Empirical Gateaux Derivatives for Causal Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study a constructive procedure that approximates Gateaux derivatives for statistical functionals by finite-differencing, with attention to causal inference functionals. We focus on the case where probability distributions are not known a priori but need also to be estimated from data, leading to empirical Gateaux derivatives, and study relationships between empirical, numerical, and analytical Gateaux derivatives. Starting with a case study of counterfactual mean estimation, we verify the exact relationship between finite-differences and the analytical Gateaux derivative. We then derive requirements on the rates of numerical approximation in perturbation and smoothing that preserve statistical benefits of the resulting estimation procedure via one-step adjustment, such as rate-double-robustness. We study more complicated functionals such as dynamic treatment regimes and the linear-programming formulation for policy optimization infinite-horizon Markov decision processes. In the case of the latter, this approach can be used to approximate bias adjustments in the presence of arbitrary constraints, illustrating the usefulness of constructive approaches for Gateaux derivatives. We find that, omitting unfavorable dimension dependence of smoothing, although rate-double robustness permits coarser rates of perturbation than implied by generic approximation analysis of finite-differences for the case of the counterfactual mean, this is not the case for the infinite-horizon MDP policy value.

1 Introduction

Inferential targets in causal machine learning are often statistical functionals of the data distribution, for example an average treatment effect or average policy value for infinite-horizon offline reinforcement learning. Estimation of these statistical functionals can be vulnerable to the first-stage bias introduced by estimating nuisance functions such as outcome regressions or transition probabilities. However, one can derive bias-adjusted estimators by leveraging the causal structure and weaken requirements on estimation of these nuisance functions. The celebrated doubly robust estimator is one such bias adjustment, and general frameworks exist to derive and verify influence functions, e.g. double robustness, semiparametric inference, Neyman orthogonality, "debiased/double" machine learning [7, 49, 44, 10, 11]. These general frameworks can also be applied to problems in offline reinforcement learning (also called offline policy evaluation/learning) [57, 35, 55, 5, 36, 54, 37, 38].

Despite extensive literature deriving bias-corrected estimators, these analytic derivations often require significant effort in case-specific analysis, making it challenging to extend to new settings. On the other hand, there are classes of estimands where an analyst may be interested in many variants. For example, in *constrained* MDPs [3], an analyst can easily computationally specify custom linear or convex constraints in the linear programming formulation. In these settings, each possible choice could require a case-specific re-analysis and deduction for deriving bias adjustment. Therefore, complementary to analytic derivations, it may be useful to also develop constructive or algorithmic approaches to obtain some bias adjustment.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

To this end, we study algorithmic and constructive procedures for obtaining bias adjustments for statistical functionals. We build on a recent line of work of [FQWD, CLvdL, IN], which proposes a constructive procedure by numerical approximation of a Gateaux derivative (in the sense of [25, 31]) by finite differences, but with a smoothed perturbation rather than a point mass. In relation to this line of work, we additionally consider interactions between statistical estimation and numerical approximation, where the probability distributions required to evaluate the statistical functional need also to be estimated.

Our contributions are as follows: We begin with a case study of counterfactual mean estimation. In this setting, we can characterize the exact relationship between finite-differences and the analytical Gateaux derivative. This helps concretize the proposal by allowing us to study rates of *numerical* approximation that preserve *statistical* performance. Finally, with these concretizations in hand, we return to the more complicated motivating examples. We illustrate with the case of policy learning in finite-state, finite-action infinite-horizon Markov decision processes how this approach can directly extend also to the case of custom constraints.

2 Problem Setup: From Numerical to Empirical Gateaux derivatives

Setup. Denote $O \sim P$ as the observations following distribution P belonging to a statistical model $\mathcal{M}$. We will focus on targeting the estimation of statistical functionals $\Psi(P)$, where $\Psi : \mathcal{M} \to \mathbb{R}^q$ is pathwise differentiable. Throughout, we generally assume conditions required for causal identification (i.e. overlap/positivity, unconfoundedness). Examples of functionals include counterfactual means.

Example 1 (Counterfactual mean.)**.** *The observation $O = (X, A, Y)$ includes covariate, treatment $A \in \{0, 1\}$, and outcome $Y \in \mathbb{R}$. The statistical functional is:*

$$\Psi(P) = \mathbb{E}[Y(1)] = \mathbb{E}_P[\mathbb{E}[Y \mid A = 1, X]].$$

Influence function and Gateaux derivative. A functional Ψ is Gateaux differentiable if $\Psi'(H; P) \triangleq \frac{d\Psi(P+\epsilon H)}{d\epsilon}|_{\epsilon=0}$ exists and is linear and continuous in H.

[25, 31] defines the *influence function* $\phi(O; P)$ as a Gateaux derivative with respect to perturbations $H = \delta_o - P$ toward a degenerate Dirac measure δ_o satisfying the identity that $\int o\delta_{o'}(o) = o'$,

$$\phi(o; P) = \frac{d\Psi\left(P + \epsilon(\delta_o - P)\right)}{d\epsilon}\Big|_{\epsilon=0}.$$

Related work: Numerical Gateaux derivatives. Although the above characterization is insufficient for continuous distributions, [CLvdL, IN] show that it suffices to restore absolute continuity in semiparametric models by taking an additional limit as a *smoothed perturbation* converges to the point mass. That is, it suffices to replace the Dirac delta measure δ_o with a smoothed distribution $\tilde{\delta}^{\lambda}_{o'}(o)$ (a smoothed delta function perturbed in the direction of an observation $\tilde{o}$, evaluated at o), dominated by P, such that $\int g(u) d\tilde{\delta}^{\lambda}_{o'}(u) = 1$. A common choice for $\tilde{\delta}^{\lambda}_{o'}$ uses kernel functions $\tilde{\delta}^{\lambda}_{o'}(o) = K_\lambda(o - o')$ and the perturbed distribution in the direction of a generic observation o_i is $P^{o_i}_{\epsilon,\lambda} = P^{i}_{\epsilon,\lambda} = (1-\epsilon)P + \epsilon\tilde{\delta}^{\lambda}_{o_i}$, where K_λ is an appropriately bandwidth-normalized kernel function and notation is overloaded to also designate product kernels. [CLvdL, Thm. 1] establishes that the Gateaux derivative is the limit as the paths and smoothing parameter, ϵ and λ, tend to 0, where $\Psi(P^*_{\epsilon,\lambda})$ is the projection of $\Psi(P_{\epsilon,\lambda})$ onto the semiparametric model:

$$\phi_P(o; P) = \lim_{\lambda \to 0} \left[\tfrac{d}{d\epsilon}\Psi\left(P_{\epsilon,\lambda}\right)\right] = \lim_{\lambda \to 0} \lim_{\epsilon \to 0} \ \epsilon^{-1}\left(\Psi(P^*_{\epsilon,\lambda}) - \Psi(P)\right).$$

They provide high-level regularity conditions for preservation of rate of convergence but generally assume all distributions are known. [CLvdL] discuss *numerical* approximation of the derivative by finite-differences: for a fixed ϵ, λ with $\epsilon \ll \lambda$, evaluate the numerical derivative $\hat{\phi}(o; P) = \epsilon^{-1}(\Psi(P_{\epsilon,\lambda}) - \Psi(P))$.

Empirical Gateaux derivatives: Plug-in estimation of the functional. In the context of statistical estimation, the underlying probability distributions also need to be estimated. Therefore, to ensure validity of the numerical derivative approach for a general class of functionals, we consider

plug-in estimation of the statistical functional by conducting probability density estimation of the collection of joint probability estimates $\tilde{p}(o)$ that comprise $\tilde{P}$. For example, a *plug-in* representation of the counterfactual mean is $\Psi(\tilde{P}) = \int\int y \frac{\tilde{p}(y, A=1, x)}{\tilde{p}(A=1, x)} dydx$.

Hence, we distinguish between the *empirical Gateaux derivative* computed by numerical approximation with estimated probability densities, the *numerical* Gateaux derivative obtained by finite-differencing but with the true probability densities, and the analytic Gateaux derivative.

$$\tilde{\phi}(O_i) = \epsilon^{-1}\left(\Psi(\tilde{P}^i_\epsilon) - \Psi(\tilde{P})\right) \quad \text{empirical derivative at smoothed and estimated distributions,}$$

$$\hat{\phi}(O_i) = \epsilon^{-1}\left(\Psi(P^i_\epsilon) - \Psi(P)\right) \quad \text{numerical derivative at smoothed and true distributions,}$$

$$\phi(O_i) = \frac{d}{d\epsilon}\Psi(P^i_\epsilon)\Big|_{\epsilon=0} \quad \text{analytical Gateaux derivative.}$$

The *one-step estimator* is an immediate way to use computationally estimated Gateaux derivatives for statistical estimators [47]; justified by the asymptotically linear representation (see Appendix B).

$$\Psi_n = \Psi(\tilde{P}) + \frac{1}{n}\sum_{i=1}^n \tilde{\phi}(O_i), \qquad \text{where } \tilde{\phi}_i(O_i) = \frac{1}{\epsilon}(\Psi(\tilde{P}^i_\epsilon) - \Psi(\tilde{P})) \tag{1}$$

Problem statement. We study these questions about the empirical, numerical, and analytical derivatives:

Q1: How exactly does the empirical Gateaux derivative approximate the analytical derivative?

Q2: What are the required rates of *numerical approximation* in perturbation and smoothing (ϵ, λ) that preserve the beneficial *statistical* properties of the constructed estimator?

In Section 3, we first address these questions in the simple case of our running example: estimation of the counterfactual mean, where we can finely compare the finite-difference estimator and the classical doubly robust estimator to characterize statistical properties. In Section 4, we study more complicated functionals to highlight more general applicability.

Example 2 (Example 1, continued.)**.** *The AIPW influence function [49, 24] for a counterfactual mean is:*

$$\phi(O) = \frac{\mathbb{I}[A=1]}{p(A=1|X)}(Y - \mathbb{E}[Y \mid A=1, X]) + \mathbb{E}[Y \mid A=1, X] - \Psi(P).$$

Hence the canonical doubly-robust estimator can be viewed as a one-step adjustment.

Remark 1. *We omit discussion of semiparametric efficiency since we focus on computing influence functions without the discussion of model restrictions and tangent spaces. Although [CLvdL, IN] focus in particular on semiparametric models, we focus on Gateaux derivatives under a nonparametric model; [CLvdL] suggest handling evaluating a Gateaux derivative and projecting onto the model space directly. Hence our analysis can be viewed as considering the case of nonparametric models; or analyzing the first step of a procedure for semiparametric models.*

3 Case study: Empirical Gateaux derivative of the counterfactual mean

To bolster confidence in the numerical procedure, we work through the running example of the counterfactual mean and concretize precisely how the *numerical differentiation* incurs error relative to the analytical derivative, i.e. the doubly robust or augmented IPW (AIPW) estimator.

3.1 Relating the empirical and analytical Gateaux derivatives

Setup. We target estimation of a counterfactual mean, Example 1. For the smoothed perturbation, we specialize to kernel functions $K_\lambda(u)$ from kernel density estimation satisfying that $\int_{-\infty}^{\infty} K(u)du = 1$ (normalization to a probability density) and $K(-u) = K(u), \forall u$ (symmetry).[1] The counterfactual mean is identified as:

$$\Psi(P) = \mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y \mid A=1, X]] = \int\int y p(y \mid A=1, x) dydx = \int\int y \frac{p(y, A=1, x)}{p(A=1, x)} dydx \tag{2}$$

[1]Examples include the Gaussian kernel with $K(u) = (2\pi)^{-\frac{1}{2}} e^{-u^2/2}$, or uniform with $K(u) = 1/2\mathbb{I}\left[|u| \leq 1\right]$. We generally consider $K_\lambda(u) = h^{-d}K(u/\lambda), K(x) = K(x_1, \ldots, x_d)$.

