p_{\mathbf{x}}(a)} \right)^\top$$

is the importance weight vector for the ℓ policies, $\pi_1, \pi_2, \ldots, \pi_\ell$, with respect to the behavior policy p. Alongside the estimating equation for $\mathbf{v}$, we also include a constraint for the importance weight vector,

$$\mathbb{E}_{\mathbf{x} \sim D, a \sim p_{\mathbf{x}}}[\mathbf{w}_a] = \mathbf{1}.$$

The reason for including this additional constraint is two-fold. While the theoretical expectation of $\mathbf{w}$ is always equal to $\mathbf{1}$, the empirical version with respect to $(\mathbf{w}, r)$ in the data space does not necessarily hold. By including this constraint in the formulation of the empirical likelihood, the average of $\mathbf{w}$ over the data space is restricted to $\mathbf{1}$. As a consequence, the support of the empirical likelihood function is bounded and is equal to the true support, $\mathbf{v} \in [0,1]^\ell$. In addition, this constraint guarantees that the empirical likelihood function is invariant to the metric used to measure the policies. Specifically, an equivalent way to assess policies is from the *regret*, $\mathbf{u} = \mathbf{1} - \mathbf{v}$, with the estimating equation, $\mathbf{u} = \mathbb{E}_{\mathbf{x} \sim D, a \sim p_{\mathbf{x}}}[\mathbf{w}_a(1 - r_a)]$. The estimating equations for the *value* and *regret* yield equivalent results if and only if the constraint, $\mathbb{E}_{\mathbf{x} \sim D, a \sim p_{\mathbf{x}}}[\mathbf{w}_a] = \mathbf{1}$, holds.

The inclusion or non-inclusion of this constraint corresponds to the analogs of two well-known importance sampling approaches. If it is not included, then the maximum empirical likelihood estimator corresponds to the pure importance sampling estimator, which takes the weighted average of the rewards directly (Horvitz and Thompson, 1952). If it is included, then the maximum empirical likelihood estimator corresponds to the self-normalized importance sampling estimator, where the empirical average of the importance weights is restricted to 1 (Swaminathan and Joachims, 2015).

Unlike the conventional empirical likelihood, we consider not only distributions with mass at the data points, $\{(\mathbf{w}_i, r_i)\}_{i=1}^n$, but all possible distributions defined over arbitrary points $(\mathbf{w}, r)$ within the hyperrectangular support,

$$\mathcal{S} = \left(\prod_{j=1}^\ell [w_{j,\min}, w_{j,\max}] \right) \times [0,1],$$

where $[w_{j,\min}, w_{j,\max}]$ is the range of importance weights inferred from policy π_j and the behavior policy p. The use of all possible points in the support $\mathcal{S}$ ensures that the empirical likelihood is well-defined for all $\mathbf{v} \in [0,1]^{\ell}$.

Formally, we start from first principles and consider all probability measures $\mathbf{Q}$ on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$. Let $Q_{\mathbf{w},r} = \mathbf{Q}(\{(\mathbf{w},r)\})$. We define the log-empirical likelihood for the policy values as the solution to the following optimization problem,

$$\log L(\mathbf{v}) = \sup_{\mathbf{Q}} \left\{ \sum_{i=1}^{n} \log\left(Q_{\mathbf{w}_i, r_i}\right) \,\middle|\, \int_{\mathcal{S}} \mathbf{w}\, d\mathbf{Q}(\mathbf{w},r) = \mathbf{1}, \int_{\mathcal{S}} \mathbf{w}r\, d\mathbf{Q}(\mathbf{w},r) = \mathbf{v} \right\}. \quad (1)$$

As a nonparametric maximum likelihood estimation problem, it suffices to consider only discrete $\mathbf{Q}$ (Lindsay, 1995). Note that the optimization variable is the entire probability measure $\mathbf{Q}$ on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ rather than a probability vector over some fixed points. Hence, the set of the points $(\mathbf{w}, r) \in \mathcal{S}$ that receive positive probability mass is not known *a priori*, and identifying this set constitutes part of the optimization problem.

By solving the Lagrange dual problem (see Appendix A for details), we get

$$\log L(\mathbf{v}) = -\sup_{(\boldsymbol{\beta}, \boldsymbol{\tau}) \in \mathcal{V}} \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\beta}^{\top}(\mathbf{w}_i - \mathbf{1}) + \boldsymbol{\tau}^{\top}(\mathbf{w}_i r_i - \mathbf{v})\right) + C,$$

where C is an immaterial constant, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are the dual variables, and

$$\mathcal{V} = \left\{ (\boldsymbol{\beta}, \boldsymbol{\tau}) \in \mathbb{R}^{\ell} \times \mathbb{R}^{\ell} \,\middle|\, 1 + \boldsymbol{\beta}^{\top}(\mathbf{w} - \mathbf{1}) + \boldsymbol{\tau}^{\top}(\mathbf{w}r - \mathbf{v}) \geq 0 \ \ \forall (\mathbf{w}, r) \in \mathcal{S} \right\}.$$

While the inequality needs to hold for all $(\mathbf{w}, r) \in \mathcal{S}$, it suffices to only consider the vertices of the hyperrectangular support $\mathcal{S}$, due to the linearity with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$.

As a consequence of considering all probability measures $\mathbf{Q}$ on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, the resulting empirical likelihood function may be flat over some region, implying that the maximum empirical likelihood estimator may not be unique (see Appendix B for details). However, this non-uniqueness does not affect the Bayesian inference that follows.

Using this joint empirical likelihood, we can derive the joint posterior distribution for $\mathbf{v}$ in the Bayesian paradigm, and subsequently construct a credible region for $\mathbf{v}$ from this distribution. Flexible inferences can be drawn with full uncertainty quantification. For example, we can compare a new policy against a baseline policy by computing $\mathbb{P}(v_{\text{new}} > v_{\text{baseline}} + \delta)$ or $\mathbb{P}(v_{\text{new}} > (1 + \delta)v_{\text{baseline}})$ for a desired margin δ, or even $\mathbb{P}(g(v_{\text{new}}, v_{\text{baseline}}) \geq 0)$ for an arbitrary function g.

For inference on a single policy, the empirical likelihood defined above is equivalent to the dual likelihood proposed by Karampatziakis et al. (2020). They used chi-squared approximation for the log-empirical likelihood ratio to construct confidence intervals for the policy value. In Section 4.1, we design an experiment to compare their confidence intervals in the frequentist framework to our credible intervals in the Bayesian paradigm.

3.2 Empirical Likelihood for Policy Value Difference

Suppose we have a baseline policy and a new policy. We are interested in evaluating the potential improvement of the new policy compared to the baseline policy. Then, we can

construct the empirical likelihood for the policy value difference, $d = \mathrm{v}_{\mathrm{new}} - \mathrm{v}_{\mathrm{baseline}}$. Let $\mathbf{t} = (-1, 1)^\top$ and $\mathbf{v} = (\mathrm{v}_{\mathrm{baseline}}, \mathrm{v}_{\mathrm{new}})^\top \in [0, 1]^2$. We define the log-empirical likelihood for the difference as the solution to the following optimization problem,

$$\log L(d) = \sup_{\mathbf{Q}} \left\{ \sum_{i=1}^{n} \log \left(Q_{\mathbf{w}_i, r_i} \right) \,\middle|\, \int_{\mathcal{S}} \mathbf{w} \, \mathrm{d}\mathbf{Q}(\mathbf{w}, r) = \mathbf{1}, \int_{\mathcal{S}} \mathbf{t}^\top \mathbf{w} r \, \mathrm{d}\mathbf{Q}(\mathbf{w}, r) = d \right\}. \quad (2)$$

From the definitions (1) and (2), it is clear that $\log L(d) = \sup_{\mathbf{v}: d = \mathbf{t}^\top \mathbf{v}} \log L(\mathbf{v})$. Note that rather than work with the vector $\mathbf{v} \in [0, 1]^2$, we directly project $\mathbf{v}$ onto a one-dimensional interval when formulating the empirical likelihood for d. This approach is more aligned with our interest and more computationally efficient.

By solving the Lagrange dual problem (see Appendix C for details), we get

$$\log L(d) = -\sup_{(\boldsymbol{\beta}, \delta) \in \mathcal{D}} \sum_{i=1}^{n} \log \left(1 + \boldsymbol{\beta}^\top (\mathbf{w}_i - \mathbf{1}) + \delta (\mathbf{t}^\top \mathbf{w}_i r_i - d) \right) + C,$$

where C is an immaterial constant, $\boldsymbol{\beta}$ and δ are the dual variables, and

$$\mathcal{D} = \left\{ (\boldsymbol{\beta}, \delta) \in \mathbb{R}^2 \times \mathbb{R} \,\middle|\, 1 + \boldsymbol{\beta}^\top (\mathbf{w} - \mathbf{1}) + \delta (\mathbf{t}^\top \mathbf{w} r - d) \geq 0 \;\; \forall (\mathbf{w}, r) \in \mathcal{S} \right\}.$$

Similar to the joint empirical likelihood for $\mathbf{v}$, the inequality in the set $\mathcal{D}$ only needs to hold for the vertices of the hyperrectangular support $\mathcal{S}$. The maximum empirical likelihood estimator also may not be unique, though, as before, this does not affect the subsequent Bayesian inference.

