Understanding Efficient Estimators in Off-policy Value Estimation [*]

Jiaxu Ren[†]
Center for Data Science
New York University

Supervisor:
Michele Santacatterina[‡]
Grossman School of Medicine. NYU

December 8, 2024

Abstract

Off-Policy learning and evaluation have been universally applied to many fields such as health care, long-term medicine and recommendation system. Learning the optimal policy assists researchers in performing decision-making that reaps awards to dynamic systems. The objective of this technique can be generalized as a process of estimating or evaluating an unknown (known) policy that has been initiated and then use it to learn a new policy. Therefore, Off-policy policy evaluation ($OPPE$) measures both observables and counterfactuals in that it evaluates the initiated current policy as an available ground truth and use the information to fuel the process of updating new policies we may deploy to target audience in the future. Whereas conventional methods are provably successful in estimating the value of policies in AI-centered tasks such as robotic control and game training, in scenarios where it is infeasible to schedule experiments due the the cost and human interactions, finding an efficient, accurate method used to estimate off-policy value needs more efforts leads to a necessary measure to use our data sufficiently. Classic parametric will fail in such areas as precision health and medicine due to the lack of domain knowledge and danger of applying partially trained model(Levine et al. 2020). Therefore, launching a method that can safely evaluate the off-policy value is urgent.

[*] **This is a technical notes for off-policy evaluation and semiparametric efficiency theory. This technical note is under active development and may not result in a formal paper**

[†] Research Associate of Center for Data Science jr5674@nyu.edu

[‡] *

0.1 Outline and Required Elements

This section will be deleted when this ”bad” draft is finished.

1. Review of off-policy should be thorough and profound.
2. This project might not be focusing on models.
3. Should be clear about each step of technical details

1 Introduction

Policy learning capitalizes research fields such as health care, online advertising and robots, which could be boiled down to a decision-making problem or sequential planning (Bennett and Kallus 2023; Jiang and Li 2016; Kallus and Uehara 2020; Levine et al. 2020). Standard offline reinforcement learning utilizes limited data or fixed data without seizing any opportunities of collecting additional data for future use, leading to limitations and that preclude us reaching better reward regions. Therefore, in this context, policy improvement these offline policy algorithms can make should be very conservative or limited as the learned policy will generate severely erroneous actions and states. Poor estimate of the current policy and the risk of deploying the learned policy on future training or test data present a challenge *distribution shift* to offline reinforcement learning. In discussion of Levine et al. 2020, offline policy evaluation becomes susceptible for several reason. Since offline policy learning is restricted to Batch Learning, it is unable to make any corrections regarding the over-pessimistic or optimistic policy that might give a unreasonable action. In this way, policy will be making counterfactual queries in the out-of-distribution(OOD) settings. Many methods have been proposed to mitigate this issue and related literature is rich. These methods provide good solutions under the contexts where appropriate data has been collected and assumptions are satisfied.

Lurking problems remain unresolved when we apply our models or algorithms to the real production, such as healthcare, precise drug and robotic control. For example. long-standing treatment regimes in medicine leads to learning a policy using observational data of patients. Clinicians and scientists complete decision-making based on the learnt policy. Lab tests may be conducted at different levels.

The article is of organization as follows: In Section 1., we will formulate off-policy value estimation and then largely review the literature on *Off-policy evaluation* and build connections between it and nature of problems in *causal inference*. Section 1.1 surveys several benchmark works in policy value estimation that are widely studied in reinforcement learning as a minimal foundation. Section 1.2, we highlight the challenges in those regular off-policy value estimation such *curse of horizon* , high-variance and easy model mis-specification and then introduce Double Robust RL in off-policy evaluation as one potential improvement that deals with some of these challenges. From part 4 of Section 1., we focus on discussion of the major problem *distribution shift*, which will be formulated from perspectives of both causal inference and reinforcement learning.