Algorithm 1 Empirical Gateaux derivatives

1: Inputs: $\tilde{P}$ probability density estimates, data $\{O_i\}_{1:n}$, perturbation and smoothing (ϵ, λ) parameters, functional Ψ
2: **for** $i = 1, \ldots n$ **do**
3: $\quad P_\epsilon^i = (1-\epsilon)P + \epsilon \tilde{\delta}_{o_i}^\lambda$
4: $\quad \tilde{\phi}(O_i) \leftarrow \epsilon^{-1}(\Psi(P_\epsilon^i) - \Psi(P))$
5: **end for**
6: Output: one-step adjusted estimate from empirical Gateaux derivatives, $\Psi(P) + \frac{1}{n}\sum_i \tilde{\phi}(o_i)$

We obtain $\Psi(\tilde{P}_\epsilon)$ by plugging in the $\tilde{p}$ probability density estimates. Note this is distinct from the plug-in estimator in empirical process theory, which plugs in the empirical CDF; we plug-in statistical estimates. See Algorithm 1 for a summary of the procedure.

Comparison of empirical Gateaux derivative and known doubly robust adjustment. We have argued that in the context of statistical estimation, approximating Gateaux derivatives requires plug-in estimation of the statistical functional with estimated probabilities. Towards resolving Q1, we observe that the perturbed probability densities induce the functional form of the nuisance functions in the analytical derivative. This is in contrast to working with the analytical derivative, which describes the estimand in terms of, e.g. conditional expectations, and allows the analyst to choose the functional form for estimating the nuisance function.

The density-induced conditional expectation is $\mathbb{E}_{\tilde{P}}[Y \mid A = 1, X = x] = \int y\tilde{p}(y \mid A = 1, x)dy$. For example, the density-induced conditional expectation of kernel denstiy estimates is exactly Nadaraya-Watson regression. Given this, we study empirical derivatives with kernel density estimation; although analogous results hold with other nonparametric density estimation. Of course, other estimates can be used: Appendix E for discussion of other choices of density estimates.

Our first proposition will show exactly how plug-in evaluation with finite differences is nearly equivalent to evaluating the doubly-robust estimator with a Nadaraya-Watson regressor as the conditional outcome regression for $\mathbb{E}[Y \mid A = 1, X]$ and a kernel density estimate of the propensity score; up to an additive bias that arises due to the smoothed perturbation distribution.

Proposition 1 (Gateaux derivative of probability density representation)**.** *Consider perturbations in the direction of $o_i = (x_i, a_i, y_i)$. Let $\tilde{p}(x), \tilde{p}(A = 1, x), \tilde{p}(y, A = 1, x)$ denote kernel density estimates. Let $\Psi(P)$ be as in eq. (2).*

$$\frac{\Psi(\tilde{P}_{\epsilon_i}) - \Psi(\tilde{P})}{\epsilon} = \int \frac{\tilde{p}(x)\mathbb{I}\left[a_i = 1\right]\tilde{\delta}_{x_i}^\lambda(x)}{\tilde{p}_\epsilon(A = 1, x)} \left\{ \left(\int y\ \tilde{\delta}_{y_i}^\lambda(y)dy \right) - \mathbb{E}_{\tilde{P}}[Y \mid A = 1, X = x] \right\} dx$$
$$+ \left(\tilde{\mathbb{E}}_{\tilde{P}_\epsilon}[Y \mid A = 1, X = x_i] - \Psi(\tilde{P}) \right).$$

A few simplifications help us directly compare our expansion of Proposition 1 to the canonical AIPW estimator in Example 2. First, assuming multivariate kernel functions are product kernels throughout, we specialize to the case of a uniform kernel for $\tilde{\delta}_{y_i}^\lambda(y)$ so that $\int y\ \tilde{\delta}_{y_i}^\lambda(y)dy = y_i$. We next observe that the convolution with respect to the smoothing kernel introduces additive bias relative to evaluating terms in the expression at the observation o_i.

Integrating the smoothed perturbation induces smoothed nuisance functions. Integrating with respect to the smoothed perturbation distribution can be interpreted as a smoothed evaluation of the resulting nuisance functions. We define the $\tilde{\delta}_{x_0}^\lambda$-smoothed conditional expectation which smooths the evaluation point, i.e. smooths evaluation around x_0 rather than precisely at x_0, as occurs with the point mass perturbation.

Definition 1 ($\tilde{\delta}_{x_0}^\lambda$-smoothed conditional expectation)**.**

$$\tilde{\mathbb{E}}_P[Y \mid A = 1, X = x_0] = \int \mathbb{E}_P[Y \mid A = 1, X = u]\ \tilde{\delta}_{x_0}^\lambda(u)\ du \tag{3}$$

Similar ideas appeared in the literature on non-regular inference, for example smoothing a non-regular target estimand, such as a dose-response curve or density evaluated at a point x_0, by conducting a smoothed evaluation locally around x_0 [58]. In that work, a function evaluated at x_0,

$g(x_0)$, is approximated with a kernel smooth over x in a neighborhood of x_0 with bandwidth h; so $\Psi_\lambda(x_0) = \int K_\lambda(u - x_0)g(u)du$. The following lemma summarizes convergence rate implications of these corresponding smoothed nuisances; these are direct consequences of typical analyses of from kernel density estimation.