With this empirical likelihood, we can derive the posterior distribution for the difference d in the Bayesian paradigm. Policy comparison is now more straightforward, since d is the quantity of interest. For example, we can determine $\mathbb{P}(\mathrm{v}_{\mathrm{new}} > \mathrm{v}_{\mathrm{baseline}} + \delta)$ by computing $\mathbb{P}(d > \delta)$ from a one-dimensional posterior distribution. Note that this probability is not strictly identical to the one computed from the joint posterior distribution for $\mathbf{v}$, since they are derived from different empirical likelihoods. However, they share similar values. See Section 4.2 for an illustration. If the only interest is the policy value difference, we suggest using the univariate posterior distribution for d to reduce the computational cost.

3.3 Computational Method for Bayesian Inference

For computing posterior probabilities, such as $\mathbb{P}(\mathrm{v}_{\mathrm{new}} > \mathrm{v}_{\mathrm{baseline}} + \delta)$, we need to integrate the posterior distribution over a specific region. Here, we present a practical method for performing the computations required for the Bayesian inference.

Since $\mathbf{v}$ is bounded and typically in a low-dimensional space, we approximate the posterior on a fine grid, by evaluating the joint density and normalizing via summing over the grid. In high-dimensional cases, we recommend reducing the dimension of $\mathbf{v}$ down to a small number of policy value comparisons. The evaluation process can be computationally intensive as each empirical likelihood value itself involves solving an optimization problem. Nevertheless, this can be done directly in parallel, which can avoid significant computational burden.

However, as the sample size increases, the empirical likelihood concentrates on an arbitrarily small region, which renders any pre-defined grid not fine enough. Hence, the key

is to define an adaptive sub-support for $\mathbf{v}$ in which the empirical likelihood is larger than some small threshold.

We define the sub-support for each dimension v_j as follows:

$$\{v_j \mid L(\mathbf{v}) \geq c^{-1}L^*\}, \tag{3}$$

where L^* is the maximum empirical likelihood over the full support, and c is a large positive number. An adaptive sub-support for $\mathbf{v}$ is formed by taking the Cartesian product of the sub-supports for all dimensions defined above. Although this sub-support may contain $\mathbf{v}$ such that $L(\mathbf{v}) < c^{-1}L^*$, it is the smallest hyperrectangular region in which $L(\mathbf{v}) \geq c^{-1}L^*$.

With this adaptive sub-support, we can define a fine grid, then compute posterior probabilities and credible intervals with high accuracy. Additionally, using such an adaptive sub-support avoids wasting time on the region in which the empirical likelihood is practically zero. For the policy value difference d, the same method is applied to determine its adaptive sub-support.

See Appendix D for more details on the connection between the adaptive support and Wilks' confidence region (or interval), how we choose c, and the computation of the sub-support for each dimension v_j via an optimization problem. See also Section 4.1 for a brief derivation of Wilks' confidence interval.

4. Experiments

In this section, we design two experiments to demonstrate the proposed Bayesian empirical likelihood inference method.

Assumptions. Consider a contextual bandit problem with $K = 10$ arms. Let $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_1, \ldots, \mathbf{x}_K)$ be the context information, where $\mathbf{x}_c$ is a common context vector for all arms and $\mathbf{x}_a$ is an arm-specific context vector for $a \in \{1, 2, \ldots, K\}$. Assume that $\mathbf{x}_c \in \mathbb{R}^d$ and $\mathbf{x}_a \in \mathbb{R}^d$, where $d = 12$ and

$$\mathbf{x}_c \sim \text{Dirichlet}(1, \ldots, 1)$$
$$\mathbf{x}_a \sim \mathcal{N}(\mathbf{0}, I_d).$$

Given a context $\mathbf{x}$, the reward r_a for each arm follows a Bernoulli distribution specified as follows:

$$\eta_a = \beta_0 + \beta_1 \mathbf{x}_c^\top \mathbf{x}_a$$
$$p_a = \frac{\exp(\eta_a)}{1 + \exp(\eta_a)}$$
$$r_a \sim \text{Ber}(p_a),$$

where $\beta_0 = 0$ and $\beta_1 = 3$. Note that the rewards r_a are correlated, since they all depend on the common context vector $\mathbf{x}_c$.

We use a completely randomized behavior policy to generate the offline data. That is, in each round, the arm is chosen uniformly at random, regardless of the context given.

Policy Learning. We design a heuristic method for policy learning. For each arm, fit a logistic regression, using the reward r_a as the response variable and $\mathbf{x}_c^\top \mathbf{x}_a$ as the covariate.

Then estimate an upper bound ub_a for each arm as follows:

$$\eta_a = b_{0a} + b_{1a}\mathbf{x}_c^\top \mathbf{x}_a$$

$$\text{s.e.}(\eta_a) = \sqrt{[1, \mathbf{x}_c^\top \mathbf{x}_a]\, \mathcal{I}_a^{-1}[1, \mathbf{x}_c^\top \mathbf{x}_a]^\top}$$

$$\text{ub}_a = \frac{\exp(\eta_a + m \times \text{s.e.}(\eta_a))}{1 + \exp(\eta_a + m \times \text{s.e.}(\eta_a))},$$

where (b_{0a}, b_{1a}) are the coefficient estimates of the logistic regression, $\mathcal{I}_a$ is the Fisher information matrix, and $m \geq 0$ controls the confidence level for the upper bound estimate.

A randomized policy π is derived based on the collection of upper bound estimates for all arms. The probability of choosing arm a is given by

$$\pi_\mathbf{x}(a) = \text{ub}_a^s \bigg/ \sum_{k=1}^{K} \text{ub}_k^s,$$

where $s \geq 0$ is a hyperparameter of the probability mass function. If $s = 0$, then π is a completely randomized policy. For larger s, more probabilities are allocated on the arms with larger upper bound estimates. As $s \to \infty$, π converges to a deterministic policy that chooses the arm with the highest upper bound estimate. We can further restrict the arms to be chosen to the best K' arms by setting ub_a to zero if their values are ranked below K'. The three hyperparameters together, (m, s, K'), control the trade-off between exploration and exploitation.

For inference purposes, we learn a baseline policy and a new policy by this heuristic method. For the baseline policy, we use a smaller sample size and set the hyperparameters for more exploration. In contrast, for the new policy, we use a larger sample size and set the hyperparameters for more exploitation. We compute the true policy values using Monte Carlo integration with one million random points. For reference, we also include the oracle policy that always chooses the best arm. See Table 1 for a detailed summary.

Policy	Size	m	s	K'	True Value
Baseline	256	1	2	3	0.75
New	1024	1	1	1	0.83
Oracle	╱	╱	╱	╱	0.84

Table 1: Policy learning information.

Bayesian Inference. Using the proposed Bayesian inference method, we first evaluate the baseline policy and the new policy separately, and then compare these two policies jointly. A flat prior is used throughout the experiments. All posterior computations are performed with a fine grid over the adaptive sub-support defined in Section 3.3. We use a grid with 10^4 points for one-dimensional posterior distributions, and a grid with 10^6 points for the two-dimensional posterior distributions.

4.1 Single Policy Inference

In the frequentist framework, confidence intervals can be used to quantify the uncertainty of a point estimate. Karampatziakis et al. (2020) constructed confidence intervals based on

Wilks' theorem. Let v_0 be the true policy value and $\hat{v}$ be its maximum likelihood estimator. Then

$$-2\log\left(\frac{L(v_0)}{L(\hat{v})}\right) \xrightarrow{\text{d}} \chi_1^2 \text{ as } n \to \infty,$$

under mild regularity conditions. From this theorem, if $L(v)$ is much smaller than $L(\hat{v})$, then the hypothesis that $v_0 = v$ is rejected. Equivalently, if $L(v)$ is larger than a certain threshold, then v is included in a confidence interval. Specifically, a $100(1-\alpha)\%$ Wilks' interval is defined as

$$\left\{v \,\middle|\, L(v) \geq L(\hat{v})\exp\left(-\frac{1}{2}F_{\chi_1^2}^{-1}(1-\alpha)\right)\right\},$$

where $F_{\chi_1^2}^{-1}(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of the χ_1^2 distribution.

A credible interval in the Bayesian paradigm is an analog of a confidence interval. Let $p_{\text{post}}(v)$ be the posterior distribution of the policy value. A $100(1-\alpha)\%$ highest posterior density (HPD) interval, the narrowest type of credible interval, is defined as

$$\{v \,|\, p_{\text{post}}(v) \geq c\},$$

where c is the largest number such that

$$\int_{v:\, p_{\text{post}}(v)\geq c} p_{\text{post}}(v)\,\mathrm{d}v = 1-\alpha.$$

Note that the HPD and the Wilks' intervals share the same form, both being defined as an interval in which a certain threshold is exceeded. However, the HPD interval is determined by the posterior probability, while the Wilks' interval relies on a chi-squared approximation for the log-empirical likelihood ratio.

A good interval estimate should have a coverage probability that closely matches the nominal level. An interval with an overcoverage probability may provide trivial information, often because it is too wide, while an interval with an undercoverage probability can provide misleading information.