2 Related Works

2.1 Overview of off-policy value estimation

Article to cite:

1. (Kennedy 2023) Target Learning: nonparametric
2. (Jiang and Li 2016) Doubly Robust policy
3. (Kallus and Uehara 2019) fundamentals for bounded RL
4. (Bennett and Kallus 2023) proximal causal inference [POMDP]
5. (Kaddour et al. 2022) Large review of causal inference open problems
6. (Lagoudakis and Parr 2003) used for small discussion: policy iterations
7. Bennett, Kallus, and Oprescu 2023 added to summary of MDP (just mention)
8. (Chakraborty and Moodie 2013) Clear definition of Reinforcement Learning

Policy value estimation is placed at the central status in standard reinforcement learning and has obtained growing attentions and popularity (Kaddour et al. 2022; Levine et al. 2020). Albeit policy value estimation has variations in many framework of both theoretical study and intelligence applications, researchers are mainly interested in two major problems. These two problems are always sequential: one is to use a batch dataset $\mathcal{D}$ generated by a behavior policy π_b to evaluate a different target policy π_e; the other is to harness the information of behavior policy to learn a better policy π that maximizes the amount reward received in this system under different decision rules. (Chakraborty and Moodie 2013; Kaddour et al. 2022; Kallus and Uehara 2019). Many reinforcement learning problems are modeled by Markov Decision Process (MDP) and then extended to some more complicated scenarios. However, not all RL problems fall inside the category of MDP because for some areas such as medical research, personalized medicine and user recommendation(Yu et al. 2021), there is no strong well-known evidence indicating that long-term outcomes will be affected by only the immediately preceding variables as opposed to the entire history(Chakraborty and Moodie 2013). Therefore, we will base our following discussion on the general case in reinforcement learning and thus do not assume MDP. For completeness, MDP is stated in short space and then a general case is proposed.

MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, d_0, r, \gamma)$ (Definition 2.1 of Levine et al. (2020)). $\mathcal{S}$ is called state space including all possible states that can be continuous or discrete. Without generalizability, each state $s \in \mathcal{S}$ can be a multi-variate vector and any features that summarize information of states. $\mathcal{A}$ is the discrete action space. d_0 is an initial state used for generating the following trajectories afterward. r and γ are rewards and discount factors that balance the distant future and recent rewards, respectively. Under MDP, the transition probability of each state s_t depends on only the immediate past state s_{t-1} and action a_{t-1} and thus is independent of those distant history, leading to an important property *memory-less*. Combining these terms and definitions, one trajectory under the behavior policy can be written as $d = (s_0, a_0, r_0, s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t)$ and its probability is

$$p_\pi(d) = d_0(s_0) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_t|s_{t-1}, a_{t-1}) \tag{2.1}$$

We use this simple factorization of the joint distribution of trajectory to introduce a more general case. As was discussed in the last paragraph, since research fields such as precision medicine or personalized recommendation does not strictly follow MDP framework, an additional variable history H encompassing all distant, previous actions and states should be considered. Since s_t, a_t and reward r are random variables, history H collecting all past information is also a random variable and it affects the joint probability of trajectories:

$$p_\pi(d) = d_0(s_0) \prod_{t=1}^{T} \pi(a_t|H_t) p(s_t|H_{t-1}, a_{t-1}) \tag{2.2}$$

wherein each H_t absorbs all information of the past states and actions and thus satisfies the sequential randomization (Chakraborty and Moodie 2013). $\pi(a_t|H_t)$ and $p(s_t|H_{t-1}, a_{t-1})$ define the action taken at time t regarding the history and the transition probability of states. The value of policy π , defined by the expectation of total rewards over the trajectory distribution p_π, is

$$\mathbb{E}_{d \sim p_\pi(d)}[\sum_{t=0}^{T} r_t] \tag{2.3}$$

One can understand this policy value from perspectives of both probability theory and functional estimation. At each time t , $r_t = r(H_t, a_t)$ as a functional with respect to the its distant history and immediate past actions. Calculating the total rewards at each t is nothing but to construct a function with regard to a trajectory $d = (s_0, a_0, s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t)$. This leads a more complex function $R(s_0, a_0, s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t) = R(d)$. It immediately follows that this quantity can be calculated through conditional expectation of rewards over the entire trajectory distribution. It is easy to justify the rule holds for the length-2 trajectory ($T = 2$).

