

---

# Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning

---

Nathan Kallus<sup>\*1</sup> Xiaojie Mao<sup>\*2</sup> Kaiwen Wang<sup>\*1</sup> Zhengyuan Zhou<sup>\*3</sup>

## Abstract

Off-policy evaluation and learning (OPE/L) use offline observational data to make better decisions, which is crucial in applications where online experimentation is limited. However, depending entirely on logged data, OPE/L is sensitive to environment distribution shifts — discrepancies between the data-generating environment and that where policies are deployed. Si et al. (2020a) proposed distributionally robust OPE/L (DROPE/L) to address this, but the proposal relies on inverse-propensity weighting, whose estimation error and regret will deteriorate if propensities are nonparametrically estimated and whose variance is suboptimal even if not. For standard, non-robust, OPE/L, this is solved by doubly robust (DR) methods, but they do not naturally extend to the more complex DROPE/L, which involves a worst-case expectation. In this paper, we propose the first DR algorithms for DROPE/L with KL-divergence uncertainty sets. For evaluation, we propose **Localized Doubly Robust DROPE** (LDR<sup>2</sup>OPE) and show that it achieves semiparametric efficiency under weak product rates conditions. Thanks to a localization technique, LDR<sup>2</sup>OPE only requires fitting a small number of regressions, just like DR methods for standard OPE. For learning, we propose **Continuum Doubly Robust DROPL** (CDR<sup>2</sup>OPL) and show that, under a product rate condition involving a continuum of regressions, it enjoys a fast regret rate of  $\mathcal{O}(N^{-1/2})$  even when unknown propensities are nonparametrically estimated. We empirically validate our algorithms in simulations and further extend our results to general  $f$ -divergence uncertainty sets.

## 1. Introduction

The vast majority of online recommendations in search engines, e-commerce, social media, streaming platforms, etc. are made by algorithms that learn from historical user interactions (Li et al., 2010; Bottou et al., 2013; Ren & Zhou, 2020; Liu et al., 2021). Even in high-stakes domains, such as healthcare (Murphy, 2003) and education (Mandel et al., 2014), the promise of cheaper and higher quality decisions, made possible by the growing abundance of user-specific data, incentivize the inclusion of automatic decision-making components into existing approaches.

This task of making good decisions from observational data is formalized by the problems of off-policy evaluation (OPE) (Foster & Syrgkanis, 2019; Kallus & Uehara, 2020a; Chernozhukov et al., 2018; Farajtabar et al., 2018; Joachims & Swaminathan, 2016; Bottou et al., 2013; Dudík et al., 2011) and off-policy learning (OPL) (Manski, 2004; Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Zhan et al., 2021; Zhou et al., 2022; Kallus & Uehara, 2020b; Swaminathan & Joachims, 2015a; Dudík et al., 2011). OPE is concerned with estimating the expected returns of a target policy given logged data, collected under a different behavior policy. OPL is concerned with learning a policy that maximizes the expected returns given this data. OPE/L assumes that the environment in which these policies are deployed is identical to the environment that generated the training data. In practice, this often is not the case. For example, in recommendation systems, user interests naturally shift with seasonality and world events, which correspond to changes in the state and reward distributions. Moreover, the environment could also be adversarially perturbed by attackers or data corruption.

Distributional robustness is a way to guard against such unknown discrepancies between training and deployment environments. Instead of estimating/maximizing the expected policy return under the training environment, we may consider estimating/maximizing the worst-case return over all environments within an uncertainty set around the unknown training environment. Si et al. (2020a;b) tackle this distributionally robust OPE/L (DROPE/L) problem using methods based on self-normalized inverse propensity scoring (SNIPS) (Swaminathan & Joachims, 2015b). The uncertainty sets of (Si et al., 2020a) and this paper are with

---

<sup>\*</sup>Alphabetical Order. <sup>1</sup>Cornell University and Cornell Tech  
<sup>2</sup>Tsinghua University <sup>3</sup>Arena Technologies and New York University. Correspondence to: Kaiwen Wang <<https://kaiwenw.github.io>>.

respect to the KL-divergence, and generally  $f$ -divergences.

However, (Si et al., 2020a) assumes that we know the behavior propensities, which are usually absent in observational datasets. One may consider simply fitting and imputing the propensities using some flexible machine learning (ML) methods, i.e. non-parametric estimators of nuisance functions. As the propensity estimates may converge at slow rates, this leads to slow rates in estimation and learning for the proposed SNIPS-based methods. Even with known propensities, the SNIPS-based estimator’s asymptotic variance for DROPE is in fact suboptimal.

In standard (non-distributionally robust) OPE/L, doubly robust (DR) is the canonical approach for improving estimation variance and for alleviating the sensitivity to estimation of nuisances, i.e. unknown functions such as propensities. In addition to fitting a propensity model, DR also fits the expected reward given state and action and combines the two models to construct an estimator with better statistical properties. A key result in OPE is that the cross-fitted DR estimator (CFDR) is  $\sqrt{N}$ -consistent, asymptotically linear and efficient (i.e. attains the lowest possible asymptotic variance), even when nuisances are estimated at slower-than- $\sqrt{N}$ -rates (Chernozhukov et al., 2018). This, however, does not immediately extend to DROPE/L, whose objective is formed as a supremum over the log of moment generating functions. It therefore remains a question how to obtain estimation-robustness guarantees for DROPE/L.

In this paper, we propose novel doubly robust algorithms for DROPE/L, ensuring robustness to *both* environment shifts and model estimation errors. Our contributions are summarized as follows:

1. For DROPE, we propose the Localized DR DROPE (LDR<sup>2</sup>OPE) estimator and show that it is  $\sqrt{N}$ -consistent, asymptotically linear, and enjoys semiparametric efficiency under weak product rates (Section 3.1). In particular, just like DR estimators for standard OPE, LDR<sup>2</sup>OPE only requires fitting a few regressions, including a propensity and two transformed outcome regressions.
2. For DROPL, we propose Continuum DR DROPL (CDR<sup>2</sup>OPL) and prove a  $\mathcal{O}(N^{-1/2})$  regret guarantee, even when propensities are nonparametrically estimated at slow rates (Section 4).
3. We empirically show that our proposals outperform benchmarks in simulation (Section 5). Code is available at <https://github.com/CausalML/doubly-robust-dropel>.
4. We further extend our methods to general  $f$ -divergence uncertainty sets (Section 6).

## 1.1. Related Literature

We work in the distributionally robust setting (Section 2.1) proposed by Si et al. (2020a), which was motivated by the distributionally robust optimization (DRO) literature (e.g., Hu & Hong, 2013; Ben-Tal et al., 2013). Unlike Si et al. (2020a), *we do not assume that the behavior policy is known*. To derive our doubly robust DROPE estimator, we propose a novel formulation of the DRO problem as a multidimensional moment equation and leverage the techniques of Kallus et al. (2019). This allows us to tackle the complex optimization formulation of the objective and still attain semiparametric efficiency under very lax conditions.

In standard (non-distributionally robust) OPL, maximizing the CFDR objective was shown to have  $\mathcal{O}(N^{-1/2})$  regret even under slow nuisance estimator by arguing the CFDR objective concentrates uniformly over a policy class of bounded complexity (Zhou et al., 2022; Athey & Wager, 2021). However, this result is for standard policy learning, without environment shifts. As such, their uniform concentration results are with respect to the best policy in the training environment. In our setting, we are aiming to learn the best policy in the worst-case testing environment, which is a different formulation and requires a new set of techniques. In particular, we show that our objective concentrates uniformly not only over policies, but also over all adversarial environments, yielding our  $\mathcal{O}(N^{-1/2})$  distributionally robust regret guarantee. Si et al. (2020a) also proved a  $\mathcal{O}(N^{-1/2})$  distributionally robust regret bound but they crucially assumed known propensities, which allowed them to estimate the DROPL objective by reweighting via SNIPS. Also, their proof strategy is different for discrete and continuous rewards, and is specialized for SNIPS. In contrast, our proof directly decomposes the DRO objective, handling all cases in a unified way.

Si et al. (2020a) do not discuss why self-normalization (SN) was used (as opposed to IPS without SN), and only referenced Swaminathan & Joachims (2015b), which proposed SNIPS for non-distributionally-robust OPE/L. As an aside, we show in Appendix B that even though non-normalized IPS is in fact theoretically well-behaved for standard OPE under overlap conditions, the unique structure of DROPE renders IPS degenerate even under such conditions, which highlight the unique importance of SN in the DRO setting.

Mo et al. (2021); Liu et al. (2019) studied distributionally robust learning in the context of state distribution shifts (covariate shift). Kido (2022) studied distributionally robust learning in the context of known covariate shift and unknown outcome distribution shift (concept shift), under the Wasserstein distance. We highlight that these problems, and the meaning of policy value/regret therein, are different from our setting, as we study *unknown* covariate and *unknown* concept shifts, under the KL-divergence and  $f$ -divergences.

## 2. Preliminaries

We use the standard data generation process of OPE/L. Our data  $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i \in [N]}$  consists of  $N$  i.i.d. draws of  $(S, A, R)$  generated as follows. The state and potential outcomes  $(S, R(a^1), \dots, R(a^{|A|})) \in \mathcal{S} \times [0, 1]^{|A|}$  are drawn from the nominal environment  $\mathbb{P}_0$ , where  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the discrete action space, and  $R(a)$  denotes the potential reward from taking an action  $a$  (Neyman, 1923; Rubin, 1974). An *unknown* behavior policy  $\pi_0$  then samples an action  $A \sim \pi_0(S)$  given the observed state, i.e.  $A = a$  with probability  $\pi(a | s)$ . Out of the potential outcomes, only the factual outcome corresponding to the chosen action  $R = R(A)$  is observed.

For a (stochastic) policy  $\pi$ , we use  $R(\pi(S))$  to denote the random reward corresponding to the action sampled from  $\pi$ . Unless stated otherwise,  $\mathbb{E}$  and  $\mathbb{P}$  are taken over  $\mathbb{P}_0$ .

**Assumption 2.1.** We posit standard assumptions from the OPE/L literature (Si et al., 2020a):

- (i) Unconfoundedness:  $(R(a^1), \dots, R(a^{|A|})) \perp\!\!\!\perp A | S$ .
- (ii) Strong overlap:  $\eta := \inf_{s \in \mathcal{S}, a \in \mathcal{A}} \pi_0(a | s) > 0$ .

Furthermore, there exists  $\omega > 0$  such that,

- 1. If  $R(a) | S$  is continuous, its PDF  $p_R(r | s, a)$  is lower bounded:  $p_R(r | s, a) \geq \omega, \forall r \in [0, 1]$ .
- 2. If  $R(a) | S$  is discrete, its PMF  $p_R(r | s, a)$  is lower bounded:  $p_R(r | s, a) \geq \omega, \forall r \in \mathbb{D}$ , where  $\mathbb{D}$  is the set of possible rewards and  $\text{WLOG } 0 \in \mathbb{D}$ .

More generally, we may require  $R(a) | S = s$  to be mutually absolutely continuous with respect to a common measure on  $[0, 1]$  for almost all states  $s \in \mathcal{S}$ .

### 2.1. Distributionally Robust Formulation of OPE/L

We now recall the KL-distributionally robust formulation of OPE/L due to Si et al. (2020a). For an alternative environment  $\mathbb{P}_1$ , the KL-divergence is a notion of how different  $\mathbb{P}_1$  is from  $\mathbb{P}_0$  and is defined as  $D_{KL}(\mathbb{P}_1 \| \mathbb{P}_0) = \mathbb{E}_{\mathbb{P}_1} \left[ \log \left( \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right) \right]$ . Let  $\delta > 0$  denote the magnitude of distribution shifts we seek to be robust to, which we take as a fixed hyperparameter. Define the uncertainty set  $\mathcal{U}(\delta) = \{\mathbb{P}_1 : \mathbb{P}_1 \ll \mathbb{P}_0 \wedge D_{KL}(\mathbb{P}_1 \| \mathbb{P}_0) \leq \delta\}$  to be the set of perturbed environments  $\mathbb{P}_1$  which are  $\delta$ -close to the nominal distribution  $\mathbb{P}_0$ , as measured by the KL-divergence. We highlight that *both* the state and reward distributions can be perturbed. For a policy  $\pi$ , the distributionally robust value  $\mathcal{V}_\delta(\pi)$  is its worst-case performance under environment shifts with magnitude at most  $\delta$ , formalized as follows.

$$\mathcal{V}_\delta(\pi) := \inf_{\mathbb{P}_1 \in \mathcal{U}(\delta)} \mathbb{E}_{\mathbb{P}_1} [R(\pi(S))] \quad (1)$$

We remark that there are data-driven, calibration methods to choose  $\delta$ , e.g. (Mo et al., 2021).

This leads to the definitions of distributionally robust off-policy evaluation and learning (DROPE/L):

**DROPE:** For a policy  $\pi$  and radius  $\delta > 0$ , estimate the worst-case value  $\mathcal{V}_\delta(\pi)$ .

**DROPL:** For a policy class  $\Pi$  and radius  $\delta > 0$ , find a near-optimally robust policy  $\hat{\pi} \in \Pi$  with small regret in worst-case policy value

$$\mathcal{R}_\delta(\pi) := \mathcal{V}_\delta(\pi^*) - \mathcal{V}_\delta(\pi),$$

$$\text{where } \pi^* \in \arg \max_{\pi \in \Pi} \mathcal{V}_\delta(\pi).$$

While the infinite-dimensional infimum in Equation (1) seems intractable, it is in fact equivalent to a supremum over a dual variable  $\alpha$ . We now recall this strong duality result from Si et al. (2020a, Lemmas 1 and A11).

**Lemma 2.2.** Suppose Assumption 2.1. The distributionally robust value  $\mathcal{V}_\delta(\pi)$  defined in Equation (1) is equivalent to,

$$\mathcal{V}_\delta(\pi) = \max_{\alpha > 0} \phi(\pi, \alpha) := -\alpha \log W(\pi, \alpha) - \alpha \delta \quad (2)$$

$$\text{where } W(\pi, \alpha) := \mathbb{E} [\exp(-R(\pi(S))/\alpha)] \quad (3)$$

Furthermore,  $\phi(\pi, \cdot)$  is strictly concave, and is maximized at a unique  $\alpha^*(\pi) \in (0, \bar{\alpha}]$ , where  $\bar{\alpha} := 1/\delta$ .

In particular, for any policy, we know that  $\alpha^*(\pi) > 0$ . For our DROPL analysis, we need this lower bound to hold uniformly over  $\Pi$ , stated in the following assumption.

**Assumption 2.3.**  $\underline{\alpha} := \inf_{\pi \in \Pi} \alpha^*(\pi) > 0$ .

We also denote  $\underline{W} := \omega \frac{\min(\underline{\alpha}, 1)}{2}$  if  $R(a) | S$  is continuous, and  $\underline{W} := \omega$  if discrete. Lemma A.2 shows that  $W(\pi, \alpha) \geq \underline{W}$  for any  $\pi \in \Pi$  and any  $\alpha \geq \underline{\alpha}$ .

Si et al. (2020a) assume that the behavior policy  $\pi_0$  is known, and propose to estimate  $W(\pi, \alpha)$  with SNIPS, based on normalizing the propensity ratios  $w_i = \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)}$ :

$$\widehat{W}^{SNIPS}(\pi, \alpha) = \sum_{i=1}^N \frac{w_i}{\sum_j w_j} \exp(-r_i/\alpha).$$

Plugging  $\widehat{W}^{SNIPS}(\pi, \alpha)$  into Equation (2) gives an estimator for the robust policy value. Assuming that the behavior policy  $\pi_0$  is known, Si et al. (2020a) show that the resulting estimator is a  $\sqrt{N}$ -consistent for DROPE, and the resulting DROPL algorithm can achieve  $\mathcal{O}(N^{-1/2})$  regret guarantee. However, in practice, the behavior policy is often unknown and needs to be estimated (often at slow rates). Moreover, inverse-propensity scoring and its self-normalized variant cannot achieve semiparametric efficiency. This motivates us to consider improved doubly robust methods.

### 3. Doubly Robust DROPE

To estimate the robust policy value  $\mathcal{V}_\delta(\pi)$  in a doubly robust way, it is natural to first consider estimating  $W(\pi, \alpha)$  in Equation (3) with a doubly robust estimator. This however requires estimating a continuum of regression functions  $\{f_0(\cdot, \cdot; \alpha) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R} : 0 < \alpha \leq \bar{\alpha}\}$ , where

$$f_0(s, a; \alpha) := \mathbb{E}[\exp(-R/\alpha) \mid S = s, A = a], \quad (4)$$

is parameterized by the dual variable  $\alpha$ . This means that we would need to fit a large number or even infinitely many regressions functions. This is in stark contrast to standard OPE where doubly robust estimation requires fitting only a single regression function  $\mathbb{E}[R \mid S = s, A = a]$ .

To overcome the challenge of fitting a continuum of regressions, we propose to leverage the Localized Debiased Machine Learning (LDML) framework recently developed for causal inference (Kallus et al., 2019). To do so, we cast the estimation of  $\alpha^*(\pi)$  and  $\mathcal{V}_\delta(\pi)$  into a joint moment estimation problem. We then develop a localized doubly robust algorithm that only fits two regressions at an initial estimate of  $\alpha^*(\pi)$ , instead of infinitely many regressions.

#### 3.1. The Localization Approach

First, since that  $\phi(\pi, \cdot)$  is strictly concave (Lemma A.1), observe that  $\alpha^* := \alpha^*(\pi)$  is the unique root to  $\frac{\partial}{\partial \alpha} \phi(\pi, \alpha) = 0$ , and satisfies

$$-\log W_0(\pi, \alpha^*) - \frac{W_1(\pi, \alpha^*)}{\alpha^* W_0(\pi, \alpha^*)} - \delta = 0 \quad (5)$$

where  $W_j(\pi, \alpha) := \mathbb{E}[R(\pi(S))^j \exp(-R(\pi(S))/\alpha)]$ .

Moreover, we know from Equation (2) that

$$\mathcal{V}_\delta(\pi) = -\alpha^* \log W_0^* - \alpha^* \delta, \quad (6)$$

where we use the shorthand  $W_j^* = W_j(\pi, \alpha^*)$ . Therefore, estimating  $\alpha^*(\pi)$  and  $\mathcal{V}_\delta(\pi)$  in Equations (5) and (6) is equivalent to estimating the root of the following moment equation with parameter  $\theta = [\alpha, W_0, W_1, \mathcal{V}_\delta]^\top$ :

$$\mathbb{E}[U(R(\pi(S)); \alpha) + V(\theta)] = \mathbf{0} \quad (7)$$

$$U(r; \alpha) = \begin{bmatrix} \exp(-r/\alpha) \\ r \exp(-r/\alpha) \\ 0 \\ 0 \end{bmatrix}, V(\theta) = \begin{bmatrix} -W_0 \\ -W_1 \\ -\delta - \log W_0 - \frac{W_1}{\alpha W_0} \\ -\mathcal{V}_\delta - \alpha \log W_0 - \alpha \delta \end{bmatrix}.$$

Since we don't observe the counterfactual  $R(\pi(S))$ , Equation (7) is infeasible for estimation. Instead, we derive the following doubly robust moment equation in terms of the observed variables, with nuisances  $\eta_1, \eta_2$  to be estimated:

$$\mathbb{E}[\psi(Z; \theta, \eta_1^*(Z; \alpha), \eta_2^*(Z; \alpha))] = \mathbf{0} \quad (8)$$

$$\begin{aligned} \psi(z; \theta, \eta_1(z; \alpha), \eta_2(z; \alpha)) &= \frac{\pi(a \mid s)}{\eta_2(s, a)} (U(r; \alpha) - \eta_1(s, a; \alpha)) \\ &+ \mathbb{E}_{a \sim \pi(s)} [\eta_1(s, a; \alpha)] + V(\theta), \end{aligned}$$

#### Algorithm 1 Localized Doubly Robust DROPE

- 1: **Input:** Data  $\mathcal{D}$ , policy  $\pi$ , uncertainty set radius  $\delta$ .
- 2: Randomly split  $\mathcal{D}$  into  $K$  (approximately) even folds, with the indices of the  $k^{\text{th}}$  fold denoted as  $\mathcal{I}_k$ .
- 3: **for**  $k = 1, \dots, K$  **do**
- 4:   Using  $\mathcal{D}[\mathcal{I}_k^C]$ , train  $\hat{\pi}_0^{(k)}$  to fit  $\pi_0$ .
- 5:   Randomly split  $\mathcal{I}_k^C$  into two halves  $\mathcal{J}_1, \mathcal{J}_2$ .
- 6:    $\hat{\alpha}_{init}^{(k)} \leftarrow \text{InitialEstimate}(\mathcal{D}[\mathcal{J}_1], \delta, \pi)$ .
- 7:   Using  $\mathcal{D}[\mathcal{J}_2]$ , train  $\hat{f}_j^{(k)}$  to fit  $f_j(\cdot; \hat{\alpha}_{init}^{(k)})$ ,  $j = 0, 1$ .
- 8: **end for**
- 9: Find  $\hat{\alpha} > 0$  that solves the estimated moment equation:

$$-\log(\widehat{W}_0(\alpha)) - \frac{\widehat{W}_1(\alpha)}{\alpha \cdot \widehat{W}_0(\alpha)} - \delta = 0 \quad \text{where,}$$

$$\widehat{W}_j(\alpha) := \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \widehat{W}_j^{(i,k)}(\alpha)$$

$$\begin{aligned} \widehat{W}_j^{(i,k)}(\alpha) &:= \sum_{a \in \mathcal{A}} \pi(a \mid s_i) \hat{f}_j^{(k)}(s_i, a) \\ &+ \frac{\pi(a_i \mid s_i)}{\hat{\pi}_0^{(k)}(a_i \mid s_i)} \left( r_i^j \exp(-r_i/\alpha) - \hat{f}_j^{(k)}(s_i, a_i) \right). \end{aligned}$$

10: Calculate  $\widehat{\mathcal{V}}_\delta \leftarrow -\hat{\alpha} \log \widehat{W}_0(\hat{\alpha}) - \hat{\alpha} \delta$ .