In this experiment, we compute the Wilks' and the HPD intervals at the nominal levels of 90% and 95%, and then compare their coverage probabilities and widths. Both the baseline policy and the new policy derived above are considered. Figure 1a shows the coverage probabilities aggregated from $10,000$ replicates. As the sample size increases, the coverage probabilities of all intervals converge to the nominal levels. For the baseline policy, both the Wilks' and the HPD intervals exhibit overcoverage when the sample size is small, with the Wilks' intervals showing a stronger degree of overcoverage than the HPD intervals. For the new policy, it is evident that the coverage probabilities of the Wilks' intervals deviate significantly from the nominal levels when the sample size is small.

In terms of the width, the Wilks' intervals are approximately 10% wider than the HPD intervals on average when the sample size is small, but this margin quickly drops to zero as the sample size increases. Figure 1b demonstrates the distributions of the interval widths for the case where the Wilks' intervals suffer from undercoverage. We observe that the widths of the Wilks' intervals spread over a larger range, with some intervals being quite wide
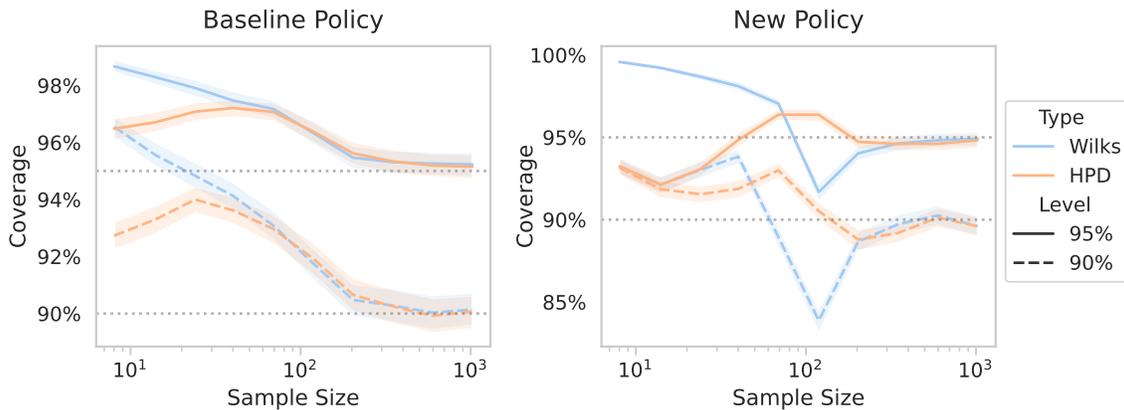
(a) Coverage probabilities for the Wilks' and the HPD intervals.
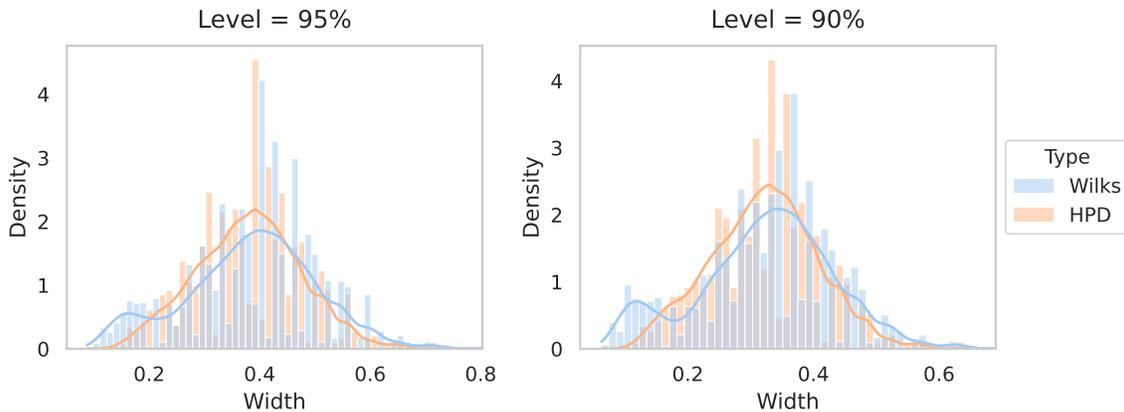
(b) Interval widths for the case where the Wilks' intervals suffer from undercoverage.

Figure 1: Comparison between the Wilks' and the HPD intervals.

and a significant proportion of intervals being too narrow. This explains the undercoverage problem.

While both the Wilks' and the HPD intervals demonstrate good asymptotic properties, the HPD intervals are clearly more robust in finite sample regimes. The mismatched coverage probabilities and the variability in the widths of the Wilks' intervals can be explained by the poor chi-squared calibration when the sample size is small. For single policy inference, we recommend using the HPD intervals in the Bayesian paradigm for more accurate uncertainty quantification.

4.2 Policy Comparison

In this experiment, we compare the new policy against the baseline policy using both the joint posterior distribution for $\mathbf{v} = (v_{\text{baseline}}, v_{\text{new}})^\top$ and the univariate posterior distribution for $d = v_{\text{new}} - v_{\text{baseline}}$. Specifically, we compute the probabilities of absolute improvement, $\mathbb{P}(v_{\text{new}} > v_{\text{baseline}} + \delta)$, and the probabilities of relative improvement, $\mathbb{P}(v_{\text{new}} > (1 + \delta)v_{\text{baseline}})$, for various margins $\delta = 0, 0.05, 0.10$.

Figure 2 demonstrates the comparison results using both methods. On the left, the colored regions illustrate the probabilities $\mathbb{P}(v_{new} > v_{baseline})$ within the joint and univariate posterior distributions when the sample size is 100. From the joint posterior distribution, it is clear that $v_{baseline}$ and v_{new} are not independent, which suggests the necessity of the joint inference, either using the policy value vector $\mathbf{v}$ or the policy value difference d.

On the right of Figure 2, the curves show the probabilities of improvements for various margins, and the shaded areas represent the 95% confidence bands based on the results aggregated from 500 replicates. Since the joint posterior distribution for $\mathbf{v}$ encompasses all information about the policy value vector, it supports arbitrary comparisons. We present the probabilities of both absolute and relative improvements; however, more flexible comparisons are possible. For example, we can compute the conditional probability $\mathbb{P}(v_{new} > v_{baseline}|v_{new}, v_{baseline} > 0.5)$. While the joint posterior distribution for $\mathbf{v}$ allows for arbitrary comparisons, the univariate posterior distribution for d provides a more direct method for comparing policies at lower computational cost. Depending on the interest, both methods offer effective policy comparisons with full quantification of uncertainty.
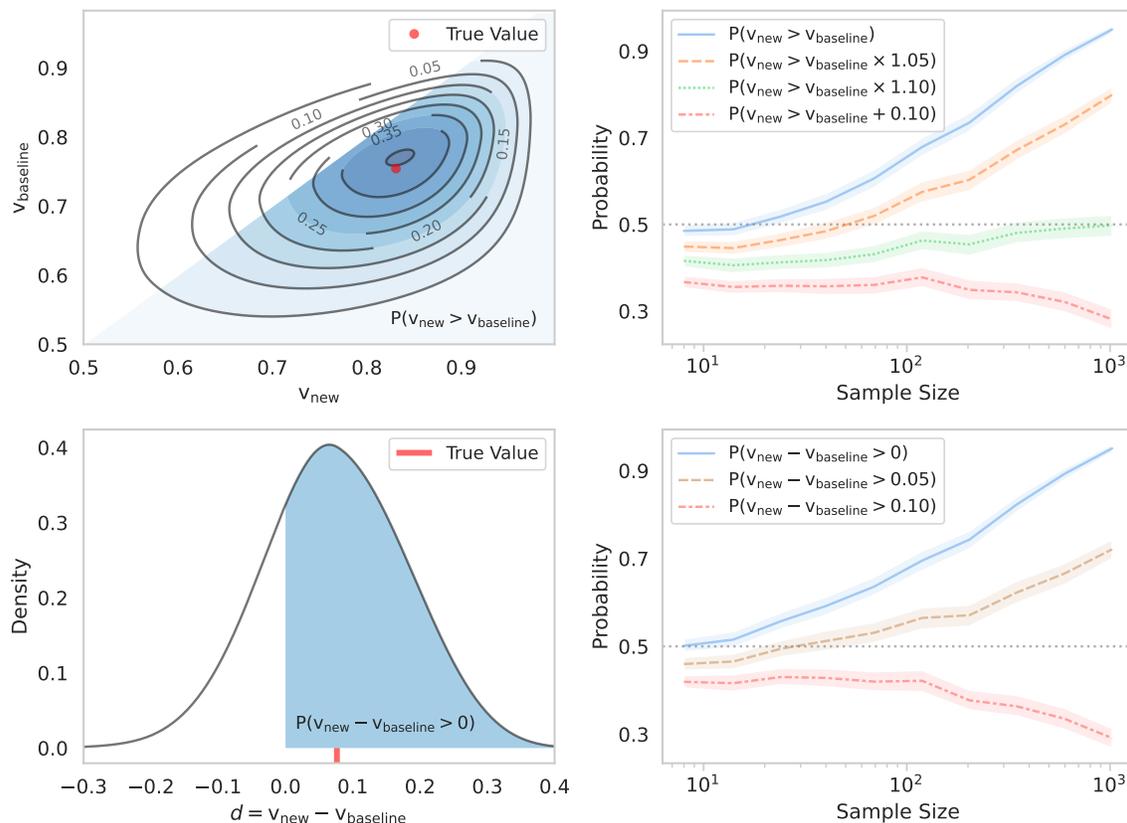

Figure 2: Policy comparison using the joint posterior distribution for $\mathbf{v} = (v_{baseline}, v_{new})^{\top}$ (top) and the univariate posterior distribution for $d = v_{new} - v_{baseline}$ (bottom).

