Batch Learning via Log-Sum-Exponential Estimator from Logged Bandit Feedback

Armin Behnamnia*
Department of Computer Engineering
Sharif University of Technology
armin.behnamnia@sharif.edu

Gholamali Aminian*
The Alan Turing Institute
g.aminian@turing.ac.uk

Alireza Aghaei
Department of Computer Engineering
Sharif University of Technology
alireza.aghaei123@sharif.edu

Chengchun Shi
Department of Statistic
London School of Economics
c.shi7@lse.ac.uk

Vincent Y. F. Tan
Department of Electrical and Computer Engineering
National University of Singapore
vtan@nus.edu.sg

Hamid R. Rabiee
Department of Computer Engineering
Sharif University of Technology
rabiee@sharif.edu

Abstract

Offline policy learning methods in batch learning aim to derive a policy from a logged bandit feedback dataset, encompassing context, action, propensity score, and feedback for each sample point. Inverse propensity score estimators are employed to minimize the cost to achieve this objective. However, this approach is susceptible to high variance and poor performance under low-quality propensity scores. In response to these limitations, we propose a novel estimator inspired by the log-sum-exponential operator, mitigating variance. Furthermore, we offer theoretical analysis, encompassing upper bounds on the bias, variance of our estimator, and an upper bound on the generalization error of the log-sum-exponential estimator—the difference between the empirical risk of the log-sum-exponential estimators and the true risk- with a convergence rate of $O(1/\sqrt{n})$ where n is the number of training samples. Additionally, we examine the performance of our estimator under limited access to clean propensity scores and an imbalanced logged bandit feedback dataset, where the number of samples per action is different. The code for our experiments is available at https://github.com/Slifer-The-Sky-Dragon/LSE_Code.

1 Introduction

Offline policy learning from logged data is an important problem in reinforcement learning theory and practice. The logged dataset represents interaction logs of a system with its environment, recording context, action, propensity score (i.e., the probability of action selection for a given context under the logging policy), and feedback (cost). The literature has considered this setting in the context of contextual bandits and partially labeled observations. It is used in many real applications, e.g., recommendation systems [Aggarwal, 2016, Li et al., 2011], personalized medical treatments [Kosorok and Laber, 2019, Bertsimas et al., 2017], and personalized advertising campaigns [Tang et al., 2013, Bottou et al., 2013]. However, there are two main obstacles to learning from this kind of logged bandit feedback (LBF) dataset: first, the observed feedback (cost) is available for the chosen action

[*] Equal Contribution.

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

only, and second, the LBF dataset is taken under the logging policy so that it could be biased. Batch learning with LBF (a.k.a. Counterfactual Risk Minimization) is a strategy for off-policy learning from LBF datasets, which has been proposed by Swaminathan and Joachims [2015a] to tackle these challenges.

Due to the bias of logging policy in the LBF dataset, the inverse propensity score (IPS) estimator is proposed [Thomas et al., 2015, Swaminathan and Joachims, 2015a]. However, this method suffers from significant variance in many cases [Rosenbaum and Rubin, 1983]. To address this, some truncated importance sampling methods have been proposed, such as the IPS estimator with the truncated ratio of policy and logging policy [Ionides, 2008a], IPS estimator with truncated propensity score [Strehl et al., 2010], self-normalizing estimator [Swaminathan and Joachims, 2015b], exponential smoothing (ES) estimator [Aouali et al., 2023], implicit exploration (IX) estimator [Neu, 2015] and power-mean (PM) estimator [Metelli et al., 2021].

In addition to the significant variance issue of IPS estimators, there are two more challenges in real problems: noisy propensity scores and imbalanced LBF dataset. In particular, in previous works such as Swaminathan and Joachims [2015a], Metelli et al. [2021], Aouali et al. [2023], it is assumed that the propensity scores in the LBF dataset are true values. However, access to the exact values of the propensity scores may not be possible, for example, when the LBF dataset is annotated by human agents. In this situation, one may settle for a qualitative estimation of the propensity score or train a model to estimate the propensity scores. In either case, the propensity score stored in the LBF dataset can be considered a noisy version of the true propensity score. In addition to noisy propensity scores, we can encounter an imbalanced LBF dataset, particularly due to the logging policy's tendency to focus on specific actions more. For example, in recommendation systems, people may be biased towards middle-range ratings and mostly avoid the highest and lowest scores in their ratings. This introduces the concept of action imbalance in bandit learning from logged data, where there is no opportunity to explore minority actions in offline policy learning. Therefore, there is a need for an estimator that can effectively manage variance, noisy propensity scores, and imbalanced LBF datasets.