Lemma 1. *Let $\tilde{\mu}_\epsilon(X) = \tilde{E}_{\tilde{P}_\epsilon}[Y \mid A = 1, X]$ and $\tilde{e}_\epsilon(X) = \tilde{p}_\epsilon(A = 1 \mid X)/\tilde{p}(X)$ denote the nuisances induced by plug-in estimation of probability distributions $\tilde{P}$. Assume that $\tilde{\delta}^\lambda$ is a product kernel, and $\int |u|^\beta |\tilde{\delta}^\lambda_u| \,\mathrm{d}u < \infty$ and $\int u^s \tilde{\delta}^\lambda_u \,\mathrm{d}u = 0$ for $s \leq \beta$. For any bounded function $g(x)$ in a Hölder class of degree β, $\int \tilde{\delta}^\lambda_x(u)g(u)\,\mathrm{d}u - g(x))^2 \,\mathrm{d}x = O(\lambda^\beta)$. And, the perturbed nuisances satisfy the following rates:*

$$\mathbb{E}[(\tilde{\mu}_\epsilon(X) - \mu(X))^2] = \mathbb{E}[(\tilde{\mu}(X) - \mu(X))^2] + O(\epsilon^2\lambda^{-d}),$$
$$\mathbb{E}[(\tilde{e}_\epsilon(X) - e(X))^2] = \mathbb{E}[(\tilde{e}(X) - e(X))^2] + O(\epsilon^2\lambda^{-d}).$$

The perturbed nuisances are asymptotically consistent, so long as $\tilde{p}$ induces asymptotically consistent nuisances and we fix ϵ as a sequence vanishing in n at appropriate rate to counteract the growth in λ (due to "roughness", ie the variability of $\tilde{\delta}^\lambda$). The difference between the function evaluation and its smoothed nuisance arises from the bias analysis of kernel density estimators and the dimension-dependence of ensuring absolute continuity, but is a non-stochastic, entirely deterministic argument. While the analysis improves mildly in the dimension dependence on λ compared to the sufficient general analysis of [CLvdL], this still highlights the generally unfavorable dependence of the numerical approach on the dimension due to smoothing.

Lemma 1 allows us to further simplify and conclude that the decomposition of Proposition 1 is close to the canonical AIPW estimator up to the $O(\lambda^\beta)$ bias induced by smoothed evaluation.

Corollary 1. *When the perturbation observation is the observation datapoint $O_i = (X_i, A_i, Y_i)$,*

$$\tilde{\phi}(O_i) = \frac{\mathbb{I}[A_i=1]}{\tilde{p}_\epsilon(A=1|X_i)}\left(Y_i - \mathbb{E}_{\tilde{P}}[Y \mid A = 1, X_i]\right) + \left(\mathbb{E}_{\tilde{P}_\epsilon}[Y \mid A = 1, X_i] - \Psi(\tilde{P})\right) + O(\lambda^\beta).$$

3.2 Implications for estimation

Next, we use our characterization of Proposition 1, Lemma 1, and Corollary 1 to study the *statistical* properties of our *computational/numerical* approximation. Corollary 1 is deterministic and allows us to study the relationship to the typical "oracle" AIPW estimator; albeit with induced ϵ-perturbed nuisances. An appealing property of bias-adjusted treatment effect estimation is as *rate-double robustness*: because we incur the product of convergence rates of the nuisances, we may enjoy parametric $n^{-\frac{1}{2}}$-rate convergence of the target functional while nuisances converge at slower rates, for example both at $n^{-\frac{1}{4}}$; see [11, 51] for more detail. We use the rate-double robustness property of the target and Lemma 1 to infer the required rate conditions on (ϵ, λ) that would retain the beneficial statistical properties of the canonical AIPW estimator.

Assumption 1 (Regularity conditions.)**.** *(i) Y is bounded.*

(ii) $p(y \mid A = 1, x), p(A = 1 \mid x)$ are in Hölder classes of minimum degree β.

(iii) The estimates $p(y, a, x), p(a, x), p(x)$ belong to a Donsker function class.

(iv) Assume $\tilde{\mu}, \tilde{e}$ satisfy the product-rate condition:

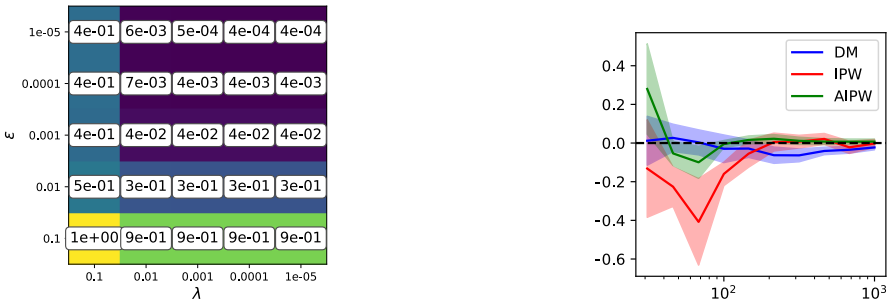
$$\mathbb{E}[(\tilde{\mu}_\epsilon(X) - \mu(X))^2] \times \mathbb{E}[(\tilde{e}_\epsilon(X) - e(X))^2] = o_p(n^{-1})$$

Equivalently, we assume $\tilde{\mu}, \tilde{e}$ are RMSE-consistent with r_μ, r_e, respectively, so that $r_\mu + r_e \geq \frac{1}{2}$.

Theorem 1 (Rate double-robustness)**.** *Consider the one step-estimator with the empirical Gateaux derivative. Under Assumption 1, when $\epsilon\lambda^{-d/2} = o(n^{-\max(r_\mu, r_e)})$, and $\lambda^\beta = o(n^{-\frac{1}{2}})$:*

$$(\Psi(\tilde{P}) + \tfrac{1}{n}\textstyle\sum_{i=1}^n \tilde{\phi}(O_i)) - \Psi(P) = O_p(n^{-\frac{1}{2}}).$$

We now interpret the qualitative estimation implications of the analysis. Theorem 1 verifies the rate conditions of *numerical approximation* in ϵ, λ that preserve the *statistical property* of rate double



(a) (b)

Figure 1: (a) Epsilon-lambda plot for counterfactual mean: MAE in approximation of one-step adjustment. (b) Error in estimation of $\mathbb{E}[Y(1)]$ over n.

robustness. Assuming that the nuisances induced by the probability density estimations satisfy a product-rate condition that the product of their RMSEs is faster than the parametric $O(n^{-\frac{1}{2}})$ rate, the form of Proposition 1 suggests that the rate of RMSE convergence of perturbed nuisances is required to be faster than the rates of the unperturbed $-$induced nuisance functions, $\max(r_\mu, r_e)$, which can be a slower rate than that implied by the generic analysis of finite differences.