12

5. Application to Body Mass Index Data

Body mass index (BMI) is the ratio of human body weight (kg) to squared height (m^2). It was devised by Quetelet back to mid-19th century, and formalized by Keys et al. (1972). We demonstrate the proposed Bayesian empirical likelihood inference method with an adolescent BMI data set, generated by Linn et al. (2015) to mimic a two-stage randomized clinical trial of BMI reduction on adolescent obesity (Berkowitz et al., 2011). Two treatments were used, meal replacement (MR) and conventional diet (CD). At both stages, patients were randomized to receive either MR or CD, each with probability 0.5. Four- and twelve-month patient BMI measurements were recorded. Four covariates are included: gender, race, parent BMI, and patient BMI measured at baseline.

Consider a contextual bandit problem with two treatments, MR and CD. We use the data from the first stage only, since there is little change in the BMI measurements at the second stage. The data can be considered as being collected with a completely randomized policy, since patients were randomized to receive either treatment with equal probability. The median of the BMI reduction is 2.6, so we set the reward to 1 if the BMI reduces by more than 2.6, and 0 otherwise.

A question of interest is what we would have observed if an alternative assignment of treatments had been applied to achieve more positive clinical outcomes. Our goal is to use the proposed method to address this question and guide subsequent decision-making with full quantification of uncertainty.

We set the current completely randomized policy as the baseline and define a new contextual bandit policy as follows. Partition data into two sets by the treatments chosen. For each treatment, fit a logistic regression with the four covariates. Then for each regression model, predict $\mathbb{P}(\text{reward} = 1)$ for all patients, and choose the treatment with the higher predicted estimate. Note that we expend no effort on tuning this heuristic policy, since our focus is on policy inference rather than on policy learning.

We evaluate this new policy and compare it against the baseline policy using the proposed Bayesian empirical likelihood inference method. The point estimate of the new policy value is 0.64, and the 95% HPD interval is $(0.55, 0.72)$. From the joint posterior distribution for the policy value vector, $\mathbb{P}(v_{\text{new}} > v_{\text{baseline}} \times 1.20) = 0.92$. From the univariate posterior distribution for the policy value difference, $\mathbb{P}(v_{\text{new}} > v_{\text{baseline}} + 0.10) = 0.92$. Both metrics suggest a significant improvement in the policy value. These results demonstrate that the proposed method can effectively quantify uncertainty in policy comparisons and support subsequent decision-making in clinical treatment.

6. Discussion

We propose a novel Bayesian empirical likelihood inference method for the joint analysis of multiple contextual bandit policies. While most existing techniques focus on policy learning and optimization, formal and effective inference on the learned policies is also necessary to complement these approaches. Our inference method provides accurate uncertainty measurements for policy value evaluation and facilitates flexible policy comparisons with full uncertainty quantification. Although it does not directly perform policy optimization, the

effective inference results can be used to choose among multiple candidate policies learned, and consequently aid in the search for an optimal policy.

By constructing the joint empirical likelihood for the policy value vector, arbitrary comparisons among potentially correlated policies are made possible. In addition, we construct the empirical likelihood for the policy value difference, which enables a more straightforward comparison at lower computational cost.

The method for the policy value difference can be further generalized to accommodate other policy comparisons. For example, if $(v_3 - v_1, v_3 - v_2)^\top$ or $v_3 - (v_1 + v_2)/2$ is of interest, then we can apply the corresponding transformation on $\mathbf{v}$ and construct the empirical likelihood in a lower-dimensional space.

For computational efficiency, we develop an adaptive grid method for the posterior computations. However, it may still be inefficient in high dimensions. One resolve for this issue is to transform $\mathbf{v}$ to a lower-dimensional space as discussed above. Alternatively, one can resort to variational inference (Blei et al., 2017; Yu and Bondell, 2024). However, a loss of accuracy arising from the variational approximation is inevitable. Further investigation is required to control the loss of accuracy.

Acknowledgments

Appendix A. Joint Empirical Likelihood for Multiple Policies

Let $\mathbf{v} \in [0,1]^\ell$ be the values of a set of ℓ policies, $\pi_1, \pi_2, \ldots, \pi_\ell$. As in (1) of Section 3.1, we define $-\log L(\mathbf{v})$ as the solution to the following optimization problem,

$$-\log L(\mathbf{v}) = \inf_{\mathbf{Q}} \left\{ -\sum_{i=1}^{n} \log \left(Q_{\mathbf{w}_i, r_i} \right) \, \middle| \, \int_{\mathcal{S}} \mathbf{w} \, \mathrm{d}\mathbf{Q}(\mathbf{w}, r) = \mathbf{1}, \int_{\mathcal{S}} \mathbf{w}r \, \mathrm{d}\mathbf{Q}(\mathbf{w}, r) = \mathbf{v} \right\},$$

where $\mathcal{S} = \left(\prod_{j=1}^{\ell} [w_{j,\min}, w_{j,\max}] \right) \times [0,1]$, $\mathbf{Q}$ is a probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, and $Q_{\mathbf{w}_i, r_i} = \mathbf{Q}(\{(\mathbf{w}_i, r_i)\})$.

As a nonparametric maximum likelihood estimation problem, it suffices to consider only discrete probability measures $\mathbf{Q}$ (Lindsay, 1995). Let $\mathcal{A}$ denote the set of points in $\mathcal{S}$ with positive probability mass,

$$\mathcal{A} = \{(\mathbf{w}, r) \in \mathcal{S} \mid \mathbf{Q}(\{(\mathbf{w}, r)\}) > 0\}.$$

Then $-\log L(\mathbf{v})$ can be rewritten as

$$
\begin{aligned}
\underset{\mathbf{Q}}{\text{minimize}} \quad & -\sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r}\log\left(Q_{\mathbf{w},r}\right) \\
\text{subject to} \quad & \sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} = 1 \\
& \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}Q_{\mathbf{w},r} = \mathbf{1} \\
& \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}rQ_{\mathbf{w},r} = \mathbf{v},
\end{aligned}
$$

where $c_{\mathbf{w},r} = \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{w}=\mathbf{w}_i, r=r_i\}}$.

The associated Lagrange dual function is given by

$$
\begin{aligned}
g(\gamma, \boldsymbol{\beta}, \boldsymbol{\tau}) = \inf_{\mathbf{Q}} \Bigg\{ & -\sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r}\log\left(Q_{\mathbf{w},r}\right) + \gamma\left(\sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} - 1\right) \\
& + \boldsymbol{\beta}^\top\left(\sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}Q_{\mathbf{w},r} - \mathbf{1}\right) \\
& + \boldsymbol{\tau}^\top\left(\sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}rQ_{\mathbf{w},r} - \mathbf{v}\right) \Bigg\} \\
= \inf_{\mathbf{Q}} \Bigg\{ & -\left(\gamma + \boldsymbol{\beta}^\top\mathbf{1} + \boldsymbol{\tau}^\top\mathbf{v}\right) \\
& + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{\left(\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r\right)Q_{\mathbf{w},r} - c_{\mathbf{w},r}\log\left(Q_{\mathbf{w},r}\right)\right\} \Bigg\}.
\end{aligned}
$$

To ensure that the Lagrange dual function is bounded, we must have

$$
\begin{aligned}
\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r &\geq 0 \quad \forall (\mathbf{w}, r) \in \mathcal{A} \subset \mathcal{S} \\
\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r &> 0 \quad \text{if } c_{\mathbf{w},r} > 0.
\end{aligned}
$$

Note that $\mathcal{A}$ is an arbitrary subset of $\mathcal{S}$. Since we require boundedness for any such $\mathcal{A} \subset \mathcal{S}$, the constraints above must therefore hold for all $(\mathbf{w}, r) \in \mathcal{S}$. Now, if $\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r = 0$, then $c_{\mathbf{w},r} = 0$, and the corresponding summand involving $Q_{\mathbf{w},r}$ vanishes from the Lagrange dual function. Hence, the only relevant allocation of $Q_{\mathbf{w},r}$ with respect to the infimum is when $\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r > 0$. Using standard calculus for minimization, we get

$$
Q_{\mathbf{w},r}^* = \frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top\mathbf{w} + \boldsymbol{\tau}^\top\mathbf{w}r}.
$$

Hence, the Lagrange dual function becomes

$$g(\gamma, \boldsymbol{\beta}, \boldsymbol{\tau}) = -\left(\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \boldsymbol{\tau}^\top \mathbf{v}\right) + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ c_{\mathbf{w},r} - c_{\mathbf{w},r} \log\left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \boldsymbol{\tau}^\top \mathbf{w}r}\right) \right\}$$

$$= -\left(\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \boldsymbol{\tau}^\top \mathbf{v}\right) + n + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ -c_{\mathbf{w},r} \log\left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \boldsymbol{\tau}^\top \mathbf{w}r}\right) \right\}.$$