In this work, we propose a novel estimator for batch learning from the LBF dataset, which is shown to have better performance under the noisy propensity score and imbalance scenarios compared to other estimators. The contributions of our work are as follows.

- We propose a novel (non-linear) estimator inspired by the Log-Sum-Exponential (LSE) operator as an LSE estimator, which mitigates variance and can be applied to unbounded cost functions.
- We provide a bias and variance analysis of our LSE estimator. We also propose bounds on the generalization error, i.e. the absolute difference between the LSE estimation and the true average cost. We provide generalization error bounds with a convergence rate of $O(1/\sqrt{n})$ where n is the number of training samples under mild assumptions.
- Motivated by our theoretical analysis, we introduce a novel regularization based on α-Rényi divergence. This regularization reduces the variance of the LSE estimator and demonstrates superior performance compared to the LSE estimator without regularization.
- We introduce a set of experiments conducted on different datasets to show the performance of the LSE and α-Rényi regularized LSE estimators in scenarios with clean and noisy propensity scores and imbalance LBF datasets based on different numbers of samples per action in comparison with other estimators.

Notation: We adopt the following convention for random variables and their distributions in the sequel. A random variable is denoted by an upper-case letter (e.g., Z), an arbitrary value of this variable is denoted with the lower-case letter (e.g., z), and its space of all possible values with the corresponding calligraphic letter (e.g., $\mathcal{Z}$). This way, we can describe generic events like $\{Z = z\}$ for any $z \in \mathcal{Z}$, or events like $\{g(Z) \leq 5\}$ for functions $g : \mathcal{Z} \to \mathbb{R}$. P_Z denotes the probability distribution of the random variable Z. The joint distribution of a pair of random variables (Z_1, Z_2) is denoted by P_{Z_1,Z_2}. We denote the set of integer numbers from 1 to n by $[n] \triangleq \{1, \cdots, n\}$. In this work, we consider the natural logarithm, i.e., $\log(x) := \log_e(x)$.

Divergences: Suppose $p(x)$ and $q(x)$ are arbitrary distributions defined on the same space, and $\alpha > 0$. If $\alpha \geq 1$, we should have $q(x) > 0$ if $p(x) > 0$. The α-Rényi Divergence [Van Erven and Harremos, 2014], between $p(x)$ and $q(x)$ is defined as, $D_\alpha(p||q) = \frac{1}{\alpha-1} \log \int_x p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x$.

For $\alpha \to 1$, D_α reduces to the KL divergence, i.e., $D_{\mathrm{KL}}(p||q) = \int_x p(x)\log(p(x)/q(x))\mathrm{d}x$. We also define power divergence as $P_\alpha(p\|q) := \exp(D_\alpha(p\|q))^{\alpha-1}$ is the power divergence with order α.

2 Problem formulation

Let $\mathcal{X}$ be the set of contexts and $\mathcal{A}$ the set of actions. We consider policies as conditional distributions over actions, given contexts. For each pair of context and action $(x,a) \in \mathcal{X} \times \mathcal{A}$ and policy $\pi \in \Pi$, where Π is the set of policies, the value $\pi(a|x)$ is defined as the conditional probability of choosing action a given context x under the policy π.

Inspired by Swaminathan and Joachims [2015a], a cost (loss) function* $c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^-$, which is unknown, defines the cost (feedback) of each observed pair of context and action. However, in the LBF setting, we only observe the cost (feedback) for the chosen action a in a given context x, under the logging policy $\pi_0(a|x)$. We have access to the LBF dataset $S = (x_i, a_i, p_i, c_i)_{i=1}^n$ where each 'data point' (x_i, a_i, p_i, c_i) contains the context x_i which is sampled from unknown distribution P_X, the action a_i which is sampled from the logging policy $\pi_0(\cdot|x_i)$, the propensity score $p_i \triangleq \pi_0(a_i|x_i)$, and the observed feedback (cost) $c_i \triangleq c(x_i, a_i)$ under logging policy $\pi_0(a_i|x_i)$.

