Balanced Off-Policy Evaluation for Personalized Pricing

Adam N. Elmachtoub
Columbia IEOR and DSI

Vishal Gupta
USC Marshall

Yunfan Zhao
Columbia IEOR

Abstract

We consider a personalized pricing problem in which we have data consisting of feature information, historical pricing decisions, and binary realized demand. The goal is to perform off-policy evaluation for a new personalized pricing policy that maps features to prices. Methods based on inverse propensity weighting (including doubly robust methods) for off-policy evaluation may perform poorly when the logging policy has little exploration or is deterministic, which is common in pricing applications. Building on the balanced policy evaluation framework of Kallus (2018), we propose a new approach tailored to pricing applications. The key idea is to compute an estimate that minimizes the worst-case mean squared error or maximizes a worst-case lower bound on policy performance, where in both cases the worst-case is taken with respect to a set of possible revenue functions. We establish theoretical convergence guarantees and empirically demonstrate the advantage of our approach using a real-world pricing dataset.

1 INTRODUCTION

Data-driven and personalized pricing has received considerable attention over the past two decades (Cohen et al., 2017; Besbes et al., 2010; Ferreira et al., 2016; Bu et al., 2022; Baardman et al., 2019; Wang and Zheng, 2021; Qi et al., 2022; Biggs, 2022). Utilizing contextual information in pricing is especially popular due to applications in online shopping (Nambiar et al., 2019; Elmachtoub et al., 2021), auto lending (Phillips et al., 2015; Ban and Keskin, 2021), air travel (Kolbeinsson et al., 2022) and beyond (Chen et al., 2022; Wang et al., 2021; Aouad et al., 2019). The increasing availability of customer data enables personalized pricing strategies. However, experimenting with a new personalized pricing policy that is potentially more profitable or fairer (Cohen et al., 2022) can be costly and difficult, motivating the use of off-policy evaluation. Specifically, we study the problem of off-policy evaluation for personalized pricing where feature information such as customer order history, demographics, and market conditions are observed alongside the offered prices and binary purchase decisions.

There is an extensive literature on off-policy evaluation. Inverse propensity weighting (IPW) and doubly robust (DR) methods are especially popular (Dudík et al., 2011; Hanna et al., 2017; Swaminathan and Joachims, 2015a; Thomas and Brunskill, 2016; Wang et al., 2017; Bottou et al., 2013; Athey and Wager, 2021). Both approaches reweight historical data to make the data look as if they were generated by the target policy that we wish to evaluate. While initial research in the area focused on finite, discrete action spaces, more recently Sondhi et al. (2020); Kallus and Zhou (2018); Cai et al. (2021) propose extensions to more general, potentially infinite, action spaces. Biggs et al. (2021) recasts IPW methods as optimizing a particular loss function and uses this insight to propose suitable generalizations.

Each of the aforementioned methods leverages an approximation of the inverse propensity score to form weights. As noted by Kallus (2018), an inherent shortcoming of such approaches is that when the overlap between the target and logging policy is limited, these methods assign large weights to a small number of data points in the overlap and assign zero weight elsewhere. This weighting scheme yields high variance estimates, especially on small datasets. In the worst-case when there is zero overlap, IPW methods are not even well-defined.

While such cases might seem pathological, they are common in pricing applications. Many real-world firms are reticent to engage in extensive randomized pricing, making limited overlap fairly prevalent. When firms price deterministically, even simple policy adjustments such as raising all prices 2% yield zero overlap. These features make the aforementioned methods less attractive.

Many authors have proposed general purpose modifications of traditional methods to address these shortcomings.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Elliott (2008); Ionides (2008); Swaminathan and Joachims (2015a,b) each propose various ways to regularize the naive IPW weights, e.g. by clipping large values, to reduce variance. These methods introduce additional bias into estimates, often in ways that are instance dependent and difficult to quantify.

Other authors attempt to circumvent the issue with IPW by focusing on policy learning – i.e., identifying a good policy – rather than policy evaluation. In cases of zero overlap, Sachdeva et al. (2020) compares three different approaches – restricting the action space, extrapolating reward, and restricting policy spaces – and argues in favor of restricting policy spaces. Kallus (2021) proposes a retargeting approach which reframes the optimal policy as the solution to an alternate off-policy problem with better overlap properties and near-optimal asymptotic variance. As stated, however, neither approach directly addresses policy evaluation. Insofar as firms are often interested in the performance of a specific, target pricing policy that may have been chosen for qualitative, business-specific reasons, there remains a need for effective policy evaluation methods that balance bias and variance and provide provable performance guarantees.

Inspired by the balanced policy evaluation method of Kallus (2018), we propose an alternate approach to off-policy evaluation for pricing applications. Like IPW and DR methods, we estimate the performance of the policy by a weighted average of the historical data points. However, unlike these methods, we use weights that either *(i)* minimize the worst-case mean squared error of our estimated revenue or *(ii)* maximize a worst-case lower bound on the unknown target revenue. In both cases, the worst-case is taken over a set of plausible revenue functions.

Our work differs from Kallus (2018) in three critical aspects: *(i)* We focus on a binary demand response variable rather than a continuous one with a homoscedastic variance. Binary demand induces a more complex form for the variance of our estimator and consequently complicates the worst-case optimization problem defining our weights. By contrast, the corresponding optimization in Kallus (2018) is an unconstrained, convex quadratic program with a closed-form solution. *(ii)* Although we treat worst-case mean squared error (MSE) (similar to Kallus (2018)), firms are also concerned with operational criteria such as a guaranteed lower bound on revenue. We show how our approach can be modified to compute such a lower bound (via Bernstein's inequality) and contrast the behavior of the resulting estimator with the MSE approach. *(iii)* Kallus (2018) focuses primarily on the case of a small number of discrete actions, while typical pricing problems involve continuous action spaces. In particular, one cannot apply Kallus (2018) "out-of-the-box" to continuous action spaces, since the approach assumes no structure across actions (prices) and would thus yields overly conservative estimates (Kallus (2020) suggests a way to address this). By contrast, we enforce smoothness of the demand function across prices by assuming this revenue function belongs to a particular reproducing kernel Hilbert space (RKHS).

Paper Outline: We start in Section 2 with the notation and setup, followed by an analysis of weighted revenue estimators in Section 3. We present our off-policy evaluation approach in Section 4. In Section 5, we establish theoretical guarantees for our approaches. We present experimental results on both synthetic datasets and a real world pricing dataset in Section 6. We describe heuristics for estimating parameters in Section 7 and conclude in Section 8.

2 NOTATION AND MODEL

We assume the following (fixed-design) data generation mechanism: We are given a set features $\boldsymbol{x}_i \in \mathcal{X}$ for $i = 1 \ldots, n$. Price-demand pairs are distributed as

$$\begin{aligned} P_i &\sim g_0(\cdot, \boldsymbol{x}_i), && i = 1, \ldots, n, \\ D_i \mid P_i &\sim \text{Bernoulli}(d(\boldsymbol{x}_i, P_i)), && i = 1, \ldots, n, \end{aligned}$$

for some unknown demand function $d(\cdot, \cdot)$ that maps features and prices to $[0, 1]$. Here the density $g_0(\cdot, \cdot)$ encodes our logging pricing policy, i.e. we draw a random price from density $g_0(\cdot, \boldsymbol{x})$ when presented with a feature $\boldsymbol{x}$. When the logging policy is deterministic, we interpret $g_0(\cdot, \boldsymbol{x})$ as a Dirac delta function.

Our dataset $\{(\boldsymbol{x}_i, p_i, d_i) \subseteq \mathcal{X} \times \mathbb{R}_+ \times \{0, 1\} : i \in [n]\}$ consists of single a realization of this process.

Loosely, our goal is to evaluate a target policy that draws a random price from the density $g_1(\cdot, \boldsymbol{x})$ when presented with feature $\boldsymbol{x}$. Formally, let

$$P_{n+i} \sim g_1(\cdot, \boldsymbol{x}_i) \qquad i = 1, \ldots, n,$$

and let $p_{n+i} \in \mathbb{R}$ for $i \in [n]$ be a corresponding realization. Then, if we define the expected revenue function $r(\boldsymbol{x}, p) := pd(\boldsymbol{x}, p)$, the expected revenue under the target policy is

$$\begin{aligned} \mathcal{R} &:= \tfrac{1}{n} \textstyle\sum_{i=1}^n p_{n+i} d(\boldsymbol{x}_i, p_{n+i}) && \text{(Target Revenue)} \\ &= \tfrac{1}{n} \textstyle\sum_{i=1}^n r(\boldsymbol{x}_i, p_{n+i}), \end{aligned}$$

which we emphasize is a constant. Our goal is to estimate and provide high confidence bounds on this constant.

We stress that, in what follows, our method does not require explicit knowledge of $g_0(\cdot, \cdot)$ or $g_1(\cdot, \cdot)$.

In keeping with the literature on doubly-robust estimators, we define a reference revenue function:

Definition 1 (Reference Revenue Function)**.** *The revenue function can be written as $r(\cdot, \cdot) = \hat{r}(\cdot, \cdot) + \Delta(\cdot, \cdot)$, for a known reference revenue $\hat{r}(\cdot, \cdot)$, and a perturbation function $\Delta(\cdot, \cdot)$.*

This decomposition is without loss of generality (take $\hat{r}(\cdot,\cdot) = 0$). In practice, we may have a good reference model $\hat{r}(\cdot,\cdot)$ that we believe reasonably captures the revenue curve. Thus, the estimators are best thought of as a perturbation to this reference.