In general, finally the dimension dependence on λ in Lemma 1 will prevent setting a slower rate for ϵ for nontrivial dimensions. As a thought experiment, we note that without the smoothing requirement, the potential improvement in the rate of ϵ could be on the order of generic rate improvements implied by a central difference scheme. That is: we would have attained the approximation benefits of generically approximation-improved finite difference schemes via the specialized *statistical* structure of the target adjustments.

4 Applications: counterfactual mean and dynamic settings

We first show an epsilon-lambda plot in the simple case of the counterfactual mean as a numerical validation. Then, with some of the tools we derived in the previous section, we analyze more complicated settings: longitudinal data (dynamic treatment regime) and the policy linear program for an infinite-horizon Markov decision with linear constraints, with attention to how different functionals have different interactions between approximation error and statistical properties.

4.1 Counterfactual mean

We first conduct a validation of our decomposition and theoretical characterization. Our data generating process includes a piecewise linear outcome model (such that kernel regression is misspecified). In Figure 1a we include an (ϵ, λ)-plot for the simple case of AIPW (see [9] for more discussion and examples). We consider a one-dimensional case with uniformly distributed X, piecewise-linear Y, and smooth propensity scores that are logistic in $\sin(X)$. We use $n = 500$ and fix the bandwidth $h = 0.05$. We use colors to denote magnitude of the mean absolute error, included in text on the heatmap. Without loss of generality, we study the estimation of a mean under missingness, $\mathbb{E}[Y(1)]$. Figure 1b illustrates the estimation error of various strategies with the comparable kernel-based estimates (DM is regression adjustment): even in this simple setting AIPW offers benefits and sample efficiency relative to DM. We include further numerical experiments in the Appendix.

4.2 Dynamic treatment regimes

We study the estimation properties of the empirical Gateaux derivative for multi-stage DTR from the probability density representation. In T-stage dynamic treatment regimes, the causal quantity of interest is the counterfactual mean outcome $\mathbb{E}\left[Y(\bar{a})\right]$, where $\bar{a} = (a_0, \ldots, a_T)$ is the (deterministic) treatment strategy. Assuming $Y(\bar{a})$ is sequentially ignorable given treatment and covariate history

$(\bar{A}_t, \bar{X}_t)$ at each time t, this causal quantity can be identified by the g-formula,

$$\mathbb{E}\left[Y(\bar{a})\right] = \int \mathbb{E}\left[Y \mid \bar{A} = \bar{a}, \bar{X} = \bar{x}\right] \prod_{t=1}^{T} \tilde{p}(x_t \mid \bar{a}_{t-1}, \bar{x}_{t-1})\, \mathrm{d}\bar{x},$$

where $\bar{A} = (A_0, \ldots, A_t), \bar{X} = (X_0, \ldots, X_t)$. To derive its influence function, we take the empirical Gateaux derivative by considering the perturbation in the direction of $o_i = (\bar{x}_i, \bar{a}_i, y_i)$, where $\bar{x}_i = (x_{i0}, \ldots, x_{it})$ and $\bar{a}_i = (a_{i0}, \ldots, a_{it})$,

$$\Psi(\tilde{P}_\epsilon^i) = \int \left(\int y \frac{\tilde{p}_\epsilon(y, \bar{A}=\bar{a}, \bar{X}=\bar{x})}{\tilde{p}_\epsilon(\bar{A}=\bar{a}, \bar{X}=\bar{x})}\, \mathrm{d}y \right) \prod_{t=1}^{T} \frac{\tilde{p}_\epsilon(x_t, \bar{a}_{t-1}, \bar{x}_{t-1})}{\tilde{p}_\epsilon(\bar{a}_{t-1}, \bar{x}_{t-1}))} \tilde{p}_\epsilon(x_0)\, \mathrm{d}\bar{x}. \tag{4}$$

Below we characterize how this empirical Gateaux derivative at the smoothed distribution differs from the one at the (unsmoothed) estimated distribution. The following result is analogous to Proposition 1 and corollary 1 but is extended to the dynamic treatment regimes.

Proposition 2.

$$\begin{aligned}
&\epsilon^{-1}(\Psi(\tilde{P}_\epsilon^i) - \Psi(\tilde{P})) \\
&= \left(\mathbb{E}_{\tilde{p}(x_{1:T})} \left[\mathbb{E}_{\tilde{P}_\epsilon} \left[Y \mid \bar{A}, \bar{X} \right] \mid \bar{x}_0 \right] - \Psi(\tilde{P}) \right) + \frac{\mathbb{I}\left[\bar{A} = \bar{a}_i\right]}{\tilde{p}_\epsilon(\bar{a} \mid \bar{x})} \left[Y_i - \mathbb{E}_{\tilde{p}(\bar{x})} \left[\mathbb{E}_{\tilde{P}} \left[Y \mid \bar{A}, \bar{X} \right] \right] \right] \\
&\quad + \sum_{s=1}^{T} \frac{\mathbb{I}\left[\bar{A} = \bar{a}_{i,T-s}\right]}{\tilde{p}_\epsilon(\bar{a}_{T-s} \mid \bar{x}_{T-s})} \left\{ \mathbb{E}_{\tilde{P}} \left[\mathbb{E}_{\tilde{P}_\epsilon} \left[Y \mid \bar{A}, \bar{X} \right] \mid \bar{a}_{i,T-s}, \bar{x}_{i,T-s+1} \right] \right. \\
&\qquad \left. - \mathbb{E}_{\tilde{P}} \left[\mathbb{E}_{\tilde{P}_\epsilon} \left[Y \mid \bar{A}, \bar{X} \right] \mid \bar{a}_{i,T-s}, \bar{x}_{i,T-s} \right] \right\} + O(\epsilon^2) + O(\lambda^\beta)
\end{aligned} \tag{5}$$

The expansion verifies that the requirements for numerical approximation are similar in ϵ as in the case of a single-timestep counterfactual mean, despite the additional complexity of the dynamic treatment regime function as a nested expectation. Again, smoothing incurs overall unfavorable dependence in the dimension, as the extension of Lemma 1 to $\mathbb{E}_{\tilde{P}_\epsilon}[Y \mid \bar{A}, \bar{X}]$ also incurs unfavorable dependence on the dimension. The history-dependent nuisances imply that the dimension additionally grows in time horizon.