The *true risk* of a learning policy π_θ is,

$$R(\pi_\theta) = \mathbb{E}_{P_X}[\mathbb{E}_{\pi_\theta(A|X)}[c(A,X)|X]]. \tag{1}$$

Our objective is to find an optimal $\pi_\theta^\star$, one which minimizes $R(\pi_\theta)$, i.e., $\pi_\theta^\star = \arg\min_{\pi_\theta \in \Pi_\theta} R(\pi_\theta)$, where Π_θ is the set of all policies parameterized by $\theta \in \Theta$. We denote the importance weighted cost function as $w_\theta(A,X)c(A,X)$, where

$$w_\theta(A,X) = \frac{\pi_\theta(A|X)}{\pi_0(A|X)}.$$

As discussed by Swaminathan and Joachims [2015b], we can apply the IPS estimator over the LBF dataset S [Rosenbaum and Rubin, 1983] to get an unbiased estimator of the risk (a *linear empirical risk*) by considering the importance weighted cost function as,

$$\hat{R}(\pi_\theta, S) = \frac{1}{n}\sum_{i=1}^n c_i w_\theta(a_i, x_i), \tag{2}$$

where $w_\theta(a_i, x_i) = \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}$. The IPS estimator is unbiased with bounded variance if the $\pi_\theta(A|X)$ is absolutely continuous with respect to $\pi_0(A|X)$ [Strehl et al., 2010, Langford et al., 2008].

To mitigate the problem of the large variance of the IPS estimator, many estimators have been proposed [Strehl et al., 2010, Ionides, 2008a, Swaminathan and Joachims, 2015b, Aouali et al., 2023, Metelli et al., 2021, Neu, 2015], including truncated IPS, exponential smoothing (ES) [Aouali et al., 2023] and power-mean (PM) [Metelli et al., 2021] estimators. Note that we can represent the linear empirical risk of these estimators as the weighted average of cost (feedback) where the weights are defined by a transformation of $w_\theta(a_i, x_i)$,

$$\hat{R}(\pi_\theta, S) = \frac{1}{n}\sum_{i=1}^n c_i g\big(w_\theta(a_i, x_i)\big), \tag{3}$$

where $g : \mathbb{R} \to \mathbb{R}$ is defined for different estimators. For example, we have $g(x) = x$ in the IPS estimator, $g(x) = \max(x, M)$ in the truncated IPS estimator, $g(x) = ((1-\lambda)x^s + \lambda)^{1/s}$ in the PM estimator and $g(x) = x^\beta$ for $\beta \in (0,1)$ in the ES estimator. It is worth mentioning that another version of the ES estimator is proposed as $g(x) = \pi_0^\alpha x$ for $\alpha \in (0,1)$. For the IX-estimator with parameter η, we have $g(x) = \frac{x}{1+\eta/\pi_0}$. The generalization error for an estimator is defined as the difference between the true risk and the empirical risk.

In the next section, we propose the LSE estimator as a non-linear estimator to learn a policy with a ~~low variance that minimizes the true~~ risk using the LBF dataset.

[*] The cost can be viewed as the opposite (negative) of the reward. Consequently, a low cost (equivalent to maximum reward) signifies user (context) satisfaction with the given action, and conversely. For the reward function, we have $r(x,a) = -c(x,a)$.

3 Log-Sum-Exponential estimator

Main idea: Inspired by the log-sum-exponential operator with applications in multinomial linear regression and naive Bayes classifiers [Calafiore et al., 2019, Murphy, 2012, Williams and Barber, 1998], we define the LSE estimator with parameter λ,

$$\mathrm{LSE}_\lambda(\mathbf{Z}) = \frac{1}{\lambda}\log\Big(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda z_i}\Big), \tag{4}$$

where $\mathbf{Z} = \{z_i\}_{i=1}^n$ are samples from the random variable Z. The key property of the LSE operator is its robustness to outliers and big errors in a limited number of data samples. Here an outlier, by intuition, is a point with abnormally large negative z_i. Such points vanish in the exponential sum as $\lim_{z_i \to -\infty} e^{\lambda z_i} = 0$. Therefore the LSE operator ignores terms with large values.