To streamline notation, we define $\boldsymbol{p} \in \mathbb{R}^{2n}$ to be the vector of prices $p_1, ..., p_{2n}$. Similarly, we define the vectors $\boldsymbol{r}, \hat{\boldsymbol{r}}, \boldsymbol{\Delta} \in \mathbb{R}^{2n}$ such that for for $i \in [2n]$,

$$r_i = r(\boldsymbol{x}_i, p_i), \;\; \hat{r}_i = \hat{r}(\boldsymbol{x}_i, p_i), \;\; \Delta_i = \Delta(\boldsymbol{x}_i, p_i).$$

We focus on the doubly robust weighted revenue estimator

$$\hat{\mathcal{R}}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} w_i(p_i D_i - \hat{r}_i) + \frac{1}{n}\sum_{i=n+1}^{2n} \hat{r}_i, \;\; \text{(Estimator)}$$

for some weights $\boldsymbol{w}$ that we will specify.

3 PROPERTIES OF WEIGHTED REVENUE ESTIMATORS

We first introduce general properties of weighted revenue estimators with honest weights, i.e., the weights are independent of demand realizations. These properties depend on the vector $\boldsymbol{r}$ which in practice is unknown. Nonetheless, these properties serve as a building block for our approach later on where we take a worst-case perspective on $\boldsymbol{r}$.

3.1 Mean Squared Error

Define

$$\begin{aligned}\text{MSE}(\boldsymbol{w}, \boldsymbol{r}) &:= \mathbb{E}\left[(\mathcal{R} - \hat{\mathcal{R}}(\boldsymbol{w}))^2\right] \\ &= \mathbb{E}\left[\left(\mathcal{R} - \frac{1}{n}\sum_{i=n+1}^{2n} \hat{r}_i - \frac{1}{n}\sum_{i=1}^{n} w_i(p_i D_i - \hat{r}_i)\right)^2\right]. \quad (1)\end{aligned}$$

Note $\mathcal{R}$ and $\mathbb{E}[p_j D_j]$ depend on the unknown revenue $\boldsymbol{r}$. In Lemma 1 below, we provide a more explicit expression for the MSE, which takes into account the binary nature of demand. (See Appendix A for proofs.)

Lemma 1 (Bias and Variance Decomposition)**.** *Let*

$$\begin{aligned}\boldsymbol{b}(\boldsymbol{w}) &:= \frac{1}{n}(w_1, \ldots, w_n, -1, \ldots, -1)^\top \in \mathbb{R}^{2n} \\ \boldsymbol{v}(\boldsymbol{w}) &:= \frac{1}{n^2}\left(w_1^2 p_1, \ldots, w_n^2 p_n, 0, \ldots, 0\right)^\top \in \mathbb{R}^{2n}\end{aligned}$$

Then, we have

$$\begin{aligned}Bias(\boldsymbol{w}, \boldsymbol{r}) &:= \mathbb{E}\left[\hat{\mathcal{R}}(\boldsymbol{w}) - \mathcal{R}\right] = \boldsymbol{b}(\boldsymbol{w})^\top(\boldsymbol{r} - \hat{\boldsymbol{r}}), \\ Var(\boldsymbol{w}, \boldsymbol{r}) &:= \mathbb{E}\left[\left(\hat{\mathcal{R}}(\boldsymbol{w}) - \mathbb{E}\left[\hat{\mathcal{R}}(\boldsymbol{w})\right]\right)^2\right] \\ &= \boldsymbol{v}(\boldsymbol{w})^\top \boldsymbol{r} - \frac{1}{n^2}\boldsymbol{r}^\top \begin{pmatrix} diag(w_1^2, \ldots, w_n^2) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} \boldsymbol{r},\end{aligned}$$

and, of course, $MSE(\boldsymbol{w}, \boldsymbol{r}) = Bias(\boldsymbol{w}, \boldsymbol{r})^2 + Var(\boldsymbol{w}, \boldsymbol{r})$.

3.2 High-Probability Bound

We next provide a high-confidence lower bound on the true revenue $\mathcal{R}$ in terms of the estimate $\hat{\mathcal{R}}(\boldsymbol{w})$. From an operational perspective, lower bounds provide "safe" guarantees on potential revenue. Although similar techniques could be used to form upper bounds, they are less useful in practice.

Define

$$\begin{aligned}Bern(\boldsymbol{w}, \boldsymbol{r}) := \boldsymbol{b}(\boldsymbol{w})^\top(\boldsymbol{r} - \hat{\boldsymbol{r}}) &+ \sqrt{2Var(\boldsymbol{w}; \boldsymbol{r})\log(1/\epsilon)} \\ &+ \frac{1}{3n}\max_{1\le i\le n} |w_i| p_i \log(1/\epsilon). \quad (2)\end{aligned}$$

Lemma 2 (Revenue Lower Bound)**.** *With probability at least* $1 - \epsilon$ *over the realization of* $(D_1, \ldots, D_n)$*, we have that* $\mathcal{R} \ge \hat{\mathcal{R}}(\boldsymbol{w}) - Bern(\boldsymbol{w}, \boldsymbol{r})$.

The lemma is a direct application of Bernstein's inequality.

Remark 1 (Convexity in $\boldsymbol{w}$)**.** *Since the expectation of a convex function is convex, Eq.* (1) *shows the map* $\boldsymbol{w} \mapsto MSE(\boldsymbol{w}, \boldsymbol{r})$ *is convex in* $\boldsymbol{w}$ *for a fixed* $\boldsymbol{r}$*. Similarly, the function* $Bern(\boldsymbol{w}, \boldsymbol{r})$ *is convex in* $\boldsymbol{w}$ *for a fixed* $\boldsymbol{r}$ *since* $\sqrt{Var(\boldsymbol{w}; \boldsymbol{r})} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} w_i^2 r_i(p_i - r_i)}$ *is a weighted* ℓ_2*-norm, and, hence,* $\boldsymbol{w} \mapsto Bern(\boldsymbol{w}, \boldsymbol{r})$ *is a sum of convex functions. We will leverage these convexity properties when formulating optimization problems to compute our weights.*

4 A BALANCED APPROACH FOR OFF-POLICY EVALUATION IN PRICING

The expression for mean squared error and the lower bound in Eq. (2) depend on the unknown revenue vector $\boldsymbol{r}$. Our approach will be to compute weights $\boldsymbol{w}$ that optimize these metrics over "plausible" worst case realizations of $\boldsymbol{r}$. To define "plausible," we make the following assumption for the remainder of the paper:

Assumption 1 (Perturbation Function is in RKHS)**.** *There exists an RKHS* $\mathcal{H}$ *with kernel* $K(\cdot,\cdot)$ *and norm* $\|\cdot\|_{\mathcal{H}}$ *such that* $\Delta(\cdot,\cdot) \in \mathcal{H}$ *and* $\|\Delta(\cdot,\cdot)\|_{\mathcal{H}} < \infty$.

Assumption 1 asserts that the unknown perturbation function is "smooth" in the sense that it has a bounded RKHS norm. By suitably choosing the kernel $K(\cdot,\cdot)$, we can enforce structural constraints on $\Delta(\cdot,\cdot)$, e.g., that $\Delta(\cdot,\cdot)$ is linear in price or Sobolev smooth in the covariates. See (Smola and Schölkopf, 2004) for details.

For notational convenience, we let $\Gamma := \|\Delta(\cdot,\cdot)\|_{\mathcal{H}}$. Define the Graham matrix $\boldsymbol{G} \in \mathbb{R}^{2n\times 2n}$ by

$$\boldsymbol{G}_{ij} := K\left((\boldsymbol{x}_i, p_i), (\boldsymbol{x}_j, p_j)\right) \qquad 1 \le i, j \le 2n.$$

Under Assumption 1, the Representer Theorem (Wahba, 1990) implies that there exists $\boldsymbol{\alpha} \in \mathbb{R}^{2n}$ such that

$$\boldsymbol{\Delta} = \boldsymbol{G}\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{\alpha}^\top \boldsymbol{G}\boldsymbol{\alpha} = \Gamma^2.$$

We further make the following common assumption.

Assumption 2. *The Graham matrix $\boldsymbol{G}$ is invertible.*

From Assumptions 1 and 2,

$$\boldsymbol{\Delta}^\top \boldsymbol{G}^{-1} \boldsymbol{\Delta} = \Gamma^2. \tag{3}$$

Since $d(x, p) \in [0, 1]$ for any x, p, we have

$$\mathbf{0} \le \boldsymbol{r} \le \boldsymbol{p}. \tag{4}$$

Combining (3) and (4), we seek weights that minimize a worst-case metrics $\phi(\boldsymbol{w}, \boldsymbol{r})$ over plausible revenue functions:

$$\begin{aligned} \boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}} \max_{\boldsymbol{r}} \quad & \phi(\boldsymbol{w}, \boldsymbol{r}) \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{r} \le \boldsymbol{p}\,, \quad (\boldsymbol{r} - \hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1} (\boldsymbol{r} - \hat{\boldsymbol{r}}) \le \hat{\Gamma}^2. \end{aligned} \tag{5}$$

Here $\phi(\boldsymbol{w}, \boldsymbol{r})$ can be MSE$(\boldsymbol{w}, \boldsymbol{r})$ or Bern$(\boldsymbol{w}, \boldsymbol{r})$. We denote the corresponding solutions by $\boldsymbol{w}^{MSE}$ and $\boldsymbol{w}^{B}$, respectively.

Since in practice, we do not know the ground truth Γ, we proxy Γ by user-specified constant $\hat{\Gamma}$ in Problem 5. (We discuss heuristics for estimating $\hat{\Gamma}$ in Section 7.)