Now consider

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} (\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \boldsymbol{\tau}^\top \mathbf{w}r) Q^*_{\mathbf{w},r}.$$

From the $Q^*_{\mathbf{w},r}$ obtained above, this sum equals

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} = n.$$

But, expanding the same sum and then applying the primal constraints yields

$$\gamma \sum_{(\mathbf{w},r)\in\mathcal{A}} Q^*_{\mathbf{w},r} + \boldsymbol{\beta}^\top \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q^*_{\mathbf{w},r} + \boldsymbol{\tau}^\top \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}r Q^*_{\mathbf{w},r} = \gamma + \boldsymbol{\beta}^\top \mathbf{1} + \boldsymbol{\tau}^\top \mathbf{v},$$

where the equality holds when the dual variables are evaluated at their optimal values. Hence, $\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \boldsymbol{\tau}^\top \mathbf{v} = n$. This identity allows us to eliminate γ. We reparameterize the Lagrange dual function as

$$\tilde{g}(\boldsymbol{\beta}, \boldsymbol{\tau}) = g(n - \boldsymbol{\beta}^\top \mathbf{1} - \boldsymbol{\tau}^\top \mathbf{v},\ \boldsymbol{\beta},\ \boldsymbol{\tau}),$$

which simplifies to

$$\tilde{g}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(n + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \boldsymbol{\tau}^\top(\mathbf{w}r - \mathbf{v})\right) - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}).$$

Let $\boldsymbol{\beta} \leftarrow n\boldsymbol{\beta}$ and $\boldsymbol{\tau} \leftarrow n\boldsymbol{\tau}$, rescaling the dual variables to maintain numerical stability as the sample size n increases. Then by strong duality, $-\log L(\mathbf{v})$ can be obtained by solving the Lagrange dual problem,

$$-\log L(\mathbf{v}) = \sup_{(\boldsymbol{\beta},\boldsymbol{\tau})\in\mathcal{V}} \tilde{g}(\boldsymbol{\beta}, \boldsymbol{\tau})$$

$$= \sup_{(\boldsymbol{\beta},\boldsymbol{\tau})\in\mathcal{V}} \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \boldsymbol{\tau}^\top(\mathbf{w}r - \mathbf{v})\right) - C$$

$$= \sup_{(\boldsymbol{\beta},\boldsymbol{\tau})\in\mathcal{V}} \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w}_i - \mathbf{1}) + \boldsymbol{\tau}^\top(\mathbf{w}_i r_i - \mathbf{v})\right) - C,$$

where $C = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}) - n\log(n)$ is an immaterial constant, and

$$\mathcal{V} = \left\{ (\boldsymbol{\beta}, \boldsymbol{\tau}) \in \mathbb{R}^\ell \times \mathbb{R}^\ell \,\middle|\, 1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \boldsymbol{\tau}^\top(\mathbf{w}r - \mathbf{v}) \geq 0 \ \ \forall (\mathbf{w}, r) \in \mathcal{S} \right\}.$$

Appendix B. Maximum Empirical Likelihood Estimator

Let $\hat{\mathbf{v}}$ be the maximum empirical likelihood estimator, i.e.,

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmax}} \log L(\mathbf{v}).$$

Recall that $\log L(\mathbf{v})$ itself is defined as the maximum of an optimization problem, where the parameter $\mathbf{v}$ is incorporated through the estimating equation, $\mathbf{v} = \int_{\mathcal{S}} \mathbf{w} r \, d\mathbf{Q}(\mathbf{w}, r)$, and the probability measure $\mathbf{Q}$ is the optimization variable. For a fixed value $\mathbf{v}$, this equation results in a constraint on the set of feasible $\mathbf{Q}$ to be considered in the optimization. This is what creates the likelihood surface as $\mathbf{v}$ varies. To compute the maximum empirical likelihood estimator, the problem becomes optimizing over all possible values of this surface, which occurs when the set of feasible $\mathbf{Q}$ is unconstrained with respect to $\mathbf{v}$. Then the maximum empirical likelihood estimator can be obtained by plugging the overall optimal $\mathbf{Q}^*$ into the estimating equation, i.e., $\hat{\mathbf{v}} = \int_{\mathcal{S}} \mathbf{w} r \, d\mathbf{Q}^*(\mathbf{w}, r)$.

Consider the optimization problem for the negative of the maximum log-empirical likelihood,

$$-\log L^* = \inf_{\mathbf{Q}} \left\{ -\sum_{i=1}^{n} \log \left(Q_{\mathbf{w}_i, r_i} \right) \,\middle|\, \int_{\mathcal{S}} \mathbf{w} \, d\mathbf{Q}(\mathbf{w}, r) = \mathbf{1} \right\},$$

where $\mathcal{S} = \left(\prod_{j=1}^{\ell} [w_{j,\min}, w_{j,\max}] \right) \times [0, 1]$, $\mathbf{Q}$ is a probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, and $Q_{\mathbf{w}_i, r_i} = \mathbf{Q}(\{(\mathbf{w}_i, r_i)\})$.

As a nonparametric maximum likelihood estimation problem, it suffices to consider only discrete probability measures $\mathbf{Q}$ (Lindsay, 1995). Let $\mathcal{A}$ denote the set of points in $\mathcal{S}$ with positive probability mass,

$$\mathcal{A} = \{(\mathbf{w}, r) \in \mathcal{S} \,|\, \mathbf{Q}(\{(\mathbf{w}, r)\}) > 0\}.$$

Then $-\log L^*$ can be rewritten as

$$
\begin{aligned}
\underset{\mathbf{Q}}{\text{minimize}} \quad & -\sum_{(\mathbf{w}, r) \in \mathcal{A}} c_{\mathbf{w}, r} \log \left(Q_{\mathbf{w}, r} \right) \\
\text{subject to} \quad & \sum_{(\mathbf{w}, r) \in \mathcal{A}} Q_{\mathbf{w}, r} = 1 \\
& \sum_{(\mathbf{w}, r) \in \mathcal{A}} \mathbf{w} Q_{\mathbf{w}, r} = \mathbf{1},
\end{aligned}
$$

where $c_{\mathbf{w}, r} = \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{w} = \mathbf{w}_i, r = r_i\}}$.

The associated Lagrange dual function is given by

$$g(\gamma, \boldsymbol{\beta}) = \inf_{\mathbf{Q}} \left\{ - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(Q_{\mathbf{w},r}\right) + \gamma \left(\sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} - 1 \right) \right.$$
$$\left. + \boldsymbol{\beta}^\top \left(\sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} - \mathbf{1} \right) \right\}$$
$$= \inf_{\mathbf{Q}} \left\{ -\left(\gamma + \boldsymbol{\beta}^\top \mathbf{1}\right) + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ (\gamma + \boldsymbol{\beta}^\top \mathbf{w}) Q_{\mathbf{w},r} - c_{\mathbf{w},r} \log\left(Q_{\mathbf{w},r}\right) \right\} \right\}.$$

To ensure that the Lagrange dual function is bounded, we must have

$$\gamma + \boldsymbol{\beta}^\top \mathbf{w} \geq 0 \quad \forall (\mathbf{w}, r) \in \mathcal{A} \subset \mathcal{S}$$
$$\gamma + \boldsymbol{\beta}^\top \mathbf{w} > 0 \quad \text{if } c_{\mathbf{w},r} > 0.$$

Note that $\mathcal{A}$ is an arbitrary subset of $\mathcal{S}$. Since we require boundedness for any such $\mathcal{A} \subset \mathcal{S}$, the constraints above must therefore hold for all $(\mathbf{w}, r) \in \mathcal{S}$. Now, if $\gamma + \boldsymbol{\beta}^\top \mathbf{w} = 0$, then $c_{\mathbf{w},r} = 0$, and the corresponding summand involving $Q_{\mathbf{w},r}$ vanishes from the Lagrange dual function. Hence, the only relevant allocation of $Q_{\mathbf{w},r}$ with respect to the infimum is when $\gamma + \boldsymbol{\beta}^\top \mathbf{w} > 0$. Using standard calculus for minimization, we get

$$Q^*_{\mathbf{w},r} = \frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w}}.$$

Hence, the Lagrange dual function becomes

$$g(\gamma, \boldsymbol{\beta}) = -\left(\gamma + \boldsymbol{\beta}^\top \mathbf{1}\right) + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ c_{\mathbf{w},r} - c_{\mathbf{w},r} \log\left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w}} \right) \right\}$$
$$= -\left(\gamma + \boldsymbol{\beta}^\top \mathbf{1}\right) + n + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ -c_{\mathbf{w},r} \log\left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w}} \right) \right\}.$$

Now consider

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} (\gamma + \boldsymbol{\beta}^\top \mathbf{w}) Q^*_{\mathbf{w},r}.$$

