Off-Policy Evaluation with General Logging Policies

YUSUKE NARITA, Yale University, USA
KYOHEI OKUMURA, Northwestern University, USA
AKIHIRO SHIMIZU, Mercari, Inc., Japan
KOHEI YATA, Yale University, USA

ACM Reference Format:
Yusuke Narita, Kyohei Okumura, Akihiro Shimizu, and Kohei Yata. 2021. Off-Policy Evaluation with General Logging Policies. In *BCIRWIS 2021: Bayesian Causal Inference for Real World Interactive Systems - (KDD 2021 Workshop), August 14–15, 2021.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

In bandit and reinforcement learning, off-policy (batch) policy evaluation attempts to estimate the performance of some counterfactual policy given data from a different logging policy.[1] Off-policy evaluation (OPE) is essential when deploying a new policy might be costly or risky, such as in education, medicine, consumer marketing, and robotics. OPE relates to other fields that study counterfactual/causal reasoning, such as statistics and economics.

Most existing OPE studies focus on stochastic logging policies, such as stochastic bandit (e.g. ϵ-greedy and Thompson Sampling) and random A/B testing. However, real-world decision-making often uses deterministic logging policies, including deterministic bandit (e.g. Upper Confidence Bound) as well as deterministic decision-making based on predictions obtained from supervised and unsupervised learning. An example in the latter group is a policy that greedily chooses the action with the largest predicted reward. OPE is difficult with a deterministic logging policy, since its log data contain no information about the reward from actions never chosen by the deterministic logging policy [17].

We provide a solution to this problem. Our proposed OPE estimator is applicable not only to stochastic logging policies but also to deterministic logging policies. We also allow for hybrid stochastic and deterministic logging policies, i.e., logging policies that choose actions stochastically for some individuals and deterministically for other individuals.

Related Work. Widely-used OPE methods include inverse probability weighting (IPW) [16, 20], self-normalized IPW [23], Doubly Robust [5], and more advanced variants [6, 21, 28]. These methods are based on importance sampling (IS) and require that the logging policy assigns a positive probability to every action potentially chosen by the counterfactual policy. This restriction makes them hard to use when the logging policy is deterministic.

There are two existing approaches to deterministic logging policies. The first approach considers a logging policy that varies over time or across individuals [20]. Viewing the sequence of varying logging policies as a single stochastic logging policy, it is possible to apply IS-based OPE methods. Unlike this approach, our approach is usable even when the logging policy is fixed. The second approach, called the Direct Method or Regression Estimator, predicts the mean

[1] Key prior studies include [1, 3, 5, 11–13, 15, 18–24, 27, 29] for bandit, and [6–9, 14, 16, 25, 26] for reinforcement learning.

reward conditional on the action and context by supervised learning and uses the prediction to estimate the performance of a counterfactual policy [2, 5]. Similar regression-based methods are proposed for reinforcement learning settings [4]. This approach is sensitive to the accuracy of the mean reward prediction. It may have a large bias if the regression model is not correctly specified. This issue is particularly severe when the logging policy is deterministic, since each action is observed only in a limited area of the context space. Our approach instead predicts the mean reward differences between actions by exploiting local subsamples near the decision boundaries without specifying the regression model. This idea relates to regression discontinuity designs in the social sciences [10].

2 FRAMEWORK AND METHOD

$\mathcal{A} = \{1, ..., m\}$ is a set of *actions* that the decision maker can choose from. Let $Y(a)$ be the potential reward when action a is chosen. Let $X \in \mathcal{X} \subset \mathbb{R}^p$ denote the *context* that the decision maker observes when picking an action, where p is the number of context variables. To simplify the exposition, we assume that X is continuously distributed. If some context variables in X are discrete, our analysis still holds conditional on the discrete variables.

We consider policies that choose actions based on individual context X. Let $ML : \mathbb{R}^p \longrightarrow \Delta(\mathcal{A})$ represent the *logging policy*, where $ML(a|x)$ is the probability of taking action a for individuals with context x. We assume that the analyst knows the logging policy and is able to simulate it. We allow for the case with deterministic policies, in which $ML(a|x) \in \{0, 1\}$ for every (a, x). Suppose we have log data $\{(Y_i, X_i, A_i)\}_{i=1}^n$ generated as follows. For each individual i, (1) $(Y_i(\cdot), X_i)$ is i.i.d. drawn from an unknown distribution; (2) Given X_i, the action A_i is randomly chosen based on the probability $ML(\cdot|X_i)$; (3) We observe the reward $Y_i = Y_i(A_i)$. We are interested in estimating the expected reward from any given *counterfactual policy* $\pi : \mathbb{R}^p \longrightarrow \Delta(\mathcal{A})$, which chooses a distribution of actions given individual context:

$$V(\pi) \equiv E\left[\sum_{a\in\mathcal{A}} Y(a)\pi(a|X)\right].$$

Proposed OPE Estimator.