Remark 2 (Unconstrained Weights)**.** *In contrast to Kallus (2018), we do not impose an additional simplex constraint on the weights. Indeed, the value of the target policy need not be on the same order of magnitude as the logging policy, e.g., when we raise price significantly. Thus, an ideal set of weights might not satisfy such a constraint. That said, our Bernstein variant ($\boldsymbol{w}^B$) does regularize away from overly large weights via the weighted ℓ_∞ norm in Eq.* (2)*. This regularization emerges naturally via the probabilistic analysis rather than being imposed via an artificial simplex, or normalizing, constraint.*

Remark 3 (Honest vs. Dishonest Weights)**.** *When $\hat{\boldsymbol{r}}$, $\hat{\Gamma}$, and the kernel $K(\cdot, \cdot)$ are specified exogenously, i.e., independently of the demand realizations, both $\boldsymbol{w}^{MSE}$ and $\boldsymbol{w}^{B}$ are honest. We study the corresponding estimators $\hat{R}(\boldsymbol{w}^{MSE})$ and $\hat{R}(\boldsymbol{w}^{B})$ theoretically in Section 5.*

In practice, we suggest fitting these parameters to the data via the heuristics in Section 7. The resulting weights are "dishonest." While it might be possible to extend our theoretical results to this setting by assuming that $(\hat{r}, \hat{\Gamma}, K(\cdot, \cdot))$ are chosen from a suitably low-complexity class, we do not pursue this theoretical analysis here. Rather, we present numerical evidence in Sec. 6 that even with dishonest weights, our estimator performs well.

4.1 Solution Approach

We next discuss how to solve (5).

For a fixed $\boldsymbol{w}$, consider the inner problem of finding the worst case (WC) revenue:

$$\begin{aligned} \boldsymbol{r}^{WC}(\boldsymbol{w}) := \operatorname*{argmax}_{\boldsymbol{r}} \quad & \phi(\boldsymbol{w}, \boldsymbol{r}) \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{r} \le \boldsymbol{p}\,, \quad (\boldsymbol{r} - \hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1} (\boldsymbol{r} - \hat{\boldsymbol{r}}) \le \hat{\Gamma}^2. \end{aligned}$$

Let $h(\boldsymbol{w}) = \phi(\boldsymbol{w}, \boldsymbol{r}^{WC}(\boldsymbol{w}))$. Since $\boldsymbol{w} \mapsto \phi(\boldsymbol{w}, \boldsymbol{r})$ is convex for each $\boldsymbol{r}$ by Remark 1, Danskin's Theorem (Bertsekas, 1997) shows that $h(\boldsymbol{w})$is in fact convex in $\boldsymbol{w}$, and, when $\boldsymbol{r}^{WC}(\boldsymbol{w})$ is the unique optimizer,

$$\nabla h(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w}, \boldsymbol{r})|_{\boldsymbol{r} = \boldsymbol{r}^{WC}(\boldsymbol{w})}.$$

Thus, we can minimize $h(\boldsymbol{w})$ using any number of gradient-based algorithms. (In our numerical experiments, we use the a first order trust-region method from scipy.optimize.) Evaluating a gradient requires determining $\boldsymbol{r}^{WC}(\boldsymbol{w})$, i.e., solving the inner problem.

That said, for large n, computing gradients in the Bernstein objective is perhaps easier than for the MSE objective. For the Bernstein objective, the inner maximization problem can be reformulated as a concave quadratic maximization problem in $\boldsymbol{r}$ (see Appendix B). By contrast, for the MSE objective, the inner problem is an in-definite quadratic programming problem. Such problems can, in the worst-case, be NP-Hard, but are often practically solvable with modern solvers for moderate sized instances. In our experiments, we use Gurobi for both computations.

5 THEORETICAL RESULTS

Recall our approach to off-policy evaluation for pricing applications is partially motivated by the observation that in typical pricing applications, the overlap between the logging and evaluation policies may be small since both policies may entail little randomization. This feature precludes the use of methods based on inverse propensity scores that require sufficient overlap, including doubly-robust methods.

In this section we establish a "sanity-check" result, i.e., that when sufficient overlap does exist, our method achieves convergence rates similar to the doubly-robust methods.

Assumption 3 (Overlap)**.** *For all $(p, \boldsymbol{x}) \in \mathbb{R}_+ \times \mathcal{X}$, if $g_0(p, \boldsymbol{x}) = 0$ then $g_1(p, \boldsymbol{x}) = 0$.*

From Assumption 3, the inverse propensity (IP) weights

$$W_i^{IP} := \frac{g_1(P_i, \boldsymbol{x}_i)}{g_0(P_i, \boldsymbol{x}_i)} \tag{6}$$

are well-defined for all $i = 1, \ldots, n$.

5.1 Mean Squared Error

We first consider $\boldsymbol{w}^{MSE}$ and the corresponding estimator $\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$. Theorem 1 shows that true (unknown) MSE of this estimator converges to zero at a rate of $\frac{1}{n}$, despite not knowing $g_0(\cdot,\cdot)$, $g_1(\cdot,\cdot)$ or Γ. (See Appendix C for proof.)

For convenience, let $Z_i = (\boldsymbol{x}_i, P_i)$ for $i = 1, \ldots, 2n$.

Theorem 1 (Convergence of MSE)**.** *Suppose that*

i) $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(W_i^{IP} - 1\right) K(Z_{n+i}, Z_{n+i})\right] = O(1)$

ii) $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[(W_i^{IP} P_i)^2\right] = O(1)$

Then, under Assumptions 1, 2, and 3, we have $\mathit{MSE}(\boldsymbol{w}^{MSE}, \boldsymbol{r}) = O_p\left(\frac{1}{n}\right)$.

For clarity, the "probability" in Theorem 1 is taken over the randomness in both $\{D_i : i \in [n]\}$ and $\{P_i : i \in [2n]\}$.

To help develop intuition around the assumptions of the above theorem, consider the case where $K(\cdot,\cdot)$ is the gaussian kernel, so that $K(Z_{n+i}, Z_{n+i})$ is almost surely a constant. Then the first condition $i)$ holds trivially since $\mathbb{E}\left[W_i^{IP}\right] = 1$ by construction. The second condition $ii)$ essentially requires that for a typical point, the inverse propensity score weights are not too large – they are $O(1)$. This requirement is analogous to requiring sufficient overlap between the logging and evaluation policies, since W^{IP} explodes as the overlap shrinks. In this sense, Theorem 1 is a "sanity-check" result.

5.2 Bernstein Bound

We next consider $\boldsymbol{w}^{B}$ and corresponding estimator $\hat{\mathcal{R}}(\boldsymbol{w}^{B})$. Recall Lemma 2 shows that, with high probability, $\hat{\mathcal{R}}(\boldsymbol{w}^{B}) - \mathrm{Bern}(\boldsymbol{w}^{B}, \boldsymbol{r})$ lower bounds the true (unknown) revenue. We will next show that this lower bound is not too loose, specifically, that $\mathrm{Bern}(\boldsymbol{w}^{B}, \boldsymbol{r}) = O_p(1/\sqrt{n})$. (See Appendix D for proof.)

Theorem 2 (Safe Guarantee)**.** *Suppose that*

i) $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(W_i^{IP} - 1\right) K(Z_{n+i}, Z_{n+i})\right] = O(1)$

ii) $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[(W_i^{IP} P_i)^2\right] = O(1)$

Then, under Assumptions 1, 2, and 3, we have $\max\left(0, \mathit{Bern}(\boldsymbol{w}^{B}, \boldsymbol{r})\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$.

In other words, the unknown true revenue cannot exceed our estimate by more than $O_p(1/\sqrt{n})$. In this sense, our estimate provides a "safe" guarantee that is not too loose.

Remark 4 (One-Sided vs. Two-Sided Bounds)**.** *In Theorem 2, we obtain a one-sided convergence result because we used a one-sided probability bound to define* $\mathit{Bern}(\boldsymbol{w}, \boldsymbol{r})$*. If one sought a stronger two-sided convergence, one could instead introduce an absolute value in Eq.* (2) *and define the corresponding estimator.*

In our numerical experiments, we found this "two-sided" estimator performs worse than our proposed one-sided estimator. Hence we have chosen to only present theoretical results for the one-sided estimator.

6 NUMERICAL RESULTS

We describe our numerical results, but please see our GitHub for for reproducibility code and documentation. [1]

6.1 Mean Squared Error

We first study $\boldsymbol{w}^{MSE}$ and corresponding estimator $\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$. We denote our corresponding method as BOPE-B for "Balanced Off-Policy Evaluation for Binary response."

We compare the performance of the following methods on synethetic and real-world datasets:

- (LASSO) A "direct" regression estimator corresponding to $\hat{\mathcal{R}}(\mathbf{0})$. This linear regression method with ℓ_1 penalty predict the demand $d(\cdot,\cdot)$, and revenue is obtained from multiplying it by the price. This serves as a baseline.

- (SPPE) Semi-parametric policy evaluation (Chernozhukov et al., 2019) which is an extension of the classical DR method to a setting where the dependence of the policy value on the treatment is known. In pricing applications, this amounts to specifying a priori how demand depends on price. In our experiments, we apply the method assuming demand is linear in price.

- (BOPE) The Balanced Off-Policy Evaluation method of Kallus (2018). This method can be seen as an instance of Problem 5 with $\phi(\boldsymbol{w}, \boldsymbol{r}) = Bias^2(\boldsymbol{w}, \boldsymbol{r}) + \frac{1}{n^2}\sigma^2 \sum_{i=1}^{n} w_i^2$ for some user-defined σ^2. Loosely, this objective is the worst-case mean squared error if $p_i D_i$ were homoscedastic random variables with variance σ^2 and mean $r(p_i, \boldsymbol{x}_i)$. Thus, this method does not exploit the binary structure of demand. We select hyperparameters according to the heuristic proposed in Kallus (2018) (see Section 7).