11: **Return:**  $\hat{\theta}^{\text{LDR}^2 \text{OPE}} = (\hat{\alpha}, \widehat{W}_0(\hat{\alpha}), \widehat{W}_1(\hat{\alpha}), \widehat{\mathcal{V}}_\delta)$ .

where  $\eta_2^*(z) = \pi_0(a \mid s)$  is the behavior propensity and

$$\begin{aligned} \eta_1^*(s, a; \alpha) &= \mathbb{E}[U(R; \alpha) \mid S = s, A = a] \\ &= [f_0(s, a; \alpha), f_1(s, a; \alpha), 0, 0]^\top, \\ f_j(s, a; \alpha) &:= \mathbb{E}[R^j \exp(-R/\alpha) \mid S = s, A = a]. \end{aligned}$$

Importantly, Equation (8) involves not only the regression function  $f_0$  in Equation (4), but also an additional regression function  $f_1$ . With this new regression function, the Gâteaux derivatives of  $\mathbb{E}[\psi(Z; \theta, \eta_1(Z; \alpha), \eta_2(Z; \alpha))]$  with respect to the functions  $(\eta_1, \eta_2)$  are zero when evaluated at  $\theta^* = (\alpha^*, W_0^*, W_1^*, \mathcal{V}_\delta(\pi))$ ,  $\eta_1(\cdot; \alpha) = \eta_1^*(\cdot; \alpha^*)$ , and  $\eta_2 = \eta_2^*$ . This property is called *Neyman Orthogonality* (Chernozhukov et al., 2018), which implies that the doubly robust moment estimation is insensitive to errors of estimating  $\eta_1^*, \eta_2^*$ . Therefore, if an initial guess  $\hat{\alpha}_{init}$  is close enough to  $\alpha^*$ , it suffices to only fit  $\eta_1^*(\cdot; \alpha)$  localized at  $\alpha = \hat{\alpha}_{init}$ , rather than the whole continuum of regressions.

We propose **Localized Doubly Robust DROPE** (LDR<sup>2</sup>OPE) in Algorithm 1. Following LDML (Kallus et al., 2019), we employ a two-level cross-fitting scheme to accommodate flexible (non-parametric) ML estimators while preserving strong theoretical guarantees. For each data fold  $k \in [K]$ , we use the out-of-fold (OOF) data to fit

the estimator  $\widehat{\pi}_0^{(k)}$  for  $\pi_0$ , and half of OOF data to fit the estimator  $\widehat{f}_j^{(k)}$ , localized at an estimate  $\widehat{\alpha}_{init}^{(k)}$  for  $\alpha^*$  based on the other half of OOF data. These estimators trained on OOF data are then evaluated at data in each corresponding fold, forming the estimated doubly robust moment equation in Line 9 of the algorithm. The final moment equation can be solved with 1D Newton-Raphson with projection to  $\mathbb{R}^+$  (see Appendix C.3). A reasonable candidate for InitialEstimate is the cross-fitted SNIPS estimator. Thus, Algorithm 1 only requires fitting propensities and two regression functions; all three regressions are amenable to flexible, black-box ML tools.

### 3.2. Asymptotic Theory

Define the estimation rates  $\rho_f, \rho_{\pi_0}, \rho_\alpha$  as random quantities corresponding to the  $L_2$  loss as follows:

$$\begin{aligned} \max_{j=0,1} \left\| \widehat{f}_j^{(k)} - f_j(\cdot; \widehat{\alpha}_{init}^{(k)}) \right\|_{L_2(\mathbb{P}_0)} &\leq \rho_f(N), \\ \left\| \widehat{\pi}_0^{(k)} - \pi_0 \right\|_{L_2(\mathbb{P}_0)} &\leq \rho_{\pi_0}(N), \quad \left| \widehat{\alpha}_{init}^{(k)} - \alpha^* \right| \leq \rho_\alpha(N). \end{aligned}$$

**Assumption 3.1** (Product Rates for LDR<sup>2</sup>OPE). We assume that  $\rho_{\pi_0}(N) \cdot (\rho_f(N) + \rho_\alpha(N)) = o_p(N^{-1/2})$ .

We now state our main result for DROPE: the asymptotic behavior and optimality of LDR<sup>2</sup>OPE. Specifically, we show that LDR<sup>2</sup>OPE converges at a  $\mathcal{O}_p(N^{-1/2})$  rate, i.e.  $\sqrt{N}$ -consistency, and is asymptotically linear. Furthermore, LDR<sup>2</sup>OPE achieves semi-parametric efficiency, as its asymptotic variance is the smallest possible variance amongst regular estimators – equivalently, LDR<sup>2</sup>OPE is locally minimax optimal in mean-squared error amongst *all* estimators. In essence, this shows that our estimator is asymptotically optimal and amenable to uncertainty quantification with confidence intervals.

**Theorem 3.2.** Suppose Assumptions 2.1 and 3.1. Let  $\theta^* = [\alpha^*, W_0^*, W_1^*, \mathcal{V}_\delta^*]^\top$  be the solution to Equation (7). Then,

$$\sqrt{N}(\widehat{\theta}^{\text{LDR}^2\text{OPE}} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^N J^{*-1} \psi^*(Z_i) + o_p(1)$$

where  $Z_i = (s_i, a_i, r_i)$ ,  $\psi$  is defined in Equation (8),  $\psi^*(Z) := \psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))$ ,

$$J^* = \begin{bmatrix} \frac{W_1^*}{(\alpha^*)^2} & -1 & 0 & 0 \\ \frac{W_2^*}{(\alpha^*)^2} & 0 & -1 & 0 \\ \frac{W_1^*}{(\alpha^*)^2 W_0^*} & -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* (W_0^*)^2} & -\frac{1}{\alpha^* W_0^*} & 0 \\ -\log W_0^* - \delta & -\frac{\alpha^*}{W_0^*} & 0 & -1 \end{bmatrix}$$

and  $\Sigma = \mathbb{E} [J^{*-1} \psi^*(Z) \psi^*(Z)^\top J^{*-1}]$  is the optimal covariance. Hence,  $\sqrt{N}(\widehat{\theta}^{\text{LDR}^2\text{OPE}} - \theta^*) \rightsquigarrow \mathcal{N}(0, \Sigma)$  and  $\widehat{\theta}^{\text{LDR}^2\text{OPE}}$  achieves the semiparametric efficiency lower bound for  $\theta^*$ .

Please see Appendix C.2 for the proof. Assumption 3.1 is the product rate condition, which has the desired multiplicative structure that allows trading off estimation rates between nuisances. If InitialEstimate is cross-fitted SNIPS, then Proposition 2 of Kallus et al. (2019) implies that  $\rho_\alpha(N) = \mathcal{O}_p(\rho_{\pi_0}(N))$ . In this case, it suffices that  $\rho_{\pi_0}(N) = o_p(N^{-1/4})$  and  $\rho_f(N) = \mathcal{O}_p(N^{-1/4})$ . We can also run LDR<sup>2</sup>OPE again where InitialEstimate is outputted  $\widehat{\alpha}$  from the last LDR<sup>2</sup>OPE run. Recursing  $M$  times, the product rate becomes  $\rho_{\pi_0}(N) (\rho_f(N) + \rho_{\pi_0}(N) (\rho_f(N) + \rho_{\pi_0}(N) (\dots))) = \mathcal{O}(\rho_{\pi_0}(N) \rho_f(N) + \rho_{\pi_0}(N)^M \rho_\alpha(N))$ . By iteratively refining localizations, we become more robust to a slower initial localization  $\rho_\alpha(N)$ , at the cost of more computation.

Theorem 3.2 is significant even when behavior propensities are known, as LDR<sup>2</sup>OPE improves over SNIPS in that LDR<sup>2</sup>OPE is efficient and has a smaller asymptotic variance. As remarked by Kallus et al. (2019); Kasy (2019), this theorem also holds uniformly over a family of nominal distributions  $\mathbb{P}_0$  under some regularity conditions, which implies a stronger finite-sample performance guarantee.

Finally, we note that while cross-fitting does require training regression models  $K$  times, in practice this does not pose a computational burden, as  $K = 2$  is sufficient for theory and in practice  $K = 5$  is a reasonable choice. Furthermore, each cross-fitting run is identical, just running on different splits of the data, so they can be done in parallel. For a complete run-time analysis, please see Appendix C.5.

## 4. Doubly Robust DROPL

We now turn to distributionally robust off-policy learning, where we aim to find a policy with high distributionally robust value. Ostensibly, DROPL involves DROPE for many policies, since to find a policy with high value, we need to be able to evaluate, or at least compare, different policies. Directly applying the localization technique from LDR<sup>2</sup>OPE does not help since an initial guess  $\widehat{\alpha}_{init}(\pi_1)$  for one policy may be far from  $\alpha^*(\pi_2)$  of another policy. Thus, estimating a continuum of regression functions appears inevitable for the more challenging DROPL task. This motivates us to directly apply doubly robust estimation to  $W(\pi, \alpha)$ , which requires estimating the continuum of regression functions  $\{f_0(\cdot, \cdot; \alpha) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R} : 0 < \alpha \leq \bar{\alpha}\}$ .

### 4.1. Estimating a Continuum of Regression Functions

We propose to estimate the continuum of regression functions  $f_0(\cdot, \cdot; \alpha)$  via a local weighting approach. Given  $N$  data points, we first learn data-driven weighting functions  $\{\widehat{\omega}_i(s, a)\}_{i \in [N]}$  such that the conditional reward distribution  $R \mid S = s, A = a$  can be approximated by  $\sum_{i=1}^N \widehat{\omega}_i(s, a) \delta_{r_i}$ , where  $\delta_{r_i}$  is the Dirac measure at  $r_i$ .

Here,  $\widehat{\omega}_i(s, a)$  roughly measures the proximity of the  $i^{\text{th}}$  datapoint to the query point  $(s, a)$ , so it is typically larger when  $(s_i, a_i)$  is closer to  $(s, a)$ . Common weight construction methods include k-nearest neighbors, kernel regressions, decision trees and various tree ensembles (Bertsimas & Kallus, 2020; Ćevic et al., 2020; Khosravi et al., 2022; Oprescu et al., 2019; Meinshausen & Ridgeway, 2006; Athey et al., 2019). With these weights, we can approximate  $f_0(s, a; \alpha)$  for any  $\alpha$  with the following continuum estimator:

$$\widehat{f}_0(s, a; \alpha) = \sum_{i=1}^N \widehat{\omega}_i(s, a) \exp(-r_i/\alpha). \quad (9)$$

In our experiments, we constructed the weights using random forests (Breiman, 2001): we first run random forest to regress  $R$  with respect to  $(S, A)$ , and then compute  $\widehat{\omega}_i(s, a)$  as the average frequency that data point  $(s_i, a_i)$  and query point  $(s, a)$  lie in the same tree leaf node. This method has been successfully applied in statistical estimation and decision making (e.g., Bertsimas & Kallus, 2020; Meinshausen & Ridgeway, 2006; Kallus & Mao, 2022).

## 4.2. Learning Algorithm

In Algorithm 2, we propose **Continuum Doubly Robust DROPL** (CDR<sup>2</sup>OPL), which targets the policy  $\widehat{\pi}^{DR}$  that maximizes the doubly robust objective. It does so by jointly optimizing the dual variable  $\alpha$  and policy (e.g., by policy gradient updates) in an alternating fashion. We fit the continuum of regressions in Line 5.

$$\widehat{\pi}^{DR} \in \arg \max_{\pi \in \Pi} \widehat{\mathcal{V}}_\delta^{DR}(\pi) \quad (10)$$

$$\begin{aligned} \widehat{\mathcal{V}}_\delta^{DR}(\pi) &:= \max_{\alpha > 0} -\alpha \log \widehat{W}^{DR}(\pi, \alpha) - \alpha \delta \\ \widehat{W}^{DR}(\pi, \alpha) &:= \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} \left( \exp(-r_i/\alpha) \right. \\ &\quad \left. - \widehat{f}_0^{(k)}(s_i, a_i; \alpha) \right) + \sum_{a \in \mathcal{A}} \pi(a | s_i) \widehat{f}_0^{(k)}(s_i, a; \alpha). \end{aligned}$$

## 4.3. Regret Bounds

We now derive a finite-sample distributionally robust regret guarantee for  $\widehat{\pi}^{DR}$  (from Equation (10)). We adopt the Hamming entropy integral  $\kappa(\Pi)$  from Si et al. (2020a) as a complexity measure for the policy class  $\Pi$ . Recall the Hamming distance between two policies is the fraction of mismatched action distributions in the dataset,

$$d_H(\pi_1, \pi_2) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\pi_1(s_i) \neq \pi_2(s_i)].$$

Then, the Hamming covering number  $\mathcal{C}(\epsilon, \Pi; \{s_i\}_{i \in [N]})$  is the cardinality of the smallest set of policies  $\widetilde{\Pi}$  such that

### Algorithm 2 Continuum Doubly Robust DROPL

- 1: **Input:** Data  $\mathcal{D}$ , policy class  $\Pi$ , uncertainty set radius  $\delta$ .
- 2: Randomly split  $\mathcal{D}$  into  $K$  (approximately) even folds, with the indices of the  $k^{\text{th}}$  fold denoted as  $\mathcal{I}_k$ .
- 3: **for**  $k = 1, \dots, K$  **do**
- 4:   Using  $\mathcal{D}[\mathcal{I}_k^C]$ , train  $\widehat{\pi}_0^{(k)}$  to fit  $\pi_0$ .
- 5:   Using  $\mathcal{D}[\mathcal{I}_k^C]$ , train  $\widehat{f}_0^{(k)}(\cdot; \alpha)$  to fit  $f_0(\cdot; \alpha)$  for all  $\alpha \in (0, \bar{\alpha})$ , e.g. using Section 4.1.
- 6: **end for**
- 7: Initialize  $\widehat{\pi}$ .
- 8: **while**  $\widehat{\pi}$  has not converged **do**
- 9:   Set  $\widehat{\alpha} \leftarrow \arg \max_{\alpha > 0} -\alpha \log \widehat{W}^{DR}(\widehat{\pi}, \alpha) - \alpha \delta$ .
- 10:   Update the policy  $\widehat{\pi}$  (e.g., take some gradient steps) to minimize  $\widehat{W}^{DR}(\pi, \widehat{\alpha})$ .
- 11: **end while**
- 12: **Return:**  $\widehat{\pi}$ .

for any  $\pi \in \Pi$ , there exists  $\widetilde{\pi} \in \widetilde{\Pi}$  with  $d_H(\pi, \widetilde{\pi}) \leq \epsilon$ . Denote the largest size over all datasets as  $\mathcal{N}(\epsilon, \Pi) := \sup_{N \geq 1} \sup_{\{s_i\}_{i \in [N]}} \left| \mathcal{C}(\epsilon, \Pi; \{s_i\}_{i \in [N]}) \right|$ .

**Definition 4.1.** The Hamming entropy integral of  $\Pi$  is

$$\kappa(\Pi) := \int_0^1 \log^{1/2} \mathcal{N}(t^2, \Pi) dt.$$

For example, if  $\Pi$  is finite, we have  $\kappa(\Pi) \leq \log^{1/2}(|\Pi|)$ .

Since CDR<sup>2</sup>OPL fits a continuum of regressions, our guarantee involves the uniform estimation rate over the continuum.

**Definition 4.2.** Suppose  $\{\widehat{f}_0^{(k)}(\cdot, \alpha), \alpha \in [\underline{\alpha}, \bar{\alpha}]\}$  is learned from a dataset of  $\frac{N(K-1)}{K}$  points. For any  $\beta \in (0, 1)$ , define  $\text{Rate}_f^\epsilon(N, \beta)$  so that w.p. at least  $1 - \beta$ , it upper bounds  $\left\| \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left| \widehat{f}_0^{(k)}(S, A; \alpha) - f_0(S, A; \alpha) \right| \right\|_{L_2(\mathbb{P}_0)}$ .

Similarly, let  $\text{Rate}_{\pi_0}(N, \beta)$  be the estimation rate for  $\widehat{\pi}_0^{(k)}$ . Unlike the rates we used for the asymptotic theory of LDR<sup>2</sup>OPE, these rates are deterministic functions of  $N$  and  $\beta$ , which is needed for our finite-sample guarantee. We now state our main guarantee for DROPL.

**Theorem 4.3.** Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1/6)$ , w.p. at least  $1 - 6\beta$ , the distributionally robust regret  $\mathcal{R}_\delta(\widehat{\pi}^{DR})$  is at most

$$\begin{aligned} &\frac{2112\bar{\alpha}\sqrt{K}}{W\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} + \log^{1/2}(K/\beta) \right) \\ &+ \frac{4\bar{\alpha}}{W\eta^2} (\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^\epsilon(N, \beta/K)), \end{aligned}$$

provided  $N$  is sufficiently large (Assumption D.1).

Please see Appendix D for the proof. This theorem shows that  $\hat{\pi}^{DR}$  achieves distributionally robust regret with a  $\mathcal{O}(N^{-1/2})$  term, plus a product rates term  $\mathcal{O}(\text{Rate}_{\pi_0}(N, \beta) \cdot \text{Rate}_f^c(N, \beta))$ . We highlight that our dependence on nuisance estimation is manifested as the product of rates, which allows for non-parametric (sub- $\sqrt{N}$ ) rates for each nuisance. For example, if the estimation rates  $\text{Rate}_{\pi_0}(N, \beta)$  and  $\text{Rate}_f^c(N, \beta)$  are both  $o(N^{-1/4})$ , the contribution of this product term is lower order, and the regret is  $\mathcal{O}(N^{-1/2})$ . If the estimated propensities  $\hat{\pi}_0^{(k)}$  are obtained by empirical risk minimization methods, then the results in Wainwright (2019); Bartlett et al. (2005) can be used to show that  $\text{Rate}_{\pi_0}(N, \beta) \leq C(\frac{1}{N^p} + \sqrt{\log(1/\beta)/N})$  where the rate  $p$  depends on the complexity of the function class, such as given by its metric entropy. The rate of convergence for the continuum nuisance  $\text{Rate}_f^c(N, \beta)$  can be argued based on analysis of Bertsimas & Kallus (2020); Belloni et al. (2017). In the proof, we use Assumption 2.3 to show that  $\alpha \mapsto \exp(-r/\alpha)$  (and the expectation variants) is Lipschitz, with Lipschitz constant at most  $1/\underline{\alpha}^2$  (Lemma D.2). This implies that the continuum estimator proposed in Section 4.1, as a convex combination of Lipschitz functions, is also Lipschitz. We remark that point-wise rates provided in Ćevic et al. (2020); Oprescu et al. (2019); Athey et al. (2019); Györfi et al. (2002) can then be translated into uniform rates, thanks to this Lipschitz property (see Lemma D.10).

## 5. Experiments

We evaluated our doubly robust algorithms for DROPE/L in a simulated setting where distributional shifts can be easily visualized. The following is our data generating process  $\mathbb{P}_0$ . The state space is two-dimensional  $\mathcal{S} = [-1, 1]^2$ , and states are sampled uniformly  $S \sim \text{Unif}([-1, 1]^2)$ . The action space is  $\mathcal{A} = \{0, 1, \dots, 4\}$ , and the behavior policy is a softmax policy  $\pi_0(a | s) \propto \exp(2s^\top \beta_a)$ , where  $\beta_a$ 's are the coordinates of the  $k$ -th fifth root of unity, i.e.  $\beta_a = (\text{Re } \zeta_a, \text{Im } \zeta_a)$  where  $\zeta_a = \exp(2\pi i/5)$ . Potential outcomes are normally distributed:  $R(a) | S = s \sim \mathcal{N}(s^\top \beta_a, \sigma_a^2)$ , where  $\sigma = [0.1, 0.2, 0.3, 0.4, 0.5]^\top$ . This setup is visualized in Figure 1. We see that the optimal policy for OPL partitions the state space in equal angles, based on which root of unity the given state is closest to, while the optimal policy for DROPL favors actions with lower variances in the reward. This connection of KL-DRO to variance regularization has been studied in the DRO literature (Lam, 2016; Duchi & Namkoong, 2019).