From the $Q^*_{\mathbf{w},r}$ obtained above, this sum equals

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} = n.$$

But, expanding the same sum and then applying the primal constraints yields

$$\gamma \sum_{(\mathbf{w},r)\in\mathcal{A}} Q^*_{\mathbf{w},r} + \boldsymbol{\beta}^\top \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q^*_{\mathbf{w},r} = \gamma + \boldsymbol{\beta}^\top \mathbf{1},$$

where the equality holds when the dual variables are evaluated at their optimal values. Hence, $\gamma + \boldsymbol{\beta}^\top \mathbf{1} = n$. This identity allows us to eliminate γ. We reparameterize the Lagrange dual function as

$$\tilde{g}(\boldsymbol{\beta}) = g(n - \boldsymbol{\beta}^\top \mathbf{1}, \ \boldsymbol{\beta}),$$

which simplifies to

$$\tilde{g}(\boldsymbol{\beta}) = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(n + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1})\right) - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}).$$

Let $\boldsymbol{\beta} \leftarrow n\boldsymbol{\beta}$, rescaling the dual variables to maintain numerical stability as the sample size n increases. Then by strong duality, $-\log L^*$ can be obtained by solving the Lagrange dual problem,

$$
\begin{aligned}
-\log L^* &= \sup_{\boldsymbol{\beta}\in\mathcal{V}'} \tilde{g}(\boldsymbol{\beta}) \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{V}'} \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1})\right) - C \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{V}'} \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w}_i - \mathbf{1})\right) - C,
\end{aligned}
$$

where $C = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}) - n\log(n)$ is an immaterial constant, and

$$\mathcal{V}' = \left\{ \boldsymbol{\beta} \in \mathbb{R}^\ell \,\middle|\, 1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) \geq 0 \ \ \forall(\mathbf{w}, r) \in \mathcal{S} \right\}.$$

After solving this dual problem, we can use the optimal dual variable, $\boldsymbol{\beta}^*$, to construct the maximum empirical likelihood estimator, $\hat{\mathbf{v}}$. Recall that the policy value vector depends on the distribution of the mass over the hyperrectangular support, i.e.,

$$\mathbf{v} = \int_{\mathcal{S}} \mathbf{w} r \, \mathrm{d}\mathbf{Q}(\mathbf{w}, r) = \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} r Q_{\mathbf{w},r}.$$

The probabilities assigned to the observed data points are given by

$$Q^*_{\mathbf{w}_i, r_i} = \frac{1}{n(1 + \boldsymbol{\beta}^{*\top}(\mathbf{w}_i - \mathbf{1}))}.$$

The remaining probabilities, if any, are allocated to the boundary points of $\mathcal{S}$ that satisfies

$$1 + \boldsymbol{\beta}^{*\top}(\mathbf{w} - \mathbf{1}) = 0.$$

Note that this condition is free of the reward, r. Hence, for any $\mathbf{w}_0$ that satisfies the condition, probabilities can be assigned to points over the interval $\{(\mathbf{w}_0, r) \,|\, r \in [0,1]\}$. This explains why the maximum empirical likelihood estimator may not be unique.

Let

$$\mathcal{W}_0 = \left\{ \mathbf{w}_0 \in \prod_{j=1}^{\ell} [w_{j,\min}, w_{j,\max}] \,\middle|\, 1 + \boldsymbol{\beta}^{*\top}(\mathbf{w}_0 - \mathbf{1}) = 0 \right\}.$$

Then

$$\hat{\mathbf{v}} = \left\{ \sum_{i=1}^{n} \mathbf{w}_i r_i Q^*_{\mathbf{w}_i,r_i} + \sum_{\mathbf{w}_0 \in \mathcal{W}_0} \mathbf{w}_0 r_{\mathbf{w}_0} Q^*_{\mathbf{w}_0} \;\middle|\; \forall r_{\mathbf{w}_0} \in [0,1] \right\},$$

where $Q^*_{\mathbf{w}_0}$ are solved from the primal conditions,

$$\sum_{\mathbf{w}_0 \in \mathcal{W}_0} Q_{\mathbf{w}_0} = 1 - \sum_{i=1}^{n} Q^*_{\mathbf{w}_i,r_i}$$

$$\sum_{\mathbf{w}_0 \in \mathcal{W}_0} \mathbf{w}_0 Q_{\mathbf{w}_0} = \mathbf{1} - \sum_{i=1}^{n} \mathbf{w}_i Q^*_{\mathbf{w}_i,r_i}.$$

Note that $\mathcal{W}_0$ may be empty, which occurs if $\sum_{i=1}^{n} Q^*_{\mathbf{w}_i,r_i} = 1$. In this case, $\hat{\mathbf{v}}$ is unique. If $\mathcal{W}_0$ is not empty, then due to the linearity with respect to $\boldsymbol{\beta}$, it suffices to consider those elements of $\mathcal{W}_0$ that are the vertices of $\prod_{j=1}^{\ell}[w_{j,\min}, w_{j,\max}]$ when constructing $\hat{\mathbf{v}}$. For the same reason, for any candidate $\mathbf{w}_0 \in \mathcal{W}_0$, it suffices to consider $(\mathbf{w}_0, 0)$ and $(\mathbf{w}_0, 1)$.

Appendix C. Empirical Likelihood for Policy Value Difference

Let $\mathbf{t} = (-1, 1)^\top$, $\mathbf{v} = (v_{\text{baseline}}, v_{\text{new}})^\top \in [0,1]^2$, and $d = \mathbf{t}^\top \mathbf{v}$. As in (2) of Section 3.2, we define $-\log L(d)$ as the solution to the following optimization problem,

$$-\log L(d) = \inf_{\mathbf{Q}} \left\{ -\sum_{i=1}^{n} \log\left(Q_{\mathbf{w}_i,r_i}\right) \;\middle|\; \int_{\mathcal{S}} \mathbf{w} \, d\mathbf{Q}(\mathbf{w}, r) = \mathbf{1}, \int_{\mathcal{S}} \mathbf{t}^\top \mathbf{w} r \, d\mathbf{Q}(\mathbf{w}, r) = d \right\},$$

where $\mathcal{S} = \left(\prod_{j=1}^{\ell}[w_{j,\min}, w_{j,\max}] \right) \times [0,1]$, $\mathbf{Q}$ is a probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, and $Q_{\mathbf{w}_i,r_i} = \mathbf{Q}(\{(\mathbf{w}_i, r_i)\})$.

As a nonparametric maximum likelihood estimation problem, it suffices to consider only discrete probability measures $\mathbf{Q}$ (Lindsay, 1995). Let $\mathcal{A}$ denote the set of points in $\mathcal{S}$ with positive probability mass,

$$\mathcal{A} = \{(\mathbf{w}, r) \in \mathcal{S} \mid \mathbf{Q}(\{(\mathbf{w}, r)\}) > 0\}.$$

Then $-\log L(d)$ can be rewritten as

$$\begin{aligned}
\underset{\mathbf{Q}}{\text{minimize}} \quad & -\sum_{(\mathbf{w},r) \in \mathcal{A}} c_{\mathbf{w},r} \log\left(Q_{\mathbf{w},r}\right) \\
\text{subject to} \quad & \sum_{(\mathbf{w},r) \in \mathcal{A}} Q_{\mathbf{w},r} = 1 \\
& \sum_{(\mathbf{w},r) \in \mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} = \mathbf{1} \\
& \sum_{(\mathbf{w},r) \in \mathcal{A}} \mathbf{t}^\top \mathbf{w} r Q_{\mathbf{w},r} = d,
\end{aligned}$$

where $c_{\mathbf{w},r} = \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{w}=\mathbf{w}_i, r=r_i\}}$.

20

The associated Lagrange dual function is given by

$$
g(\gamma, \boldsymbol{\beta}, \delta) = \inf_{\mathbf{Q}} \left\{ - \sum_{(\mathbf{w},r) \in \mathcal{A}} c_{\mathbf{w},r} \log \left(Q_{\mathbf{w},r} \right) + \gamma \left(\sum_{(\mathbf{w},r) \in \mathcal{A}} Q_{\mathbf{w},r} - 1 \right) \right.
$$
$$
+ \boldsymbol{\beta}^\top \left(\sum_{(\mathbf{w},r) \in \mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} - \mathbf{1} \right)
$$
$$
\left. + \delta \left(\sum_{(\mathbf{w},r) \in \mathcal{A}} \mathbf{t}^\top \mathbf{w} r Q_{\mathbf{w},r} - d \right) \right\}
$$
$$
= \inf_{\mathbf{Q}} \left\{ - \left(\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \delta d \right) \right.
$$
$$
\left. + \sum_{(\mathbf{w},r) \in \mathcal{A}} \left\{ \left(\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r \right) Q_{\mathbf{w},r} - c_{\mathbf{w},r} \log \left(Q_{\mathbf{w},r} \right) \right\} \right\}.
$$

To ensure that the Lagrange dual function is bounded, we must have

$$
\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r \geq 0 \quad \forall (\mathbf{w}, r) \in \mathcal{A} \subset \mathcal{S}
$$
$$
\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r > 0 \quad \text{if } c_{\mathbf{w},r} > 0.
$$

Note that $\mathcal{A}$ is an arbitrary subset of $\mathcal{S}$. Since we require boundedness for any such $\mathcal{A} \subset \mathcal{S}$, the constraints above must therefore hold for all $(\mathbf{w}, r) \in \mathcal{S}$. Now, if $\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r = 0$, then $c_{\mathbf{w},r} = 0$, and the corresponding summand involving $Q_{\mathbf{w},r}$ vanishes from the Lagrange dual function. Hence, the only relevant allocation of $Q_{\mathbf{w},r}$ with respect to the infimum is when $\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r > 0$. Using standard calculus for minimization, we get

$$
Q_{\mathbf{w},r}^* = \frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r}.
$$