- (BOPE-B) Our proposed Balanced Off-Policy Evaluation estimator for Binary response, $\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$, with hyper-parameters chosen according to heuristics described in Section 7.

[1] https://github.com/yzhao3685/pricing-evaluation

For each of BOPE-B, BOPE, and LASSO, we use a LASSO linear regression to estimate $\hat{r}(\cdot,\cdot)$.

Before delving into the details of the experiments, we summarize our main findings:

- By exploiting the binary nature of demand, the BOPE-B estimator generally has an advantage over the BOPE estimator, and substantive advantage of the SPPE estimator.
- When the baseline LASSO, itself, has small MSE, there is little room for improvement and both BOPE and BOPE-B perform comparably. When the baseline estimate is poor, both BOPE and BOPE-B peform substantively better than baseline.
- Generally, the improvements in the BOPE-B estimator over the BOPE estimator are driven by improvements in *both* bias and variance, but in many cases, the improvement in variance is the dominant factor.
- The SPPE method can perform quite poorly when the assumption on the apriori structure of demand does not hold.

6.1.1 Synthetic Datasets

We present results for two different demand functions.

(a) A Simple Demand Function

The features $\boldsymbol{x}_i$ are generated uniformly random from the square $[-1,1]^2$. The logging pricing policy is $P_i = \frac{1}{2}\boldsymbol{x}_i^\top[1,-1] + 7 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,2)$ are i.i.d. noise. The target pricing policy is $P_i = \frac{1}{2}\boldsymbol{x}_i^\top[1,-1] + b + \epsilon_i$, where b is chosen from $\{2,3,4\}$ and then fixed throughout each experiment. We present results for each value of b.

The demand function is

$$d(\boldsymbol{x},p) = \frac{1}{4} + \frac{3}{4}\sigma\left(5 - \frac{1}{2}p - \boldsymbol{x}^\top[-1,1]\right),$$
$$\text{where } \sigma(y) = \frac{1}{1+e^{-y}}.$$

The sigmoid function $\sigma(y)$ is used to ensure *(i)* demand is within [0,1] *(ii)* demand decreases while price increases.

We fix the sample size to be $n = 50$ throughout the experiment. We use the ground truth to simulate realizations of the binary demand vector corresponding to these 50 sample points. We repeat the procedure 100 times to obtain the bias, variance, and MSE of the four estimators. We perform the experiment for 30 different random seeds and report the average results in Table 1. Notice for each random seed, we sample a different set of features and prices.

(b) A Different Demand Function

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy has $b = 2$.				
MSE	1.63	1.83	1.71	1.08
Bias^2	0.22	0.27	0.25	0.17
Variance	1.41	1.56	1.46	0.91
Target Policy has $b = 3$.				
MSE	1.73	1.95	1.80	1.92
Bias^2	0.33	0.40	0.35	0.17
Variance	1.40	1.55	1.45	1.75
Target Policy has $b = 4$.				
MSE	1.50	1.81	1.57	1.60
Bias^2	0.31	0.38	0.33	0.16
Variance	1.19	1.43	1.24	1.44

Table 1: Decomposition of the mean squared error. Synthetic dataset setting (a).

We consider a different demand function

$$d(\boldsymbol{x},p) = \frac{1}{4} + \frac{3}{4}\sigma\left(5 - \frac{1}{2}p - \arctan(\boldsymbol{x}_1/\boldsymbol{x}_2)\right).$$

Notice this demand function is more complicated than that in setting (a). In the sigmoid function, we now have a nonlinear function $\arctan(\boldsymbol{x}_1/\boldsymbol{x}_2)$ instead of the linear function $\boldsymbol{x}^\top[-1,1]$.

The rest of the set up is the same as in part (a). We repeat the experiment for 30 different random seeds and report the average results in Table 2.

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy has $b = 2$.				
MSE	1.18	1.47	1.21	1.43
Bias^2	0.20	0.25	0.22	0.14
Variance	0.98	1.22	0.99	1.29
Target Policy has $b = 3$.				
MSE	2.09	2.30	2.13	2.22
Bias^2	0.52	0.57	0.54	0.46
Variance	1.57	1.73	1.59	1.76
Target Policy has $b = 4$.				
MSE	1.99	2.18	2.05	2.42
Bias^2	0.38	0.45	0.41	0.27
Variance	1.61	1.73	1.64	2.15

Table 2: Decomposition of the mean squared error. Synthetic dataset setting (b).

6.1.2 A Real World Dataset

We conduct experiments on a real world dataset of auto loan applications collected by a major auto lender in North America. The dataset was first studied by Phillips et al. (2015) and later used to evaluate personalized pricing algorithms by Ban and Keskin (2021). The dataset includes data collected over a period of several years. We present results for 5 different subsets of the Nomis dataset. To train the models, we use two covariates: FICO score and requested loan amount. We use the offered interest rate as price. We consider four target policies that take the original prices and increase/decrease them by 5 or 10%.

We impute counterfactuals, including the expected demand, using XGBoost trained on the entire subset to represent the ground truth model. We choose $n = 50$ and sample these points randomly from the dataset. We use the ground truth to simulation 100 realizations of the demand vector corresponding to these 50 sample points, which we use to obtain the bias and variance of the different estimators. We repeat the experiment 30 times (with a different training set each time) and report the average results.

In Table 3 and 4, we present results obtained from 2 different subsets of the Nomis dataset. In Appendix E, we provide results obtained from 3 other subsets of the Nomis dataset.

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy is 5% increase.				
MSE	0.11	0.13	0.17	0.80
Bias^2	0.03	0.04	0.06	0.34
Variance	0.08	0.09	0.11	0.46
Target Policy is 5% decrease.				
MSE	0.03	0.05	0.10	0.20
Bias^2	0.01	0.02	0.05	0.10
Variance	0.02	0.03	0.05	0.10
Target Policy is 10% increase.				
MSE	0.37	0.41	0.44	1.23
Bias^2	0.12	0.13	0.15	0.49
Variance	0.25	0.28	0.29	0.74
Target Policy is 10% decrease.				
MSE	0.009	0.007	0.018	0.034
Bias^2	0.003	0.003	0.008	0.012
Variance	0.006	0.004	0.010	0.022

Table 3: For each target policy and for each method, we present the MSE, bias squared, and variance. Results obtained from a subset of the Nomis dataset with Year = 2003, Tier = 1, Car Type = Used, Term = 60, and Partner Bin = 1. There are 1,065 datapoints in the subset.

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy is 5% increase.				
MSE	0.39	0.46	0.45	0.80
Bias^2	0.09	0.11	0.11	0.15
Variance	0.30	0.35	0.34	0.65
Target Policy is 5% decrease.				
MSE	0.69	0.75	0.77	0.58
Bias^2	0.25	0.29	0.30	0.14
Variance	0.44	0.46	0.47	0.44
Target Policy is 10% increase.				
MSE	0.78	0.92	0.86	0.80
Bias^2	0.26	0.30	0.30	0.13
Variance	0.52	0.62	0.56	0.67
Target Policy is 10% decrease.				
MSE	0.38	0.56	0.36	0.38
Bias^2	0.10	0.10	0.10	0.10
Variance	0.28	0.46	0.26	0.28

Table 4: For each target policy and for each method, we present the MSE, bias squared, and variance. Results obtained from a subset of the Nomis dataset with Year from 2002 to 2004, Tier = 3, Car Type = Used, Term = 48, and Partner Bin = 3. There are 578 datapoints in the subset.

6.2 Bernstein Bounds

We next consider $\boldsymbol{w}^B$ and the corresponding estimator $\hat{\mathcal{R}}(\boldsymbol{w}^B)$. We denote the corresponding method BOPE-Bern. Since the primary motivation of BOPE-Bern was to provide high-quality safe guarantees on the revenue, we focus our experiments on such safe guarantees, and specifically comparisons to BOPE.

Recall Lemma 2 provides a safe guarantee for *any* set of honest weights. Hence, to form a safe guarantee for BOPE, we take the weights computed by BOPE, and then solve the inner maximization problem in (5) with the Bernstein bound objective for those weights. Since the revenue must be non-negative, we take the positive part of the optimal value. If weights computed by BOPE were honest, this procedure would yield a theoretically valid safe guarantee. Insofar as we specify hyperparameters in BOPE in a "dishonest" fashion, the resulting safe guarantee is only heuristically valid. (The same criticism holds for our own method, BOPE-B, making it a fair comparison.)

Our experiments suggest BOPE-B yields much better safe guarantees than BOPE, while providing comparably good estimates of the actual revenue.

In Tables 5 and 6, we present results on the two synthetic datasets described in Section 6.1.1. In Tables 7 and 8, we present results on subsets of the Nomis dataset. The

Target Policy	$\mathcal{R}$	BOPE-Bern		BOPE-B	
		$Bern(\boldsymbol{w}^B, r^{WC}(\boldsymbol{w}^B))$	$\hat{\mathcal{R}}(\boldsymbol{w}^B)$	$Bern(\boldsymbol{w}^{MSE}, r^{WC}(\boldsymbol{w}^{MSE}))$	$\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$
b = 2	3.51	1.07 ± 0.050	3.03 ± 0.054	0.10 ± 0.016	3.08 ± 0.053
b = 3	5.10	1.97 ± 0.098	4.51 ± 0.102	0.10 ± 0.024	4.53 ± 0.099
b = 4	4.95	1.81 ± 0.064	4.91 ± 0.048	0.04 ± 0.014	4.75 ± 0.054

Table 5: We present average and standard error of revenue bounds, computed from 100 demand realizations. The bounds in *BOPE-B* are the worst-case Bernstein bounds with *BOPE-B* weights. Results obtained from synthetic dataset (a) described in Section 6.1.1.