Motivating example: We provide an example to investigate the behaviour of LSE as a general estimator and its difference from Monte-Carlo estimator (a.k.a. simple average). Suppose that Z is distributed as a Pareto distribution* with scale x_m and shape ζ. Let $\zeta = 1.5$ and x_m be such that $\mathbb{E}[Z] = -\frac{\zeta x_m}{\zeta - 1} = -1$. The objective is to estimate $\mathbb{E}[Z]$ with independent samples drawn from the Pareto distribution. We set $n = 10, 50, 100, 1000, 10,000$ and compute the Monte-Carlo (a.k.a. simple average) and LSE estimation of the expectation of Y. Table 1 shows that LSE effectively keeps the variance low without significant side-effects on bias.

Table 1: Bias, variance, and MSE of LSE and Monte-Carlo estimators. We run the experiment 10,000 times and find the variance, bias, and MSE of the estimations.

	Estimator	$n = 10$	$n = 50$	$n = 100$	$n = 1000$	$n = 10000$
Bias	Monte-Carlo	0.276	0.079	0.167	0.05	0
	LSE	0.797	0.594	0.518	0.312	0.167
Variance	Monte-Carlo	23.726	62.891	12.66	10.332	8.928
	LSE	0.206	0.176	0.165	0.105	0.071
MSE	Monte-Carlo	23.802	62.897	12.688	10.334	8.928
	LSE	0.841	0.529	0.433	0.202	0.098

Risk functions: The LSE estimator has a tunable parameter λ which helps us to recover the IPS estimator for $\lambda \to 0$. Furthermore, the LSE estimator is an increasing function with respect to λ. We study the LSE estimator properties in App.C. The LSE empirical risk (LSER) is defined as

$$\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S, \pi_\theta) := \mathrm{LSE}_\lambda(S) = \frac{1}{\lambda}\log\Big(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda c_i w_\theta(a_i, x_i)}\Big), \tag{5}$$

which is supposed to estimate the true risk. As previous works consider the deviation of the empirical risk from the true risk, we also examine the deviation of the LSER from the true risk. For this purpose, we define the *generalization error*, as the difference between the true risk and the LSER, i.e.,

$$\mathrm{gen}_\lambda(\pi_\theta) := R(\pi_\theta) - \hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S, \pi_\theta). \tag{6}$$

Furthermore, we are interested in providing high probability upper and lower bounds on $\mathrm{gen}_\lambda(\pi_\theta)$,

$$P(\mathrm{gen}_\lambda(\pi_\theta) > g_u(\delta, n, \lambda)) \le \delta, \quad \text{and,} \quad P(\mathrm{gen}_\lambda(\pi_\theta) < g_l(\delta, n, \lambda)) \le \delta.$$

where $0 < \delta < 1$ and n is the number of samples. The derivative of the LSER can be represented as,

$$\nabla_\theta \hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S, \pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} c_i e^{\lambda(c_i w_\theta(a_i, x_i) - \hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S, \pi_\theta))} \nabla_\theta w_\theta(a_i, x_i). \tag{7}$$

Note that, in (7), we have a weighted average of the gradient of the propensity-weighted cost samples. In contrast to the linear empirical risk for which the gradient is a uniform mean of cost samples, in the LSE estimator, the gradient for large values of $c_i w_\theta(a_i, x_i)$, $\forall i \in [n]$ (small absolute value), contributes more to the final gradient. It can be interpreted as the robustness of the LSE estimator with respect to the very large absolute values of $c_i w_\theta(a_i, x_i)$ (i.e. high $w_\theta(a, x)$), $\forall i \in [n]$.

*If $Z \sim \mathrm{Pareto}(x_m, \zeta)$, we have $f_Z(z) = \frac{\zeta x_m^\zeta}{z^{\zeta+1}}$.

4 Related works

Direct method: The direct method for off-policy learning from the LBF datasets is based on the estimation of the cost function, followed by the application of a supervised learning algorithm to the problem [Dudík et al., 2014]. However, this approach does not generalize well, as shown by Beygelzimer and Langford [2009]. A different approach based on policy optimization and boosted base learner is proposed to improve the performance in direct methods [London et al., 2023]. Our approach differs from this area, as we do not estimate the cost function.