First, we compared our DROPE proposal, LDR<sup>2</sup>OPE (Algorithm 1), to the SNIPS-based evaluation baseline (Si et al., 2020a, Algorithm 1). The target policy we seek to evaluate is  $\pi_{\text{target}}(a | s) \propto \exp(s^\top \beta_a)$ , which is like the behavior policy  $\pi_0$  but with a different softmax temperature. We

conducted experiments under three uncertainty set radii  $\delta = 0.1, 0.2, 0.3$ , and in two settings, where propensities  $\pi_0$  were known and unknown. If propensities were known, both LDR<sup>2</sup>OPE and SNIPS used ground truth propensities  $\pi_0$ . If propensities were unknown, both methods used estimated propensities obtained from Gradient Boosted Trees using the LightGBM package (Ke et al., 2017). We also used LightGBM for regressing LDR<sup>2</sup>OPE's outcome functions  $\hat{f}_j^{(k)}$  for  $j = 0, 1$ . We self-normalized the propensity weights for our proposed doubly robust methods as we found it beneficial in the small  $N$  regime. All models were fitted with  $K = 5$  fold cross-fitting, and we repeated this over 30 seeds. Shaded regions in plots are 90% confidence intervals computed with the bootstrap in Seaborn (Waskom, 2021).

Figure 2 shows the results of the DROPE experiments. In all experimental setups, as long as  $N$  is large enough, we observe that LDR<sup>2</sup>OPE outperforms SNIPS. Importantly, LDR<sup>2</sup>OPE has a faster rate of MSE decrease. However, in the setting when  $N$  is small (non-asymptotic regime), propensities are unknown, and  $\delta$  is large, LDR<sup>2</sup>OPE may be less stable than SNIPS; that is, doubly robust appears to suffer when all three challenges arise, but as long as one challenge is mitigated, doubly robust offers a significant improvement over baseline. While performance of both methods deteriorated, as expected, when  $\pi_0$  was not known had to be estimated, we see that whenever  $N \geq 10^4$ , LDR<sup>2</sup>OPE with estimated propensities is actually competitive with the algorithms with access to the ground truth  $\pi_0$  *a priori*. This empirically reinforces our theory that LDR<sup>2</sup>OPE is asymptotically optimal, even when propensities are estimated with flexible, non-parametric ML methods. Overall, except in the setting with small  $N$ , unknown propensities and large  $\delta$ , LDR<sup>2</sup>OPE offers a significant benefit over SNIPS.

Next, we compared our DROPL proposal, CDR<sup>2</sup>OPL (Algorithm 2), to maximizing the SNIPS objective (Si et al., 2020a, Algorithm 2). In CDR<sup>2</sup>OPL, the continuum of regression functions  $\{\hat{f}_0(s, a; \alpha)\}$  was estimated according to Section 4.1, with weights  $\hat{\omega}_i(s, a)$  derived from fitting a Random Forest with 25 trees. Our policies were neural network softmax policies with a hidden layer of 32 neurons and ReLU activation. For Line 10, we minimized  $\widehat{W}^{DR}(\cdot, \alpha)$  using Adam with a learning rate of 0.01. Following Dudík et al. (2011), we repeated each policy update ten times with perturbed starting weights and picked the best weights based on training objective, since the doubly robust estimate  $\widehat{W}^{DR}(\cdot, \alpha)$  is non-convex in the policy weights.

Figure 3 shows the results of the DROPL experiments. When  $\delta = 0.1$ , CDR<sup>2</sup>OPL consistently learns policies that improve over the baseline distributionally robust value by about 1%, but this benefit from double robustness becomes less significant as  $\delta$  grows, a trend we also saw in the DROPE experiments. Here, this decrease in improvement

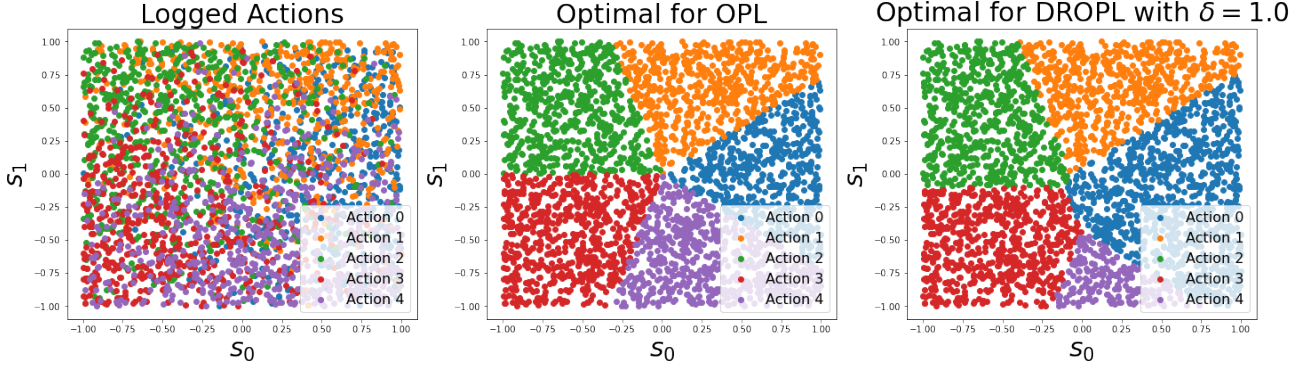


Figure 1. Scatter plots of (Left) the behavior policy, (Center) the optimal policy for expected reward, and (Right) the optimal policy for distributionally robust value with  $\delta = 1.0$ . Notice that in the right-most plot, the distributionally robust policy prefers Action 0 more, since its conditional reward has the lowest variance, and so choosing it is more robust.

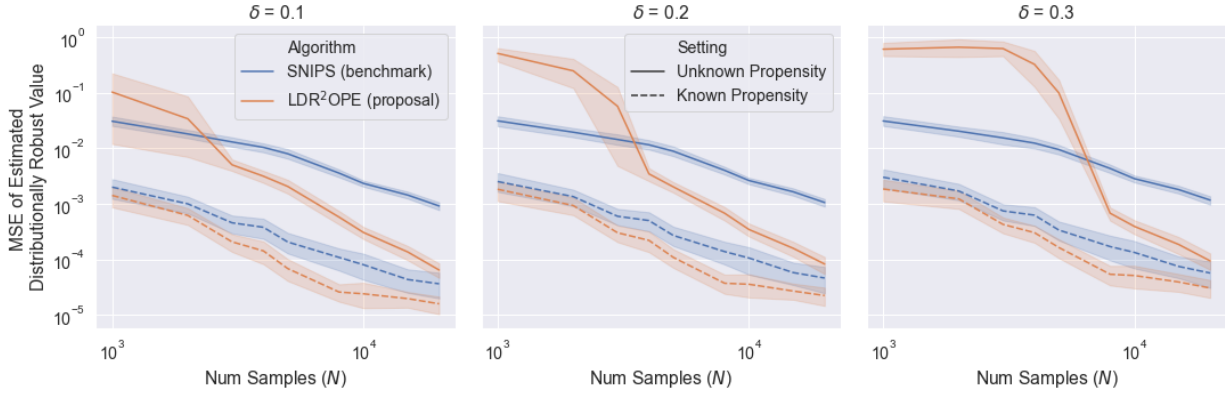


Figure 2. Comparison of our proposal LDR<sup>2</sup>OPE to the baseline SNIPS in the DROPE task, repeated for  $\delta = 0.1, 0.2, 0.3$ . Solid and dashed lines denote settings when behavior propensities are unknown and known, respectively. The x-axis is the number of samples  $N$  used by the evaluation algorithm, and the y-axis is the mean squared error (MSE) of the DROPE estimator, so lower is better. When  $N$  is large enough, we see that LDR<sup>2</sup>OPE has lower MSE than SNIPS in all cases.

may be due to the fact that as  $\delta$  increases to infinity, the distributionally robust value of all policies converge to the minimum reward, and so the policy improvement becomes less noticeable. We also highlight that, while CDR<sup>2</sup>OPL offers a performance improvement at least for smaller  $\delta$ , it comes at a computational cost. This is because each call to the estimator  $\hat{f}_0(s, a; \alpha)$  requires a weighted sum over the training dataset, rendering the overall running time for CDR<sup>2</sup>OPL to be  $\mathcal{O}(N^2)$ , while it is  $\mathcal{O}(N)$  for SNIPS. Further, the necessity of restarting policy optimization at many random starting weights to combat non-convexity of  $\hat{W}^{DR}(\cdot, \alpha)$  also increases computational cost by a constant factor. Given the computational and optimization challenges of CDR<sup>2</sup>OPL, resulting in the potentially marginal improvement for large  $\delta$ , this investigation reveals that (cross-fitted) SNIPS still remains an attractive choice in many practical situations. Our recommendation is to try both learning algorithms, and then select the better one by evaluating with

LDR<sup>2</sup>OPE. Finding a more computationally efficient and stable algorithm for DROPL with unknown propensities is an interesting direction for future work.

## 6. Extension to $f$ -divergences

Now, we generalize our results to uncertainty sets generated by the  $f$ -divergence  $D_f$ . Recall that for any convex function  $f : \mathbb{R}^+ \rightarrow \mathbb{R}$  satisfying  $f(1) = 0$ , the  $f$ -divergence is defined as  $D_f(P \parallel Q) := \mathbb{E}_Q[f(dP/dQ)]$  (Sason & Verdú, 2016), and recall that  $f^*(z) := \sup_{x \geq 0} \langle z, x \rangle - f(x)$  is the Fenchel conjugate of  $f$ . Strong duality gives a variational form for the distributionally robust value, now with a second dual variable  $\lambda$  (Namkoong & Duchi, 2016),

$$\mathcal{V}_\delta^f(\pi) = \sup_{\alpha \geq 0, \lambda \in \mathbb{R}} \phi^f(\pi, \alpha, \lambda) \quad (11)$$

$$\phi^f(\pi, \alpha, \lambda) = -\alpha \mathbb{E}_{\mathbb{P}_0} \left[ f^* \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \right] - \alpha \delta - \lambda.$$

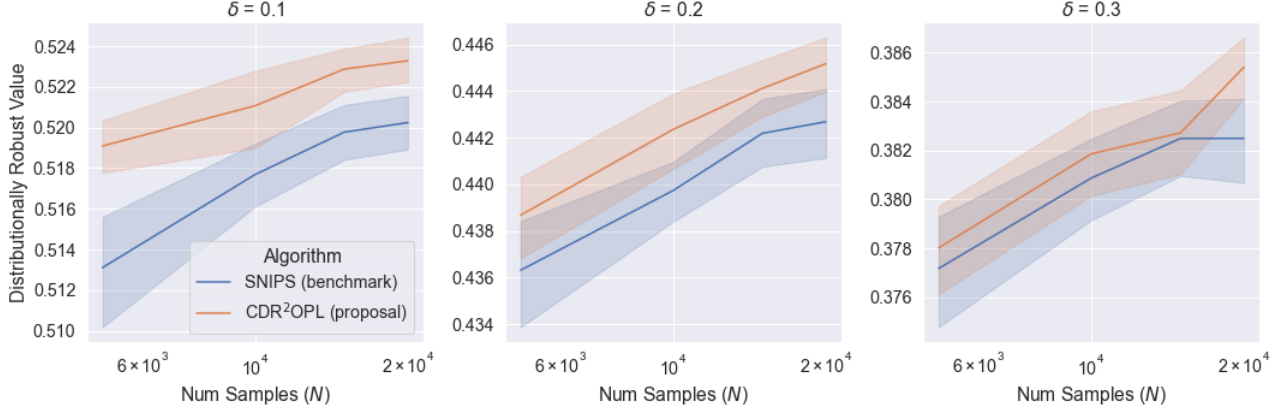


Figure 3. Comparison of our proposal CDR<sup>2</sup>OPL to the baseline SNIPS in the DROPL task, repeated for  $\delta = 0.1, 0.2, 0.3$ . The x-axis is the number of samples  $N$  used by the algorithm. The y-axis is the distributionally robust value  $\mathcal{V}_\delta$  of the learned policy, so higher is better.

Akin to the KL-DRO setting, Equation (11) has a unique solution  $(\alpha^*, \lambda^*)$ , which is the root to  $\nabla \phi^f = \mathbf{0}$ . To extend LDR<sup>2</sup>OPE to  $f$ -divergence DROPE, we can solve the doubly robust moment equation in Equation (8) with  $\theta = [\alpha, \lambda, W_0^f, W_1^f, W_2^f, \mathcal{V}_\delta^f]^\top$  and  $U, V$  as

$$U(r; \alpha, \lambda) = \begin{bmatrix} f^* \left( \frac{-r-\lambda}{\alpha} \right) \\ (f^*)' \left( \frac{-r-\lambda}{\alpha} \right) \cdot \alpha \\ (f^*)' \left( \frac{-r-\lambda}{\alpha} \right) \cdot (r + \lambda) \\ 0 \\ 0 \\ 0 \end{bmatrix}, V(\theta) = \begin{bmatrix} -W_0^f \\ -W_1^f \\ -W_2^f \\ W_1^f - 1 \\ -W_0^f - W_2^f / \alpha - \delta \\ -\mathcal{V}_\delta - \alpha W_0^f - \alpha \delta - \lambda \end{bmatrix}.$$

By a similar argument to Theorem 3.2, the resulting localized doubly robust estimator is asymptotically linear and enjoys semiparametric efficiency. Note that we could have solved the KL problem by setting  $f(x) = x \log(x)$  in Equation (11), and solving a supremum over  $\alpha$  and  $\lambda$  jointly. This would also be efficient and thus have the same asymptotic variance as Theorem 3.2. But since the supremum over  $\lambda$  can be solved in a closed-form way that recovers Equation (2) (see Appendix A.3), our direct analysis for KL should yield better empirical results since we don't need to optimize over  $\lambda$ . In Appendix A.5, we also discuss a direct analysis of the Cressie-Read divergences, which has a closed form solution for the supremum over  $\alpha$ .

## 7. Concluding Remarks

In this paper, we present LDR<sup>2</sup>OPE and CDR<sup>2</sup>OPL, the first doubly robust methods for distributionally robust off-policy evaluation (DROPE) and learning (DROPL), respectively. By virtue of being both distributionally robust and doubly robust, our methods are robust to environment shifts and slow nuisance estimations. By leveraging a localization technique, LDR<sup>2</sup>OPE only needs to fit two outcome functions, instead of a continuum of outcome functions. We prove that LDR<sup>2</sup>OPE is  $\sqrt{N}$ -consistent, asymptotically lin-

ear and enjoys semiparametric efficiency for DROPE, and empirical showed that it offers significant benefits over the SNIPS baseline. Our learning method CDR<sup>2</sup>OPL fits a continuum of outcome functions using a data-driven weighting approach. Under a product rate condition, we prove that CDR<sup>2</sup>OPL achieves  $\mathcal{O}(N^{-1/2})$  regret, via a uniform coupling over the dual variables that generalizes previous results Zhou et al. (2022). Given additional computational overhead of CDR<sup>2</sup>OPL, the simplicity and stability of SNIPS renders it still quite attractive. Our suggestion for practitioners is to try several DROPL methods, e.g. CDR<sup>2</sup>OPL and cross-fitted SNIPS, and select the best one using LDR<sup>2</sup>OPE. Developing a more computationally efficient and stable algorithm for DROPL with non-parametrically estimated propensities is an interesting direction for future work. Another promising next step is to develop methods to deal with unknown Wasserstein environment shifts, which could be a more intuitive metric when contexts are images. Finally, we are interested in generalizing our techniques to distributionally robust reinforcement learning, for which an emerging line of work Zhou et al. (2021); Panaganti & Kalathil (2022); Smirnova et al. (2019); Liu et al. (2022) has been devoted to studying the simulator access model, which is hence no harder than the known data collection setting. Generalizing our techniques here to the RL setting would be worthwhile and challenging, which we leave for future work.

## ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1939704 and CCF-2106508, a JP Morgan research grant, and a Cornell University Fellowship. We thank Ban Kawas, Nian Si, and the anonymous reviewers for useful discussions and feedback.

## References

- Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.
- Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.
- Ben-Tal, A., Den Hertog, D., De Waegenare, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- Bertsimas, D. and Kallus, N. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.
- Bottou, L., Peters, J., Quiñero-Candela, J., Charles, D. X., Chikering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR*, 14(65):3207–3260, 2013. URL <http://jmlr.org/papers/v14/bottou13a.html>.
- Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.
- Čevid, D., Michel, L., Meinshausen, N., and Bühlmann, P. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458*, 2020.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL <https://doi.org/10.1111/ectj.12097>.
- Cressie, N. and Read, T. R. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.
- Duchi, J. C. and Namkoong, H. Variance-based regularization with convex objectives. *JMLR*, 20:68:1–68:55, 2019. URL <http://jmlr.org/papers/v20/17-750.html>.
- Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.
- Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *ICML*, pp. 1097–1104, 2011.
- Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *ICML*, pp. 1447–1456. PMLR, 2018.
- Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.
- Györfi, L., Kohler, M., Krzyżak, A., Walk, H., et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.
- Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.
- Joachims, T. and Swaminathan, A. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1199–1201, 2016.
- Kallus, N. and Mao, X. Stochastic optimization forests. *Management Science (Forthcoming)*, 2022.
- Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *JMLR*, 21:167–1, 2020a.
- Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. In *ICML*, pp. 5089–5100. PMLR, 2020b.
- Kallus, N., Mao, X., and Uehara, M. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*, 2019.
- Kasy, M. Uniformity and the delta method. *Journal of Econometric Methods*, 8(1), 2019.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.
- Khosravi, K., Lewis, G., and Syrgkanis, V. Non-parametric inference adaptive to intrinsic dimension. In *CLeaR*, 2022. URL <https://openreview.net/forum?id=59PvrMnEZm7>.
- Kido, D. Distributionally robust policy learning with wasserstein distance. *arXiv preprint arXiv:2205.04637*, 2022.
- Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

- Lam, H. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pp. 661–670, 2010.
- Liu, A., Liu, H., Anandkumar, A., and Yue, Y. Triply robust off-policy evaluation. *arXiv preprint arXiv:1911.05811*, 2019.
- Liu, Y., Chen, Z., Virochsiri, K., Wang, J., Wu, J., and Liang, F. Reinforcement learning-based product delivery frequency control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15355–15361, 2021.
- Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributional robust q-learning. In *ICML*. PMLR, 2022.
- Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.
- Manski, C. F. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.
- Meinshausen, N. and Ridgeway, G. Quantile regression forests. *JMLR*, 7(6), 2006.
- Mo, W., Qi, Z., and Liu, Y. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674, 2021.
- Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.
- Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pp. 2208–2216, 2016.
- Neyman, J. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.
- Oprescu, M., Syrgkanis, V., and Wu, Z. S. Orthogonal random forest for causal inference. In *ICML*, pp. 4932–4941. PMLR, 2019.
- Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *AISTATS*, pp. 9582–9602. PMLR, 2022.
- Ren, Z. and Zhou, Z. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*, 2020.
- Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- Sason, I. and Verdú, S.  $f$ -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.
- Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributional robust batch contextual bandits. *arXiv preprint arXiv:2006.05630*, 2020a.
- Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributionally robust policy evaluation and learning in offline contextual bandits. In *ICML*, pp. 8884–8894. PMLR, 2020b.
- Smirnova, E., Dohmatob, E., and Mary, J. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.
- Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, pp. 814–823. PMLR, 2015a.
- Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015b.
- Tsatis, A. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
- Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, 1996. ISBN 9781475725476. doi: 10.1007/978-1-4757-2545-2. URL <http://link.springer.com/10.1007/978-1-4757-2545-2>.
- Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Waskom, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
- Zhan, R., Ren, Z., Athey, S., and Zhou, Z. Policy learning with adaptively collected data. *arXiv preprint arXiv:2105.02344*, 2021.
- Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *AISTATS*, pp. 3331–3339. PMLR, 2021.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 2022.

# Appendices

## A. DRO Calculations

In this section, we list some useful calculations for distributionally robust optimization (DRO).