Target Policy	$\mathcal{R}$	BOPE-Bern		BOPE-B	
		$Bern(\boldsymbol{w}^B, r^{WC}(\boldsymbol{w}^B))$	$\hat{\mathcal{R}}(\boldsymbol{w}^B)$	$Bern(\boldsymbol{w}^{MSE}, r^{WC}(\boldsymbol{w}^{MSE}))$	$\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$
b = 2	3.86	1.29 ± 0.051	3.29 ± 0.054	0.08 ± 0.017	3.36 ± 0.054
b = 3	4.59	1.76 ± 0.082	4.30 ± 0.073	0.05 ± 0.021	4.29 ± 0.075
b = 4	5.23	1.57 ± 0.081	4.86 ± 0.072	0.01 ± 0.008	4.85 ± 0.075

Table 6: We present average and standard error of revenue bounds, computed from 100 demand realizations. The bounds in *BOPE-B* are the worst-case Bernstein bounds with *BOPE-B* weights. Results obtained from synthetic dataset (b) described in Section 6.1.1.

Target Policy	$\mathcal{R}$	BOPE-Bern		BOPE-B	
		$Bern(\boldsymbol{w}^B, r^{WC}(\boldsymbol{w}^B))$	$\hat{\mathcal{R}}(\boldsymbol{w}^B)$	$Bern(\boldsymbol{w}^{MSE}, r^{WC}(\boldsymbol{w}^{MSE}))$	$\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$
+5%	4.22	1.99 ± 0.020	4.13 ± 0.017	1.52 ± 0.031	4.24 ± 0.014
-5%	4.16	1.72 ± 0.007	3.87 ± 0.005	1.17 ± 0.009	4.05 ± 0.004
+10%	3.85	1.55 ± 0.034	3.70 ± 0.031	1.06 ± 0.043	3.84 ± 0.028
-10%	4.05	1.97 ± 0.003	3.82 ± 0.004	1.46 ± 0.005	4.08 ± 0.003

Table 7: We present average and standard error of revenue bounds, computed from 100 demand realizations. The bounds in *BOPE-B* are the worst-case Bernstein bounds with *BOPE-B* weights. Results obtained from a subset of the Nomis dataset with year = 2003, Tier = 1, Car Type = Used, Term = 60, and Partner Bin = 1. There are 1,065 datapoints in the subset.

Target Policy	$\mathcal{R}$	BOPE-Bern		BOPE-B	
		$Bern(\boldsymbol{w}^B, r^{WC}(\boldsymbol{w}^B))$	$\hat{\mathcal{R}}(\boldsymbol{w}^B)$	$Bern(\boldsymbol{w}^{MSE}, r^{WC}(\boldsymbol{w}^{MSE}))$	$\hat{\mathcal{R}}(\boldsymbol{w}^{MSE})$
+5%	2.77	0.45 ± 0.021	2.42 ± 0.031	0.00 ± 0.000	2.52 ± 0.032
-5%	3.35	0.53 ± 0.028	2.64 ± 0.033	0.00 ± 0.000	2.75 ± 0.030
+10%	2.87	0.21 ± 0.018	2.45 ± 0.042	0.00 ± 0.000	2.48 ± 0.045
-10%	3.67	0.55 ± 0.027	2.81 ± 0.034	0.00 ± 0.000	2.93 ± 0.034

Table 8: We present average and standard error of revenue bounds, computed from 100 demand realizations. The bounds in *BOPE-B* are the worst-case Bernstein bounds with *BOPE-B* weights. Results obtained from a subset of the Nomis dataset with year from 2002 to 2004, Tier = 3, Car Type = Used, Term = 48, and Partner Bin = 3. There are 578 datapoints in the subset.

experiment details are the same as described in Section 6. For each method, we present the one-sided 90% confidence lower bound on revenue (i.e. we choose $\epsilon = 0.1$). For all experiments in this subsection, we use sample size $n = 50$.

7 HYPER-PARAMETER HEURISTICS

Our heuristics for fitting hyper-parameters are inspired by the heuristics of Kallus (2018) for BOPE.

Define the revenue random variable $R_i := p_i D_i$. Loosely, Kallus (2018) assumes that the R_i are homoscedastic with

variance σ^2 and mean $r(p_i, \boldsymbol{x}_i)$ for each $i \in [n]$. They then compute the worst-case MSE of the weighted doubly robust estimator over a suitable RKHS ball. It turns out the resulting expression is identical to the expected MSE of this same estimator assuming the unknown expected revenue function was drawn from the following Gaussian Process Prior:

$$r(\cdot,\cdot) \sim \mathcal{GP}(\hat{\boldsymbol{r}}(\cdot,\cdot), \hat{\Gamma}^2 K(\cdot,\cdot)). \tag{7}$$

Said differently, the worst-case MSE is equal to an expected MSE under a suitable prior.

Thus, Kallus (2018) proposes to fit any hyperparameters needed for BOPE by using standard marginal likelihood techniques (Williams and Rasmussen, 2006, Chapt. 5) to instead fit the above Gaussian Process prior and then "read off" the parameters needed for BOPE.

We follow this same strategy in our experiments. For the kernel, we adopt a Gaussian kernel but standardize each component by its variance. Specifically, we take

$$K(\boldsymbol{z}, \overline{\boldsymbol{z}}) := \exp\left(-(\boldsymbol{z}-\overline{\boldsymbol{z}})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\overline{\boldsymbol{z}})\right),$$

where $\boldsymbol{z} = (p, \boldsymbol{x}) \in \mathbb{R}_+ \times \mathcal{X}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix.

We then optimize the choice of $\boldsymbol{\Sigma}$, σ^2 and $\hat{\Gamma}^2$ to maximize the marginal likelihood of the data under the prior Eq. (7) assuming the likelihood $R_i \mid \boldsymbol{r}(\cdot,\cdot) \sim \mathcal{N}(r(p_i, \boldsymbol{x}_i), \sigma^2)$. Because the Gaussian process prior and Gaussian likelihood are conjugate, the resulting marginal likelihood has a nice closed-form expression and the entire optimization can be represented tractably. (Again, see Williams and Rasmussen (2006) for details.)

Unfortunately, for the case of BOPE-B, our expression for the worst-case MSE does not seem to match the expected MSE under a simple prior. Hence, we heuristically seek parameters that maximize the marginal likelihood of the data under the model Eq. (7), but now assuming that $D_i | P_i = p_i \sim$ Bernoulli$(d(p_i, \boldsymbol{x}_i))$ and $R_i = p_i D_i$. In other words, we adjust the previous heuristic to account for the binary nature of demand. For this binary likelihood, we do not have conjugacy, and so there is no simple closed-form expression for the marginal likelihood. Instead, we follow Flaxman et al. (2015) and employ a Laplace approximation to the marginal likelihood. The resulting approximate likelihood does admit a simple form and the resulting maximal marginal likelihood optimization is tractable.

For our BOPE-B method, we optimize this approximate marginal likelihood to fit Eq. (7), and read off the necessary hyper-parameters.

8 CONCLUSION

In this paper, we have proposed a new approach for policy evaluation tailored to pricing applications. Our approaching uses special structures of pricing problems, including: *(i)* demand observations are binary; *(ii)* revenue per customer is nonnegative and no greater than the price offered; *(iii)* revenue equals demand times price; *(iv)* the value of the target policy can be very different from that of the logging policy, and thus weights do not need to sum to n. We compute weights to optimize either *(i)* the worst-case mean squared error of our estimate or *(ii)* a worst-case lower bound on the unknown revenue of the target policy. In both cases, the worst-case is taken over a set of plausible revenue functions described by an RKHS ball. We establish theoretical guarantees showing our weighted revenue estimator converges under overlap assumptions and empirically demonstrate the advantage of our approach using a real-world pricing dataset where there is little overlap. Future work might consider specialized algorithms for computing the weights in our method given its special structure, e.g., adapting the Mirror Prox algorithm of (Nemirovski, 2004), the primal-dual method in (Nesterov, 2007), or various algorithms for saddle point problems (Juditsky et al., 2011; Mertikopoulos et al., 2019).

Acknowledgements

The authors are listed in alphabetical order. We acknowledge the support of NSF grants CMMI-1763000, CMMI-1944428, and IIS-2147361

References

Aouad, A., Elmachtoub, A. N., Ferreira, K. J., and McNellis, R. (2019). Market segmentation trees. *arXiv preprint arXiv:1906.01174*.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Baardman, L., Cohen, M. C., Panchamgam, K., Perakis, G., and Segev, D. (2019). Scheduling promotion vehicles to boost profits. *Management Science*, 65(1):50–70.

Ban, G.-Y. and Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9):5549–5568.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Besbes, O., Phillips, R., and Zeevi, A. (2010). Testing the validity of a demand model: An operations perspective. *Manufacturing & Service Operations Management*, 12(1):162–183.

Biggs, M. (2022). Convex loss functions for contextual pricing with observational posted-price data. *arXiv preprint arXiv:2202.10944*.

Biggs, M., Gao, R., and Sun, W. (2021). Loss functions for discrete contextual pricing with observational data. *arXiv preprint arXiv:2111.09933*.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11).

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bu, J., Simchi-Levi, D., and Wang, L. (2022). Offline pricing and demand learning with censored data. *Management Science*.

Cai, H., Shi, C., Song, R., and Lu, W. (2021). Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300.

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. (2022). A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937.

Chernozhukov, V., Demirer, M., Lewis, G., and Syrgkanis, V. (2019). Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems*, 32.

Cohen, M. C., Elmachtoub, A. N., and Lei, X. (2022). Price discrimination with fairness constraints. *Management Science*, 68(12):8536–8552.