Estimation of propensity scores: We can estimate the propensity score using different methods, e.g., logistic regression [D'Agostino Jr, 1998, Weitzen et al., 2004], generalized boosted models [McCaffrey et al., 2004], neural networks [Setoguchi et al., 2008], parametric modeling [Xie et al., 2019a] or classification and regression trees [Lee et al., 2010, 2011]. Note that, as discussed in [Tsiatis, 2006, Shi et al., 2016], under the estimated propensity scores (noisy propensity score), the variance of the IPS estimator is reduced. In this work, we consider both clean and noisy propensity scores.

Imbalanced batch learning: Dai et al. [2023] used an ensemble method to avoid the bias of the model towards the minority classes. They train different resampled subsets of the training data and classifier models and resampling techniques during training and validate each combination by the samples that were left out of the sampled subset. Then they use the model with the highest performance on the test data. [Hong et al., 2023] handle the imbalanced dataset in offline RL by reweighting the samples during the training to fit to the high-return samples instead of many low-return samples in the dataset. Although the notion of imbalance in this work is different from ours, the same idea of reweighting samples from minority classes can be used with appropriate objectives instead of simple return values. In addition, Zhu et al. [2024] incorporates the median-of-means estimator to estimate the mean of the heavy-tailed distribution of cost (reward) in offline RL that shares the same challenge of the high variance In our work, we model the imbalance LBF dataset as non-equal samples per action.

5 Theoretical foundations of the LSE estimator

In this section, assuming the LSE estimator, we present the bias-variance, and generalization error analyses. Based on our theoretical results, we propose the α-Rényi divergence as a regularization for the LSE estimator. All the proof details are deferred to App.D. In this section, the following assumptions are made.

Assumption 5.1 (Bounded true risk)**.** The learning policy $\pi_\theta(A|X)$, cost function $c(A,X)$ and P_X are such that the expected true risk satisfies $\mathbb{E}_{P_X\otimes\pi_\theta(A|X)}[c(A,X)] = R_\theta < \infty$.

Assumption 5.2 (Bounded Variance)**.** The variance of the weighted cost function is bounded,

$$\mathbb{V}(w_\theta(A,X)c(A,X)) \le M. \tag{8}$$

In comparison with bounded cost function assumption in literature, [Metelli et al., 2021, Aouali et al., 2023], our theoretical results can be applied to unbounded cost function satisfying Assumption 5.1 and Assumption 5.2. Moreover, our assumptions are weaker with respect to the uniform overlap assumption [*].

5.1 Bias-variance

One of the evaluation metrics for an estimator is the mean squared error (MSE) which is decomposed into squared bias and the variance of the estimator. For the LSE estimator, we consider the following MSE decomposition,

$$\mathrm{MSE}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) = \mathbb{B}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta))^2 + \mathbb{V}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)),$$

where,

$$\begin{aligned}\mathbb{B}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) &= \mathbb{E}[w_\theta(A,X)c(A,X)] - \mathbb{E}[\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)],\\ \mathbb{V}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) &= \mathbb{E}[(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta) - \mathbb{E}[\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)])^2].\end{aligned}$$

In this section, we will use the following helpful lemma to prove some results.

[*] In the uniform coverage (overlap) assumption, it is assumed that $\sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} w_\theta(a,x) = U_c < \infty$.

Lemma 5.3. *Given bounded variance on $X < 0$, $\mathbb{V}(X) < \infty$, the following upper bound holds on the variance of $e^{\lambda X}$ for $\lambda > 0$,*

$$\mathbb{V}\left(e^{\lambda X}\right) \le \lambda^2 \mathbb{V}(X). \tag{9}$$

In the following proposition, we provide a bias-variance analysis of the LSER.