4.3 Perturbations for (constrained) infinite-horizon off-policy evaluation

We show how this framework can be applied to derive generic bias corrections for infinite-horizon off-policy optimization in offline reinforcement learning, via the canonical *linear programming* characterization of [48, 14]. Our application is relevant to a recent line of work on off-policy evaluation and learning (OPE/L) [35, 55, 5, 36, 54, 37, 38].

Setup. We recall the classical linear programming formulation of finding an optimal policy in a tabular (finite-state, finite-action) Markov decision process, due to [48, 14]. Recent work has revisited this formulation with interest in developing primal-dual algorithms [59, 42, 42, 1, 52]. The dynamics follow an infinite-horizon Markov decision process with discount factor γ. The offline dataset comprises of trajectories of state, action, next state observations: $\{(s_i^0, a_i^0, \ldots, s_i^t, \ldots, s_i^T)\}_{1:N}$. We derive our bias adjustment under a statistical model where the stationary distribution factorizes as $p(s, a, s') = d(s, a)P(s \mid s, a)$. From the joint state-action-state occupancy distribution p(s,a,s'), we estimate the marginalization $d(s, a)$, the stationary state-action occupancy distribution, which we use to estimate the transition probability $P(s' \mid a, s) = \frac{p(s,a,s')}{d(s,a)}$. Let $\mu_0(s)$ denote the initial state distribution, estimated from offline data and assumed to be equivalent in the original confounded and evaluation setting. Because we focus on the tabular case with discrete distributions, we do not require smoothing.

The optimal policy linear program. The *optimal* value function solves the following linear program in a finite-state and finite-action setting; the objective value at optimality is the policy value. Here, $P_a \in \mathbb{R}^{|S| \times |S|}$ is transition matrix at action a, with $P_a(s, s') = P(s' \mid s, a)$, $\mu_0(s)$ is the marginal occupancy distribution of the initial state. We assume, as is fairly common in theoretical analyses of reinforcement learning, that the reward function is a known function of state and action, $r(s, a)$. Let $\mathbf{r}_a = \{r(s, a)\}_{s \in \mathcal{S}}$ be the vector of reward values per state, for fixed action a.

$$\Psi_D(P) = \min_V \{(1-\gamma)\mu_{\mathbf{0}}{}^\top \mathbf{V} \colon\ (I - \gamma P_a)\, \mathbf{V} - \mathbf{r}_a \geq 0, \quad \forall a \in \mathcal{A}\} \tag{6}$$

The dual formulation is well-known to parametrize the stationary occupancy probabilities of the optimal policy (at optimality):

$$\Psi_P(P) = \max_{\mu} \left\{ \sum_{a\in\mathcal{A}} \mu_a^\top \mathbf{r_a} \colon \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\mu_a = (1-\gamma)\mu_{\mathbf{0}}, \quad \mu_a \geq 0, \quad \forall a \in \mathcal{A} \right\} \tag{7}$$

We verify the bias correction from the Gateaux derivative $\frac{d}{d\epsilon}\Psi_D(P_\epsilon)\Big|_{\epsilon=0}$. Note that evaluating this functional at the perturbed distribution corresponds to evaluating the linear program with perturbed left-hand side coefficients, i.e. perturbing the constraints of the coefficient matrix as well as the objective (marginalizing over a perturbed initial state distribution. We evaluate the Gateaux derivative in the direction of $(\tilde{s}, \tilde{a}, \tilde{s}')$. We assume nondegeneracy (unique optimal solution); this can be relaxed by using more sophisticated perturbation analysis.

Assumption 2. *The optimal solution is unique; there is no degeneracy.*

Proposition 3. *Perturb in the direction of a generic observation* $(\tilde{s}, \tilde{a}, \tilde{s}')$.

$$\frac{d}{d\epsilon}\Psi_D(P_\epsilon)\Big|_{\epsilon=0} = (1-\gamma)V^*(\tilde{s}) - \frac{\mu^*(\tilde{s},\tilde{a})}{d(\tilde{s},\tilde{a})}\left(r(\tilde{s},\tilde{a}) + \gamma V^*(\tilde{s}') - V^*(\tilde{s})\right) - \Psi_D(P) \tag{8}$$

Sketch of argument. The statement of proposition 3 is not surprising per se, since we can immediately verify the mean-zero sample version of the Bellman residual, and so one might plausibly derive this form from double robustness. Rather, we provide an alternative argument that applies perturbation analysis of optimization programs with the specific perturbations from the finite difference. We build on the sensitivity analysis characterization of [18] which studies perturbations of linear programs with respect to constraint matrix perturbations. For ϵ small enough, the *active basis* remains the same. When the perturbation matrix is $P + \epsilon G$ for some matrix G, [18] notes that the derivative of the optimal value is $d'(\epsilon) = -\overline{\pi} G \overline{x}$ (where $(\overline{\pi}, \overline{x})$ are the dual and primal optimal solutions, respectively) can be obtained from the derivative of a matrix inverse, and obtains higher order expansions. We generalize this argument for the specific form of the perturbed transition probabilities.

We next compute the approximation error that arises from evaluating finite differences, so long as ϵ is small enough to maintain the same active basis, which is empirically verifiable.