Cohen, M. C., Leung, N.-H. Z., Panchamgam, K., Perakis, G., and Smith, A. (2017). The impact of linear optimization on promotion planning. *Operations Research*, 65(2):446–468.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, page 1097–1104. PMLR.

Elliott, M. R. (2008). Model averaging methods for weight trimming. *Journal of official statistics*, 24(4):517.

Elmachtoub, A. N., Gupta, V., and Hamilton, M. L. (2021). The value of personalized pricing. *Management Science*, 67(10):6055–6070.

Ferreira, K. J., Lee, B. H. A., and Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & service operations management*, 18(1):69–88.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *International Conference on Machine Learning*, pages 607–616. PMLR.

Hanna, J. P., Stone, P., and Niekum, S. (2017). Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Kallus, N. (2018). Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31.

Kallus, N. (2020). Comment: Entropy learning for dynamic treatment regimes. *arXiv preprint arXiv:2004.02778*.

Kallus, N. (2021). More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658.

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR.

Kolbeinsson, A., Shukla, N., Gupta, A., Marla, L., and Yellepeddi, K. (2022). Galactic air improves ancillary revenues with dynamic personalized pricing. *INFORMS Journal on Applied Analytics*.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR 2019-7th International Conference on Learning Representations*, pages 1–23.

Nambiar, M., Simchi-Levi, D., and Wang, H. (2019). Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nesterov, Y. (2007). Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344.

Phillips, R., Şimşek, A. S., and Van Ryzin, G. (2015). The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science*, 61(8):1741–1759.

Qi, Z., Tang, J., Fang, E., and Shi, C. (2022). Offline personalized pricing with censored demand. *Available at SSRN*.

Sachdeva, N., Su, Y., and Joachims, T. (2020). Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 965–975.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Sondhi, A., Arbour, D., and Dimmery, D. (2020). Balanced off-policy evaluation in general action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 2413–2423. PMLR.

Swaminathan, A. and Joachims, T. (2015a). Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR.

Swaminathan, A. and Joachims, T. (2015b). The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, Y., Chen, X., Chang, X., and Ge, D. (2021). Uncertainty quantification for demand prediction in contextual dynamic pricing. *Production and Operations Management*, 30(6):1703–1717.

Wang, Y. and Zheng, Z. (2021). Measuring policy performance in online pricing with offline data. *Available at SSRN 3729003*.

Wang, Y.-X., Agarwal, A., and Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

A Proof of Properties of Weighted Revenue Estimators

Proof of Lemma 1. By definition of $\mathcal{R}$ and r_j, we have that $\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n} r_{n+i}$. Since $\mathbb{E}\left[p_i D_i\right] = r_i$, it follows from the definitions of $\hat{\mathcal{R}}(\boldsymbol{w})$ that $Bias(\boldsymbol{w}, \boldsymbol{r}) = \mathbb{E}\left[\hat{\mathcal{R}}(\boldsymbol{w}) - \mathcal{R}\right] = \boldsymbol{b}(\boldsymbol{w})^\top (\boldsymbol{r} - \hat{\boldsymbol{r}})$.

For the variance, we see that

$$\begin{aligned} Var(\boldsymbol{w}, \boldsymbol{r}) &= \text{Var}(\hat{\mathcal{R}}(\boldsymbol{w})) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} w_i p_i D_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 p_i^2 \text{Var}\left(D_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 p_i^2 d(\boldsymbol{x}_i, p_i)(1 - d(\boldsymbol{x}_i, p_i)) = \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 r_i (p_i - r_i) = \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 p_i r_i - \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 (r_i)^2 \\ &= \boldsymbol{v}(\boldsymbol{w})^\top \boldsymbol{r} - \frac{1}{n^2}\boldsymbol{r}^\top \begin{pmatrix} \text{diag}(w_1^2, \ldots, w_n^2) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{r}. \end{aligned}$$

The expression for MSE follows from the usual bias-variance decomposition. □

Proof of Lemma 2. Write

$$\hat{\mathcal{R}}(\boldsymbol{w}) - \mathcal{R} = \boldsymbol{b}(\boldsymbol{w})^\top (\boldsymbol{r} - \hat{\boldsymbol{r}}) + \sum_{i=1}^{n} w_i \left(p_i D_i - r_i\right)$$

The first term is the bias of our estimator, evaluated in Lemma 1. The second term is a sum of mean-zero independent random variables. From (Boucheron et al., 2013, Thm. 2.10) and surrounding discussion (i.e. Bernstein's inequality), we have that with probability at least $1 - \epsilon$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i (p_i D_i - r_i) \leq \sqrt{2 Var(\boldsymbol{w}; \boldsymbol{r}) \log(1/\epsilon)} - \frac{\max_{1 \leq i \leq n} |w_i| p_i \log(1/\epsilon)}{3n}.$$

Combining completes the proof. □

B Reformulation of the Worst-Case Bernstein Inner Problem

For the Bernstein objective, the inner maximization problem can be reformulated as follows:

$$\begin{aligned} \boldsymbol{r}^{WC}(\boldsymbol{w}) \in \underset{\boldsymbol{r}, t}{\text{argmax}} \quad & \boldsymbol{w}^\top (\boldsymbol{r} - \hat{\boldsymbol{r}}) + \sqrt{2\log(1/\epsilon)} \cdot t \\ \text{s.t.} \quad & 0 \leq \boldsymbol{r} \leq \boldsymbol{p} \\ & t^2 \leq \boldsymbol{v}(\boldsymbol{w})^\top \boldsymbol{r} - \boldsymbol{r}^\top \boldsymbol{Q}(\boldsymbol{w}) \boldsymbol{r}, \\ & (\boldsymbol{r} - \hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1} (\boldsymbol{r} - \hat{\boldsymbol{r}}) \ \leq \ \Gamma^2, \end{aligned}$$

where $\boldsymbol{Q}(\boldsymbol{w}) := \text{diag}\left(\boldsymbol{w}_1^2, \ldots, \boldsymbol{w}_n^2, 0, \ldots, 0\right) \in \mathbb{R}^{2n \times 2n}$.

C Proof of Theorem 1

Recall the following classical fact about inverse propensity score weights:

Lemma 3. *For any function $f : \mathbb{R} \mapsto \mathbb{R}$ and any $i = 1, \ldots, n$ such that the expectations exist, we have the following identity:*

$$\mathbb{E}\left[W_i^{IP} f(P_i)\right] = \mathbb{E}\left[f(P_{n+i})\right].$$

Proof. Simply write the integrals:

$$\mathbb{E}\left[W_i^{IP} f(P_i)\right] \ = \ \int_{p \in \mathbb{R}} f(p) \frac{g_1(p, \boldsymbol{x}_i)}{g_0(p, \boldsymbol{x}_i)} \, g_0(p, \boldsymbol{x}_i) dp \ = \ \int_{p \in \mathbb{R}} f(p) g_1(p, \boldsymbol{x}_i) dp \ = \ \mathbb{E}\left[f(P_{n+i})\right].$$

□

In particular, the lemma implies that $\mathbb{E}\left[\mathcal{R}(\boldsymbol{w}^{IP}/n)\right] = \mathbb{E}\left[\mathcal{R}\right]$, i.e., using the (scaled) IP weights yields an unbiased estimator.

Finally, for convenience, define the function

$$\begin{aligned}\text{WCMSE}(\boldsymbol{w};\hat{\Gamma}) \;:=\; \max_{\boldsymbol{r}} \quad & \boldsymbol{\Delta}^\top \boldsymbol{b}(\boldsymbol{w})\boldsymbol{b}(\boldsymbol{w})^\top \boldsymbol{\Delta} + \text{Var}(\boldsymbol{w};\hat{\boldsymbol{r}}+\boldsymbol{\Delta}).\\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{r} \le \boldsymbol{\pi}, \quad (\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}}) \le \hat{\Gamma}^2.\end{aligned}$$

A challenge in our analysis that $\hat{\Gamma}$ might be misspecified, i.e., it might be much smaller than Γ. Hence, WCMSE$(\boldsymbol{w};\hat{\Gamma})$ may not upper bound $MSE(\boldsymbol{w},\boldsymbol{r})$.

The next lemma shows we can cover such misspecification by inflating the worst-case MSE by a constant.

Lemma 4. *For any* $\boldsymbol{w}$, *MSE*$(\boldsymbol{w},\boldsymbol{r}) \;\le\; \max\left(1, \frac{\Gamma^2}{\hat{\Gamma}^2}\right) \cdot$ *WCMSE*$(\boldsymbol{w};\hat{\Gamma})$.