Hence, the Lagrange dual function becomes

$$
g(\gamma, \boldsymbol{\beta}, \boldsymbol{\tau}) = - \left(\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \delta d \right) + \sum_{(\mathbf{w},r) \in \mathcal{A}} \left\{ c_{\mathbf{w},r} - c_{\mathbf{w},r} \log \left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r} \right) \right\}
$$
$$
= - \left(\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \delta d \right) + n + \sum_{(\mathbf{w},r) \in \mathcal{A}} \left\{ - c_{\mathbf{w},r} \log \left(\frac{c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r} \right) \right\}.
$$

Now consider

$$
\sum_{(\mathbf{w},r) \in \mathcal{A}} (\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \delta \mathbf{t}^\top \mathbf{w} r) Q_{\mathbf{w},r}^*.
$$

From the $Q_{\mathbf{w},r}^*$ obtained above, this sum equals

$$
\sum_{(\mathbf{w},r) \in \mathcal{A}} c_{\mathbf{w},r} = n.
$$

But, expanding the same sum and then applying the primal constraints yields

$$\gamma \sum_{(\mathbf{w},r)\in\mathcal{A}} Q^*_{\mathbf{w},r} + \boldsymbol{\beta}^\top \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}Q^*_{\mathbf{w},r} + \delta \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{t}^\top \mathbf{w}rQ^*_{\mathbf{w},r} = \gamma + \boldsymbol{\beta}^\top \mathbf{1} + +\delta d,$$

where the equality holds when the dual variables are evaluated at their optimal values. Hence, $\gamma + \boldsymbol{\beta}^\top \mathbf{1} + \delta d = n$. This identity allows us to eliminate γ. We reparameterize the Lagrange dual function as

$$\tilde{g}(\boldsymbol{\beta}, \delta) = g(n - \boldsymbol{\beta}^\top \mathbf{1} - \delta d, \ \boldsymbol{\beta}, \ \delta),$$

which simplifies to

$$\tilde{g}(\boldsymbol{\beta}, \delta) = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(n + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \delta(\mathbf{t}^\top \mathbf{w}r - d)\right) - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}).$$

Let $\boldsymbol{\beta} \leftarrow n\boldsymbol{\beta}$ and $\delta \leftarrow n\delta$, rescaling the dual variables to maintain numerical stability as the sample size n increases. Then by strong duality, $-\log L(d)$ can be obtained by solving the Lagrange dual problem,

$$\begin{aligned}
-\log L(d) &= \sup_{(\boldsymbol{\beta},\delta)\in\mathcal{D}} \tilde{g}(\boldsymbol{\beta}, \delta) \\
&= \sup_{(\boldsymbol{\beta},\delta)\in\mathcal{D}} \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \delta(\mathbf{t}^\top \mathbf{w}r - d)\right) - C \\
&= \sup_{(\boldsymbol{\beta},\delta)\in\mathcal{D}} \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\beta}^\top(\mathbf{w}_i - \mathbf{1}) + \delta(\mathbf{t}^\top \mathbf{w}_i r_i - d)\right) - C,
\end{aligned}$$

where $C = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}) - n\log(n)$ is an immaterial constant, and

$$\mathcal{D} = \left\{(\boldsymbol{\beta}, \delta) \in \mathbb{R}^2 \times \mathbb{R} \ \middle| \ 1 + \boldsymbol{\beta}^\top(\mathbf{w} - \mathbf{1}) + \delta(\mathbf{t}^\top \mathbf{w}r - d) \geq 0 \ \ \forall(\mathbf{w},r) \in \mathcal{S}\right\}.$$

Appendix D. Adaptive Sub-Support

To enable efficient Bayesian computations with a fine grid, we need an adaptive sub-support for $\mathbf{v}$ in which the empirical likelihood is larger than some small threshold. As in (3) of Section 3.3, we define the sub-support for each dimension v_j as follows:

$$\{\mathrm{v}_j \mid L(\mathbf{v}) \geq c^{-1} L^*\},$$

where L^* is the maximum empirical likelihood over the full support, and c is a large positive number. Taking their Cartesian product yields the adaptive sub-support for $\mathbf{v}$.

Note that this sub-support is similar to Wilks' confidence region in that both are defined by the condition that the likelihood exceeds a certain threshold. The difference lies in their construction. Wilks' confidence region applies the threshold condition jointly to the entire vector, whereas the adaptive sub-support applies this condition separately to each dimension and then takes their Cartesian product.

Let $F_{\chi_k^2}^{-1}(q)$ denote the q-th quantile of the χ_k^2 distribution. We choose c as summarized in Table 2. The resulting adaptive sub-support corresponds to the 99.99% Wilks' confidence interval or the smallest rectangle that encompasses the 99.99% Wilks' confidence region. In all cases, the adaptive sub-support obtained from such choices of c covers most of the relevant points in a practical sense.

Parameter	c
$\mathbf{v} \in [0,1]$	$\exp\left(\frac{1}{2}F_{\chi_1^2}^{-1}(0.9999)\right)$
$\mathbf{v} \in [0,1]^2$	$\exp\left(\frac{1}{2}F_{\chi_2^2}^{-1}(0.9999)\right)$
$d \in [-1,1]$	$\exp\left(\frac{1}{2}F_{\chi_1^2}^{-1}(0.9999)\right)$

Table 2: Choices of c for the adaptive sub-support.

Now we show how the sub-support for each dimension can be obtained by solving an optimization problem. Recall that

$$\mathbf{v} = \int_{\mathcal{S}} \mathbf{w}r \, d\mathbf{Q}(\mathbf{w}, r) = \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w}r Q_{\mathbf{w},r},$$

where $\mathcal{S} = \left(\prod_{j=1}^{\ell}[w_{j,\min}, w_{j,\max}]\right) \times [0,1]$, $\mathbf{Q}$ is a probability measure on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, $\mathcal{A} = \{(\mathbf{w},r)\in\mathcal{S} \mid \mathbf{Q}(\{(\mathbf{w},r)\}) > 0\}$, and $Q_{\mathbf{w},r} = \mathbf{Q}(\{(\mathbf{w},r)\})$.

Let $\mathbf{e}_j$ be the unit vector with a one in the j-th component. Then the sub-support for $\mathbf{v}_j$ can be reformulated as

$$\left\{ \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{e}_j^\top \mathbf{w}r Q_{\mathbf{w},r} \;\middle|\; \sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} = 1, \; \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} = \mathbf{1}, \; \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(Q_{\mathbf{w},r}) \geq \phi \right\},$$

where $c_{\mathbf{w},r} = \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{w}=\mathbf{w}_i, r=r_i\}}$ and $\phi = \log L^* - \log c$.

We can then identify the sub-support for $\mathbf{v}_j$ by finding the infimum and supremum of this set with the corresponding optimization problems. We show the derivation for the infimum. The derivation for the supremum is similar and is omitted here. Consider

$$\underset{\mathbf{Q}}{\text{minimize}} \quad \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{e}_j^\top \mathbf{w}r Q_{\mathbf{w},r}$$

$$\text{subject to} \quad \sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} = 1$$

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} = \mathbf{1}$$

$$\sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(Q_{\mathbf{w},r}) \geq \phi.$$

23

The associated Lagrange dual function is given by

$$
\begin{aligned}
g(\gamma, \boldsymbol{\beta}, \kappa) = \inf_{\mathbf{Q}} \Bigg\{ & \sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{e}_j^\top \mathbf{w} r Q_{\mathbf{w},r} + \gamma \left(\sum_{(\mathbf{w},r)\in\mathcal{A}} Q_{\mathbf{w},r} - 1 \right) \\
& + \boldsymbol{\beta}^\top \left(\sum_{(\mathbf{w},r)\in\mathcal{A}} \mathbf{w} Q_{\mathbf{w},r} - \mathbf{1} \right) \\
& + \kappa \left(\phi - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(Q_{\mathbf{w},r}\right) \right) \Bigg\} \\
= \inf_{\mathbf{Q}} \Bigg\{ & -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \kappa\phi \\
& + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ \left(\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r \right) Q_{\mathbf{w},r} - \kappa c_{\mathbf{w},r} \log\left(Q_{\mathbf{w},r}\right) \right\} \Bigg\}.
\end{aligned}
$$

To ensure that the Lagrange dual function is bounded, we must have

$$
\begin{aligned}
\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r \geq 0 & \quad \forall (\mathbf{w}, r) \in \mathcal{A} \subset \mathcal{S} \\
\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r > 0 & \quad \text{if } \kappa c_{\mathbf{w},r} > 0.
\end{aligned}
$$

Note that $\mathcal{A}$ is an arbitrary subset of $\mathcal{S}$. Since we require boundedness for any such $\mathcal{A} \subset \mathcal{S}$, the constraints above must therefore hold for all $(\mathbf{w}, r) \in \mathcal{S}$. Hence, the first constraint simplifies to

$$
\gamma + \boldsymbol{\beta}^\top \mathbf{w} \geq 0 \quad \forall (\mathbf{w}, r) \in \mathcal{S}.
$$