Two estimates of interest are defined beyond the value of policy π for the purpose of modelling and optimization: state value $V^d(s_j)$ and action-state value $Q^d(s_j, a_j)$. V value is defined as the total expected rewards received by the agent starting at a particular state s_j and taking actions according to the trajectory distribution. Without assuming MDP, s_k is replaced with an entire history at time k:

$$V_k^d(H_k = h_k) = \mathbb{E}[\sum_{t=k}^{T} r(H_k, A_k)|H_k = h_k] \tag{2.4}$$

To avoid confusion for the random variable H and a fixed value h, we write here explicitly. This value is actually an expectation of total regards starting off at time $t = j$ onward. Therefore, this quantity excludes rewards received at much earlier stages $t \leq j - 1$. It is straightforward to conclude that, if the starting time is set to $t = 0$, we have

$$V_0^d(H_k = h_0) = \mathbb{E}[\sum_{t=0}^{T} r(H_k, A_k)|H_k = h_0] = \mathbb{E}[\sum_{t=0}^{T} r(H_k, A_k)|H_k = s_0] \tag{2.5}$$

Similarly, Q value is defined as V by adding one more condition for a fixed action $A_k = a_k$:

$$Q_k^d(H_k = h_k, A_k = a_k) = \mathbb{E}[\sum_{t=k}^{T} r(H_k, A_k)|H_k = h_k, A_k = a_k] \tag{2.6}$$

Q value seems to be opaque to understand because it it fixes two values for the starting state. After some expansions in Q, one can find that randomness included in Q function comes from both the starting h_k and a_k but that randomness included in V comes only from h_k. Therefore, by the marginal distribution, they satisfy the following relation:

$$V_k^d(H_k = h_k) = \mathbb{E}_{a_k \sim \pi}[Q_k^d(H_k = h_k, A_k = a_k)] \tag{2.7}$$

Equation 2.7 is useful in that is reveals the underlying logic to use Bell Equation to find an optimal policy given a hypothetical policy space $\mathcal{P}$. Expanding it with some algebra and law of conditional expectation, we can recursively define the V value as follows:

$$V_k^d(H_k = h_k) = \mathbb{E}[\sum_{t=k}^{T} r(H_k, A_k)|H_k = h_k] = \mathbb{E}[R_k(H_k, A_k) + V_{k+1}^d(H_{k+1})] \tag{2.8}$$

Equation 2.8 can be expressed in a form of Bellman Equation System (see Sutton and Barto

(2018) for details) . The optimal policy can be derived through solving this linear equation system. This method is as known as Dynamic Programming but will be extremely computational costly. An approximation version of Bellman and actor-critic methods that combine the ideas of both Dynamic Programming and Policy Gradient have been proposed to mitigate those issues. Our work plans to focus only on the methods in off-policy estimation and evaluation and thus those topics are sitting beyond the scope of our discussion. However, this article shall emphasize that even if without clear presentation of those optimization methods used for policy update, the solutions derived in the corresponding framework mainly aim at solving problems in a parametric manner. This is because both Dynamic Programming and Actor-critic employ algorithms that target a parameterized model of V and Q value based on which the policy value can be more accurately estimated.

2.2 Methods in Off-policy estimation

Albeit there exists abundant literature that discuss a variety of methods for off-policy evaluation. I will mainly follow works of Jiang and Li (2016) and Kallus and Uehara (2020) who give a classification of mostly widely used methods in policy evaluation.

1. Direct Method

2. Inverse Propensity Score

3. Double Robust

4. Doubly distributional robust to environments

5. Efficient Influence function-free methods(pretty new 2023)

6. β source conditions in Linear Inverse Problems

2.3 Challenges in off-policy value estimation

Even if some scholars have not noticed that off-policy policy estimation is closely related to any fundamental problem in causal inference, our work aims at launching a unified framework for the two problems. They share affinities with each other but one remarkable difference is that they differ for two independent purposes, well-known as estimating treatment effect in the former case and estimating the expected value of total outcome (or reward) in the latter one.