Proof. If $\Gamma \le \hat{\Gamma}$, then the unknown revenue function $\boldsymbol{r} = \hat{\boldsymbol{r}} + \boldsymbol{\Delta}$ is feasible in the inner maximization defining $\boldsymbol{w}^{MSE}(\hat{\Gamma})$, so that MSE$(\boldsymbol{w}^{MSE}(\hat{\Gamma});\hat{\boldsymbol{r}}+\boldsymbol{\Delta}) \le$ WCMSE$(\boldsymbol{w}^{MSE}(\hat{\Gamma});\hat{\Gamma})$. We thus focus on the case when $\Gamma > \hat{\Gamma}$. Then,

$$\begin{aligned}\text{MSE}(\boldsymbol{w},\hat{\boldsymbol{r}}+\boldsymbol{\Delta}) \;&=\; \boldsymbol{\Delta}^\top \boldsymbol{b}(\boldsymbol{w})\boldsymbol{b}(\boldsymbol{w})^\top \boldsymbol{\Delta} + \text{Var}(\boldsymbol{w};\hat{\boldsymbol{r}}+\boldsymbol{\Delta})\\ &=\; \frac{\Gamma^2}{\hat{\Gamma}^2}\left(\frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta}^\top \boldsymbol{b}(\boldsymbol{w})\boldsymbol{b}(\boldsymbol{w})^\top \boldsymbol{\Delta}\frac{\hat{\Gamma}}{\Gamma}\right) + \text{Var}(\boldsymbol{w};\hat{\boldsymbol{r}}+\boldsymbol{\Delta}).\end{aligned}$$

Now consider the variance term. From the proof of Lemma 1,

$$\text{Var}(\boldsymbol{w},\hat{\boldsymbol{r}}+\boldsymbol{\Delta}) \;=\; \frac{1}{n^2}\sum_{i=1}^n w_i^2(\hat{\boldsymbol{r}}_i+\Delta_i)(p_i-\hat{\boldsymbol{r}}_i-\Delta_i).$$

Since $\hat{\boldsymbol{r}}_i \ge 0$, and $\hat{\boldsymbol{r}}_i \le p_i$,

$$\hat{\boldsymbol{r}}_i + \Delta_i \le \frac{\Gamma}{\hat{\Gamma}}\left(\hat{\boldsymbol{r}}_i + \frac{\hat{\Gamma}}{\Gamma}\Delta_i\right), \quad \text{and} \quad p_i - \hat{\boldsymbol{r}}_i - \Delta_i \;\le\; \frac{\Gamma}{\hat{\Gamma}}\left(p_i - \hat{\boldsymbol{r}}_i - \frac{\hat{\Gamma}}{\Gamma}\Delta_i\right).$$

Substituting above shows that $\text{Var}(\boldsymbol{w},\hat{\boldsymbol{r}}+\boldsymbol{\Delta}) \;\le\; \frac{\Gamma^2}{\hat{\Gamma}^2}\text{Var}(\boldsymbol{w},\hat{\boldsymbol{r}}+\frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta})$. In summary, we have shown that

$$\text{MSE}(\boldsymbol{w};\hat{\boldsymbol{r}}+\boldsymbol{\Delta}) \;\le \frac{\Gamma^2}{\hat{\Gamma}^2}\text{MSE}(\boldsymbol{w};\hat{\boldsymbol{r}}+\frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta}).$$

To complete the proof, note that $\hat{\boldsymbol{r}} + \frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta}$ is feasible in the optimization defining WCMSE$(\boldsymbol{w};\hat{\Gamma})$. □

Proof of Theorem 1. From Lemma 4, it suffices to show that WCMSE$(\boldsymbol{w}^{MSE}(\hat{\Gamma});\hat{\Gamma}) = O_p(1/n)$. We show this latter claim by relating $\boldsymbol{w}^{MSE}(\hat{\Gamma})$ with the scaled inverse propensity weights $\boldsymbol{W}^{IP}/n$.

Specifically, since $\boldsymbol{W}^{IP}/n$ is feasible in the outer optimization problem defining $\boldsymbol{w}^{MSE}(\hat{\Gamma})$ we have that

$$\begin{aligned}&\text{WCMSE}(\boldsymbol{w}^{MSE}(\hat{\Gamma});\hat{\Gamma})\\ &\le\; \text{WCMSE}(\boldsymbol{W}^{IP}/n;\hat{\Gamma})\\ &\le\; \max_{\boldsymbol{r}:(\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}})\le\hat{\Gamma}^2} \text{MSE}(\boldsymbol{W}^{IP}/n;\boldsymbol{r})\\ &\le\; \max_{\boldsymbol{r}:(\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}})\le\hat{\Gamma}^2} (\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{b}(\boldsymbol{W}^{IP}/n)\boldsymbol{b}(\boldsymbol{W}^{IP}/n)^\top(\boldsymbol{r}-\hat{\boldsymbol{r}}) + \frac{1}{4n^2}\sum_{i=1}^n (W_i^{IP}P_i)^2,\end{aligned}$$

where the second to last inequality follows by expanding the feasible region and the last by upper bounding the variance since $d_i(1-d_i) \le \frac{1}{4}$. We evaluate the maximization in closed form and round the constants up to 1 yielding

$$\text{WCMSE}(\boldsymbol{w}^{MSE}(\hat{\Gamma});\hat{\Gamma}) \;\le\; \hat{\Gamma}^2 \boldsymbol{b}(\boldsymbol{W}^{IP})^\top \boldsymbol{G} \boldsymbol{b}(\boldsymbol{W}^{IP}) + \frac{1}{n^2}\sum_{i=1}^{n}(W_i^{IP}P_i)^2. \tag{8}$$

We tackle the first term by upper bounding its expectation and applying Markov's inequality. Using the definition of $\boldsymbol{G}$ and $\boldsymbol{b}(\boldsymbol{W}^{IP})$, write

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{b}(\boldsymbol{W}^{IP})^\top \boldsymbol{G} \boldsymbol{b}(\boldsymbol{W}^{IP})\right] \;=\; & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[W_i^{IP}W_j^{IP}K(Z_i,Z_j)\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[K(Z_{n+i},Z_{n+j})\right] \\ & - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[W_i^{IP}K(Z_i,Z_{n+j})\right],\end{aligned}$$

where for convenience $Z_i = (\boldsymbol{x}_i, P_i)$ for $i = 1, \ldots, 2n$.

Next fix some (i,j) with $i \neq j$. By Lemma 3,

$$\mathbb{E}\left[W_i^{IP}W_j^{IP}K(Z_i,Z_j)\right] \;=\; \mathbb{E}\left[W_j^{IP}K(Z_{n+i},Z_j)\right] \;=\; \mathbb{E}\left[K(Z_{n+i},Z_{n+j})\right].$$

Similarly,

$$\mathbb{E}\left[W_i^{IP}K(Z_i,Z_{n+i})\right] \;=\; \mathbb{E}\left[K(Z_{n+i},Z_{n+i})\right].$$

Hence, substituting above, we see that all terms with $i \neq j$ drop out and we have that

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{b}(\boldsymbol{W}^{IP})^\top \boldsymbol{G} \boldsymbol{b}(\boldsymbol{W}^{IP})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(W_i^{IP})^2K(Z_i,Z_i)\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[K(Z_{n+i},Z_{n+i})\right] - \frac{2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[W_i^{IP}K(Z_i,Z_{n+i})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[w_i^{IP}K(Z_{n+i},Z_{n+i})\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[K(Z_{n+i},Z_{n+i})\right] - \frac{2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[K(Z_{n+i},Z_{n+i})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(W_i^{IP}-1)K(Z_{n+i},Z_{n+i})\right],\end{aligned}$$

by applying Lemma 3 again.

By assumption *i)*, this last term is $O(1/n)$. Thus, by Markov's inequality, the first term of Eq. (8) is $O_p(1/n)$.

For the second term of Eq. (8), observe that

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(W_i^{IP}P_j\right)^2\right] \;=\; \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(W_i^{IP}P_i\right)^2\right]\right) \;=\; O(1/n),$$

by assumption *ii)*. Thus, by Markov's inequality, the second term of Eq. (8) is also $O_p(1/n)$.

Combining these two pieces completes the proof. □

D Proof of Theorem 2

For convenience, define the functions

$$\begin{aligned}q_{\max}(\boldsymbol{w}) &:= \frac{1}{n}\max_{1\le i\le n}|w_i|\,p_i \\ \text{WCBern}(\boldsymbol{w};\hat{\Gamma}) &:= \max_{\boldsymbol{r}} \quad \boldsymbol{b}(\boldsymbol{w})^\top(\boldsymbol{r}-\hat{\boldsymbol{r}}) + \sqrt{2\text{Var}(\boldsymbol{w};\boldsymbol{r})\log(1/\epsilon)} + \frac{q_{\max}(\boldsymbol{w})\log(1/\epsilon)}{3}. \\ &\qquad\text{s.t.} \quad \mathbf{0} \le \boldsymbol{r} \le \boldsymbol{\pi}, \quad (\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}}) \le \hat{\Gamma}^2.\end{aligned}$$

Our proof technique follows Theorem 1 closely.

Lemma 5. *For any* $\boldsymbol{w}$, $\max(\textit{Bern}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}), 0) \;\le\; \max\left(1, \frac{\Gamma}{\hat{\Gamma}}\right) \cdot \textit{WCBern}(\boldsymbol{w}; \hat{\Gamma})$.

Proof. Notice by considering the feasible solution $\boldsymbol{r} = \hat{\boldsymbol{r}}$ that $\text{WCBern}(w, \hat{\Gamma}) \;\ge\; \sqrt{2\log(1/\epsilon)\text{Var}(\boldsymbol{w}, \hat{\boldsymbol{r}})} + \frac{q_{\max}(\boldsymbol{w})\log(1/\epsilon)}{3} \ge 0$. Hence, when $\text{Bern}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \le 0$, the inequality is trivially satisfied.

Similarly, when $\Gamma \le \hat{\Gamma}$, $\boldsymbol{r}$ feasible in the inner maximization defining $\boldsymbol{w}^B(\hat{\Gamma})$, so that $\text{Bern}(\boldsymbol{w}^B(\hat{\Gamma}); \boldsymbol{r}) \le \text{WCBern}(\boldsymbol{w}^B(\hat{\Gamma}); \hat{\Gamma})$.