(1) For a small bandwidth δ, compute the *Approximate Propensity Score* (APS):

$$p_\delta^{ML}(a|X_i) \equiv \frac{\int_{B(X_i,\delta)} ML(a|x^*)dx^*}{\int_{B(X_i,\delta)} dx^*},$$

where $B(X_i, \delta)$ is a p-dimensional ball with radius δ centered at X_i.[2]

(2) Compute $q_\delta^{ML}(a|X_i) \equiv \frac{p_\delta^{ML}(a|X_i)}{p_\delta^{ML}(a|X_i)+p_\delta^{ML}(1|X_i)}$. For each $a = 2, ..., m$, minimize the sum of squared errors on the subsample $\mathcal{I}(a; \delta) \equiv \{i : A_i \in \{1, a\}, q_\delta^{ML}(a|X_i) \in (0, 1)\}$:

$$(\hat{\alpha}_a, \hat{\beta}_a, \hat{\gamma}_a) = \underset{(\alpha_a,\beta_a,\gamma_a)}{\operatorname{argmin}} \sum_{i\in\mathcal{I}(a;\delta)} \left(Y_i - \alpha_a - \beta_a 1\{A_i = a\} - \gamma_a q_\delta^{ML}(a|X_i)\right)^2,$$

where $1\{\cdot\}$ is the indicator function.

(3) For any given counterfactual policy π, define our OPE estimator for $V(\pi)$ as:

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i + \sum_{a=2}^{m} \hat{\beta}_a\big(\pi(a|X_i) - ML(a|X_i)\big)\right). \tag{1}$$

[2] $p_\delta^{ML}(a|X_i)$ may be difficult to compute analytically if ML is complex. In such a case, we propose to approximate it using brute force simulation. We draw a value of x from the uniform distribution on $B(X_i, \delta)$ a number of times, compute $ML(a|x)$ for each draw, and take the average of $ML(a|x)$ over the draws.

Table 1. Simulation results: RMSE of estimators of $V(\pi)$

	Our Proposed Method with APS Controls				Method with Mean Differences		Direct
	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$	$\delta = 2.5$	A/B Test Sample	Full Sample	Method
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Experiment 1: Mix of A/B Test and Deterministic Logging Policy							
RMSE	.115	.113	.112	.113	.118	.128	—
Avg. N	1862	6362	12502	33122	500	50000	—
Experiment 2: Upper Confidence Bound Logging Policy							
RMSE	.058	.056	.055	.055	—	—	.342
Avg. N	3397	17344	31107	47601	—	—	50000

Notes: This table shows the root mean squared error (RMSE) of the estimators of the reward from the counterfactual policy $V(\pi)$ in the two simulation experiments. We use $1,000$ simulations of a size $50,000$ simulated sample to compute these statistics. Columns (1)–(4) report estimates from our method with several choices of δ used to compute APS. Each APS is computed by averaging 100 simulation draws of the ML value. In columns (5)–(6), we compute the mean reward differences between actions a and 1 in the A/B test segment or the full sample, and them plug them into $\hat{\beta}_a$ of Eq. (1). In column (7), we use the Direct Method with a linear model. The bottom two rows of each panel show the average number of observations used for estimation.

APS of action a at context x is the average probability that the logging policy chooses action a over a shrinking neighborhood around x (Step (1)). If APS at x is nonzero for a pair of actions, the logging policy chooses both actions locally around x. This enables us to estimate the difference in the mean reward between the two actions by exploiting the local subsample around x (Step (2)). When the logging policy is deterministic, the subsample consists of individuals near the decision boundary between the two actions. We then use the estimated reward differences to construct an estimator for the performance of any given counterfactual policy (Step(3)).

Under the assumptions stated in Appendix A, $\hat{V}(\pi)$ is shown to be a consistent estimator of the true expected reward from a counterfactual policy, that is, $\hat{V}(\pi)$ converges in probability to $V(\pi)$ as $n \longrightarrow \infty$. This result holds whether the logging policy is stochastic or deterministic.

3 SIMULATION EXPERIMENTS

We validate our method with two simulation experiments. The first considers a mix of stochastic and deterministic policies as the logging policy. Actions are randomly chosen for a small A/B test segment of the population and are chosen by a deterministic supervised learning algorithm for the rest of the population. The second experiment considers a situation in which we have a batch of data generated by Upper Confidence Bound, a deterministic bandit algorithm. In each experiment, we use our method and benchmark methods to evaluate the value of a counterfactual policy.