2.3.1 Distribution shift and concept drift

Data Shifts, distribution mismatch and domain adaptation refer to the similar problem that manifest in reinforcement learning. Subtle difference exists among them but the central idea is concluded as good performance of models only at the target domain. Data shift is the most generalized term for other two items. Data Shift happens when we apply our model or algorithm to a new dataset that differs from that we used to train our model. This type of data shift occurs when training data does not match testing data, which leads machines to learn some wrong results and our model loses interpretability. Compared to data shift, distribution shift or mismatch highlights the role of distribution that plays in policy learning. Therefore, distribution shift looms large in that it decodes the underlying probabilistic mechanism of data shift. Amongst all types of distribution shift, covariates shift is crucial to policy learning problems especially in healthcare and medicine in which features of patients are gear that drives the whole system of causal relationships.

Although the idea for distribution shift is straightforward in the real life applications, its realization and explanation is directly related to the rigorous, systematic theory existing in measure theory. We refer to it as one most acceptable theorem *Theory of Change in Measure.*

One simple example is the estimation methods of causal effect under potential outcome framework. To estimate the average treatment effect in a study, distribution shift indeed occurs because the conditional expectation over covariates X is calculated by the outside domain $x \sim P(X|T=1)$ that is only obtained in our observational data. Inverse propensity score re-weighting plays a pivotal role as described in *proposition 1.* This argument is useful in causal effect identification.

Theorem 2.1 (Change of measure). *A space of distributions that have a measure $\mathcal{V}$ in a σ-algebra*

Proposition 2.1.1. *A true propensity score $p(T=1|X)$ reweighs the imbalanced outcome distribution so that $\mathbb{E}_{p\sim p(x)}[\mathbb{E}[Y_1|X]]$ can be calculated by $\mathbb{E}_{p\sim p(x|T=1)}[\frac{p(T=1)}{p(T=1|x)}\mathbb{E}[Y|X,T=1]]$*

Propensity score weighting works well when the model of propensity score is correctly specified.

2.3.2 Density ratio estimation

To apply any of importance weighting methods discussed in the last section, it is important to get weights , ratios of $p(x)$ and $q(x)$. There are two major families of density ratio estimating methods. One is to empirically estimate the shifted distribution density and then compute the ratio as we desire. The other family is to avoid estimating any one of distributions but directly estimate the ratio through fitting a model or solving an optimization problem.

This section introduces some widely used density ratio estimation methods regarding the challenge *distribution shift(domain adaptation)* described in the last section. Our work summarizes all density ratio estimation methods as two subdomains: optimization problems and generic machine learning methods. Although this problem important in reinforcement meaning, it does not always hurt the model performance. Therefore, we will disregard special cases in which the effects of domain adaptation on policy learning can be ignored. Just for completeness, we will link the related discussion to the last part of this article and propose some non-technical suggestion.

2.3.3 Importance Weighting with Constraints

Aside basic methods that just use importance sampling and weights, other more complicated strategies such as *Policy Constraint* have been developed. By these constraints, researchers control any potentially learnt π_a to be close to the initiated behavior policy π_β .

3 Efficient Estimators of Off-policy Value

Articles to cite :

1. (Fernholz 1983) Von Mises calculus for statistical functionals
2. (Hampel 1974)The Influence Curve and its Role in Robust Estimation
3. (Fisher and Kennedy 2021) Teaching and visulization
4. (Oliver Hines and Vansteelandt 2022) Demystifying Influence Function
5. (Levy 2019) Tutorial
6. (Tsybakov 2009) Introduction to non-parametric estimation (not yet decide which part to cite)
7. (Vaart 1998) Chapter 25

3.1 Lower bound of off-policy estimators

We have examined three types of methods for estimating the ground truth behavior policy value and also discussed their advantages and shortcomings. Finding the best estimator among the space of all possible candidates is a hardcore problem and sometimes may be impossible. A benchmark for the lower bound of an estimator is Cramer-Rao bounds and the most recent work has connected this lower bound to a more general framework in semi-parametric and nonparametric estimation (Casella and Berger 2001) . Kallus and Uehara (2020) and Jiang and Li (2016) studied the attractive properties of Double Robustness and derived the lower bound of Minimax risk that can be achieved.

3.2 Parametric submodels to non-parametric estimator

Once the target estimand or the true parameter that echos our scientific question is selected. All we need is to construct an estimator (or model) for the problem-solving procedure. We shall emphasize that under some well-defined contexts, finding statistical models is to find a family of distributions. Therefore, we will abuse the term usage in the remaining texts.