### A.1. KL-divergence DRO

Recall from Equations (2) and (3)

$$\begin{aligned} W(\pi, \alpha) &:= \mathbb{E} [\exp(-R(\pi(S))/\alpha)] \\ \phi(\pi, \alpha) &:= -\alpha \log W(\pi, \alpha) - \alpha\delta \end{aligned}$$

1-st derivatives w.r.t.  $\alpha$ :

$$\begin{aligned} \frac{\partial}{\partial \alpha} W(\pi, \alpha) &= \frac{1}{\alpha^2} \mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)] \\ \frac{\partial}{\partial \alpha} \phi(\pi, \alpha) &= -\log W(\pi, \alpha) - \alpha \frac{\frac{\partial}{\partial \alpha} W(\pi, \alpha)}{W(\pi, \alpha)} - \delta \\ &= -\log W(\pi, \alpha) - \frac{\mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)]}{\alpha \cdot \mathbb{E} [\exp(-R(\pi(S))/\alpha)]} - \delta \end{aligned}$$

2-nd derivatives w.r.t.  $\alpha$ :

$$\begin{aligned} \frac{\partial^2}{\partial \alpha^2} W(\pi, \alpha) &= -\frac{2}{\alpha^3} \mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)] \\ &\quad + \frac{1}{\alpha^4} \mathbb{E} [R(\pi(S))^2 \exp(-R(\pi(S))/\alpha)] \\ \frac{\partial^2}{\partial \alpha^2} \phi(\pi, \alpha) &= -2 \frac{\frac{\partial}{\partial \alpha} W(\pi, \alpha)}{W(\pi, \alpha)} - \frac{\alpha}{W(\pi, \alpha)^2} \left( \left( \frac{\partial^2}{\partial \alpha^2} W(\pi, \alpha) \right) \cdot W(\pi, \alpha) - \left( \frac{\partial}{\partial \alpha} W(\pi, \alpha) \right)^2 \right) \\ &= \frac{1}{\alpha^3 \mathbb{E} [\exp(-R(\pi(S))/\alpha)]} \left( \frac{(\mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)])^2}{\mathbb{E} [\exp(-R(\pi(S))/\alpha)]} - \mathbb{E} [R(\pi(S))^2 \exp(-R(\pi(S))/\alpha)] \right) \end{aligned}$$

1-st derivatives w.r.t.  $\pi$ 's parameters  $\theta$ :

$$\frac{\partial}{\partial \pi} \phi(\pi, \alpha) = -\frac{\alpha}{W(\pi, \alpha)} \mathbb{E} [\exp(-R(\pi(S))/\alpha) \cdot \nabla_{\theta} \log \pi(A|S)]$$

### A.2. $f$ -divergence DRO

Recall

$$\phi^f(\pi, \alpha, \lambda) = -\alpha \mathbb{E} \left[ f^* \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \right] - \alpha\delta - \lambda$$

Then,

$$\begin{aligned}
 \frac{\partial}{\partial \lambda} \phi(\alpha, \lambda) &= -\alpha \mathbb{E} \left[ (f^*)' \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \cdot \left( \frac{-1}{\alpha} \right) \right] - 1 \\
 &= \mathbb{E} \left[ (f^*)' \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \right] - 1 \\
 \frac{\partial}{\partial \alpha} \phi(\alpha, \lambda) &= -W(\alpha, \lambda) - \alpha \mathbb{E} \left[ (f^*)' \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \cdot \left( \frac{R(\pi(S)) + \lambda}{\alpha^2} \right) \right] - \delta \\
 &= -W(\alpha, \lambda) - \frac{1}{\alpha} \mathbb{E} \left[ (f^*)' \left( \frac{-R(\pi(S)) - \lambda}{\alpha} \right) \cdot (R(\pi(S)) + \lambda) \right] - \delta
 \end{aligned}$$

### A.3. Recovering KL duality

Recall that the KL divergence is an  $f$ -divergence, where  $f_{KL}(x) = x \log(x)$  with the dual  $f_{KL}^*(x) = \exp(x - 1)$ .

$$\begin{aligned}
 \phi^{KL}(\pi, \alpha, \lambda) &= -\alpha \exp(-(\lambda + 1)/\alpha) \mathbb{E}_{\mathbb{P}_0} [\exp(-R(\pi(S))/\alpha)] - \alpha \delta - \lambda \\
 \frac{\partial}{\partial \lambda} \phi^{KL}(\pi, \alpha, \lambda) &= \exp(-(\lambda + 1)/\alpha) \mathbb{E}_{\mathbb{P}_0} [\exp(-R(\pi(S))/\alpha)] - 1
 \end{aligned}$$

Setting this to 0, yields:

$$\begin{aligned}
 \exp((\lambda^* + 1)/\alpha) &= \mathbb{E}_{\mathbb{P}_0} [\exp(-R(\pi(S))/\alpha)] \\
 \lambda^* &= \alpha \log(\mathbb{E} [\exp(-R(\pi(S))/\alpha)]) - \alpha
 \end{aligned}$$

Plugging this into the original expression, we get the same equation as Equation (2),

$$\begin{aligned}
 & -\alpha \exp(-(\lambda^* + 1)/\alpha) \mathbb{E} [\exp(-R(\pi(S))/\alpha)] - \alpha \delta - \lambda^* \\
 &= -\alpha - \alpha \delta - \alpha \log(\mathbb{E} [\exp(-R(\pi(S))/\alpha)]) + \alpha \\
 &= \phi^{KL}(\pi, \alpha)
 \end{aligned}$$

### A.4. KL DRO Lemmas

In this section, we prove useful lemmas about the KL DROPE objective  $\phi$  (Equation (2)).

First, we show that  $\phi(\pi, \alpha)$  is strictly concave in  $\alpha$ , except when  $R(\pi(S))$  is almost surely a constant. This corner case implies that  $\mathbb{E}_{\mathbb{P}_0} [f(R(\pi(S)))] = f(R(\pi(S)))$  for any measurable function  $f$ , and hence the distributionally robust objective simplifies to the constant  $R(\pi(S))$ , since  $\sup_{\alpha > 0} \phi(\pi, \alpha) = \sup_{\alpha > 0} R(\pi(S)) - \alpha \delta = R(\pi(S))$ .

**Lemma A.1** ( $\phi$  is strictly concave).  $\forall \alpha > 0 : \frac{\partial^2}{\partial \alpha^2} \phi(\pi, \alpha) \leq 0$ , with strict inequality iff  $R(\pi(S))$  is not almost surely a constant.

*Proof.* By Cauchy-Schwartz in  $L_2$ ,

$$\mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)] \leq \sqrt{\mathbb{E} [\exp(-R(\pi(S))/\alpha)] \mathbb{E} [R(\pi(S))^2 \exp(-R(\pi(S))/\alpha)]}$$

with equality iff  $R(\pi(S))$  and  $\exp(-R(\pi(S))/\alpha)$  are colinear, which happens iff  $R(\pi(S))$  is almost surely constant. By calculations in Appendix A, we have

$$\frac{\partial^2}{\partial \alpha^2} \phi(\pi, \alpha) = \frac{1}{\alpha^3 \mathbb{E} [\exp(-R(\pi(S))/\alpha)]} \left( \frac{(\mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)])^2}{\mathbb{E} [\exp(-R(\pi(S))/\alpha)]} - \mathbb{E} [R(\pi(S))^2 \exp(-R(\pi(S))/\alpha)] \right) < 0$$

Thus,  $\phi(\pi, \alpha)$  is concave in  $\alpha$ , and strictly concave unless  $R(\pi(S))$  is almost surely a constant.  $\square$

Next, we show that under the reward coverage assumption (in Assumption 2.1), we can lower bound  $W$  (Equation (3)).

**Lemma A.2** (Lower bound of  $W$ ). *Under Assumption 2.1, we can lower bound  $W(\pi, \alpha)$  as follows:*

(i) *If  $R(a) \mid S$  is continuous,  $W(\pi, \alpha) \geq \frac{\omega}{2} \min(\alpha, 1)$ .*

(ii) *If  $R(a) \mid S$  is discrete,  $W(\pi, \alpha) \geq \omega$ .*

*Proof of Lemma A.2.* Let  $\mu_\pi$  represent the distribution of over  $\mathcal{S} \times \mathcal{A}$ , of  $s \sim \mathbb{P}_0, a \sim \pi(s)$ .

**Proof of discrete case**

First, it's easier to see the discrete case,

$$W(\pi, \alpha) = \int_{\mathcal{S} \times \mathcal{A}} \sum_{r \in \mathbb{D}} p_R(r \mid s, a) \exp(-r/\alpha) d\mu_\pi \geq \int_{\mathcal{S} \times \mathcal{A}} \sum_{r \in \mathbb{D}} \omega \exp(-0/\alpha) d\mu_\pi = \omega$$

since  $0 \in \mathbb{D}$  by Assumption 2.1.

**Proof of continuous case**

In the continuous case,

$$W(\pi, \alpha) = \int_{\mathcal{S} \times \mathcal{A}} \int_0^1 f(r) \exp(-r/\alpha) dr d\mu_\pi \geq \omega \alpha \int_{\mathcal{S} \times \mathcal{A}} \int_0^{1/\alpha} \exp(-r) dr d\mu_\pi = \omega \alpha (1 - \exp(-1/\alpha))$$

To remove the exponentiation, observe that  $\alpha(1 - \exp(-1/\alpha))$  is increasing and concave, so we can lower-bound the first part by a line with an appropriately chosen slope, and the second part by the intersection point with the line. For some slope  $m$  which we'll set later, the two points of intersection between  $\alpha(1 - \exp(-1/\alpha))$  and the line  $\alpha m$  are  $(x_1, y_1) = (0, 0)$  and  $(x_2, y_2) = (\frac{1}{-\log(1-m)}, \frac{1}{-\log(1-m)} m)$ , which can be seen by solving the following:

$$\alpha(1 - \exp(-1/\alpha)) = \alpha m$$

From  $\alpha \in [0, x_2]$ , we have  $\alpha(1 - \exp(-1/\alpha)) \geq \alpha m$ , and for  $\alpha \geq x_2$ , we have  $\alpha(1 - \exp(-1/\alpha)) \geq y_2$ . Hence, we have

$$\alpha(1 - \exp(-1/\alpha)) \geq \min \left( \alpha m, \frac{m}{-\log(1-m)} \right)$$

We can choose  $m_c$  so that  $c = \frac{m_c}{-\log(1-m_c)}$  for some chosen constant  $0 < c < 1$  (sufficient and necessary to be less than 1, since  $\sup_{\alpha > 0} \alpha(1 - \exp(-1/\alpha)) = 1$ ). For example setting  $c = 0.5$  gives  $m_c \approx 0.797$  implies

$$\alpha(1 - \exp(-1/\alpha)) \geq \min(0.797\alpha, 0.51) \geq \frac{\min(\alpha, 1)}{2}.$$

Thus,  $W(\pi, \alpha) \geq \omega \alpha(1 - \exp(-1/\alpha)) \geq \omega \frac{\min(\alpha, 1)}{2}$ . □

## A.5. Cressie-Read divergence DRO

For  $k > 1$ , the  $k$ -Cressie-Read divergence is the  $f$ -divergence where  $f_k(t) = \frac{1}{k} - \frac{t}{k-1} + \frac{1}{k-1} \frac{t^k}{k}$  (Cressie & Read, 1984). While KL had a close form solution for  $\lambda^*$  in Equation (11), Cressie-Read divergences have a close form solution for  $\alpha^*$ , as shown in the Appendix of (Duchi & Namkoong, 2021). Shown in Equation (12), the dual expression for Cressie-Read divergences is a supremum over just  $\lambda$ .

$$\mathcal{V}_\delta^k(\pi) = \sup_{\lambda \in \mathbb{R}} \phi^k(\pi, \lambda) \tag{12}$$

$$\text{where } \phi^k(\pi, \lambda) = -c_k(\delta) \mathbb{E} \left[ (-R(\pi(S)) - \lambda)_+^{k_*} \right]^{1/k_*} - \lambda$$

By concavity,  $\lambda^*$  is the unique solution to  $\nabla_\lambda \phi^k = \mathbf{0}$ . Thus, the Cressie-Read LDR<sup>2</sup>OPE is to run Algorithm 1, with  $\theta, U, V$  defined by

$$\theta = \{\lambda, W_0, W_1, Q\}, \quad (13)$$

$$U(r; \lambda) = \begin{bmatrix} (-r - \lambda)_+^{k_*} \\ (-r - \lambda)_+^{k_*-1} \\ 0 \\ 0 \end{bmatrix} \quad V(\theta) = \begin{bmatrix} -W_0 \\ -W_1 \\ -c_k(\delta)W_0^{1/k_*-1} \cdot W_1 + 1 \\ -Q - c_k(\delta)W_0^{1/k_*} - \lambda \end{bmatrix}.$$

The algorithm is asymptotically linear and enjoys semiparametric efficiency.

## B. Degeneracy of Weighted DROPE Estimators

While not crucial to the development of our DR estimators, we now digress to describe and characterize a blow-up phenomenon arising from the non-linear and supremum structure of the DROPE objective. Prior work found self-normalized IPS for DROPE to be empirically more stable than IPS (Si et al., 2020a). When the propensity ratios were small, we actually found IPS to explode and have infinite estimation error! From the point of view of OPE, this is surprising since the difference between IPS and SNIPS would never be as extreme as infinite. Theorem B.1 theoretically characterizes when this explosion occurs for any weighted estimator for  $W$ . For non-negative weights  $\{w_i, i \in [N]\}$ , define weight-mean  $S_w := \frac{1}{N} \sum_{i=1}^N w_i$  and min-reward weight-mean  $S_w^m := \frac{1}{N} \sum_{r_i=m} w_i$ , where  $m = \min_i r_i$ .

The weighted estimator we consider is

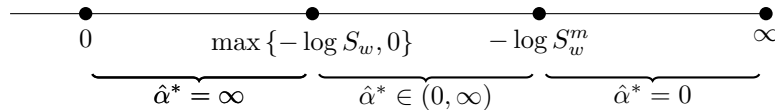
$$\begin{aligned} \hat{\phi}(\pi, \alpha) &= -\alpha \log \left( \frac{1}{N} \sum_{i=1}^N w_i \exp(-r_i/\alpha) \right) - \alpha \delta \\ \widehat{\mathcal{V}}_\delta(\pi) &= \sup_{\alpha > 0} \hat{\phi}(\pi, \alpha) \end{aligned} \quad (14)$$

If  $w_i = \pi(a_i | s_i) / \pi_0(a_i | s_i)$ , then this is IPS. If  $w_i = \frac{\pi(a_i | s_i) / \pi_0(a_i | s_i)}{\frac{1}{N} \sum_{i=1}^N \pi(a_i | s_i) / \pi_0(a_i | s_i)}$ , then this is SNIPS. Observe that for SNIPS, we have the mean of the weights is  $S_w = 1$ . This property turns out to be important in the characterization below.

**Theorem B.1.** *Let  $\delta > 0$ , and let  $\hat{\alpha}$  be the empirical solution to Equation (14). Then, under Assumption 2.1,*

- (i) *If  $S_w = 1$  (as in SNIPS), then  $\widehat{\mathcal{V}}_\delta(\pi) \leq 1$ .*
- (ii) *If  $S_w < 1$ , then  $\delta < -\log S_w$  if and only if  $\widehat{\mathcal{V}}_\delta = \infty$  (hence also  $\hat{\alpha} = \infty$ ).*
- (iii) *If  $S_w^m < 1$ , then  $-\log(S_w^m) < \delta$  if and only if  $\hat{\alpha} = 0$  (hence also  $\widehat{\mathcal{V}}_\delta = m$ ).*

Graphically, the number of line for  $\delta$  looks like:



Case (i) implies that self-normalization is sufficient to avoid blow-up, which is why SNIPS seems to be more stable in practice. The degenerate case of (ii) occurs when  $\delta$  or the propensity ratios are small, and estimation error becomes infinity. Case (iii), while less degenerate than (ii), is also a degenerate case since we know  $\alpha^* > 0$ , so  $\hat{\alpha} = 0$  is not even feasible and provides no useful information about the true value of  $\alpha^*$ .

*Proof of Theorem B.1.* First, note that  $\hat{\phi}(\pi, \alpha)$  is concave in  $\alpha$ . As shown in Appendix A.1, one can calculate the second derivative to be

$$\frac{1}{\alpha^3 \left( \frac{1}{n} \sum_i w_i \exp(-r_i/\alpha) \right)} \left[ \frac{\left( \frac{1}{n} \sum_i w_i r_i \exp(-r_i/\alpha) \right)^2}{\frac{1}{n} \sum_i w_i \exp(-r_i/\alpha)} - \left( \frac{1}{n} \sum_i w_i r_i^2 \exp(-r_i/\alpha) \right) \right]$$

Without changing the sign, give a  $\frac{1}{S_w}$  factor to the quantity inside the brackets so that  $\frac{1}{n} \sum w_i$  becomes  $\frac{1}{n S_w} \sum w_i$ . Then the same Cauchy-Schwarz reasoning from Lemma 2 of (Si et al., 2020a) concludes that the whole quantity is non-negative, and strictly positive iff there are two different  $r_i$ 's.

**Proof of (i):**

Since  $S_w = 1$ , the  $w_i$  form an empirical distribution, which is bounded by Jensen's inequality

$$\hat{\phi}(\pi, \alpha) \leq \sup_{\alpha > 0} -\alpha \left( \frac{1}{n} \sum_{i=1}^n w_i (-r_i/\alpha) \right) - \alpha \delta \leq \sum_{i=1}^n w_i r_i \leq 1$$

**Proof of (ii):**

If  $S_w \geq 1$ , the claim is vacuous, so let  $S_w < 1$ . By concavity,  $\lim_{\alpha \rightarrow \infty} \hat{\phi}(\pi, \alpha) = \infty$  is equivalent to  $\lim_{\alpha \rightarrow \infty} \frac{\partial}{\partial \alpha} \hat{\phi}(\pi, \alpha) > \epsilon$  for some  $\epsilon > 0$ . The limit of the derivative can be calculated explicitly to be  $-\log S_w - \delta$ :

$$\lim_{\alpha \rightarrow \infty} -\log \left( \frac{1}{N} \sum_{i=1}^N w_i \exp(-r_i/\alpha) \right) - \frac{\sum_{i=1}^N w_i (r_i/\alpha) \exp(-r_i/\alpha)}{\sum_{i=1}^N w_i \exp(-r_i/\alpha)} - \delta = -\log S_w - 0 - \delta \quad (15)$$

To see the forward direction, if  $\delta < -\log S_w$ , then Equation (15) is at least  $\epsilon = \frac{-\log S_w - \delta}{2} > 0$ , implying  $\hat{\alpha} = \infty$ .

For the converse, suppose  $\hat{\alpha} = \infty$ , which implies Equation (15) is at least some  $\epsilon > 0$ . Then clearly  $\delta < -\log S_w - \epsilon < -\log S_w$ .

**Proof of (iii):**

Again leveraging concavity, the idea is that  $\hat{\phi}(\pi, \alpha)$  achieves sup at  $\hat{\alpha} = 0$  if and only if the gradient w.r.t.  $\alpha$  at 0 is negative. We can calculate the limit explicitly to be  $-\log S_w^m - \delta$ .

Concretely, consider the limit of  $\alpha \rightarrow 0^+$ .

$$\lim_{\alpha \rightarrow 0^+} -\log \left( \frac{1}{N} \sum_{i=1}^N w_i \exp(-r_i/\alpha) \right) - \frac{\sum_{i=1}^N w_i (r_i/\alpha) \exp(-r_i/\alpha)}{\sum_{i=1}^N w_i \exp(-r_i/\alpha)} - \delta$$

Let  $m = \min_i r_i \geq 0$  be the minimum logged reward. Then we have

$$\begin{aligned} & \lim_{\alpha \rightarrow 0^+} \log \left( \frac{1}{N} \sum_{i=1}^N w_i \exp(-r_i/\alpha) \right) \\ &= \lim_{\alpha \rightarrow 0^+} -\frac{m}{\alpha} + \log \left( \frac{1}{N} \left( \sum_{r_i=m} w_i + \sum_{r_i>m} w_i \exp((m-r_i)/\alpha) \right) \right) \\ &= \lim_{\alpha \rightarrow 0^+} -\frac{m}{\alpha} + \log(S_w^m), \end{aligned}$$

and

$$\begin{aligned} & \lim_{\alpha \rightarrow 0^+} \frac{\sum_{i=1}^N w_i (r_i/\alpha) \exp(-r_i/\alpha)}{\sum_{i=1}^N w_i \exp(-r_i/\alpha)} \\ &= \lim_{\alpha \rightarrow 0^+} \frac{\sum_{r_i=m} w_i (m/\alpha) + \sum_{r_i>m} w_i (r_i/\alpha) \exp((m-r_i)/\alpha)}{\sum_{r_i=m} w_i + \sum_{r_i>m} w_i \exp((m-r_i)/\alpha)} \\ &= \lim_{\alpha \rightarrow 0^+} \frac{m}{\alpha} \end{aligned}$$

since  $m - r_i < 0$  and so  $\lim_{\alpha \rightarrow 0^+} \exp((m - r_i)/\alpha) = \exp(-\infty) = 0$ . Putting these two together, we get

$$\lim_{\alpha \rightarrow 0^+} \frac{\partial}{\partial \alpha} \hat{\phi}(\pi, \alpha) = -\log(S_w^m) - \delta$$

Thus, the limit is negative if and only if  $\delta > -\log(S_w^m)$ , as desired.  $\square$

## C. Proofs for LDR<sup>2</sup>OPE

### C.1. Generic Bandit Moment Equations

Recall the proposed target equation for bandit feedback by (Kallus et al., 2019) (see Equation (5)),

$$\mathbb{E}[U(Y(1); \theta_1) + V(\theta_2)] = 0, \quad (16)$$

where  $U(\cdot; \theta_1)$  and  $V(\theta_2)$  were arbitrary functions that satisfied the conditions of Theorem 3 of (Kallus et al., 2019). Note that  $V$  only depends on  $\theta_2$ . While Equation (16) captured Quantile Treatment Effect (QTE) and Conditional Value at Risk (CVaR) (which is equivalent to DRO under  $\|\cdot\|_\infty$ ), it is not expressive enough to capture the DROPE objective for KL or  $f$ -divergences.