Table 1 reports the root mean squared error (RMSE) of our proposed estimator with several choices of δ and that of alternative estimators in each experiment. In Experiment 1, although both our estimator and the alternative estimator using the A/B test sample are consistent, our estimator outperforms the alternative in terms of RMSE. This is because the alternative uses only the A/B test sample while our method additionally uses the local subsample near the decision boundary of the deterministic policy. In Experiment 2, while the Direct Method suffers from large RMSE due to model misspecification, our proposed estimator has a small RMSE. This also suggests the effectiveness of the use of local subsample near decision boundaries.

4 REAL-WORLD APPLICATION

We empirically apply our method to evaluate and optimize coupon targeting policies. Our application is based on proprietary data provided by Mercari Inc., a major e-commerce company running online C2C marketplaces in Japan and the US. This company uses a deterministic policy based on uplift modeling to decide whether they offer a promotional coupon to each target customer. We use the data produced by their policy and our method to evaluate a counterfactual policy that offers the coupon to more customers. Our method predicts that the counterfactual policy would increase revenue more than the cost of coupon offers, suggesting that redesigning the current policy is profitable.

REFERENCES

[1] Andrew Bennett and Nathan Kallus. 2019. Policy Evaluation with Latent Confounders via Optimal Balance. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 32 (2019), 4827–4837.

[2] Alina Beygelzimer and John Langford. 2009. The Offset Tree for Learning with Partial Labels. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2009), 129–138.

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research* 14, 65 (2013), 3207–3260.

[4] Yaqi Duan, Zeyu Jia, and Mengdi Wang. 2020. Minimax-optimal off-policy evaluation with linear function approximation. *Proceedings of the 37th International Conference on Machine Learning* 119 (2020), 2701–2709.

[5] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly robust policy evaluation and optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[6] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. 2018. More Robust Doubly Robust Off-policy Evaluation. *Proceedings of the 35th International Conference on Machine Learning* 80 (2018), 1447–1456.

[7] Alex Irpan, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibarz, and Sergey Levine. 2019. Off-Policy Evaluation via Off-Policy Classification. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019), 5437–5448.

[8] Nan Jiang and Lihong Li. 2016. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning* 48 (2016), 652–661.

[9] Nathan Kallus and Masatoshi Uehara. 2020. Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes. *Journal of Machine Learning Research* 21, 167 (2020), 1–63. http://jmlr.org/papers/v21/19-827.html

[10] David S. Lee and Thomas Lemieux. 2010. Regression Discontinuity Designs in Economics. *Journal of Economic Literature* 48, 2 (June 2010), 281–355. https://doi.org/10.1257/jel.48.2.281

[11] Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. 2012. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *Journal of Machine Learning Research: Workshop and Conference Proceedings* 26 (2012), 19–36.

[12] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web (WWW)* (2010), 661–670.

[13] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining* (2011), 297–306.

[14] Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. 2018. Representation balancing mdps for off-policy policy evaluation. *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (2018), 2644–2653.

[15] Yusuke Narita, Shota Yasui, and Kohei Yata. 2019. Efficient counterfactual learning from bandit feedback. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence* (2019), 4634–4641.

[16] Doina Precup. 2000. Eligibility traces for off-policy policy evaluation. *Proceedings of the Seventeenth International Conference on Machine Learning* (2000), 759–766.

[17] Sachdeva, Noveen and Su, Yi and Joachims, Thorsten. 2020. Off-Policy Bandits with Deficient Support. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2020), 965–975.

[18] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. Open Bandit Dataset and Pipeline: towards Realistic and Reproducible Off-Policy Evaluation. (2021). arXiv:2008.07146 [cs.LG] Unpublished Manuscript, Tokyo Institute of Technology.

[19] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (2016), 1670–1679.

[20] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. 2010. Learning from logged implicit exploration data. *Proceedings of the 23rd International Conference on Neural Information Processing Systems* (2010), 2217–2225.

[21] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudik. 2020. Doubly robust off-policy evaluation with shrinkage. *Proceedings of the 37th International Conference on Machine Learning* 119 (2020), 9167–9176.

[22] Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged vandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research* 16 (2015), 1731–1755.

[23] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *Proceedings of the 28th International Conference on Neural Information Processing Systems* (2015), 3231–3239.

[24] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), 3635–3645.

[25] Philip Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. *Proceedings of The 33rd International Conference on Machine Learning* (2016), 2139–2148.

