Offline Contextual Bandits with Covariate Shift

Yingying Fan
University of Southern California
fanyingy@marshall.usc.edu

Yuxuan Han
New York University
yh6061@stern.nyu.edu

Jinchi Lv
University of Southern California
jinchilv@marshall.usc.edu

Xiaocong Xu
University of Southern California
xuxiaoco@marshall.usc.edu

Zhengyuan Zhou
New York University
zzhou@stern.nyu.edu

Abstract

Offline policy learning aims to optimize decision-making policies using historical data and plays a central role in many real-world applications, such as personalized advertising, medical treatment recommendation, and pricing decisions. A major challenge in this setting is the potential mismatch between the training environment—where the data were collected—and the test environment—where the learned policy is evaluated. This challenge has motivated extensive research on distributionally robust methods, which aim to maintain performance under worst-case distribution shifts. However, such approaches can be overly conservative when the environment changes in more structured ways. In this paper, we focus on the offline learning setting where the only difference between the training and test environments lies in the distribution of the context variables. Adopting the concept of transfer exponents from the transfer learning literature to model such covariate shift, we establish minimax-optimal sample complexity bounds for offline decision-making with general nonparametric reward functions. We further show that a pessimism-based algorithm attains these optimal rates.

1 Introduction

In many real-world applications, rich historical data are readily available, whereas online experimentation—such as randomized trials or active data collection—can be costly or impractical. Examples include personalized healthcare, digital advertising, and dynamic pricing. In such contexts, offline policy learning—the problem of learning effective decision-making policies from logged data without further environment interaction—has become a central paradigm in data-driven decision-making. A common and general formalism for this problem is the contextual bandit model, where the decision maker observes a context, selects an action, and then observes the reward only for the chosen action. This model captures a wide range of real-world scenarios and has motivated extensive algorithmic and theoretical developments [13, 28, 20, 9, 3].

While significant progress has been made in recent years on offline learning algorithms for both bandit and reinforcement learning settings [1, 39, 10, 21, 33, 29, 38, 37, 14, 25, 16, 34, 22], a growing line of work highlights a fundamental challenge in offline policy learning: distribution shift, where the distribution under which the training data were collected may differ from the distribution in which the learned policy is eventually deployed [2, 27, 30, 23, 6, 36, 19, 18, 26, 24, 15]. For

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML×OR Workshop.

example, in medical decision-making, data may be collected from large urban hospitals with relatively homogeneous patient populations, while the learned policy may need to generalize to rural or underserved populations with different demographic profiles or treatment access. In such scenarios, even if the conditional relationship between covariates and outcomes remains stable, the distribution of covariates shifts across environments.

A popular approach to addressing this challenge is distributionally robust offline policy learning, which seeks to ensure reliable performance under covariate shifts by optimizing for the worst-case scenario within a specified uncertainty set over the joint distribution of the observed data [27, 30, 23, 6]. However, optimizing over worst-case joint distributions can be overly conservative in practice. Returning to the clinical example, if the covariate distribution changes while the conditional relationship between treatments and outcomes remains stable, modeling shifts at the joint level may ignore this structure and lead to unnecessarily pessimistic solutions.

Motivated by these limitations, this paper focuses on a structured and practically relevant setting: offline policy learning under covariate shift. Here, we assume that the change between training and deployment environments lies only in the marginal distribution of covariates, while the conditional outcome model remains unchanged. We formalize the setting in Section 2. Our goal is to learn a policy $\widehat{\pi}$ from logged data whose expected reward under a target covariate distribution is close to optimal. To quantify the shift between the source and target covariate distributions, we adopt the notion of transfer exponent from the transfer learning literature [5, 3, 4, 12], as formally defined in Definition 1. Within this framework we obtain minimax-optimal guarantees: a unified upper bound under partial coverage and a matching minimax lower bound (up to logarithmic factors), achieved by a simple pessimism-based algorithm.