We first state a generic moment condition that slightly generalizes Equation (16) in two ways: (1) we will allow  $V$  to also depend on  $\theta_1$ , and (b) we will allow for stochastic multi-action policies, rather than restricting to binary, deterministic policies. Our target moment equation is

$$\mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)] = 0. \quad (17)$$

Since we only have access to  $Z = (S, A, R)$ , the corresponding orthogonal  $\psi$  is,

$$\begin{aligned} \psi(z; \theta, \eta_1(z; \theta_1), \eta_2(z)) &= \frac{\pi(a|s)}{\eta_2(s, a)} (U(r; \theta_1) - \eta_1(s, a; \theta_1)) + \mathbb{E}_{a \sim \pi(s)} [\eta_1(s, a; \theta_1)] + V(\theta) \\ \text{where } \eta_1^*(s, a; \theta_1) &= \mathbb{E}[U(R; \theta_1) \mid S = s, A = a] \\ \eta_2^*(s, a) &= \pi_0(a \mid s) \end{aligned} \quad (18)$$

where  $\theta^*$  is the solution to Equation (17),  $\eta_1$  is the outcome function and  $\eta_2$  is the behavior policy. This is analogous to Equation 9 of (Kallus et al., 2019), which is the orthogonalized version of Equation (16). It is standard to check that Equation (18) satisfies universal orthogonality (see Equation 9 of (Kallus et al., 2019), or Equation 21 of (Foster & Syrgkanis, 2019)). Denote the Jacobian and covariance matrices as follows,

$$\begin{aligned} J(\theta') &:= \partial_{\theta'} \mathbb{E}[\psi(Z; \theta, \eta_1^*(Z; \theta_1), \eta_2^*(Z))] \big|_{\theta=\theta'} & J^* &:= J(\theta^*) \\ \psi^*(Z) &:= \psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z)) \\ \Sigma &:= \mathbb{E}_{\mathbb{P}_0} [J^{*-1} \psi^*(Z) \psi^*(Z)^\top J^{*-1}] \end{aligned}$$

By replacing  $V(\theta_2)$  by  $V(\theta)$  and changing  $Y(1)$  for  $R(\pi(S))$ , we arrive at an exact analog of Theorem 3 of (Kallus et al., 2019), which we state for completeness.

Let  $\mathcal{P}_N$  denote a sequence of models for the data generating distribution. Let  $\mathcal{T}_N$  be the set of possible nuisance realizations. We use  $x_j$  to denote the  $j$ -th component of a vector  $x$ . For example,  $\eta_{1,j}$  denotes the  $j$ -th component of  $\eta_1$ .

**Assumption C.1** (Regularity of Estimating Equations). Assume there exist positive constants  $c_1$  to  $c_4$  such that the following conditions hold for all  $\mathbb{P}_0 \in \mathcal{P}_N$ :

- (i)  $\Theta$  is a compact set and it contains a ball of radius  $c_1 N^{-1/2} \log N$  centered at  $\theta^*$ .
- (ii) The map  $(\theta, \eta_1(\cdot; \theta_1'), \eta_2) \mapsto \mathbb{E}_{\mathbb{P}_0} [\psi(Z; \theta, \eta_1(Z; \theta_1'), \eta_2(Z))]$  is twice continuously Gateaux-differentiable.

- (iii)  $\Sigma$  satisfies  $c_2 \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq c_3$ . The lower bound is for invertibility, while the upper bound is for bounded variance.
- (iv) The nuisance realization set  $\mathcal{T}_N$  contains the true nuisance parameters  $(\eta_1^*(\cdot; \theta_1^*), \eta_2^*(\cdot))$ . Moreover, the parameter space  $\Theta$  is bounded and for each  $(\eta_1(\cdot; \theta_1'), \eta_2(\cdot)) \in \mathcal{T}_N$ , the function class  $\mathcal{F}_{\eta, \theta_1'} = \{\psi_j(Z; \theta, \eta_1(Z; \theta_1'), \eta_2(Z)), j \in [d], \theta \in \Theta\}$  is suitably measurable and its uniform covering entropy satisfies the following: for positive constants  $a, v$  and  $q > 2$ ,

$$\sup_{\mathbb{Q}} \log N \left( \epsilon \|F_{\eta, \theta_1'}\|_{\mathbb{Q}, 2}, \mathcal{F}_{\eta, \theta_1'}, \|\cdot\|_{\mathbb{Q}, 2} \right) \leq v \log(a\epsilon), \forall \epsilon \in [0, 1]$$

where  $F_{\eta, \theta_1'}$  is a measurable envelope for  $\mathcal{F}_{\eta, \theta_1'}$  that satisfies  $\|F_{\eta, \theta_1'}\|_{\mathbb{P}, q} \leq c_4$ .

Note, if  $\mathcal{F}_{\eta, \theta_1'}$  are Donsker classes, then this condition is satisfied (Van der Vaart, 2000).

**Assumption C.2** (Nuisance Estimation Rates). Let  $\rho_{\mu, N}, \rho_{\pi, N}, \rho_{\theta, N}$  denote the converge rates. Suppose there exists sequence of constants  $\Delta_N \rightarrow 0$  s.t. for any  $\mathbb{P}_0 \in \mathcal{P}_N$ , w.p.  $1 - \Delta_N$ , the estimates  $(\hat{\eta}_1^{(k)}(\cdot; \hat{\theta}_{1, \text{init}}^{(k)}), \hat{\pi}_0^{(k)})$  belong to  $\mathcal{T}_N$ , and every  $j \in [d]$ ,

$$\begin{aligned} \left\| \hat{\eta}_{1, j}^{(k)}(S, A; \hat{\theta}_{1, \text{init}}^{(k)}) - \eta_{1, j}^*(S, A; \hat{\theta}_{1, \text{init}}^{(k)}) \right\|_{L_2(\mathbb{P}_0)} &\leq \rho_{\eta_1, N} \\ \left\| \hat{\pi}_0^{(k)}(S, A) - \pi_0(S, A) \right\|_{L_2(\mathbb{P}_0)} &\leq \rho_{\pi_0, N} \\ \left\| \hat{\theta}_{1, \text{init}}^{(k)} - \theta^* \right\| &\leq \rho_{\theta, N} \end{aligned}$$

**Theorem C.3.** Let  $\hat{\theta}$  be given by applying LDML to Equation (18). Suppose Assumption C.2. Suppose there exists positive constants  $c_1$  to  $c_{10}$  s.t. for any  $\mathbb{P}_0 \in \mathcal{P}_N$ , the following holds:

- (i) Assumption C.1 with constants  $c_1$  to  $c_4$
- (ii) The estimating equation solution approximation error satisfies  $\varepsilon_N = \delta_N N^{-1/2}$ , where  $d_N \rightarrow 0$ .
- (iii) Let  $\theta \in \Theta$  be arbitrary. For each  $j \in [d]$ , the map  $\theta \mapsto \mathbb{E}[U_j(R(\pi(S)); \theta_1) + V_j(\theta)]$  is differentiable, and each component of its gradient is Lipschitz continuous at  $\theta^*$  with Lipschitz constant  $c_5$ . Moreover, if  $\|\theta - \theta^*\| \geq \frac{c_6}{2\sqrt{d}c_5}$ , then  $2\|\mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)]\| \geq c_7$ .
- (iv)  $J^* = \partial_{\theta^T} \mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)]|_{\theta=\theta^*}$  satisfies that  $c_8 \leq \sigma_{\min}(J^*) \leq \sigma_{\max}(J^*) \leq c_9$ .
- (v) For any  $\theta \in \mathcal{B}(\theta^*; \frac{4c_{10}\sqrt{d}\rho_{\pi, N}}{\delta_N \eta}) \cap \Theta$ ,  $r \in (0, 1)$  and for  $j \in [d]$ , there exists functions  $h_1, h_2$  s.t.  $\mathbb{E}[h_i(S, A, \theta_1)] < \infty$  for  $i \in [2]$ , and almost surely

$$\begin{aligned} |\partial_r \eta_{1, j}^*(S, A; \theta_1^* + r(\theta_1 - \theta_1^*))| &\leq h_1(S, A, \theta_1) \\ |\partial_r^2 \eta_{1, j}^*(S, A; \theta_1^* + r(\theta_1 - \theta_1^*))| &\leq h_2(S, A, \theta_1) \end{aligned}$$

- (vi) For  $j \in [d]$ :

$$\begin{aligned} \left( \mathbb{E}[\eta_{1, j}^*(S, A; \theta_1)]^2 \right)^{1/2} &\leq c_{10} \\ \left\| \left( \mathbb{E}[\partial_{\theta_1} \eta_{1, j}^*(X, t, \theta_1)]^2 \right)^{1/2} \right\| &\leq c_{10} \\ \sigma_{\max} \left( \mathbb{E}[\partial_{\theta_1} \partial_{\theta_1^T} \eta_{1, j}^*(S, A; \theta_1)] \right) &\leq c_{10} \\ \sigma_{\max} (\mathbb{E}[\partial_{\theta} \partial_{\theta^T} V_j(\theta)]) &\leq c_{10} \end{aligned}$$

and for any  $\theta \in \mathcal{B}(\theta^*; \max\{\frac{4c_{10}\sqrt{d}\rho_{\pi, N}}{\delta_N \eta}, \rho_{\theta, N}\}) \cap \Theta$ ,

$$\left( \mathbb{E}[\eta_{1, j}^*(S, A; \theta_1) - \eta_{1, j}^*(S, A; \theta_1^*)]^2 \right)^{1/2} \leq c_{10} \|\theta_1 - \theta_1^*\|$$

$$(vii) \quad \rho_{\pi,N}(\rho_{\mu,N} + c_{10}\rho_{\theta,n}) \leq \frac{\eta^3}{3}\delta_N N^{-1/2}, \quad \rho_{\pi,N} \leq \frac{\delta_N^3}{\log N}, \quad \text{and} \quad \rho_{\mu,N} + c_{10}\rho_{\theta,n} \leq \frac{\delta_N^2}{\log N}, \quad \delta_N \leq \frac{4c_{10}^2\sqrt{d}+2\eta}{\eta^2}, \quad \text{and} \quad \delta_N \leq \min\left\{\frac{\eta^2}{8c_{10}^2d} \log N, \sqrt{\frac{\eta^3}{2c_{10}\sqrt{d}}} \log^{1/2} N\right\}$$

Then, uniformly over  $\mathbb{P}_0 \in \mathcal{P}_N$ ,

$$\sqrt{N}\Sigma^{-1/2}(\hat{\theta} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \Sigma^{-1/2} J^{*-1} \psi(Z_i; \theta^*, \eta_1^*(Z_i, \theta_1^*), \eta_2^*(Z_i)) + \mathcal{O}_P(\rho_N) \rightsquigarrow N(0, I_d)$$

where  $\rho_N = o_{\mathbb{P}_0}(1)$  is given in Theorem 1 of (Kallus et al., 2019). Furthermore,  $\Sigma$  is the best possible covariance matrix for regular and asymptotic linear (RAL) estimators; that is, for every RAL estimator with  $\Sigma'$  covariance matrix,  $\Sigma' - \Sigma$  is positive semi-definite (Tsiatis, 2007).

*Proof of Theorem C.3.* The proof is the same as the proof of Theorem 3 in (Kallus et al., 2019), except we replace  $Y(t)$  by  $R(\pi(S))$  and  $V(\theta_2)$  by  $V(\theta)$ .  $\square$

## C.2. Efficiency for DROPE

We now prove Theorem 3.2 by showing that the specific choice of  $U, V$  in Equation (7) is well-behaved, and satisfies the assumptions of Theorem C.3. Note that Theorem C.3 is a uniform guarantee over a family of nominal distributions  $\mathcal{P}_N$ . For simplicity, we will take the family of models as a singleton with the nominal data generating process  $\mathcal{P}_N = \{\mathbb{P}_0\}$ . This simplifies many of the regularity assumptions, as we will remark in the proof below. Under these additional regularity conditions (which are standard), our proof is easily extendable to hold uniformly over a family of nominal distributions, which may be beneficial from a finite-sample perspective (Kasy, 2019).

**Theorem 3.2.** Suppose Assumptions 2.1 and 3.1. Let  $\theta^* = [\alpha^*, W_0^*, W_1^*, \mathcal{V}_\delta^*]^\top$  be the solution to Equation (7). Then,

$$\sqrt{N}(\hat{\theta}^{\text{LDR}^2 \text{ OPE}} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^N J^{*-1} \psi^*(Z_i) + o_p(1)$$

where  $Z_i = (s_i, a_i, r_i)$ ,  $\psi$  is defined in Equation (8),  $\psi^*(Z) := \psi(Z; \theta^*, \eta_1^*(Z; \theta_1^*), \eta_2^*(Z))$ ,

$$J^* = \begin{bmatrix} \frac{W_1^*}{(\alpha^*)^2} & -1 & 0 & 0 \\ \frac{W_2^*}{(\alpha^*)^2} & 0 & -1 & 0 \\ \frac{W_1^*}{(\alpha^*)^2 W_0^*} & -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* (W_0^*)^2} & -\frac{1}{\alpha^* W_0^*} & 0 \\ -\log W_0^* - \delta & -\frac{\alpha^*}{W_0^*} & 0 & -1 \end{bmatrix}$$

and  $\Sigma = \mathbb{E}[J^{*-1} \psi^*(Z) \psi^*(Z)^\top J^{*- \top}]$  is the optimal covariance. Hence,  $\sqrt{N}(\hat{\theta}^{\text{LDR}^2 \text{ OPE}} - \theta^*) \rightsquigarrow \mathcal{N}(0, \Sigma)$  and  $\hat{\theta}^{\text{LDR}^2 \text{ OPE}}$  achieves the semiparametric efficiency lower bound for  $\theta^*$ .

*Proof of Theorem 3.2.* First, we will list some useful calculations. Then, we will verify the assumptions of Theorem C.3, with slight simplifications since we are showing convergence for a single distribution  $\mathbb{P}_0$ , rather than a set of distributions  $\mathcal{P}_N$ .

The function  $\alpha \mapsto \mathbb{E}[\exp(-R(\pi(S))/\alpha)]$  is three-times differentiable, and w.p. 1, the random function  $\alpha \mapsto \mathbb{E}[\exp(-R(\pi(S))/\alpha)|S, A]$  is three-times differentiable. This follows from Dominated Convergence Theorem, since  $R(\pi(S))^j \exp(-R(\pi(S))/\alpha), j \in [4]$  is bounded, and so we can pass limits into the expectation. Let us denote

$$\theta^* = [\alpha^*, W_0^*, W_1^*, \mathcal{V}_\delta^*].$$

$$\begin{aligned} \mathbb{E} [\psi(Z; \theta, \eta_1^*(Z; \theta_1), \eta_2^*)] &= \mathbb{E} [U(R(\pi(S)); \theta_1) + V(\theta)] \\ &= \begin{bmatrix} \mathbb{E} [\exp(-R(\pi(S))/\alpha)] - W_0 \\ \mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)] - W_1 \\ -\delta - \log W_0 - \frac{W_1}{\alpha W_0} \\ -\mathcal{V}_\delta - \alpha \log W_0 - \alpha \delta \end{bmatrix}, \end{aligned}$$

and, the Jacobian is the following (recall the order of  $\theta = [\alpha, W_0, W_1, \mathcal{V}_\delta]$ ):

$$J(\theta) = \begin{bmatrix} \frac{1}{\alpha^2} \mathbb{E} [R(\pi(S)) \exp(-R(\pi(S))/\alpha)] & -1 & 0 & 0 \\ \frac{1}{\alpha^2} \mathbb{E} [R(\pi(S))^2 \exp(-R(\pi(S))/\alpha)] & 0 & -1 & 0 \\ \frac{W_1}{\alpha^2 W_0} & -\frac{1}{W_0} + \frac{W_1}{\alpha W_0^2} & -\frac{1}{\alpha W_0} & 0 \\ -\log W_0 - \delta & -\frac{\alpha}{W_0} & 0 & -1 \end{bmatrix}. \quad (19)$$

Hence, substituting in the optimal value, we have

$$J^* = J(\theta^*) = \begin{bmatrix} \frac{W_1^*}{\alpha^{*2}} & -1 & 0 & 0 \\ \frac{W_2^*}{\alpha^{*2}} & 0 & -1 & 0 \\ \frac{W_1^*}{\alpha^{*2} W_0^*} & -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* W_0^{*2}} & -\frac{1}{\alpha^* W_0^*} & 0 \\ -\log W_0^* - \delta & -\frac{\alpha^*}{W_0^*} & 0 & -1 \end{bmatrix}.$$

By Cramer's rule, we now show that  $\det J^* = -\phi''(\pi, \alpha^*)$ :

$$\begin{aligned} -\det J^* &= \det \begin{bmatrix} \frac{W_1^*}{\alpha^{*2}} & -1 & 0 \\ \frac{W_2^*}{\alpha^{*2}} & 0 & -1 \\ \frac{W_1^*}{\alpha^{*2} W_0^*} & -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* W_0^{*2}} & -\frac{1}{\alpha^* W_0^*} \end{bmatrix} \\ &= \frac{W_1^*}{\alpha^{*2}} \cdot \left( -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* W_0^{*2}} \right) - \frac{W_2^*}{\alpha^{*2}} \cdot \frac{1}{\alpha^* W_0^*} + \frac{W_1^*}{\alpha^{*2} W_0^*} \\ &= \frac{1}{\alpha^{*3} W_0^*} \left( \frac{W_1^{*2}}{W_0^*} - W_2^* \right) \\ &= \phi''(\alpha^*) \end{aligned}$$

Since  $\phi$  is strictly concave, we have  $\det J^* > 0$  and so  $J^*$  is invertible. We can compute the inverse by querying “invert  $\{\{A, -1, 0, 0\}, \{B, 0, -1, 0\}, \{C, D, E, 0\}, \{F, G, 0, -1\}\}$ ” on Wolfram Alpha,

$$M = \begin{bmatrix} A & -1 & 0 & 0 \\ B & 0 & -1 & 0 \\ C & D & E & 0 \\ F & G & 0 & -1 \end{bmatrix} \quad M^{-1} = \frac{1}{S} \begin{bmatrix} D & E & 1 & 0 \\ -(BE + C) & AE & A & 0 \\ BD & -(AD + C) & B & 0 \\ DF - G(BE + C) & E(AG + F) & AG + F & -S \end{bmatrix}$$

where  $S = AD + BE + C = \phi''(\pi, \alpha^*)$  since

$$\begin{aligned} AD + BE + C &= \frac{W_1^*}{\alpha^{*2}} \left( -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* W_0^{*2}} \right) + \frac{1}{\alpha^{*2} W_0^*} \left( -\frac{W_2^*}{\alpha^*} + W_1^* \right) \\ &= \frac{1}{\alpha^{*3} W_0^*} \left( \frac{W_1^{*2}}{W_0^*} - W_2^* \right) \\ &= \phi''(\pi, \alpha^*) \end{aligned}$$

Thus

$$J^{\star-1} = \frac{1}{\phi''(\pi, \alpha^*)} \begin{bmatrix} -\frac{1}{W_0^*} + \frac{W_1^*}{\alpha^* W_0^{*2}} & -\frac{1}{\alpha^* W_0^*} & 1 & 0 \\ \frac{W_0^*/\alpha^* - W_1^*}{\alpha^{*2} W_0^*} & -\frac{W_1^*}{\alpha^{*3} W_0^*} & \frac{W_1^*}{\alpha^{*2}} & 0 \\ \frac{W_2^*}{\alpha^{*3} W_0^*} \left( \frac{W_1^*}{W_0^*} - \alpha^* \right) & -\frac{W_1^{*2}}{\alpha^{*3} W_0^{*2}} & \frac{W_2^*}{\alpha^{*2}} & 0 \\ X_1 & \frac{X_2}{\alpha^* W_0^*} & -X_2 & -\phi''(\pi, \alpha^*) \end{bmatrix}$$

where  $X_1 = \left( \frac{W_1^*}{\alpha^* W_0^{*2}} - \frac{1}{W_0^*} \right) (-\log W_0^* - \delta) - \frac{1}{\alpha^* W_0^{*2}} (W_1^* - W_2^*/\alpha^*)$  and  $X_2 = W_1^*/(\alpha^* W_0^*) + \log W_0^* + \delta$ .