[26] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. 2020. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. *Proceedings of the 37th International Conference on Machine Learning* 119 (2020), 9659–9668.

[27] Masatoshi Uehara, Masahiro Kato, and Shota Yasui. 2020. Off-policy evaluation and learning for external validity under a covariate shift. *Proceedings of the 34th International Conference on Neural Information Processing Systems* 33 (2020), 49–61.

[28] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[29] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. 2017. Optimal and adaptive off-policy evaluation in contextual bandits. *Proceedings of the 34th International Conference on Machine Learning* (2017), 3589–3597.

A ASSUMPTIONS FOR CONSISTENCY

Assumption 1 (Constant Conditional Mean Differences). *There exists a function* $\beta : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ *such that* $E[Y(a)|X] - E[Y(a')|X] = \beta(a, a')$.

Our consistency result uses the following assumptions for the subsample assigned to one of the actions a and 1, for every $a \in \{2, ..., m\}$. Let $\mathcal{X}_{a,1} \equiv \{x \in \mathcal{X} : ML(a|x) > 0 \text{ or } ML(1|x) > 0\}$, $\widetilde{ML}(a|x) \equiv \Pr(A_i = a|A_i \in \{1, a\}, X_i = x) = \frac{ML(a|x)}{ML(a|x)+ML(1|x)}$, $\mathcal{X}^a_{a,1} \equiv \{x \in \mathcal{X} : \widetilde{ML}(a|x) = 1\}$, and $\mathcal{X}^1_{a,1} \equiv \{x \in \mathcal{X} : \widetilde{ML}(a|x) = 0\}$. In other words, $\mathcal{X}_{a,1}$ is the set of context values for which action 1 or a can be taken, $\widetilde{ML}(a|x)$ is the probability of choosing action a conditional on $A_i \in \{1, a\}$ and $X_i = x$, and $\mathcal{X}^a_{a,1}$ and $\mathcal{X}^1_{a,1}$ are the set of context values for which the conditional probability is 1 and 0, respectively.

Assumption 2. *The following holds for all* $a \in \{2, ..., m\}$.

(a) (Existence of Subsample) $\Pr(A_i \in \{1, a\}) > 0$.

(b) (Almost Everywhere Continuity of ML) $ML(a|\cdot)$ *and* $ML(1|\cdot)$ *are continuous almost everywhere on* $\mathcal{X}_{a,1}$ *with respect to the Lebesgue measure.*

(c) (Measure Zero Boundaries of $\mathcal{X}^a_{a,1}$ and $\mathcal{X}^1_{a,1}$). *For* $a' \in \{1, a\}$, $\mathcal{L}^p(\mathcal{X}^{a'}_{a,1}) = \mathcal{L}^p(\text{int}(\mathcal{X}^{a'}_{a,1}))$, *where* $\mathcal{L}^p$ *is the Lebesgue measure on* $\mathbb{R}^p$.

(d) (Finite Moments) $E[Y_i^2] < \infty$.

(e) (Nonzero Conditional Variance) *If* $\Pr(\widetilde{ML}(a|X_i) \in (0, 1)|A_i \in \{1, a\}) > 0$, *then* $\text{Var}(\widetilde{ML}(a|X_i)|\widetilde{ML}(a|X_i) \in (0, 1), A_i \in \{1, a\}) > 0$.

If $\Pr(\widetilde{ML}(a|X_i) \in (0, 1)|A_i \in \{1, a\}) = 0$, *then the following conditions (f)–(i) additionally hold.*

(f) (Deterministic ML) *For all* $x \in \mathbb{R}^p$, *either* $ML(a|x) = 1$ *or* $ML(a|x) = 0$.

(g) (C^2 Boundary of Ω^*_a) *There exists a partition* $\{\Omega^*_{a,1}, ..., \Omega^*_{a,K}\}$ *of* $\Omega^*_a = \{x \in \mathbb{R}^p : ML(a|x) = 1\}$ *(the set of the context values for which the probability of choosing action a is one) such that*

(1) $\text{dist}(\Omega^*_{a,k}, \Omega^*_{a,l}) > 0$ *for any* $k, l \in \{1, ..., K\}$ *such that* $k \neq l$. *Here* $\text{dist}(S, T) = \inf_{x \in S, y \in T} \|x - y\|$ *is the distance between two sets* S *and* $T \subset \mathbb{R}^p$;

(2) $\Omega^*_{a,k}$ *is nonempty, bounded, open, connected and twice continuously differentiable for each* $k \in \{1, ..., K\}$.[3]

(h) (Regularity of Deterministic ML)

[3] We say that a bounded open set $S \subset \mathbb{R}^p$ is *twice continuously differentiable* if for every $x \in S$, there exists a ball $B(x, \epsilon)$ and a one-to-one mapping ψ from $B(x, \epsilon)$ onto an open set $D \subset \mathbb{R}^p$ such that ψ and ψ^{-1} are twice continuously differentiable, $\psi(B(x, \epsilon) \cap S) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p > 0\}$ and $\psi(B(x, \epsilon) \cap \partial S) \subset \{(x_1, ..., x_p) \in \mathbb{R}^p : x_p = 0\}$, where ∂S is the boundary of S.