Proposition 4 (Error analysis from finite-differencing)**.** *Perturb in the direction of a generic observation* $(\tilde{s}, \tilde{a}, \tilde{s}')$. *Then:*

$$\epsilon^{-1}(\Psi(P_\epsilon) - \Psi(P)) = (1-\gamma)V_\epsilon^*(\tilde{s}) - \frac{\mu^*(\tilde{s},\tilde{a})}{d_\epsilon(\tilde{s},\tilde{a})}\left(r(\tilde{s},\tilde{a}) + \gamma V(\tilde{s}') - V_\epsilon(\tilde{s})\right) - \Psi_D(P) + O(\epsilon) \tag{9}$$

In this setting, unlike the earlier case of the treatment mean in Corollary 1 and Theorem 1, even in a favorable discrete-state case without smoothing requirements, it does not appear that rate-double robustness would admit weaker numerical requirements on ϵ.

Generalization to constraints. More importantly, we argue the main benefit of this approach is that by computing these finite differences, we can obtain generic bias corrections for generalizations building upon this formulation. For example, consider a relevant subclass of constrained Markov decision processes comprising of additional constraints on the policy variables: the linear programming formulation is particularly appealing because linear constraints can be directly added [3]. An advantage is that bias correction can be obtained by evaluating the functional at feasible points (i.e. optimizing with respect to feasible probability distributions).

Example 3 (Constrained policy optimization)**.**

$$\Psi_P(P) = \max \left\{ \sum_{a\in\mathcal{A}} \mu_a^\top \mathbf{r_a} \colon \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\mu_a = (1-\gamma)\mu_{\mathbf{0}}, \;\; \mu_a \geq 0, \;\; \forall a \in \mathcal{A}, \; \mu \in \mathcal{P} \right\}$$

The additional constraint $\mu \in \mathcal{P}$ reflects linearly representable constraints on state-occupancy under the optimal policy. For example, relevant constraints could require the occupancy in certain risky states to be bounded. [46] develops a framework with a "caution function", a convex function on the state-action occupancy measures; this analysis applies to linear caution functions.

Comparison to related work. Double robustness has been studied (for policy evaluation) in [54] while this formulation has been built upon for policy learning in [43]. The first work focuses on policy evaluation. Unsurprisingly, our derivation via perturbation analysis is similar to [54, Thm. 4.1]'s illustration via the Lagrangian (except they also consider self-normalization). This is unsurprising since the Lagrangian provides the dual characterization that also underlies perturbation analysis. Doubly robust evaluation has been studied in [36] (although efficiency is subtle due to regularity issues). Relative to the doubly robust estimators derived elsewhere, this influence function may be higher variance. Note this approach attains doubly robust estimation of the optimal policy's *value*.

5 Related Work

In this section, we discuss the most closely related directions but do not attempt to provide a survey or overview of influence functions/semiparametric statistics; see [7, 56, 16, 40] for a broader overview or the appendix for more detailed discussions.

Numerical approximation of Gateaux derivatives. Recent pedagogical reviews [16, 27, 40] emphasize the Gateaux derivative and its use as a heuristic tool for deriving influence functions. [9] derive high-level sufficient conditions for approximation error via general results for finite-difference approximations. Relative to those general sufficient conditions, we focus on specific improvements in approximation error in different functionals. For the specific case of the counterfactual mean/average treatment effect [17] use additional structural knowledge. Namely, they change orders of integration to obtain functional evaluations in terms of the *conditional expectation*. In general, it may be difficult to deduce the form of the nuisance functions without closely using knowledge of the estimand. The approach in Appendix B of [8] considers the specific case of debiasing moment conditions with respect to first stages but also implicitly leverages this change in order of integration.

Other work on computerizing causal inference with different approaches. Other recent work in causal inference builds towards "automatic" or "computerized" semiparametric inference via different approaches: generalized regression residuals [12, 13, 28, 15, 39], or graphical criteria when the graph is known) [53, 50, 4, 21, 2]. Our last application to optimization is not a generalized regression residual.

Numerical derivatives and statistical machine learning. Independently of causal inference, other works study the use of numerical gradients in the context of machine learning and stochastic gradients [45], and statistical estimation and inference [30, 29]. More broadly in machine learning, influence functions have been used for qualitative sensitivity analysis or bias correction [41, 19, 20, 60]; further connections could be interesting, although the causal inference setting is different because one needs to compute influence functions of the statistical functional with respect to perturbations in the distribution (rather than perturbing an estimator). [22] builds on the finite difference method of [33] for debiasing data-driven optimization with causal outcome coefficients. [23] also considers finite differencing for debiasing linear programs.

6 Conclusion

Limitations and negative impacts. We discuss some limitations. Throughout, we have assumed the functional is pathwise differentiable. This does warrant caution for a completely general approach, since it may be difficult to verify pathwise differentiability generically without analytically analyzing/deriving the influence function. Further, establishing that the remainder term is second-order is a problem-specific exercise. Causal inference remains a highly context-dependent analysis, so that empirical Gateaux derivatives are intended to augment but not supplant or replace analytical expertise.

Concluding remarks. Our analysis suggests that the current major technical limitations include the curse of dimensionality that arises due to smoothing. On the other hand, regarding Q2, our analysis of different functionals surfaced fine-grained differences in opportunities for statistical properties to weaken requirements of computational approximation: Theorem 1 and Proposition 4 have different, estimand-dependent implications. It would be interesting to generalize this further, such as to Z- and M-estimators. Our analysis also relied on a priori knowledge of beneficial statistical properties: it would be interesting to verify these adaptively. Generalizing our analysis to an approximation error calculus would also be an interesting direction.