We thus focus on the case when $\Gamma > \hat{\Gamma}$ and $\text{Bern}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \ge 0$. Then,

$$\begin{aligned}\text{Bern}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \;&=\; \boldsymbol{b}(\boldsymbol{w})^\top \boldsymbol{\Delta} + \sqrt{2\log(1/\epsilon)\text{Var}(\boldsymbol{w}; \hat{\boldsymbol{r}} + \boldsymbol{\Delta})} + \frac{q_{\max}(\boldsymbol{w})\log(1/\epsilon)}{3} \\ &\le\; \frac{\Gamma}{\hat{\Gamma}}\left(\boldsymbol{b}(\boldsymbol{w})^\top \boldsymbol{\Delta}\frac{\hat{\Gamma}}{\Gamma} + \frac{q_{\max}(\boldsymbol{w})\log(1/\epsilon)}{3}\right) + \sqrt{2\log(1/\epsilon)\text{Var}(\boldsymbol{w}; \hat{\boldsymbol{r}} + \boldsymbol{\Delta})},\end{aligned}$$

since $\Gamma/\hat{\Gamma} > 1$ and $q_{\max} \ge 0$ by construction.

Now consider the variance term. From the proof of Lemma 1,

$$\text{Var}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \;=\; \frac{1}{n^2}\sum_{i=1}^n w_i^2(\hat{r}_i + \Delta_i)(p_i - \hat{r}_i - \Delta_i).$$

Since $\hat{r}_i \ge 0$, and $\hat{r}_i \le p_i$,

$$\hat{r}_i + \Delta_i \le \frac{\Gamma}{\hat{\Gamma}}\left(\hat{r}_i + \frac{\hat{\Gamma}}{\Gamma}\Delta_i\right), \quad \text{and} \quad p_i - \hat{r}_i - \Delta_i \;\le\; \frac{\Gamma}{\hat{\Gamma}}\left(p_i - \hat{r}_i - \frac{\hat{\Gamma}}{\Gamma}\Delta_i\right).$$

Substituting above shows that $\text{Var}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \;\le\; \frac{\Gamma^2}{\hat{\Gamma}^2}\text{Var}(\boldsymbol{w}, \hat{\boldsymbol{r}} + \frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta})$. In summary, we have shown that

$$\text{Bern}(\boldsymbol{w}; \hat{\boldsymbol{r}} + \boldsymbol{\Delta}) \;\le \frac{\Gamma}{\hat{\Gamma}}\text{Bern}(\boldsymbol{w}; \hat{\boldsymbol{r}} + \frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta}).$$

To complete the proof, note that $\hat{\boldsymbol{r}} + \frac{\hat{\Gamma}}{\Gamma}\boldsymbol{\Delta}$ is feasible in the optimization defining $\text{WCBern}(\boldsymbol{w}; \hat{\Gamma})$. □

We can now prove our main result.

Proof of Theorem 2. From Lemma 5, it suffices to show that $\text{WCBern}(\boldsymbol{w}^B(\hat{\Gamma}); \hat{\Gamma}) = O_p(1/\sqrt{n})$. We show this latter claim by relating $\boldsymbol{w}^B(\hat{\Gamma})$ with the scaled inverse propensity weights $\boldsymbol{W}^{IP}/n$.

Specifically, since $\boldsymbol{W}^{IP}/n$ is feasible in the outer optimization problem defining $\boldsymbol{w}^B(\hat{\Gamma})$ we have that

$$\begin{aligned}&\text{WCBern}(\boldsymbol{w}^B(\hat{\Gamma}); \hat{\Gamma}) \\ &\le\; \text{WCBern}(\boldsymbol{W}^{IP}/n; \hat{\Gamma}) \\ &\le \max_{\boldsymbol{r}:(\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}}) \le \hat{\Gamma}^2} \text{Bern}(\boldsymbol{W}^{IP}/n; \boldsymbol{r}) \\ &\le \max_{\boldsymbol{r}:(\boldsymbol{r}-\hat{\boldsymbol{r}})^\top \boldsymbol{G}^{-1}(\boldsymbol{r}-\hat{\boldsymbol{r}}) \le \hat{\Gamma}^2} \boldsymbol{b}(\boldsymbol{W}^{IP}/n)^\top(\boldsymbol{r} - \hat{\boldsymbol{r}}) + \frac{\sqrt{2\log(1/\epsilon)}}{4}\sqrt{\frac{1}{n^2}\sum_{i=1}^n (W_i^{IP}P_i)^2} + \frac{q_{\max}(\boldsymbol{W}^{IP}/n)\log(1/\epsilon)}{3},\end{aligned}$$

where the second to last inequality follows by expanding the feasible region and the last by upper bounding the variance since $d_j(1-d_j) \le \frac{1}{4}$. We evaluate the maximization in closed form and round the constants up to 1 yielding

$$\text{WCBern}(\boldsymbol{w}^B(\hat{\Gamma}); \hat{\Gamma}) \;\le\; \hat{\Gamma}\sqrt{\boldsymbol{b}(\boldsymbol{W}^{IP})^\top \boldsymbol{G}\boldsymbol{b}(\boldsymbol{W}^{IP})} + \sqrt{\log(1/\epsilon)}\sqrt{\frac{1}{n^2}\sum_{i=1}^n (W_i^{IP}P_i)^2} + q_{\max}(\boldsymbol{W}^{IP}/n)\log(1/\epsilon). \quad (9)$$

We tackle the first term by upper bounding its expectation and applying Markov's inequality. Specifically, $\mathbb{E}\left[\sqrt{\boldsymbol{b}(\boldsymbol{W}^{IP})\top \boldsymbol{G}\boldsymbol{b}(\boldsymbol{W}^{IP})}\right] \leq \sqrt{\mathbb{E}\left[\boldsymbol{b}(\boldsymbol{W}^{IP})\top \boldsymbol{G}\boldsymbol{b}(\boldsymbol{W}^{IP})\right]}$ by Jensen's inequality.

Following an identical argument to that in Theorem 1 which uses assumption *i)*, we have that $\mathbb{E}\left[\boldsymbol{b}(\boldsymbol{W}^{IP})^\top \boldsymbol{G}\boldsymbol{b}(\boldsymbol{W}^{IP})\right] = O(1/n)$. Thus, by Markov's inequality, the first term of Eq. (9) is $O_p(1/\sqrt{n})$.

For the second term of Eq. (9), observe again by Jensen's inequality that

$$\mathbb{E}\left[\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\left(W_i^{IP}P_i\right)^2}\right] \leq \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(W_i^{IP}P_i\right)^2\right]} = O(1/\sqrt{n}),$$

again by assumption *ii)*. Thus, by Markov's inequality, the second term of Eq. (9) is also $O_p(1/\sqrt{n})$.

Finally, for the last term, observe that

$$q_{\max}(\boldsymbol{W}^{IP}/n) = \frac{1}{n}\max_i \left|W_i^{IP}P_i\right| \leq \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(W_i^{IP}P_i\right)^2},$$

since the ℓ_2-norm bounds the ℓ_∞-norm. Taking expectations and applying the above inequality with Markov's inequality shows the last term is also $O_p(1/\sqrt{n})$.

Combining these three pieces completes the proof. □

E Additional Experiments

We present results for 3 more subsets of the Nomis dataset. Apart from the subset of data used, the experiment set up are same as that in Table 3.

Metrics	**BOPE-B**	**BOPE**	**LASSO**	**SPPE**
Target Policy is 5% increase.				
MSE	0.60	0.69	0.90	3.15
Bias^2	0.28	0.32	0.43	1.51
Variance	0.32	0.37	0.47	1.64
Target Policy is 5% decrease.				
MSE	0.09	0.09	0.18	0.71
Bias^2	0.04	0.04	0.09	0.34
Variance	0.05	0.05	0.09	0.37
Target Policy is 10% increase.				
MSE	1.77	1.94	2.22	5.78
Bias^2	0.84	0.92	1.06	2.79
Variance	0.93	1.02	1.16	2.99
Target Policy is 10% decrease.				
MSE	0.02	0.02	0.03	0.08
Bias^2	0.01	0.01	0.01	0.03
Variance	0.01	0.01	0.02	0.05

Table 9: Decomposition of the mean squared error. Tier = 2, Car Type = Used, Term = 60, Partner Bin = 1, and year 2003. There are 609 datapoints in the subset.

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy is 5% increase.				
MSE	0.54	0.62	0.61	0.64
Bias^2	0.12	0.14	0.15	0.06
Variance	0.42	0.48	0.46	0.59
Target Policy is 5% decrease.				
MSE	0.47	0.54	0.48	0.47
Bias^2	0.10	0.10	0.11	0.08
Variance	0.37	0.44	0.37	0.39
Target Policy is 10% increase.				
MSE	0.98	1.13	1.10	1.23
Bias^2	0.31	0.37	0.37	0.22
Variance	0.67	0.76	0.73	1.01
Target Policy is 10% decrease.				
MSE	0.55	0.58	0.51	0.90
Bias^2	0.11	0.09	0.10	0.25
Variance	0.44	0.49	0.41	0.65

Table 10: Decomposition of the mean squared error. Tier = 3, Car Type = Used, Term = 72, Partner Bin = 3, and year 2002-2004. There are 667 datapoints in the subset.

Metrics	BOPE-B	BOPE	LASSO	SPPE
Target Policy is 5% increase.				
MSE	0.39	0.52	0.49	1.11
Bias^2	0.08	0.12	0.12	0.24
Variance	0.31	0.40	0.37	0.87
Target Policy is 5% decrease.				
MSE	0.66	0.81	0.78	0.38
Bias^2	0.24	0.29	0.30	0.06
Variance	0.42	0.52	0.48	0.32
Target Policy is 10% increase.				
MSE	0.64	0.88	0.75	1.08
Bias^2	0.20	0.25	0.25	0.17
Variance	0.44	0.63	0.50	0.91
Target Policy is 10% decrease.				
MSE	0.50	0.56	0.50	0.46
Bias^2	0.14	0.15	0.15	0.10
Variance	0.36	0.41	0.35	0.36

Table 11: Decomposition of the mean squared error. Tier = 3, Car Type = Used, Term = 60, Partner Bin = 3, and year 2002-2004. There are 1851 datapoints in the subset.