Finally, in addition to the technical contributions, we would like to highlight a conceptual perspective presented by this work—namely, the use of tools from transfer learning to address distributional robustness in offline policy learning. Our approach provides a solution to distributional robustness by leveraging tools from the transfer learning literature (see Section 2 for a detailed discussion of the similarities and differences between the two settings). We hope this perspective will inspire further research at the intersection of these two areas, as already explored in recent works [35, 7, 31].

2 Problem Setup and Assumptions

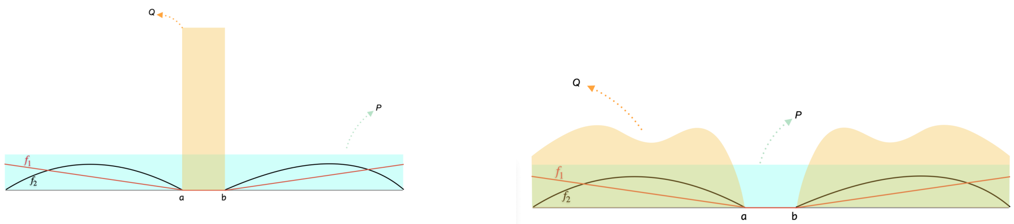
We consider a two-armed contextual bandit problem, where the action space is $\{0, 1\}$[1]. For each arm $a \in \{0, 1\}$, there exists an unknown reward function $f_a : \mathcal{X} \to \mathbb{R}$, where the covariate space is taken to be $\mathcal{X} = [0, 1]^d$ throughout this paper. Given a covariate $X \in \mathcal{X}$, the potential reward associated with arm a is modeled as $Y(a) = f_a(X) + \varepsilon$, where ε is a zero-mean sub-Gaussian random variable independent of both X and a. In the offline learning setting, the learner has access to a dataset of the form $\mathcal{D}_n \equiv \left\{\left(X_t, a_t, Y_t(a_t)\right)\right\}_{t\in[n]}$, where the $(X_t, a_t, Y_t(a_t))$ is generated i.i.d. according to the following joint distribution: for each t, the context X_t is sampled from some source distribution P, the action a_t is then sampled from a randomized logging policy $\pi_{\text{off}}(\cdot \mid X_t)$, and finally the reward $Y_t(a_t)$ is observed. The goal is to use this observed data to learn a policy $\widehat{\pi} : \mathcal{X} \to \{0, 1\}$ that performs well under a potentially different target distribution Q over $\mathcal{X}$. Specifically, we aim to minimize the *sub-optimality gap*

$$\mathrm{SubOpt}(\widehat{\pi}; Q) \equiv \mathbb{E}_{x\sim Q}\left[f_{a^*(x)}(x) - f_{\widehat{\pi}(x)}(x)\right], \text{ where } a^*(x) = \arg\max_{a\in\{0,1\}} f_a(x). \quad (1)$$

To describe the covariate shifts, we introduce the following concept of transfer exponent between P and Q, originally proposed in [12] for classification tasks. For any $x \in \mathbb{R}^d$ and $h > 0$, we let $\mathcal{B}(x, h) \equiv \{x' \in \mathbb{R}^d : \|x - x'\| \le h\}$.

Definition 1 (Transfer Exponent)**.** *We say that P and Q have a transfer exponent upper bound α for some $\alpha \ge 0$, or equivalently $\kappa(P, Q) \le \alpha$, if there exists a constant $C_\alpha > 0$ such that $P(\mathcal{B}(x, h)) \ge C_\alpha h^\alpha Q(\mathcal{B}(x, h)), \forall x \in \mathrm{supp}(Q), \forall h \in (0, 1]$.*

[1]For simplicity of presentation, we focus on the two-armed setting in this paper, all results can be extended straightforwardly to the general K-armed case.



(a) Worst-case distribution of (3) for any π.

(b) Reasonable adversarial distribution of (2).