### Verifying condition (i) of Theorem C.3:

We now verify Assumption C.1 (i)-(iv). For (i), the parameter values of  $\theta = [\alpha, W_0, W_1, \mathcal{V}_\delta]$  are bounded, and by considering the closure of the set, we have a compact  $\Theta$ . Since  $\alpha^* > 0$ , we can set  $c_1$  small enough so that the ball of radius  $c_1$  exists at  $\theta^*$ . As shown already,  $\mathbb{E}[\psi]$  is three-times continuously differentiable, giving (ii). Since entries of  $J^{\star-1}$  and  $\psi^*(Z)$  are upper-bounded, this implies that  $\sigma_{\max}(\Sigma)$  is upper-bounded. Since  $\mathcal{P}_N$  is a singleton, we actually do not need that  $\Sigma$  be invertible and directly apply Central Limit Theorem (Theorem C.3 needed to invert  $\Sigma$  in the statement to make the target distribution fixed as a standard normal). Finally, observe that  $\{r \mapsto \exp(-r/\alpha) - W_0 \mid \alpha > 0, W_0 \in \mathbb{R}\}, \{r \mapsto r \exp(-r/\alpha) - W_1 \mid \alpha > 0, W_1 \in \mathbb{R}\}$  are Donsker classes. This implies the metric entropy condition in (iv), see (van der Vaart & Wellner, 1996; Van der Vaart, 2000).

### Verifying condition (ii) of Theorem C.3:

The moment condition can be solved exactly, for instance using Newton's method for an initial estimate  $\alpha_0$  sufficiently close to  $\alpha^*$ . Hence,  $\epsilon_N = 0$ . In practice, we found a good heuristic to seed  $\alpha_0$  as the average of the  $K$  localized estimates  $\hat{\alpha}_{init}^{(k)}$ .

### Verifying condition (iii) of Theorem C.3:

By visual inspection of the  $J(\theta)$  matrix (Equation (19)), and by differentiability of  $\alpha \mapsto \mathbb{E}[R(\pi(S))^j \exp(-R(\pi(S))/\alpha)]$ ,  $j = 1, 2$ , we have that each component of  $J(\theta)$  is Lipschitz, with Lipschitz constant at most  $L := \frac{1}{\alpha^{*3} W_0^{*3}}$ . We also have consistency, i.e. if  $\theta \neq \theta^*$ , then  $\|\mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)]\| > 0$ , since  $\phi(\pi, \cdot)$  is strictly concave. Indeed, if  $\alpha \neq \alpha^*$ , we have  $|\phi'(\alpha)| > 0$ , which is the third component of  $\mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)]$ . And if  $\alpha = \alpha^*$  but  $W_0 \neq W_0^*$ , then  $|\mathbb{E}[\exp(-R/\alpha^*)] - W_0| = |W_0^* - W_0| > 0$ , which is the first component of  $\mathbb{E}[U(R(\pi(S)); \theta_1) + V(\theta)]$ . The same reasoning applies for second and fourth components.

### Verifying condition (iv) of Theorem C.3:

Here, we want to show that singular values of  $J^*$  are lower and upper bounded. The maximum of the entries of  $J^*$  is an upper bound for  $\sigma_{\max}(J^*)$  and the inverse of the maximum of all the entries for  $J^{\star-1}$  is a lower bound for  $\sigma_{\min}(J^*)$ . Since both  $\alpha^*$  and  $\underline{W}$  are positive and finite, we have that the lower bound is positive, and the upper bound is positive and finite.

### Verifying conditions (v) and (vi) of Theorem C.3:

For (vi),

$$\mathbb{E}[\eta_{1,1}^*(S, A; \theta_1)]^2 = \mathbb{E}[\exp(-R/\alpha)]^2 \leq 1 \quad (20)$$

$$\mathbb{E}[\eta_{1,2}^*(S, A; \theta_1)]^2 = \mathbb{E}[R \exp(-R/\alpha)]^2 \leq 1 \quad (21)$$

The following calculations will be useful for both (v) and (vi). Let  $r \in [0, 1]$ ,

$$\begin{aligned}
 |\partial_r \eta_{1,1}^*(s, a; \theta_1^* + r(\theta_1 - \theta_1^*))| &= |\partial_r \mathbb{E} [\exp(-R/(\alpha^* + r(\alpha - \alpha^*))) | S = s, A = a]| \\
 &= \left| \frac{\mathbb{E} [R \exp(-R/(\alpha^* + r(\alpha - \alpha^*))) | S = s, A = a]}{(\alpha^* + r(\alpha - \alpha^*))^2} (\alpha - \alpha^*) \right| \\
 &\leq \left| \frac{1}{(\alpha^* + r(\alpha - \alpha^*))^2} \right| \cdot |\alpha - \alpha^*| \\
 |\partial_r^2 \eta_{1,1}^*(s, a; \theta_1^* + r(\theta_1 - \theta_1^*))| &= \left| \frac{\mathbb{E} [R^2 \exp(-R/(\alpha^* + r(\alpha - \alpha^*))) | S = s, A = a]}{(\alpha^* + r(\alpha - \alpha^*))^4} \right. \\
 &\quad \left. + \frac{2\mathbb{E} [R \exp(-R/(\alpha^* + r(\alpha - \alpha^*))) | S = s, A = a]}{(\alpha^* + r(\alpha - \alpha^*))^3} \right| \cdot |\alpha - \alpha^*|^2 \\
 &\leq \left( \frac{1}{(\alpha^* + r(\alpha - \alpha^*))^4} + \frac{2}{(\alpha^* + r(\alpha - \alpha^*))^3} \right) |\alpha - \alpha^*|^2
 \end{aligned}$$

$\eta_{1,2}^*$  has an additional  $R$  multiplied, but since  $R \in [0, 1]$ , the bound is the same. So,

$$\begin{aligned}
 |\partial_r \eta_{1,2}^*(s, a; \theta_1^* + r(\theta_1 - \theta_1^*))| &\leq \left| \frac{1}{(\alpha^* + r(\alpha - \alpha^*))^2} \right| \cdot |\alpha - \alpha^*| \\
 |\partial_r^2 \eta_{1,2}^*(s, a; \theta_1^* + r(\theta_1 - \theta_1^*))| &\leq \left( \frac{1}{(\alpha^* + r(\alpha - \alpha^*))^4} + \frac{2}{(\alpha^* + r(\alpha - \alpha^*))^3} \right) |\alpha - \alpha^*|^2
 \end{aligned}$$

If  $\theta$  is sufficiently close to  $\theta^*$  (when  $\rho_{\pi,N}$  is small enough, i.e. when  $N$  is large enough), we have that  $\alpha > \alpha^*/2 > 0$ . Hence,  $\partial_r \eta_{1,j}^*$  and  $\partial_r^2 \eta_{1,j}^*$  are upper bounded by  $\frac{3 \cdot 2 \cdot 4}{\alpha^{*2}}$ . This fully verifies (v), as well as most of (vi).

Let  $N$  be sufficiently large s.t.  $\max\{\rho_{\pi,N}, \rho_{\theta,N}\} < \alpha^*/2$ , so for any  $\theta$  close enough to  $\theta^*$  (so that  $\alpha > \alpha^*/2$ ), we have

$$\mathbb{E} [\eta_{1,1}^*(S, A; \theta_1) - \mu_1^*(S, A; \theta_1^*)]^2 = \mathbb{E} [\exp(-R/\alpha) - \exp(-R/\alpha^*)]^2 \leq \frac{4}{\alpha^{*2}} |\alpha - \alpha^*|. \quad (22)$$

The  $\eta_{1,2}^*$  case is analogous, which concludes all of (vi).

### Semiparametric efficiency for $\mathcal{V}_\delta$ and $\alpha$ :

Since we've verified all the Assumptions of Theorem C.3, we have that  $\hat{\theta}$  achieves semiparametric efficiency. Then, by Theorems 25.20, 25.21 of (Van der Vaart, 2000), and the fact that indexing is a cone-shaped function, we also have semiparametric efficiency for each index of  $\theta^*$ , in particular  $\mathcal{V}_\delta^*$  and  $\alpha^*$ .  $\square$

### C.3. Newton-Raphson Method

In this section, we use Newton-Raphson to with projections to  $\mathbb{R}^+$  to solve the moment equation in Algorithm 1, which recall is

$$M(\alpha) := -\delta - \log(\widehat{W}_0(\alpha)) - \frac{\widehat{W}_1(\alpha)}{\alpha \cdot \widehat{W}_0(\alpha)} = 0,$$

where  $\widehat{W}_j$  is defined in Algorithm 1.

First, initialize  $\alpha_0 = \frac{1}{K} \sum_{k=1}^K \alpha_{init}^{(k)}$  to be the average of the outputs of the subroutine calls to cross-fitted SNIPS (since Newton's method should be seeded with something close to  $\alpha^*$ ). Then, take the following update steps until convergence (i.e.  $|\alpha_{t+1} - \alpha_t| < \epsilon$ ),

$$\begin{aligned}
 \alpha_{t+1} &= \alpha_t - M(\alpha_t)/M'(\alpha_t) \\
 \text{where } M'(\alpha) &= -\frac{\widehat{W}_0'(\alpha)}{\widehat{W}_0(\alpha)} - \frac{\widehat{W}_1'(\alpha) \cdot \alpha \widehat{W}_0(\alpha) - \widehat{W}_1(\alpha) \cdot (\widehat{W}_0(\alpha) + \alpha \widehat{W}_0'(\alpha))}{(\alpha \widehat{W}_0(\alpha))^2},
 \end{aligned}$$

where the derivatives only include the  $\alpha$ -dependent IPS part of  $\widehat{W}_j$ , so

$$\widehat{W}'_0(\alpha) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} \frac{\pi(a_i | s_i)}{\widehat{\pi}_0(s_i, a_i)} \exp(-r_i/\alpha) \frac{r_i}{\alpha^2}, \quad (23)$$

$$\widehat{W}'_1(\alpha) = \frac{1}{N} \sum_{k=1}^K \sum_{i \in D_k} \frac{\pi(a_i | s_i)}{\widehat{\pi}_0(s_i, a_i)} \exp(-r_i/\alpha) \frac{r_i^2}{\alpha^2}. \quad (24)$$

If the update takes  $\alpha_{t+1}$  outside the feasible region  $[0, 1/\delta]$ , then project it back.

#### C.4. Multidimensional Newton's Method

Instead of thinking about the moment condition as a function of  $\alpha$ , we can think about it as a function of  $\theta$ , and perform multidimensional Newton's method. This is the formulation that is most natural from applying LDML, with the following multidimensional condition,

$$\psi(\theta) = \begin{bmatrix} -W_0 + \widehat{W}_0(\alpha) \\ -W_1 + \widehat{W}_1(\alpha) \\ -\delta - \log W_0 - \frac{W_1}{\alpha W_0} \\ -\mathcal{V}_\delta - \alpha \log W_0 - \alpha \delta \end{bmatrix} = \mathbf{0}.$$

We'll need to calculate the Jacobian matrix. The only difference from Equation (19) is that the entries with  $\mathbb{E}[\cdot]$  are replaced with IPS estimates. In other words,

$$\widehat{J}(\theta) := \begin{bmatrix} \widehat{W}'_0(\alpha) & -1 & 0 & 0 \\ \widehat{W}'_1(\alpha) & 0 & -1 & 0 \\ \frac{W_1}{\alpha^2 W_0} & -\frac{1}{W_0} + \frac{W_1}{\alpha W_0^2} & -\frac{1}{\alpha W_0} & 0 \\ -\log W_0 - \delta & -\frac{\alpha}{W_0} & 0 & -1 \end{bmatrix},$$

where  $\widehat{W}'_j, j = 0, 1$  are calculated in Equations (23) and (24). Now, we can apply the following updates, until convergence (i.e.  $\|\theta_{t+1} - \theta_t\| < \epsilon$ ):

$$\theta_{t+1} = \theta_t - \widehat{J}(\theta_t)^{-1} \psi(\theta_t).$$

*Remark C.4.* We empirically tested both (1D) Newton-Raphson and the multidimensional Newton's method and found no significant difference in the MSE or final values of  $\alpha$ . There may be some small sample differences but when  $N \geq 1024$ , both approaches essentially gave the exact same solutions.

#### C.5. Runtime Analysis

In this section, we analyze the total runtime (a.k.a. work) and parallelized runtime (a.k.a. span) of Localized Doubly Robust Algorithm 1.

Let  $\mathcal{T}_{\pi_0}(N), \mathcal{T}_f(N), \mathcal{T}_{init}(N)$  respectively denote the work of fitting  $\pi_0, f_j$  (for both  $j = 0, 1$ ), and running InitialEstimate, on an input dataset of size  $N$ . Let  $\mathcal{S}_{\pi_0}(N), \mathcal{S}_f(N), \mathcal{S}_{init}(N)$  denote the span analogs of the above. Suppose these are non-decreasing functions; that is, having more data will only increase training work/span.

Note that, assuming inference of  $\widehat{\pi}_0^{(k)}, \widehat{f}_j^{(k)}$  takes constant time on a single sample, solving the moment equation takes  $\mathcal{O}(N)$  work and  $\mathcal{O}(\log(N))$  span.

A single run of LDR<sup>2</sup>OPE has work/span bounded by

$$\begin{aligned}\mathcal{T}_{\text{LDR}^2\text{OPE}}(N) &= \mathcal{O}\left(K\left(\mathcal{T}_{\pi_0}\left(\frac{(K-1)N}{K}\right) + \mathcal{T}_f\left(\frac{(K-1)N}{2K}\right) + \mathcal{T}_{init}\left(\frac{(K-1)N}{2K}\right)\right)\right) \\ \mathcal{S}_{\text{LDR}^2\text{OPE}}(N) &= \mathcal{O}\left(\max\left(\mathcal{S}_{\pi_0}\left(\frac{(K-1)N}{K}\right), \mathcal{S}_f\left(\frac{(K-1)N}{2K}\right) + \mathcal{S}_{init}\left(\frac{(K-1)N}{2K}\right)\right)\right)\end{aligned}$$

where the work expression follows directly from examining the sizes of the datasets on each iteration. The span expression doesn't have a  $K$  multiplier since each cross-fitting step can be parallelized. Also, fitting  $\pi_0$  can be done in parallel when running InitialEstimate and then fitting  $f_j$  (which depends on the output of InitialEstimate).

Now, we analyze the work/span of  $m$ -recursive runs of LDR<sup>2</sup>OPE. They satisfy the following recurrences:

$$\begin{aligned}\mathcal{T}_{\text{LDR}^2\text{OPE},m}(N) &= \mathcal{O}\left(K\left(\mathcal{T}_{\pi_0}\left(\frac{(K-1)N}{K}\right) + \mathcal{T}_f\left(\frac{(K-1)N}{2K}\right) + \mathcal{T}_{\text{LDR}^2\text{OPE},m-1}\left(\frac{(K-1)N}{2K}\right)\right)\right) \\ \mathcal{S}_{\text{LDR}^2\text{OPE},m}(N) &= \mathcal{O}\left(\mathcal{S}_{\pi_0}\left(\frac{(K-1)N}{K}\right) + \mathcal{S}_f\left(\frac{(K-1)N}{2K}\right) + \mathcal{S}_{\text{LDR}^2\text{OPE},m-1}\left(\frac{(K-1)N}{2K}\right)\right)\end{aligned}$$

where we upper bounded max by + for span to simplify the solution. The recurrences solve to,

$$\begin{aligned}\mathcal{T}_{\text{LDR}^2\text{OPE},m}(N) &= \mathcal{O}\left(\sum_{t=1}^m K^t \mathcal{T}_{\pi_0}\left(\frac{(K-1)^t N}{2^{t-1} K^t}\right) + \sum_{t=1}^m K^t \mathcal{T}_f\left(\frac{(K-1)^t N}{2^t K^t}\right) + K^m \mathcal{T}_{init}\left(\frac{(K-1)^m N}{2^m K^m}\right)\right) \\ \mathcal{S}_{\text{LDR}^2\text{OPE},m}(N) &= \mathcal{O}\left(\sum_{t=1}^m \mathcal{S}_{\pi_0}\left(\frac{(K-1)^t N}{2^{t-1} K^t}\right) + \sum_{t=1}^m \mathcal{S}_f\left(\frac{(K-1)^t N}{2^t K^t}\right) + \mathcal{S}_{init}\left(\frac{(K-1)^m N}{2^m K^m}\right)\right)\end{aligned}$$

Since fitting the nuisances  $\pi_0, f_j$  are standard regression tasks, there are many poly-time (many of which are linear) learning algorithms. For example, linear regression, neural nets trained with SGD, and XGBoost can all take  $\mathcal{O}(N)$  work to train. To keep analysis generic, suppose that  $\mathcal{T}_{\pi_0}(N), \mathcal{T}_f(N), \mathcal{S}_{\pi_0}(N), \mathcal{S}_{\pi_0}(N) = \tilde{\mathcal{O}}(N^p)$  for some  $p \geq 1$ . Then cross-fitted SNIPS has work and span

$$\begin{aligned}\mathcal{T}_{\text{xfit-snips}}(N) &= \mathcal{O}\left(K\left(\mathcal{T}_{\pi_0}\left(\frac{(K-1)N}{K}\right) + \mathcal{T}_f\left(\frac{(K-1)N}{K}\right)\right)\right) = \tilde{\mathcal{O}}(KN^p) \\ \mathcal{S}_{\text{xfit-snips}}(N) &= \mathcal{O}\left(\mathcal{S}_{\pi_0}\left(\frac{(K-1)N}{K}\right) + \mathcal{S}_f\left(\frac{(K-1)N}{K}\right)\right) = \tilde{\mathcal{O}}(N^p)\end{aligned}$$

So starting with InitialEstimate being cross-fitted SNIPS, and recursively running LDR<sup>2</sup>OPE  $m$  times has work and span

$$\begin{aligned}\mathcal{T}_{\text{LDR}^2\text{OPE},m}(N) &= \tilde{\mathcal{O}}(K^{m+1}N^p) \\ \mathcal{S}_{\text{LDR}^2\text{OPE},m}(N) &= \tilde{\mathcal{O}}(mN^p)\end{aligned}$$

## D. Proof of Regret Guarantees for DROPL

In our analysis, we assume that the estimated nuisances fall into their appropriate ranges:  $\hat{f}_0^{(k)}(s, a; \alpha) \in (0, 1], \hat{\pi}_0^{(k)}(s, a) \in [\eta, 1]$ . This is without loss of generality since it can always be satisfied by clipping.

**Assumption D.1.** We suppose that  $N$  is sufficiently large. Specifically, for the  $\beta$  from Theorem 4.3 Let  $\beta \in (0, 1)$ , we need the following to hold

$$\begin{aligned}\frac{W}{2} &\geq \frac{288}{\eta\sqrt{N}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{4}{\eta} \log^{1/2}(1/\beta), \\ \frac{W}{4} &\geq \frac{384\sqrt{K}}{\eta\sqrt{N}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{8\sqrt{K} \log^{1/2}(K/\beta)}{\eta} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2}\end{aligned}$$

To satisfy both, it suffices to take,

$$\sqrt{N} \geq \frac{4}{\underline{W}} \left( \frac{384\sqrt{K}}{\eta} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{8\sqrt{K} \log^{1/2}(K/\beta)}{\eta} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^{\epsilon}(N, \beta/K) \sqrt{N}}{\eta^2} \right),$$

Provided that  $\text{Rate}_{\pi_0}(N, \beta) \cdot \text{Rate}_f^{\epsilon}(N, \beta) \leq o(N^{-1/2})$ , it will not be part of the dominant term.

**Theorem 4.3.** *Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1/6)$ , w.p. at least  $1 - 6\beta$ , the distributionally robust regret  $\mathcal{R}_{\delta}(\hat{\pi}^{DR})$  is at most*

$$\begin{aligned} & \frac{2112\bar{\alpha}\sqrt{K}}{\underline{W}\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} + \log^{1/2}(K/\beta) \right) \\ & + \frac{4\bar{\alpha}}{\underline{W}\eta^2} (\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^{\epsilon}(N, \beta/K)), \end{aligned}$$

provided  $N$  is sufficiently large (Assumption D.1).