(1) $\mathcal{H}^{p-1}(\partial\Omega_a^*) < \infty$, $\int_{\partial\Omega_a^* \cap \partial\mathcal{X}_{a,1}} d\mathcal{H}^{p-1}(x) = 0$, *and* $\int_{\partial\Omega_a^* \cap \mathcal{X}_{a,1}} f_X(x) d\mathcal{H}^{p-1}(x) > 0$, *where* ∂S *denotes the boundary of a set* $S \subset \mathbb{R}^p$, f_X *is the probability density function of* X_i, *and* $\mathcal{H}^k$ *is the* k*-dimensional Hausdorff measure on* $\mathbb{R}^p$.[4]

(2) There exists $\delta > 0$ *such that* $ML(a|x) = 1$ *or* $ML(1|x) = 1$ *for almost every* $x \in N(\mathcal{X}_{a,1}, \delta) \cap N(\partial\Omega_a^*, \delta)$, *where* $N(S, \delta) = \{x \in \mathbb{R}^p : \|x - y\| < \delta$ *for some* $y \in S\}$ *for a set* $S \subset \mathbb{R}^p$ *and* $\delta > 0$.

(i) (Conditional Moments and Density near $\partial\Omega_a^*$) *There exists* $\delta > 0$ *such that*

(1) $E[Y_i(a)|X_i]$, $E[Y_i(1)|X_i]$, *and* f_X *are continuous and bounded on* $N(\partial\Omega_a^*, \delta)$*;*

(2) $E[Y_i(a)^2|X_i]$ *and* $E[Y_i(1)^2|X_i]$ *are bounded on* $N(\partial\Omega_a^*, \delta)$.

Assumption 3 (Convergence Rate of Bandwidth). $\delta \to 0$ *and* $n\delta \to \infty$ *as* $n \to \infty$.

Here we only discuss a few key assumptions. Note first that Assumption 2 (b) allows the function ML to be discontinuous on a set of points with the Lebesgue measure zero. For example, ML is allowed to be a step function.

When the logging policy ML is deterministic, $\partial\Omega_a^*$ corresponds to the decision boundary for action a in the context space. Assumption 2 (g) imposes the differentiability of the boundary. The conditions are satisfied if, for example, $\Omega_a^* = \{x \in \mathbb{R}^p : f(x) \geq 0\}$ for some twice continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ such that the gradient $\nabla f(x)$ is nonzero for all $x \in \mathbb{R}^p$ with $f(x) = 0$. Furthermore, Assumption 2 (h) (1) assumes that $\partial\Omega_a^*$ is $(p-1)$ dimensional and has nonzero density.

Our consistency result requires that δ goes to zero slower than n^{-1}. The rate condition ensures that, when ML is deterministic, we have sufficiently many observations in the δ-neighborhood of the boundary of Ω_a^*. Importantly, the rate condition does not depend on the dimension of X_i. This is because we use all the observations in the δ-neighborhood of the boundary, and the number of those observations is of order $n\delta$ regardless of the dimension of X_i if the boundary is $(p-1)$ dimensional. Our estimator $\hat{V}(\pi)$ is therefore expected to perform well even if X_i is high dimensional.

[4] The k-dimensional Hausdorff measure on $\mathbb{R}^p$ is defined as follows. Let Σ be the Lebesgue σ-algebra on $\mathbb{R}^p$ (the set of all Lebesgue measurable sets on $\mathbb{R}^p$). For $S \in \Sigma$ and $\delta > 0$, let $\mathcal{H}_\delta^k(S) = \inf\{\sum_{j=1}^\infty d(E_j)^k : S \subset \cup_{j=1}^\infty E_j, d(E_j) < \delta, E_j \subset \mathbb{R}^p$ for all $j\}$, where $d(E) = \sup\{\|x - y\| : x, y \in E\}$. The k-dimensional Hausdorff measure of A on $\mathbb{R}^p$ is $\mathcal{H}^k(S) = \lim_{\delta \to 0} \mathcal{H}_\delta^k(S)$.