Figure 1: Comparison of objective functions (2) and (3) in Example 1.

Connection to Distributionally Robust Policy Learning. With the covariate shift assumption, minimizing (1) can be interpret as pursuing a worst-case guarantee

$$\max_{Q:\kappa(P,Q)\leq\alpha} \mathrm{SubOpt}(\widehat{\pi};Q) = \max_{Q:\kappa(P,Q)\leq\alpha} \mathbb{E}_{x\sim Q}\left[f_{\pi^*(x)}(x) - f_{\widehat{\pi}(x)}(x)\right] \tag{2}$$

whereas in distributionally robust policy learning, the goal is to minimize a sub-optimality gap defined with respect to *the worst-case expected reward*

$$\mathrm{SubOpt}^{\mathrm{DRO}}(\widehat{\pi}) \equiv \max_{\pi} \min_{Q:\kappa(P,Q)\leq\alpha} \mathbb{E}_{x\sim Q}[f_{\pi(x)}(x)] - \min_{Q:\kappa(P,Q)\leq\alpha} \mathbb{E}_{x\sim Q}[f_{\widehat{\pi}(x)}(x)]. \tag{3}$$

However, we clarify that optimizing objective functions (3) and (2) generally lead to different policies. In particular, when using a transfer-exponent-based uncertainty set, objective (3) can be overly pessimistic and may fail to guide informative policy learning in certain scenarios, as shown in the following example:

Example 1 (Failure of DRO Objective Function)**.** *Consider a two-armed bandit instance with $\mathcal{X} = [0,1]$ and $P = \mathrm{Unif}([0,1])$. If the reward functions f_1, f_2 are non-negative and satisfy $f_1(x) = f_2(x) = 0$ on some interval $[a,b] \subset [0,1]$, then any policy π minimizes (3) for all $\alpha \geq 0$.*

Proof of Example 1. Simply note that for any $\alpha \geq 0$, the distribution $Q_0 \equiv \mathrm{Unif}([a,b])$ satisfies $\kappa(P,Q_0) \leq \alpha$. As a result, any policy π attains worst-case expected reward $\mathbb{E}_{x\sim Q_0}[f_{\pi(x)}(x)] = 0$ and thus minimizes the objective (3). □

While part of the DRO objective's limitations may come from the choice of the uncertainty measure κ used in our analysis, one might suggest replacing it with more standard metrics—such as Wasserstein distance or information-theoretic divergences like KL—as commonly done in prior DRPL works [27, 23, 11]. However, these alternatives may be less effective in capturing a broad class of covariate shifts under which optimal statistical efficiency is still achievable. For instance, when $P = \mathrm{Unif}([0,1])$ and $Q = \mathrm{Unif}([0,0.5])$, we have $\kappa(P,Q) = 0$, indicating that the shift does not affect the learning rate, as shown in Theorem 1. In contrast, we have $\mathrm{KL}(P\|Q) = +\infty, W_2(P,Q) = \frac{1}{2\sqrt{3}}$, which means an ambiguity set with at least a constant radius is required to capture such distribution. Consequently, the corresponding algorithms must account for worst-case distributions that are significantly different from the empirical data, potentially leading to slower learning rates even when no harmful shift is present. In other words, they guard against *unnecessarily adversarial distributions*.

Finally, we state our assumptions on the observed data for our theoretical results.

Assumption 1 (Unconfoundedness)**.** *For any $x \in \mathcal{X}$, the potential outcomes $Y(0)$ and $Y(1)$ are independent of a_t conditional on $X_t = x$, that is, $(Y(0), Y(1)) \perp\!\!\!\perp a_t \mid X_t = x$.*

Assumption 2 (Policy Coverage)**.** *The logging policy π_{off} satisfies $\inf_{x\in\mathcal{X}} \pi_{\mathrm{off}}(a^*(x) \mid x) \geq \zeta$.*