*Proof of Theorem 4.3.* The steps for bounding regret are inspired by uniform coupling arguments bounding OPL regret (Athey & Wager, 2021; Zhou et al., 2022). First, define the *infeasible* CFDR values  $W^{DR}$  and  $\mathcal{V}_{\delta}^{DR}$  (without the hat), with the *true* nuisances; that is, replace  $\hat{\pi}_0^{(k)}$  and  $\hat{f}_0^{(k)}(\cdot; \alpha)$  in Equation (10) by the true  $\pi_0$  and  $f_0(\cdot; \alpha)$  respectively. Then, we show two uniform concentrations (with rate  $\mathcal{O}(N^{-1/2})$ ) simultaneously over  $\Pi$  and  $\alpha$ ; Lemma D.5 concentrates  $\mathcal{V}_{\delta}$  to  $\mathcal{V}_{\delta}^{DR}$ , and Lemma D.6 concentrates  $\mathcal{V}_{\delta}^{DR}$  to  $\hat{\mathcal{V}}_{\delta}^{DR}$ . So,

$$\begin{aligned} \mathcal{R}_{\delta}(\hat{\pi}^{DR}) &= \mathcal{V}_{\delta}(\pi^*) - \hat{\mathcal{V}}_{\delta}^{DR}(\pi^*) + \hat{\mathcal{V}}_{\delta}^{DR}(\pi^*) - \mathcal{V}_{\delta}(\hat{\pi}^{DR}) \\ &\leq \mathcal{V}_{\delta}(\pi^*) - \hat{\mathcal{V}}_{\delta}^{DR}(\pi^*) + \hat{\mathcal{V}}_{\delta}^{DR}(\hat{\pi}^{DR}) - \mathcal{V}_{\delta}(\hat{\pi}^{DR}) \\ &\leq 2 \sup_{\pi \in \Pi} |\mathcal{V}_{\delta}(\pi) - \hat{\mathcal{V}}_{\delta}^{DR}(\pi)| \\ &\leq 2 \sup_{\pi \in \Pi} |\mathcal{V}_{\delta}(\pi) - \mathcal{V}_{\delta}^{DR}(\pi)| + 2 \sup_{\pi \in \Pi} |\mathcal{V}_{\delta}^{DR}(\pi) - \hat{\mathcal{V}}_{\delta}^{DR}(\pi)| \\ &\leq \frac{2\bar{\alpha}}{\underline{W}} \left( \frac{288}{\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} \right) + \frac{4}{\eta\sqrt{N}} \log^{1/2}(1/\beta) \right) \\ &\quad + \frac{4\bar{\alpha}}{\underline{W}} \left( \frac{384}{\eta\sqrt{N/K}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} \right) + \frac{8 \log^{1/2}(K/\beta)}{\eta\sqrt{N/K}} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^{\epsilon}(N, \beta/K)}{\eta^2} \right) \\ &\leq \frac{2112\bar{\alpha}\sqrt{K}}{\underline{W}\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} \right) + \frac{40\bar{\alpha}\sqrt{K} \log^{1/2}(K/\beta)}{\underline{W}\eta\sqrt{N}} + \frac{4\bar{\alpha}}{\underline{W}\eta^2} (\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^{\epsilon}(N, \beta/K)) \end{aligned}$$

w.p. at least  $1 - 6\beta$ , where we invoked Lemmas D.5 and D.6 to bound the two supremum terms.  $\square$

We now build towards the proofs for Lemmas D.5 and D.6. First, we show that assuming  $\underline{\alpha} > 0$ , we can uniformly bound the Lipschitz constant of the functions  $\{\alpha \mapsto \exp(-r/\alpha), r \in [0, 1]\}$  by  $L := 1/\underline{\alpha}^2$ .

**Lemma D.2.** *Suppose Assumptions 2.1 and 2.3. Let  $t \in (0, 1)$ , and let  $s, a, r$  be fixed (not random variables). Then, the following deterministic functions of  $\alpha$ , restricted to  $[\underline{\alpha}, \infty)$ , are Lipschitz with Lipschitz constant upper bounded by  $1/\underline{\alpha}^2$ .*

$$\begin{aligned} & \alpha \mapsto \exp(-r/\alpha) \\ & \alpha \mapsto \mathbb{E}[\exp(-R/\alpha) \mid S = s, A = a] \\ & \alpha \mapsto \mathbb{E}[\exp(-R/\alpha) \mid S = s, A = \pi(s)] \end{aligned}$$

*Proof.* The first function  $\alpha \mapsto \exp(-r/\alpha)$  is continuous and has derivative  $\frac{r}{\alpha^2} \exp(-r/\alpha)$ . Since we're restricting to  $[\underline{\alpha}, \infty)$ , the derivative is upper bounded by  $1/\underline{\alpha}^2$ , which implies that the Lipschitz constant is also bounded by  $1/\underline{\alpha}^2$ . For

the second and third functions, limits can be passed into the expectation using Dominated Convergence Theorem, since the derivative of the random variable is bounded. Hence, the same reasoning shows that their Lipschitz constant is also upper bounded by  $1/\underline{\alpha}^2$ .  $\square$

Note the above lemma also implies that the estimated continuum nuisance in Section 4.1, as a function of  $\alpha$ , is also Lipschitz, with Lipschitz constant upper bounded by the same quantity. This is because the estimated nuisance, as a function of  $\alpha$ , is a convex combination of functions whose Lipschitz constant is upper bounded by  $1/\underline{\alpha}^2$ .

Now we show a key lemma that uniformly concentrates over both  $\Pi$  and  $[\underline{\alpha}, \bar{\alpha}]$ .

**Lemma D.3.** *Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1)$ , w.p.  $1 - \beta$  we have,*

$$\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W_{DR}(\pi, \alpha) - W(\pi, \alpha)| \leq \frac{288}{\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{2\underline{\alpha}^2} \vee 1 \right) + \frac{4}{\eta\sqrt{N}} \log^{1/2}(1/\beta).$$

*Proof.* It is sufficient (and necessary, see Page 108 of (Wainwright, 2019)) to bound the Rademacher complexity of

$$\mathcal{F}_{\Pi, \alpha} = \left\{ w_{\pi, \alpha}(s, a, r) \mapsto \frac{\pi(a | s)}{\pi_0(a | s)} (\exp(-r/\alpha) - \mathbb{E}[\exp(-R/\alpha) | S = s, A = a]) + \mathbb{E}[\exp(-R/\alpha) | S = s, A = \pi(s)] \mid \pi \in \Pi, \alpha \in [\underline{\alpha}, \bar{\alpha}] \right\}$$

This class is strictly larger than what was considered in (Athey & Wager, 2021; Zhou et al., 2022), since it is also indexed by the dual variable  $\alpha$ .

First, notice that these functions are uniformly bounded, since  $\exp(-r/\alpha) \in (0, 1]$ :

$$\begin{aligned} & \left| \frac{\pi(a | s)}{\pi_0(a | s)} (\exp(-r/\alpha) - \mathbb{E}[\exp(-R/\alpha) | S = s, A = a]) + \mathbb{E}[\exp(-R/\alpha) | S = s, A = \pi(s)] \right| \\ & \leq \eta^{-1} |\exp(-r/\alpha) - \mathbb{E}[\exp(-R/\alpha) | S = s, A = a]| + \mathbb{E}[\exp(-R/\alpha) | S = s, A = \pi(s)] \\ & \leq 2\eta^{-1} \end{aligned}$$

We now construct covers in  $\|\cdot\|_{L_2(\mathbb{P}_N)}$  to bound the Rademacher complexity. Let  $\pi, \tilde{\pi} \in \Pi$  and  $\alpha, \tilde{\alpha} \in [\underline{\alpha}, \bar{\alpha}]$ . Two useful bounds that we'll use are:

(a) We can bound the  $L_2(\mathbb{P}_N)$  distance between policies by the hamming distance:

$$\begin{aligned} \|\pi(a | s) - \tilde{\pi}(a | s)\|_{L_2(\mathbb{P}_N)}^2 &= \frac{1}{N} \sum_{i=1}^N (\pi(a | s) - \tilde{\pi}(a | s))^2 \\ &\leq \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\pi(s) \neq \tilde{\pi}(s)] \\ &= d_H(\pi, \tilde{\pi}) \end{aligned}$$

(b) By Lemma D.2, we can bound the  $L_2(\mathbb{P}_N)$  distance between  $\exp(-r/\alpha)$  functions

$$\|\exp(-r/\alpha) - \exp(-r/\tilde{\alpha})\|_{L_2(\mathbb{P}_N)} \leq L|\alpha - \tilde{\alpha}|, \text{ where } L := 1/\underline{\alpha}^2.$$

By triangle inequality, we can separately consider three terms:

$$\begin{aligned}
 & \|w_{\pi, \alpha} - w_{\tilde{\pi}, \tilde{\alpha}}\|_{L_2(\mathbb{P}_N)} \\
 & \leq \frac{1}{\eta} \left( \|\pi(a_i | s_i) \exp(-r_i/\alpha) - \tilde{\pi}(a_i | s_i) \exp(-r_i/\tilde{\alpha})\|_{L_2(\mathbb{P}_N)} \right. \\
 & \quad \left. + \|\pi(a_i | s_i) \mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a_i] - \tilde{\pi}(a_i | s_i) \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = a_i]\|_{L_2(\mathbb{P}_N)} \right) \\
 & \quad + \|\mathbb{E} [\exp(-R/\alpha) | S = s_i, A = \pi(s_i)] - \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = \tilde{\pi}(s_i)]\|_{L_2(\mathbb{P}_N)}
 \end{aligned}$$

Bound the first term:

$$\begin{aligned}
 & \|\pi(a_i | s_i) \exp(-r_i/\alpha) - \tilde{\pi}(a_i | s_i) \exp(-r_i/\tilde{\alpha})\|_{L_2(\mathbb{P}_N)} \\
 & \leq \|(\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)) \exp(-r_i/\alpha)\|_{L_2(\mathbb{P}_N)} + \|\tilde{\pi}(a_i | s_i) (\exp(-r_i/\alpha) - \exp(-r_i/\tilde{\alpha}))\|_{L_2(\mathbb{P}_N)} \\
 & \leq \|\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)\|_{L_2(\mathbb{P}_N)} + \|\exp(-r_i/\alpha) - \exp(-r_i/\tilde{\alpha})\|_{L_2(\mathbb{P}_N)} \\
 & \leq \sqrt{d_H(\pi, \tilde{\pi})} + L|\alpha - \tilde{\alpha}|
 \end{aligned}$$

Bound the second term:

$$\begin{aligned}
 & \|\pi(a_i | s_i) \mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a_i] - \tilde{\pi}(a_i | s_i) \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = a_i]\|_{L_2(\mathbb{P}_N)} \\
 & \leq \|(\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)) \mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a_i]\|_{L_2(\mathbb{P}_N)} \\
 & \quad + \|\tilde{\pi}(a_i | s_i) (\mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a_i] - \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = a_i])\|_{L_2(\mathbb{P}_N)} \\
 & \leq \|\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)\|_{L_2(\mathbb{P}_N)} + \|\mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a_i] - \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = a_i]\|_{L_2(\mathbb{P}_N)} \\
 & \leq \sqrt{d_H(\pi, \tilde{\pi})} + L|\alpha - \tilde{\alpha}|
 \end{aligned}$$

Bound the third term:

$$\|\mathbb{E} [\exp(-R/\alpha) | S = s_i, A = \pi(s_i)] - \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = \tilde{\pi}(s_i)]\|_{L_2(\mathbb{P}_N)}$$

Since  $L_2(\mathbb{P}_N)$  is bounded by  $L_\infty(\mathbb{P}_N)$ , and apply triangle inequality to each action,

$$\begin{aligned}
 & \leq \max_{i \in [N]} \sum_{a \in \mathcal{A}} |\pi(a | s_i) - \tilde{\pi}(a | s_i)| \cdot \|\mathbb{E} [\exp(-R/\alpha) | S = s_i, A = a] - \mathbb{E} [\exp(-R/\tilde{\alpha}) | S = s_i, A = a]\| \\
 & \leq L|\alpha - \tilde{\alpha}| \max_{i \in [N]} \|\pi(s_i) - \tilde{\pi}(s_i)\|_1 \\
 & \leq 2L|\alpha - \tilde{\alpha}|.
 \end{aligned}$$

Altogether, we have that,

$$\begin{aligned}
 \|w_{\pi, \alpha} - w_{\tilde{\pi}, \tilde{\alpha}}\|_{L_2(\mathbb{P}_N)} & \leq \frac{2}{\eta} \left( \sqrt{d_H(\pi, \tilde{\pi})} + L|\alpha - \tilde{\alpha}| \right) + 2L|\alpha - \tilde{\alpha}| \\
 & \leq \frac{3}{\eta} \left( \sqrt{d_H(\pi, \tilde{\pi})} + L|\alpha - \tilde{\alpha}| \right)
 \end{aligned}$$

To bound it by  $t$ , we can take  $d_H(\pi, \tilde{\pi}) \leq \left(\frac{t\eta}{6}\right)^2$  and  $|\alpha - \tilde{\alpha}| \leq \frac{t\eta}{6L}$ . Since  $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ , the covering for  $\alpha$  can be done in

$\frac{3L(\bar{\alpha}-\underline{\alpha})}{t\eta}$  points. By Dudley's chaining (see (5.48) of (Wainwright, 2019)), we have

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}_{\Pi,\alpha}) &\leq \frac{24}{\sqrt{N}} \int_0^{4\eta^{-1}} \log^{1/2} \left( \mathcal{N}_H((t\eta/6)^2, \Pi) \cdot \frac{3L(\bar{\alpha}-\underline{\alpha})}{t\eta} \right) dt \\ &\leq \frac{144}{\eta\sqrt{N}} \left( \int_0^1 \log^{1/2} \mathcal{N}_H(t^2, \Pi) + \log^{1/2} \frac{L(\bar{\alpha}-\underline{\alpha})}{2t} dt \right) \\ &\leq \frac{144}{\eta\sqrt{N}} (\kappa(\Pi) + L\bar{\alpha} \vee 1)\end{aligned}$$

By Theorem 4.10 of (Wainwright, 2019), w.p. at least  $1 - \beta$ ,

$$\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W_{DR}(\pi, \alpha) - W(\pi, \alpha)| \leq \frac{288}{\eta\sqrt{N}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{4}{\eta\sqrt{N}} \log^{1/2}(1/\beta). \quad (25)$$

□

Both Lemmas D.5 and D.6 will start with the following lemma,

**Lemma D.4.** *Let  $f, g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  be functions, then,*

$$\left| \sup_{\alpha} \{-\alpha \log f(\alpha) - \alpha \delta\} - \sup_{\alpha} \{-\alpha \log g(\alpha) - \alpha \delta\} \right| \leq \sup_{\alpha} \left| \alpha \log \left( 1 + \frac{f(\alpha) - g(\alpha)}{g(\alpha)} \right) \right|. \quad (26)$$

*Proof.* Merge the two sup's together,

$$\left| \sup_{\alpha} \{-\alpha \log f(\alpha) - \alpha \delta\} - \sup_{\alpha} \{-\alpha \log g(\alpha) - \alpha \delta\} \right| \leq \left| \sup_{\alpha} -\alpha \log f(\alpha) + \alpha \log g(\alpha) \right| \leq \sup_{\alpha} \left| \alpha \log \left( \frac{f(\alpha)}{g(\alpha)} \right) \right|$$

□

Compared to the non-distributionally robust setting studied by Zhou et al. (2022); Athey & Wager (2021), the distributionally robust objective has two additional challenges:

1. The empirical process term is not simply the reward, but the log of the moment generating function.
2. There is an additional supremum over  $\alpha$ .

We now show that DR with oracle nuisances, i.e.  $W^{DR}$ , is uniformly close to the ground truth, i.e.  $W$ .

**Lemma D.5.** *Suppose Assumptions 2.1, 2.3 and D.1. Then, for any  $\beta \in (0, 1)$ , w.p.  $1 - \beta$ , we have*

$$\sup_{\pi \in \Pi} |\mathcal{V}_{\delta}^{DR}(\pi) - \mathcal{V}_{\delta}(\pi)| \leq \frac{\bar{\alpha}}{\underline{W}} \left( \frac{288}{\eta\sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{2\underline{\alpha}^2} \right) + \frac{4}{\eta\sqrt{N}} \log^{1/2}(1/\beta) \right).$$

*Proof of Lemma D.5.* By Lemma D.4,

$$\sup_{\pi \in \Pi} |\mathcal{V}_{\delta}^{DR}(\pi) - \mathcal{V}_{\delta}(\pi)| \leq \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha \log(1 + X(\pi, \alpha))|, \text{ where } X(\pi, \alpha) := \frac{W_{DR}(\pi, \alpha) - W(\pi, \alpha)}{W(\pi, \alpha)}$$

First, due to Assumption D.1, w.p. at least  $1 - \beta$ , we have  $\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |X(\pi, \alpha)| < 1/2$ . This is because the denominator is lower bounded by  $\underline{W}$  by Lemma A.2. Then, Lemma D.3 implies the numerator is bounded w.h.p. by  $\underline{W}/2$ . Hence, under this high probability event, the above expression is well-defined.

Finally, conditioning on this high-probability event, and using  $|\log(1+x)| \leq |x|$  if  $|x| < 0.8$ , we have,

$$\begin{aligned} \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha \log(1 + X(\pi, \alpha))| &\leq \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha X(\pi, \alpha)| \\ &\leq \frac{\bar{\alpha}}{\underline{W}} \left( \frac{288}{\eta \sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{2\underline{\alpha}^2} \vee 1 \right) + \frac{4}{\eta \sqrt{N}} \log^{1/2}(1/\beta) \right). \end{aligned}$$

□

**Lemma D.6.** Suppose Assumptions 2.1, 2.3 and D.1. Then, for any  $\beta \in (0, 1)$ , w.p.  $1 - 5\beta$ , we have

$$\sup_{\pi \in \Pi} \left| \mathcal{V}_\delta^{DR}(\pi) - \widehat{\mathcal{V}}_\delta^{DR}(\pi) \right| \leq \frac{2\bar{\alpha}}{\underline{W}} \left( \frac{384}{\eta \sqrt{N/K}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{\underline{\alpha}^2} \right) + \frac{8 \log^{1/2}(K/\beta)}{\eta \sqrt{N/K}} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2} \right).$$

*Proof of Lemma D.6.* By Lemma D.4,

$$\sup_{\pi \in \Pi} \left| \mathcal{V}_\delta^{DR}(\pi) - \widehat{\mathcal{V}}_\delta^{DR}(\pi) \right| \leq \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha \log(1 + Y(\pi, \alpha))|, \text{ where } Y(\pi, \alpha) := \frac{W^{DR}(\pi, \alpha) - \widehat{W}^{DR}(\pi, \alpha)}{W^{DR}(\pi, \alpha)}$$

Decompose the numerator of  $Y$  as follows, which is only possible due to the doubly robust structure,

$$\begin{aligned} &\widehat{W}^{DR}(\pi, \alpha) - W^{DR}(\pi, \alpha) \\ &= \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left( \mathbb{E}_{a \sim \pi(s_i)} \left[ \widehat{f}_0^{(k)}(s_i, a; \alpha) - f_0(s_i, a; \alpha) \right] - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \left( \widehat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right) \right) \\ &+ \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) (\exp(-r_i/\alpha) - f_0(s_i, a_i; \alpha)) \\ &+ \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) \left( f_0(s_i, a_i; \alpha) - \widehat{f}_0^{(k)}(s_i, a_i; \alpha) \right) \\ &= \frac{1}{K} \sum_{k=1}^K \mathcal{E}_1(\pi, \alpha, k) + \mathcal{E}_2(\pi, \alpha, k) + \mathcal{E}_3(\pi, \alpha, k) \end{aligned}$$

where

$$\mathcal{E}_1(\pi, \alpha, k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \mathbb{E}_{a \sim \pi(s_i)} \left[ \widehat{f}_0^{(k)}(s_i, a; \alpha) - f_0(s_i, a; \alpha) \right] - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \left( \widehat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right) \right) \quad (27)$$

$$\mathcal{E}_2(\pi, \alpha, k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) (\exp(-r_i/\alpha) - f_0(s_i, a_i; \alpha)) \quad (28)$$

$$\mathcal{E}_3(\pi, \alpha, k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) \left( f_0(s_i, a_i; \alpha) - \widehat{f}_0^{(k)}(s_i, a_i; \alpha) \right) \quad (29)$$

A key observation is that  $\widehat{f}_0^{(k)}(\cdot, \cdot; \alpha)$  and  $\widehat{\pi}_0^{(k)}(\cdot | \cdot)$  are constant on the fold they are evaluated. In other words, in  $\mathcal{D}[\mathcal{I}_k]$ ,  $\widehat{f}_0^{(k)}(s_i, a_i; \alpha)$  and  $\widehat{\pi}_0^{(k)}(a_i | s_i)$  are only functions of the current points  $s_i, a_i$ , and are independent from every other summand in the current fold (but they are not independent from the summands on folds that they were fitted on!). Then, each  $\mathcal{E}_i(\pi, \alpha, k)$  is a sum of i.i.d. random variables, and in particular zero mean random variables. This is why cross-fitting is crucial — if we didn't cross-fit,  $\widehat{f}_0^{(k)}(s_i, a_i; \alpha)$  and  $\widehat{\pi}_0^{(k)}(a_i | s_i)$  would also be functions of the rest of the dataset, which precludes the convenient independence property. Lemmas D.7 and D.8 provide bounds for  $\mathcal{E}_1, \mathcal{E}_2$  using similar Rademacher

complexity arguments. The error term  $\mathcal{E}_3$  is a product of estimation errors, which we bound directly in Lemma D.9 with the estimation rates. Putting this together gives a bound on the numerator of  $Y$ : w.p. at least  $1 - 4\beta$ :

$$\begin{aligned} & \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left| \widehat{W}^{DR}(\pi, \alpha) - W^{DR}(\pi, \alpha) \right| \\ & \leq \frac{1}{K} \sum_{k=1}^K |\mathcal{E}_1(\pi, \alpha, k)| + |\mathcal{E}_2(\pi, \alpha, k)| + |\mathcal{E}_3(\pi, \alpha, k)| \\ & \leq \frac{384}{\eta \sqrt{N/K}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{8 \log^{1/2}(K/\beta)}{\eta \sqrt{N/K}} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2} \end{aligned}$$

where we assumed  $K$  divides  $N$ , so  $|\mathcal{I}_k| = N/K$  for convenience.