Now, if $\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r = 0$, then $\kappa c_{\mathbf{w},r} = 0$, and the corresponding summand involving $Q_{\mathbf{w},r}$ vanishes from the Lagrange dual function. Hence, the only relevant allocation of $Q_{\mathbf{w},r}$ with respect to the infimum is when $\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r > 0$. Using standard calculus for minimization, we get

$$
Q_{\mathbf{w},r}^* = \frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r}.
$$

Hence, the Lagrange dual function becomes

$$
\begin{aligned}
g(\gamma, \boldsymbol{\beta}, \kappa) &= -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \kappa\phi + \sum_{(\mathbf{w},r)\in\mathcal{A}} \left\{ \kappa c_{\mathbf{w},r} - \kappa c_{\mathbf{w},r} \log\left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right) \right\} \\
&= -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \kappa \left(\phi + n - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log\left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right) \right).
\end{aligned}
$$

Now we arrive at the Lagrange dual problem,

$$
\sup_{(\gamma,\boldsymbol{\beta},\kappa)\in\mathcal{W}_\kappa} g(\gamma,\boldsymbol{\beta},\kappa)
$$

$$
= \sup_{(\gamma,\boldsymbol{\beta},\kappa)\in\mathcal{W}_\kappa} \left\{ -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \kappa \left(\phi + n - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log \left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right) \right) \right\},
$$

where $\mathcal{W}_\kappa = \left\{ (\gamma,\boldsymbol{\beta},\kappa) \in \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R}^+ \,\middle|\, \gamma + \boldsymbol{\beta}^\top \mathbf{w} \geq 0 \ \forall (\mathbf{w},r) \in \mathcal{S} \right\}$.

Note that $\kappa > 0$. Instead of solving the Lagrange dual problem directly, we can first eliminate the dual variable κ using standard calculus for maximization,

$$
\frac{\partial g}{\partial \kappa} = \left(\phi + n - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log \left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right) \right) + \kappa \left(- \sum_{(\mathbf{w},r)\in\mathcal{A}} \frac{c_{\mathbf{w},r}}{\kappa} \right)
$$

$$
= \phi - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log \left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right)
$$

$$
\frac{\partial^2 g}{\partial \kappa^2} = - \sum_{(\mathbf{w},r)\in\mathcal{A}} \frac{c_{\mathbf{w},r}}{\kappa} = -\frac{n}{\kappa} < 0.
$$

Hence, we have

$$
\phi = \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log \left(\frac{\kappa c_{\mathbf{w},r}}{\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r} \right). \tag{$\star$}
$$

Rearranging gives

$$
\kappa = \exp \left(\frac{\phi - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r})}{n} \right) \prod_{(\mathbf{w},r)\in\mathcal{A}} \left(\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r \right)^{c_{\mathbf{w},r}/n}.
$$

Now we reparameterize the Lagrange dual function as

$$
\tilde{g}(\gamma,\boldsymbol{\beta}) = g(\gamma,\ \boldsymbol{\beta},\ \kappa(\gamma,\boldsymbol{\beta})).
$$

Then the Lagrange dual problem simplifies to

$$
\sup_{(\gamma,\boldsymbol{\beta})\in\mathcal{W}} \tilde{g}(\gamma,\boldsymbol{\beta})
$$

$$
= \sup_{(\gamma,\boldsymbol{\beta})\in\mathcal{W}} \left\{ -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \kappa(\gamma,\boldsymbol{\beta})\, n \right\} \qquad \text{(using Identity } \star\text{)}
$$

$$
= \sup_{(\gamma,\boldsymbol{\beta})\in\mathcal{W}} \left\{ -\gamma - \boldsymbol{\beta}^\top \mathbf{1} + \exp \left(\frac{\phi - \sum_{(\mathbf{w},r)\in\mathcal{A}} c_{\mathbf{w},r} \log(c_{\mathbf{w},r}) + n \log(n)}{n} \right) \right.
$$

$$
\left. \times \prod_{(\mathbf{w},r)\in\mathcal{A}} \left(\gamma + \boldsymbol{\beta}^\top \mathbf{w} + \mathbf{e}_j^\top \mathbf{w} r \right)^{c_{\mathbf{w},r}/n} \right\},
$$

where $\mathcal{W} = \left\{ (\gamma,\boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^\ell \,\middle|\, \gamma + \boldsymbol{\beta}^\top \mathbf{w} \geq 0 \ \forall (\mathbf{w},r) \in \mathcal{S} \right\}$.

Finally, by strong duality, the infimum of the sub-support for $\mathbf{v}_j$ can be obtained by solving this simplified Lagrange dual problem.

References

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375.

Robert I. Berkowitz, Thomas A. Wadden, Christine A. Gehrman, Chanelle T. Bishop-Gilyard, Reneé H. Moore, Leslie G. Womble, Joanna L. Cronquist, Natalie L. Trumpikas, Lorraine E. Levitt-Katz, and Melissa S. Xanthopoulos. Meal Replacements in the Treatment of Adolescent Obesity: A Randomized Controlled Trial. *Obesity*, 19(6):1193–1199, 2011. doi: 10.1038/oby.2010.288.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Sanjay Chaudhuri and Malay Ghosh. Empirical likelihood for small area estimation. *Biometrika*, 98(2):473–480, 2011. doi: 10.1093/biomet/asr004.

Sanjay Chaudhuri, Debashis Mondal, and Teng Yin. Hamiltonian Monte Carlo Sampling in Bayesian Empirical Likelihood Computation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):293–320, 2017. doi: 10.1111/rssb.12164.

Miroslav Dudik, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 1097–1104, 2011. doi: 10.48550/arXiv.1103.4601.

Marian Grendár and George Judge. Asymptotic equivalence of empirical likelihood and Bayesian MAP. *The Annals of Statistics*, 37(5A):2445–2457, 2009. doi: 10.1214/08-AOS645.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Nan Jiang and Lihong Li. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 652–661. PMLR, 2016. doi: 10.48550/arXiv.1511.03722.

Nathan Kallus and Masatoshi Uehara. Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 3325–3334. Curran Associates, Inc., 2019. doi: 10.48550/arXiv.1906.03735.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539, 2007. doi: 10.1214/07-STS227.

Nikos Karampatziakis, John Langford, and Paul Mineiro. Empirical Likelihood for Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 9597–9607. Curran Associates, Inc., 2020. doi: 10.48550/arXiv.1906.03323.

Ancel Keys, Flaminio Fidanza, Martti J. Karvonen, Noboru Kimura, and Henry L. Taylor. Indices of relative weight and obesity. *Journal of Chronic Diseases*, 25(6):329–343, 1972. doi: 10.1016/0021-9681(72)90027-6.

John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *Advances in Neural Information Processing Systems*, volume 20, pages 817–824. Curran Associates, Inc., 2007. URL `https://papers.nips.cc/paper_files/paper/2007/hash/4b04a686b0ad13dce35fa99fa4161c65-Abstract.html`.

Nicole A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90(2):319–326, 2003. doi: 10.1093/biomet/90.2.319.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670. Association for Computing Machinery, 2010. doi: 10.1145/1772690.1772758.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, Haywood CA and Alexandria VA, 1995. doi: 10.1214/cbms/1462106013.

Kristin A. Linn, Eric B. Laber, and Leonard A. Stefanski. iqLearn: Interactive Q-Learning in R. *Journal of Statistical Software*, 64:1–25, 2015. doi: 10.18637/jss.v064.i01.

Kerrie L. Mengersen, Pierre Pudlo, and Christian P. Robert. Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4):1321–1326, 2013. doi: 10.1073/pnas.1208827110.

Yusuke Narita, Shota Yasui, and Kohei Yata. Efficient Counterfactual Learning from Bandit Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4634–4641, 2019. doi: 10.1609/aaai.v33i01.33014634.

Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18 (1):90–120, 1990. doi: 10.1214/aos/1176347494.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. doi: 10.1093/biomet/75.2.237.

Jin Qin and Jerry Lawless. Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, 22(1):300–325, 1994. doi: 10.1214/aos/1176325370.

J. N. K. Rao and Changbao Wu. Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4): 533–544, 2010. doi: 10.1111/j.1467-9868.2010.00747.x.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. doi: 10.1090/S0002-9904-1952-09620-8.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, second edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.

Adith Swaminathan and Thorsten Joachims. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3231–3239. Curran Associates, Inc., 2015. URL `https://papers.nips.cc/paper/2015/hash/39027dfad5138c9ca0c474d71db915c3-Abstract.html`.

Ambuj Tewari and Susan A. Murphy. From Ads to Interventions: Contextual Bandits in Mobile Health. In James M. Rehg, Susan A. Murphy, and Santosh Kumar, editors, *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-51394-2_25.

Sofía S. Villar and William F. Rosenberger. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics*, 74(1):49–57, 2018. doi: 10.1111/biom.12738.

Yunwen Yang and Xuming He. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131, 2012. doi: 10.1214/12-AOS1005.

Weichang Yu and Howard D. Bondell. Variational Bayes for Fast and Accurate Empirical Likelihood Inference. *Journal of the American Statistical Association*, 119(546):1089–1101, 2024. doi: 10.1080/01621459.2023.2169701.