Assumption 3 (β-Hölder Reward Function)**.** *There exist constants $L > 0$ and $\beta \in (0,1]$ such that for $a \in \{0,1\}$, the reward function $f_a : \mathcal{X} \to \mathbb{R}$ satisfies $|f_a(x) - f_a(x')| \leq L\|x - x'\|^{\beta}, \forall x, x' \in \mathcal{X}$.*

Assumption 1 is a standard identification condition in offline policy learning [1, 39, 27]. It ensures that, conditional on covariates, the action assignment mechanism does not depend on unobserved factors that influence the potential outcomes. Assumption 2 is a standard partial coverage condition in offline policy learning, relaxing the uniform coverage requirement used in [3]. Roughly speaking, ζn

provides a lower bound on the number of samples corresponding to optimal actions under the target distribution. As shown in Theorem 1, $\zeta > 0$ is also necessary for consistent learning. Assumption 3 places f_a in a β-Hölder class; β governs the smoothness-driven rate.

3 Main results

In this section, we introduce a pessimistic algorithm for offline nonparametric contextual bandits, which combines local reward estimation with a lower confidence bound (LCB) decision rule. We provide theoretical guarantees on its sub-optimality under covariate shift.

Algorithm 1 Nadaraya–Watson estimator based Lower Confidence Bound (NW-LCB)

Require: Dataset $\mathcal{D}_n = \{(X_t, a_t, Y_t)\}_{t\in[n]}$, smoothness β

1: **for** each $x \in \mathcal{X}$ **do**

2:     **for** each $a \in \{0, 1\}$ **do**

3:         Define index set $\mathcal{I}_a \leftarrow \{t \in [n] : a_t = a\}$ and set $h_{a,x}$ as the minimizer of RHS of (4)

4:         Compute:

$$\widehat{f}_{a;h_{a,x}}(x) \leftarrow \frac{\sum_{i\in\mathcal{I}_a} Y_i \cdot \mathbf{1}\{X_i \in \mathcal{B}(x, h_{a,x})\}}{\sum_{i\in\mathcal{I}_a} \mathbf{1}\{X_i \in \mathcal{B}(x, h_{a,x})\}} \cdot \mathbf{1}\Big\{ \sum_{i\in\mathcal{I}_a} \mathbf{1}\{X_i \in \mathcal{B}(x, h_{a,x})\} > 0 \Big\}$$

5:     **end for**

6:     Define pessimistic policy: $\widehat{\pi}(x) \leftarrow \arg\max_{a\in\{0,1\}} f_a^{\mathrm{LCB}}(x)$ with f_a^{LCB} defined as in (5)

7: **end for**

8: **return** Learned policy $\widehat{\pi}$

NW Estimator for Non-Parametric Regression. In the first block (line 1 to line 4) of Algorithm 1, we use the Nadaraya–Watson estimator [17, 32] to perform the estimation of functions associated with each action a. The following lemma provides a point-wise error bound for the NW estimator.

Lemma 1. *Suppose Assumptions 1 and 3 hold. Then, for any given $x \in \mathcal{X}$, $h > 0$, and $a \in \{0, 1\}$, there exists some constant $C > 0$ independent of n such that with probability at least $1 - Cn^{-2}$,*

$$|\widehat{f}_{a;h}(x) - f_a(x)| \le \Big(h^\beta + C\sqrt{\tfrac{\log n}{N_a(x)}}\Big)\mathbf{1}\{\mathcal{G}_a(x)\} + \|f_a\|_\infty \mathbf{1}\{\mathcal{G}_a^c(x)\}, \tag{4}$$

where $\mathcal{G}_a(x) \equiv \{\sum_{t\in\mathcal{I}_a} \mathbf{1}\{X_t \in \mathcal{B}(x, h)\} \ge 1\}$ and $N_a(x) \equiv \sum_{t\in\mathcal{I}_a} \mathbf{1}\{X_t \in \mathcal{B}(x, h)\}$.