As we have discussed in the last section, finding an efficient, consistent estimator for off-policy value is hard. Consequently, this difficulty tends to be exacerbated if any non-parametric estimators that are more complex and unpredictable are considered. This is because under parametric statistics, a distribution P indexed by a finite many parameters (i.e. a vector $\theta \in \mathbb{R}^d$ containing d parameters) can be efficiently estimated by a class of Maximum Likelihood Estimation models. However, this estimator would break down if the estimator does not actually come from the presumed distribution $\mathcal{P}_\theta$. In other word, it cannot be indexed by finite many parameters and turns out to be susceptible to data-adaptive algorithms.

[CITE] considers a set of possible distributions $\mathcal{P}$ by perturbing the target true estimator. It is sufficient to just use a one-dimension distribution[1] for the simple demonstration in that we just care A smart way is to perturb it in one-dimension space with magnitude t in a direction. This perturbation process can be explicitly expressed as

$$\lim_{\epsilon \downarrow 0} \frac{\hat{\mathbb{P}} - \mathbb{P}}{\epsilon} \text{ or } \lim_{\epsilon \downarrow 0} \frac{\epsilon \mathbb{P}_\epsilon + (1-\epsilon)\mathbb{P} - \mathbb{P}}{\epsilon} \tag{3.1}$$

[1] We omit high-dimensional cases because in most liturature of causal inference and off-policy reinforcement learning, the estimand of interest is just a one-dimension description. Howver, we shall emphasize that if the target estimand is a vector, the purturbation should also be adjusted to the same dimension

One can easily verify the limit of the ratio turns to be $\frac{\partial}{\partial \epsilon}\mathbb{P}_\epsilon$ using product rule of derivatives and definition of limits. If we can show that this derivative exists for any path through the true parameter $\mathbb{P}$, we can conclude that this perturbation process is reasonable. This conclusion leads to a crucial property, *pathwise differentiable*. Pulling from [cite], we know this derivative is called Gateaux Derivative.

3.3 Tangent space

Deriving an efficient influence function poses serious challenges to evaluating a target parameter. This is because in framework of parametric estimation, there exists more than one influence functions and the efficient one should be found with some added constraints

3.4 Path-wise differentiable

4 Applications in Precision Medicine and other RL fields

Articles to cite in Healthcare Application:

1. (Yu et al. 2021) read this article: Healthcare RL
2. (Raghu et al. 2017) Application to Sepsis Treatment
3. (Tseng et al. 2017) Cancer
4. (Iván Díaz and Schenck 2023)

5 Experiment and Case Study

A family of generic weighting methods mitigate the problem of distribution shift if not many extreme observations show up in our data and if the density ratio is properly estimated through any of models aforementioned. Simply plugging in density ratio accumulates estimation error even if the density ratio model is correctly specified. Weight-Clipping and other derivative methods may help with dealing with this issue but we speculate that it suffices for explanation and interpretability of models.

Driven by this, the current state of research in distribution shift embarks on finding some invariant features that render causal relationship. Scientists have been working on isolating invariant features from those dynamic features trying to derive a better representative of the stably-behaving covariates. This open research area is *Causal Representation Learning.*(Kaddour et al. 2022)

Policy cannot always be learned in a static way. We have discussed that one immediate transformation of causal inference problem is *contextual bandit*. An optimal policy is learned by a batch dataset without updating new policy for each round. This type of learning has been used in areas where systematic change of features is trivial.

6 Discussion

7 Tables and Figures

Figures and illustrations should be incorporated directly into the manuscript, and the size of a figure should be commensurate with the amount and value of the information conveyed by the figure.

Figure 1: Sample figure with preferred style for labeling parts.

Table 1: Sample Table

One	Two	Three
Eins	Zwei	Drei
Un	Deux	Trois
Jeden	Dvě	Tři

No more than three figures should generally be included in the paper. Place figures as close as possible to where they are mentioned in the text. No part of a figure should extend beyond text width, and text should not wrap around figures. Please provide permission and attribution for any trademarked or copyright images.