We now lower bound the worst-case denominator of  $Y$ ,

$$\begin{aligned} \inf_{\pi \in \Pi} \inf_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W^{DR}(\pi, \alpha)| & \geq \inf_{\pi \in \Pi} \inf_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W(\pi, \alpha)| - |W^{DR}(\pi, \alpha) - W(\pi, \alpha)| \\ & \geq \inf_{\pi \in \Pi} \inf_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W(\pi, \alpha)| - \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |W^{DR}(\pi, \alpha) - W(\pi, \alpha)| \\ & \geq \underline{W} - \left( \frac{288}{\eta \sqrt{N}} \left( \kappa(\Pi) + \frac{\bar{\alpha}}{2\underline{\alpha}^2} \vee 1 \right) + \frac{4}{\eta \sqrt{N}} \log^{1/2}(1/\beta) \right), \end{aligned}$$

where the last inequality holds w.p. at least  $1 - \beta$ , due to Lemmas A.2 and D.3. Our assumption on  $N$  being sufficiently large (Assumption D.1) implies that the subtracted term is at most  $\underline{W}/2$ . So, the worst-case denominator of  $Y$  is lower bounded by  $\underline{W}/2$ .

Putting the two bounds together, we can bound the worst-case  $Y$ : w.p. at least  $1 - 5\beta$ ,

$$\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |Y(\pi, \alpha)| \leq \frac{2}{\underline{W}} \left( \frac{384}{\eta \sqrt{N/K}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{8 \log^{1/2}(K/\beta)}{\eta \sqrt{N/K}} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2} \right),$$

which is at most  $1/2$  when  $N$  is sufficiently large (Assumption D.1). Since  $|\log(1+x)| \leq |x|$  when  $|x| < 0.8$ , we have that,

$$\begin{aligned} & \sup_{\pi \in \Pi} \left| \mathcal{V}_\delta^{DR}(\pi) - \widehat{\mathcal{V}}_\delta^{DR}(\pi) \right| \\ & \leq \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha \log(1 + Y(\pi, \alpha))| \\ & \leq \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\alpha Y(\pi, \alpha)| \\ & \leq \frac{2\bar{\alpha}}{\underline{W}} \left( \frac{384}{\eta \sqrt{N/K}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{8 \log^{1/2}(K/\beta)}{\eta \sqrt{N/K}} + \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2} \right) \end{aligned}$$

which concludes the proof.  $\square$

**Lemma D.7.** Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1)$ , w.p.  $1 - \beta$ , we have,

$$\forall k \in [K] : \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_1(\pi, \alpha, k)| \leq \frac{192}{\eta \sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{4 \log^{1/2}(K/\beta)}{\eta \sqrt{|\mathcal{I}_k|}},$$

where  $\mathcal{E}_1$  is defined in Equation (27).

*Proof.* Let  $k \in [K]$  be fixed for now. Each summand of  $\mathcal{E}_1(\cdot, \cdot, k)$  is zero-mean, since importance sampling is unbiased. We now bound the Rademacher complexity of

$$\mathcal{F} := \left\{ (s, a) \mapsto \mathbb{E}_{\bar{a} \sim \pi(s)} \left[ \widehat{f}_0^{(k)}(s, \bar{a}; \alpha) - f_0(s, \bar{a}; \alpha) \right] - \frac{\pi(a | s)}{\pi_0(a | s)} \left( \widehat{f}_0^{(k)}(s, a; \alpha) - f_0(s, a; \alpha) \right) \mid \pi \in \Pi, \alpha \in [\underline{\alpha}, \bar{\alpha}] \right\}$$

First, we bound the envelope,

$$\begin{aligned} & \left| \mathbb{E}_{\bar{a} \sim \pi(s)} \left[ \hat{f}_0^{(k)}(s, \bar{a}; \alpha) - f_0(s, \bar{a}; \alpha) \right] - \frac{\pi(a | s)}{\pi_0(a | s)} \left( \hat{f}_0^{(k)}(s, a; \alpha) - f_0(s, a; \alpha) \right) \right| \\ & \leq 1 + \eta^{-1} \leq 2\eta^{-1} \end{aligned}$$

Now we cover in  $L_2(\mathbb{P}_{\mathcal{I}_k})$  (empirical distribution on  $\mathcal{D}[\mathcal{I}_k]$ ). So let  $\pi, \tilde{\pi} \in \Pi$  and  $\alpha, \tilde{\alpha} \in [\underline{\alpha}, \bar{\alpha}]$ , then

$$\begin{aligned} & \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \mathbb{E}_{a \sim \pi(s_i)} \left[ \hat{f}_0^{(k)}(s_i, a; \alpha) - f_0(s_i, a; \alpha) \right] - \mathbb{E}_{a \sim \tilde{\pi}(s_i)} \left[ \hat{f}_0^{(k)}(s_i, a; \tilde{\alpha}) - f_0(s_i, a; \tilde{\alpha}) \right] \right)^2} \\ & \leq \max_{i \in \mathcal{I}_k} \sum_{a \in \mathcal{A}} |\pi(a | s_i) - \tilde{\pi}(a | s_i)| \left| \hat{f}_0^{(k)}(s_i, a; \alpha) - f_0(s_i, a; \alpha) - (\hat{f}_0^{(k)}(s_i, a; \tilde{\alpha}) - f_0(s_i, a; \tilde{\alpha})) \right| \\ & \leq \max_{i \in \mathcal{I}_k} \|\pi(s_i) - \tilde{\pi}(s_i)\|_1 \cdot L|\alpha - \tilde{\alpha}| \\ & \leq 2L|\alpha - \tilde{\alpha}|, \end{aligned}$$

and

$$\begin{aligned} & \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \left( \hat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right) - \frac{\tilde{\pi}(a_i | s_i)}{\pi_0(a_i | s_i)} \left( \hat{f}_0^{(k)}(s_i, a_i; \tilde{\alpha}) - f_0(s_i, a_i; \tilde{\alpha}) \right) \right)^2} \\ & \leq \eta^{-1} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( (\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)) (\hat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha)) \right)^2} \\ & + \eta^{-1} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \tilde{\pi}(a_i | s_i) \left( \hat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) - (\hat{f}_0^{(k)}(s_i, a_i; \tilde{\alpha}) - f_0(s_i, a_i; \tilde{\alpha})) \right) \right)^2} \\ & \leq \eta^{-1} \sqrt{d_H(\pi, \tilde{\pi})} + \eta^{-1} 2L|\alpha - \tilde{\alpha}| \end{aligned}$$

Combining the two bounds, we get that the total bound is at most  $\eta^{-1} \sqrt{d_H(\pi, \tilde{\pi})} + 4\eta^{-1}|\alpha - \tilde{\alpha}|$ , so for any  $t$ , we can make  $d_H(\pi, \tilde{\pi}) \leq (t/2\eta)^2$  and  $|\alpha - \tilde{\alpha}| \leq t/8L\eta$  to bound by  $t$ . By (5.48) of (Wainwright, 2019), we have

$$\begin{aligned} \mathcal{R}_N(\mathcal{F}) & \leq \frac{24}{\sqrt{|\mathcal{I}_k|}} \int_0^{4\eta^{-1}} \log^{1/2} \left( \mathcal{N}_H((t/2\eta)^2, \Pi) \cdot \frac{4L(\bar{\alpha} - \underline{\alpha})}{t\eta} \right) dt \\ & \leq \frac{96}{\eta\sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) \end{aligned}$$

By Theorem 4.10 of (Wainwright, 2019), w.p.  $1 - \beta$ ,

$$\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_1(\pi, \alpha, k)| \leq \frac{192}{\eta\sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\bar{\alpha} \vee 1) + \frac{4 \log^{1/2}(1/\beta)}{\eta\sqrt{|\mathcal{I}_k|}}.$$

Union bound over  $k$  yields the result.  $\square$

**Lemma D.8.** Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1)$ , w.p.  $1 - \beta$ , we have,

$$\forall k \in [K] : \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_2(\pi, \alpha, k)| \leq \frac{192}{\eta\sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\alpha \vee 1) + \frac{4 \log^{1/2}(K/\beta)}{\eta\sqrt{|\mathcal{I}_k|}},$$

where  $\mathcal{E}_2$  is defined in Equation (28).

*Proof.* Let  $k \in [K]$  be fixed for now. Each summand of  $\mathcal{E}_2$  is zero-mean due to the definition of  $f_0$ . We now bound the Rademacher complexity of

$$\mathcal{F} = \left\{ (s, a, r) \mapsto \left( \frac{\pi(a | s)}{\widehat{\pi}_0^{(k)}(a | s)} - \frac{\pi(a | s)}{\pi_0(a | s)} \right) (\exp(-r/\alpha) - f_0(s, a; \alpha)) \mid \pi \in \Pi, \alpha \in [\underline{\alpha}, \bar{\alpha}] \right\}$$

First, we bound the envelope,

$$\left| \left( \frac{\pi(a | s)}{\widehat{\pi}_0^{(k)}(a | s)} - \frac{\pi(a | s)}{\pi_0(a | s)} \right) (\exp(-r/\alpha) - f_0(s, a; \alpha)) \right| \leq 2\eta^{-1}.$$

Now, we cover in  $L_2(\mathbb{P}_{\mathcal{I}_k})$  (empirical distribution on  $\mathcal{D}[\mathcal{I}_k]$ ). So let  $\pi, \tilde{\pi} \in \Pi, \alpha, \tilde{\alpha} \in [\underline{\alpha}, \bar{\alpha}]$ , then

$$\begin{aligned} & \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) \exp(-r_i/\alpha) - \left( \frac{\tilde{\pi}(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\tilde{\pi}(a_i | s_i)}{\pi_0(a_i | s_i)} \right) \exp(-r_i/\tilde{\alpha}) \right)^2} \\ & \leq 2\eta^{-1} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} (\pi(a_i | s_i) \exp(-r_i/\alpha) - \tilde{\pi}(a_i | s_i) \exp(-r_i/\tilde{\alpha}))^2} \\ & \leq 2\eta^{-1} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} ((\pi(a_i | s_i) - \tilde{\pi}(a_i | s_i)) \exp(-r_i/\alpha))^2} \\ & + 2\eta^{-1} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} (\tilde{\pi}(a_i | s_i) (\exp(-r_i/\alpha) - \exp(-r_i/\tilde{\alpha})))^2} \\ & \leq 2\eta^{-1} \sqrt{d_H(\pi, \tilde{\pi})} + 2\eta^{-1} L |\alpha - \tilde{\alpha}| \end{aligned}$$

Replacing  $\exp(-r_i/\alpha)$  by  $f_0(s_i, a_i; \alpha)$  in the above arguments yields the same bound, since  $f_0$  is also  $L$ -Lipschitz in  $\alpha$  (Lemma D.2). Thus, the total distance bound is  $4\eta^{-1}(\sqrt{d_H(\pi, \tilde{\pi})} + L|\alpha - \tilde{\alpha}|)$ . So for any  $t$ , we can make  $d_H(\pi, \tilde{\pi}) \leq (t/8\eta)^2$  and  $|\alpha - \tilde{\alpha}| \leq t/8L\eta$  to bound by  $t$ . By (5.48) of (Wainwright, 2019), we have

$$\begin{aligned} \mathcal{R}_N(\mathcal{F}) & \leq \frac{24}{\sqrt{|\mathcal{I}_k|}} \int_0^{4\eta^{-1}} \log^{1/2} \left( \mathcal{N}_H((t/8\eta)^2, \Pi) \cdot \frac{4L(\bar{\alpha} - \underline{\alpha})}{t\eta} \right) dt \\ & \leq \frac{96}{\eta\sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\bar{\alpha} \vee 1). \end{aligned}$$

By Theorem 4.10 of (Wainwright, 2019), w.p.  $1 - \beta$ ,

$$\sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_2(\pi, \alpha, k)| \leq \frac{192}{\eta\sqrt{|\mathcal{I}_k|}} (\kappa(\Pi) + L\alpha \vee 1) + \frac{4 \log^{1/2}(1/\beta)}{\eta\sqrt{|\mathcal{I}_k|}}.$$

Union bound over  $k$  yields the result.  $\square$

**Lemma D.9.** Suppose Assumptions 2.1 and 2.3. Then, for any  $\beta \in (0, 1/2)$ , w.p.  $1 - 2\beta$ , we have,

$$\forall k \in [K] : \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_3(\pi, \alpha, k)| \leq \frac{\text{Rate}_{\pi_0}(N, \beta/K) \cdot \text{Rate}_f^c(N, \beta/K)}{\eta^2},$$

where  $\mathcal{E}_3$  is defined in Equation (29).

*Proof.* Let  $k \in [K]$  be fixed first.

$$\begin{aligned}
 & \mathbb{E} \left[ \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} |\mathcal{E}_3(\pi, \alpha, k)| \right] \\
 &= \mathbb{E} \left[ \sup_{\pi \in \Pi} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left| \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{\pi(a_i | s_i)}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{\pi(a_i | s_i)}{\pi_0(a_i | s_i)} \right) \left( \widehat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right) \right| \right] \\
 &\leq \mathbb{E} \left[ \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{1}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{1}{\pi_0(a_i | s_i)} \right)^2} \cdot \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \sqrt{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \widehat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right)^2} \right] \\
 &\leq \sqrt{\mathbb{E} \left[ \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \frac{1}{\widehat{\pi}_0^{(k)}(a_i | s_i)} - \frac{1}{\pi_0(a_i | s_i)} \right)^2 \right]} \cdot \sqrt{\mathbb{E} \left[ \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left( \widehat{f}_0^{(k)}(s_i, a_i; \alpha) - f_0(s_i, a_i; \alpha) \right)^2 \right]}
 \end{aligned}$$

Since  $\mathbb{E} \left[ \sup \frac{1}{N} \sum (\cdot) \right] \leq \frac{1}{N} \sum \mathbb{E} [\sup(\cdot)]$ ,

$$\leq \sqrt{\mathbb{E} \left[ \left( \frac{1}{\widehat{\pi}_0^{(k)}(A | S)} - \frac{1}{\pi_0(A | S)} \right)^2 \right]} \cdot \sqrt{\mathbb{E} \left[ \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left( \widehat{f}_0^{(k)}(S, A; \alpha) - f_0(S, A; \alpha) \right)^2 \right]}$$

Using definition of estimation rates, and the fact that  $\widehat{\pi}_0^{(k)}, \widehat{f}_0^{(k)}(\cdot; \alpha)$  were trained on  $N - |\mathcal{I}_k| = N(1 - 1/K)$  data points (due to cross-fitting), we have w.p.  $1 - 2\beta$ ,

$$\leq \frac{1}{\eta^2} \text{Rate}_{\pi_0}(N, \beta) \cdot \text{Rate}_f^c(N, \beta)$$

Finally apply union bound over  $k$ . □

### D.1. Point-wise rate to uniform rate for Lipschitz regressions

We now show that when the target function is Lipschitz on a compact domain, point-wise rates can be translated into uniform rates. Let  $\widehat{f}(x; \alpha)$  be estimates of  $f(x; \alpha)$ , where  $\alpha \in [0, b]$  for some  $b \in \mathbb{R}$ . Supposing that  $\widehat{f}$  is learned on a random sample of  $N$  datapoints, we define the point-wise convergence rate such that for any  $\alpha \in [0, b]$ , for any  $\beta \in (0, 1)$ , w.p. at least  $1 - \beta$ , we have

$$\|\widehat{f}(x; \alpha) - f(x; \alpha)\|_{L_2(\mathbb{P}_0)} \leq \text{Rate}_{point}(N, \beta).$$

Define the uniform rate so that for any  $\beta$ , w.p.  $1 - \beta$ , we have

$$\left\| \sup_{\alpha \in [0, b]} \widehat{f}(x; \alpha) - f(x; \alpha) \right\|_{L_2(\mathbb{P}_0)} \leq \text{Rate}_{unif}(N, \beta).$$

**Lemma D.10.** Suppose  $\widehat{f}, f$  are both  $L$ -Lipschitz. Then, for any  $\beta$ , w.p.  $1 - \beta$ , we have

$$\text{Rate}_{unif}(N, \beta) \leq \inf_{0 < d \leq b} 2dL + \text{Rate}_{point}(N, 2d\beta/(b + 2d))$$

*Proof.* The idea is to ensure we have point-wise guarantees at the crucial grid points, placed at distance  $2d$  apart from each other from  $[0, b]$ , similar to Lemma C.1 of (Oprescu et al., 2019). So there are at most  $\lceil b/2d \rceil$  points. Let  $n(\alpha)$  denote the closest grid point, so we have  $|\alpha - n(\alpha)| \leq d$  for any  $\alpha$ . Also at each grid point, we have  $\|\widehat{f}(x; \alpha) - f(x; \alpha)\|_{L_2(\mathbb{P}_0)} \leq$

$\text{Rate}_{point}(N, \beta / \lceil b/2d \rceil) \leq \text{Rate}_{point}(N, 2d\beta/(b+2d))$ . Hence, w.p.  $1 - \beta$ ,

$$\begin{aligned}
 & \left\| \sup_{\alpha \in [0, b]} \widehat{f}(x; \alpha) - f(x; \alpha) \right\|_{L_2(\mathbb{P}_0)} \\
 & \leq \left\| \sup_{\alpha \in [0, b]} \left| \widehat{f}(x; \alpha) - \widehat{f}(x; n(\alpha)) \right| + \left| \widehat{f}(x; n(\alpha)) - f(x; n(\alpha)) \right| + |f(x; n(\alpha)) - f(x; \alpha)| \right\|_{L_2(\mathbb{P}_0)} \\
 & \leq \left\| \sup_{\alpha \in [0, b]} \widehat{f}(x; \alpha) - \widehat{f}(x; n(\alpha)) \right\|_{L_2(\mathbb{P}_0)} + \left\| \sup_{\alpha \in [0, b]} \widehat{f}(x; n(\alpha)) - f(x; n(\alpha)) \right\|_{L_2(\mathbb{P}_0)} + \left\| \sup_{\alpha \in [0, b]} f(x; n(\alpha)) - f(x; \alpha) \right\|_{L_2(\mathbb{P}_0)} \\
 & \leq 2dL + \text{Rate}_{point}(N, 2d\beta/(b+2d))
 \end{aligned}$$

□

Typically, the point-wise rate guarantees are of the form  $\text{Rate}_{point}(N, \beta) = C(\frac{1}{N^p} + \sqrt{\log(1/\beta)/N})$  (Wainwright, 2019; Bartlett et al., 2005). In this case, setting  $d = \frac{1}{N^p}$  gives the guarantee that

$$\text{Rate}_{unif}(N, \beta) \leq \frac{C + 2L}{N^p} + \sqrt{\frac{\log(N^p(b + 2/N^p)/2\beta)}{N}},$$

which is  $\mathcal{O}(\sqrt{\log(N)/N})$  if  $p = 1/2$  and  $\mathcal{O}(N^{-p})$  if  $p < 1/2$ . Hence, when the regression target is Lipschitz, the uniform guarantee has the same rate as pointwise if the rate is non-parametric (slower than  $\sqrt{N}$ ). And when the pointwise rate is a parametric  $\sqrt{N}$  rate, then the uniform rate only incurs an extra  $\sqrt{\log(N)}$  factor.