References

[1] Y. Abbasi-Yadkori, P. L. Bartlett, X. Chen, and A. Malek. Large-scale markov decision problems via the linear programming dual. *arXiv preprint arXiv:1901.01992*, 2019. (Cited on 7)

[2] A. Alaa and M. Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201. PMLR, 2019. (Cited on 9, 15)

[3] E. Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999. (Cited on 1, 8)

[4] R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020. (Cited on 9, 15)

[5] A. F. Bibaut, I. Malenica, N. Vlassis, and M. J. van der Laan. More efficient off-policy evaluation through regularized targeted learning. *arXiv preprint arXiv:1912.06292*, 2019. (Cited on 1, 7)

[6] P. J. Bickel and Y. Ritov. Nonparametric estimators which can be" plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003. (Cited on 31)

[7] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993. (Cited on 1, 9)

[8] F. Bravo, J. C. Escanciano, and I. Van Keilegom. Two-step semiparametric empirical likelihood inference. *The Annals of Statistics*, 48(1):1–26, 2020. (Cited on 9)

[9] M. Carone, A. R. Luedtke, and M. J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 2018. (Cited on 2, 3, 5, 6, 9)

[10] V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016. (Cited on 1)

[11] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018. (Cited on 1, 5)

[12] V. Chernozhukov, W. Newey, R. Singh, and V. Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020. (Cited on 9, 15)

[13] V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021. (Cited on 9, 15)

[14] D. P. De Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003. (Cited on 7)

[15] M. H. Farrell, T. Liang, and S. Misra. Deep learning for individual heterogeneity: an automatic inference framework. 2021. (Cited on 9, 15)

[16] A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172, 2021. (Cited on 9)

[17] C. E. Frangakis, T. Qian, Z. Wu, and I. Diaz. Deductive derivation and turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4):867–874, 2015. (Cited on 2, 9, 15)

[18] R. M. Freund. Postoptimal analysis of a linear program under simultaneous changes in matrix coefficients. In *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, pages 1–13. Springer, 1985. (Cited on 8, 28, 29)

[19] R. Giordano, M. I. Jordan, and T. Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019. (Cited on 9)

[20] R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019. (Cited on 9)

[21] F. R. Guo, E. Perković, and A. Rotnitzky. Variable elimination, graph reduction and efficient g-formula. *arXiv preprint arXiv:2202.11994*, 2022. (Cited on 9, 15)

[22] W. Guo, M. I. Jordan, and A. Zhou. Off-policy evaluation with policy-dependent optimization response. *arXiv preprint arXiv:2202.12958*, 2022. (Cited on 9)

[23] V. Gupta, M. Huang, and P. Rusmevichientong. Debiasing in-sample policy performance for small-data, large-scale optimization. *arXiv preprint arXiv:2107.12438*, 2021. (Cited on 9)

[24] J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998. (Cited on 3)

[25] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974. (Cited on 2)

[26] L. A. Hannah, D. M. Blei, and W. B. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(6), 2011. (Cited on 31)

[27] O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, pages 1–13, 2022. (Cited on 9, 15)

[28] D. A. Hirshberg and S. Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, 2021. (Cited on 9, 15)

[29] H. Hong and J. Li. The numerical delta method. *Journal of Econometrics*, 206(2):379–394, 2018. (Cited on 9)

[30] H. Hong, A. Mahajan, and D. Nekipelov. Extremum estimation and numerical derivatives. *Journal of Econometrics*, 188(1):250–263, 2015. (Cited on 9)

[31] P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004. (Cited on 2)

[32] H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*, 2015. (Cited on 2, 3, 31)

[33] S. Ito, A. Yabe, and R. Fujimaki. Unbiased objective estimation in predictive optimization. In *International Conference on Machine Learning*, pages 2176–2185. PMLR, 2018. (Cited on 9)

[34] A. Jesson, S. Mindermann, Y. Gal, and U. Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021. (Cited on 31)

[35] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*, 2016. (Cited on 1, 7)

[36] N. Kallus and M. Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 2022. (Cited on 1, 7, 9)

[37] N. Kallus and A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020. (Cited on 1, 7)

[38] N. Kallus and A. Zhou. Stateful offline contextual policy evaluation and learning. In *International Conference on Artificial Intelligence and Statistics*, pages 11169–11194. PMLR, 2022. (Cited on 1, 7)

[39] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems*, 28, 2015. (Cited on 9, 15)

[40] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022. (Cited on 9, 15)

[41] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. (Cited on 9, 15)

[42] D. Lee and N. He. Stochastic primal-dual q-learning. *arXiv preprint arXiv:1810.08298*, 2018. (Cited on 7)

[43] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019. (Cited on 9)

[44] W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994. (Cited on 1)

[45] D. Oktay, N. McGreivy, J. Aduol, A. Beatson, and R. P. Adams. Randomized automatic differentiation. *arXiv preprint arXiv:2007.10412*, 2020. (Cited on 9)

[46] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on 8)

[47] J. Pfanzagl. Estimation in semiparametric models. In *Estimation in Semiparametric Models*, pages 17–22. Springer, 1990. (Cited on 3)

[48] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (Cited on 7)

[49] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994. (Cited on 1, 3)

[50] A. Rotnitzky and E. Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.*, 21:188–1, 2020. (Cited on 9, 15)

[51] A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021. (Cited on 5)

[52] J. B. Serrano and G. Neu. Faster saddle-point optimization for solving large-scale markov decision processes. In *Learning for Dynamics and Control*, pages 413–423. PMLR, 2020. (Cited on 7)

[53] E. Smucler, F. Sapienza, and A. Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020. (Cited on 9, 15)

[54] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019. (Cited on 1, 7, 9)

[55] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *Journal of Machine Learning Research*, 2016. (Cited on 1, 7)

[56] A. A. Tsiatis. Semiparametric theory and missing data. 2006. (Cited on 9, 15)

[57] M. J. Van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*, volume 5. Springer, 2003. (Cited on 1)

[58] M. J. van der Laan, A. Bibaut, and A. R. Luedtke. Cv-tmle for nonpathwise differentiable target parameters. In *Targeted Learning in Data Science*, pages 455–481. Springer, 2018. (Cited on 4)

[59] M. Wang. Randomized linear programming solves the discounted markov decision problem in nearly-linear running time. 2017. (Cited on 7)

[60] A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020. (Cited on 9)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See limitations and potential negative impacts paragraph.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See limitations and potential negative impacts paragraph.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are typically stated before they are used.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are in appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes, included in supplemental materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]