References

Bennett, Andrew, and Nathan Kallus. 2023. *Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes.* arXiv: 2110.15332 `[cs.LG]`.

Bennett, Andrew, Nathan Kallus, and Miruna Oprescu. 2023. *Low-Rank MDPs with Continuous Action Spaces.* arXiv: 2311.03564 `[cs.LG]`.

Casella, George, and Roger Berger. 2001. *Statistical Inference.* Duxbury Resource Center, June.

Chakraborty, Bibhas, and Erica E. M. Moodie. 2013. "Statistical Reinforcement Learning." In *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine,* 31–52. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7428-9_3.

Fernholz, Luisa Turrin. 1983. *Von Mises calculus for statistical functionals.* Lecture notes in statistics 19. New York [u.a.]: Springer.

Fisher, Aaron, and Edward H. Kennedy. 2021. "Visually Communicating and Teaching Intuition for Influence Functions." *The American Statistician* 75 (2): 162–172. https://doi.org/10.1080/00031305.2020.1717620. eprint: https://doi.org/10.1080/00031305.2020.1717620.

Hampel, Frank R. 1974. "The Influence Curve and its Role in Robust Estimation." *Journal of the American Statistical Association* 69 (346): 383–393. https://doi.org/10.1080/01621459.1974.10482962. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1974.10482962.

Iván Díaz, Katherine L. Hoffman, Nicholas Williams, and Edward J. Schenck. 2023. "Nonparametric Causal Effects Based on Longitudinal Modified Treatment Policies." *Journal of the American Statistical Association* 118 (542): 846–857. https://doi.org/10.1080/01621459.2021.1955691. eprint: https://doi.org/10.1080/01621459.2021.1955691.

Jiang, Nan, and Lihong Li. 2016. *Doubly Robust Off-policy Value Evaluation for Reinforcement Learning.* arXiv: 1511.03722 [cs.LG].

Kaddour, Jean, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. *Causal Machine Learning: A Survey and Open Problems.* arXiv: 2206.15475 [cs.LG].

Kallus, Nathan, and Masatoshi Uehara. 2019. *Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for Reinforcement Learning.* arXiv: 1906.03735 [cs.LG].

———. 2020. "Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes." *J. Mach. Learn. Res.* 21, no. 1 (January).

Kennedy, Edward H. 2023. *Semiparametric doubly robust targeted double machine learning: a review.* arXiv: 2203.06469 [stat.ME].

Lagoudakis, Michail G., and Ronald E. Parr. 2003. “Least-Squares Policy Iteration.” *J. Mach. Learn. Res.* 4:1107–1149.

Levine, Sergey, Aviral Kumar, G. Tucker, and Justin Fu. 2020. “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.” *ArXiv* abs/2005.01643.

Levy, Jonathan. 2019. *Tutorial: Deriving The Efficient Influence Curve for Large Models.* arXiv: 1903.01706 `[math.ST]`.

Oliver Hines, Karla Diaz-Ordaz, Oliver Dukes, and Stijn Vansteelandt. 2022. “Demystifying Statistical Learning Based on Efficient Influence Functions.” *The American Statistician* 76 (3): 292–304. https://doi.org/10.1080/00031305.2021.2021984. eprint: https://doi.org/10.1080/00031305.2021.2021984.

Raghu, Aniruddh, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. *Deep Reinforcement Learning for Sepsis Treatment.* arXiv: 1711.09602 `[cs.AI]`.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

Tseng, Huan-Hsin, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K. Ten Haken, and Issam El Naqa. 2017. “Deep reinforcement learning for automated radiation adaptation in lung cancer.” *Medical Physics* 44 (12): 6690–6705. https://doi.org/https://doi.org/10.1002/mp.12625. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12625.

Tsybakov, Alexandre B. 2009. “Lower bounds on the minimax risk.” In *Introduction to Nonparametric Estimation,* 77–135. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-79052-7_2.

Vaart, A. W. van der. 1998. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

Yu, Chao, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. “Reinforcement Learning in Healthcare: A Survey.” *ACM Comput. Surv.* (New York, NY, USA) 55, no. 1 (November). https://doi.org/10.1145/3477600.

A Summary of Semi-parametric Efficiency

B Proofs