Note that given the smoothness factor β, RHS of (4) can be access to the learner, thus its minimizer, required in line 3 of Algorithm 1 is computable.

Pessimistic Decision Making via LCB. Based on (4), we define the pessimistic estimator for f_a as

$$f_a^{\mathrm{LCB}}(x) \equiv \widehat{f}_{a;h_{a,x}}(x) - \Big(h_{a,x}^\beta + C\sqrt{\tfrac{\log n}{N_a(x)}}\Big)\mathbf{1}\{\mathcal{G}_a(x)\} - \|f_a\|_\infty \mathbf{1}\{\mathcal{G}_a^c(x)\} \tag{5}$$

and set the pessimistic policy via taking the action point-wisely with maximal LCB values.

We now present a sub-optimality bound for Algorithm 1 and a matching minimax lower bound up to logarithmic factors.

Theorem 1. *Suppose Assumptions 1–3 hold and $\kappa(P, Q) \le \alpha$ for some $\alpha > 0$. Then, the output of Algorithm 1 satisfies*

$$\mathrm{SubOpt}(\widehat{\pi}; Q) = \widetilde{\mathcal{O}}\big((\zeta n)^{-\frac{\beta}{2\beta+\alpha+d}}\big).$$

Additionally, for any policy π, there exists an instance satisfying Assumptions 1–3 such that

$$\mathrm{SubOpt}(\pi; Q) = \Omega\big((\zeta n)^{-\frac{\beta}{2\beta+\alpha+d}}\big).$$

We note that exact knowledge of neither ζ nor α is required to achieve the rate in Theorem 1. Thus, the algorithm is fully adaptive to the observed offline data distribution, including both its coverage and its shift. However, to keep the same sub-optimality rate guarantee, adaptivity with respect to the smoothness factor β is generally impossible unless additional conditions, such as self-similarity, are imposed, as studied in [8].

References

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

[2] Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Changxiao Cai, T Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024.

[4] T Tony Cai and Hongming Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*, 2024.

[5] T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification. *The Annals of Statistics*, 49(1):100–128, 2021.

[6] Aldo Gael Carranza and Susan Athey. Robust offline policy learning with observational data from multiple sources. *arXiv*, 2024.

[7] Yu Gui, Rina Foygel Barber, and Cong Ma. Distributionally robust risk evaluation with an isotonic constraint. *arXiv preprint arXiv:2407.06867*, 2024.

[8] Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022.

[9] Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. *Operations Research*, 70(6):3261–3281, 2022.

[10] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[11] Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR, 2022.

[12] Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

[13] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[14] Zhihan Liu, Miao Lu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Welfare maximization in competitive equilibrium: Reinforcement learning for markov exchange economy. In *International Conference on Machine Learning*, pages 13870–13911. PMLR, 2022.

[15] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

[16] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[18] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

[19] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

[20] VIANNEY PERCHET and PHILIPPE RIGOLLET. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

[21] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[22] Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline RL with general function approximation via augmented lagrangian. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] Yi Shen, Pan Xu, and Michael M Zavlanos. Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *arXiv*, 2023.

[24] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

[25] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pages 19967–20025. PMLR, 2022.

[26] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36:79903–79917, 2023.

[27] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, 2023.

[28] Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[29] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

[30] Jingyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. Distributionally robust policy learning under concept drifts. *arXiv preprint arXiv:2412.14297*, 2024.

[31] Zitao Wang, Ziyuan Wang, Molei Liu, and Nian Si. Knowledge-guided wasserstein distributionally robust optimization. *arXiv preprint arXiv:2502.08146*, 2025.

[32] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[33] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[34] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. In *International Conference on Learning Representations (ICLR)*, 2023.

[35] Xin Xiong, Zijian Guo, and Tianxi Cai. Distributionally robust transfer learning. *arXiv preprint arXiv:2309.06534*, 2023.

[36] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

[37] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

[38] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pages 27117–27142. PMLR, 2022.

[39] Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.