Proposition 5.4 (Bias bound)**.** *Given Assumptions 5.1 and 5.2, the following bound holds on the bias of the LSER,*

$$\frac{n-1}{2n\lambda}\mathbb{V}(e^{\lambda c(A,X) w_\theta(A,X)}) \le \mathbb{B}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \le \frac{\lambda}{2}(M + R_\theta^2) - \frac{1}{2n\lambda}\mathbb{V}(e^{\lambda w_\theta(A,X) c(A,X)}).$$

Remark 5.5 (Asymptotically Unbiased). By selecting λ as a function of n, which tends to zero as $n \to \infty$, e.g. $\lambda(n) = \frac{1}{\sqrt{n}}$, the aforementioned upper bound in Proposition 5.4 becomes asymptotically zero. Since the lower bound is always positive, this forces both the upper and lower bounds to converge to zero. Consequently, we can prove that the LSE is asymptotically unbiased.

For the variance of LSER, we can provide the following upper bound.

Proposition 5.6 (Variance Bound)**.** *Given Assumptions 5.1 and 5.2 and assuming that $0 < \lambda < -2\frac{R_\theta}{M+R_\theta^2}$, the variance of the LSER satisfies,*

$$\mathbb{V}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \le \frac{M}{n} + \frac{\lambda}{2}\left(-2R_\theta - \frac{\lambda}{2}(M + R_\theta^2)\right)(M + R_\theta^2). \tag{10}$$

Remark 5.7 (Convergence rate of the variance bound). Selecting $\lambda = \frac{1}{n}$, the abovementioned proposition gives an upper bound of $O(\frac{1}{n})$ for the variance of LSE. Generally when $\lambda \to 0$ we have,

$$\lim_{\lambda \to 0} \mathbb{V}(\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \le \frac{\mathbb{V}(w_\theta(A,X)c(A,X))}{n}. \tag{11}$$

5.2 Generalization error bounds

In this section, we provide lower and upper bounds on the generalization error of the LSE estimator. Note that $\mathrm{gen}\lambda(\pi\theta)$ is a non-linear estimator with respect to the weighted cost, i.e., $w_\theta(A,X)c(A,X)$, which is different from linear estimators. Therefore, the previous techniques for generalization error analysis under linear estimators are not applicable. A Regret bound is also provided in the App. F.

Theorem 5.8 (Generalization Bound)**.** *Given Assumptions 5.1 and 5.2, then with probability at least $1-\delta$ we have,*

$$\mathfrak{L}(\gamma, n, \lambda, R_\theta, M, \delta) \le \mathrm{gen}_\lambda(\pi_\theta) \le \mathfrak{U}(\gamma, n_u, \lambda, R_\theta, M, \delta),$$

where $0 < \gamma < 1$ and $n_u = n$ for $n_u \ge \frac{\left(2\lambda^2 M + \frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2 \exp(2\lambda R_\theta)}$,

$$\begin{aligned} \mathfrak{L}(\gamma, n, \lambda, R_\theta, M, \delta) &:= \frac{-\lambda}{2}(M + R_\theta^2) - \frac{2}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} - \sqrt{\frac{2M\log\frac{2}{\delta}}{n\exp(2\lambda R_\theta)}}, \\ \mathfrak{U}(\gamma, n_u, \lambda, R_\theta, M, \delta) &:= \frac{2}{3(1-\gamma)}\frac{\log\frac{2}{\delta}}{n_u\lambda\exp(\lambda R_\theta)} + \frac{1}{1-\gamma}\sqrt{\frac{2M\log\frac{2}{\delta}}{n_u\exp(2\lambda R_\theta)}}. \end{aligned} \tag{12}$$

Remark 5.9 (Uniform Coverage Assumption). Theorem 5.8 does not require any assumption regarding the coverage or overlap, such as the uniform coverage assumption[*], of the learning policy with respect to the logging policy.

The generalization bound depends on λ. With an appropriate selection of λ, we can have an upper bound on generalization error with a convergence rate of $\frac{1}{\sqrt{n}}$.

Proposition 5.10 (Selection of λ)**.** *Given Assumptions 5.1 and 5.2 and assuming $n \ge \max\left(\frac{8\log\frac{2}{\delta}}{3(M+R_\theta^2)\exp(R_\theta)}, \frac{\left(8M+\frac{8}{3}\right)\log\frac{2}{\delta}}{\exp(2R_\theta)}\right)$, and setting $\lambda = \sqrt{\frac{8\log\frac{2}{\delta}}{3n(M+R_\theta^2)\exp(R_\theta)}}$, then with probability of at least $1-\delta$ ($\delta \in (0,1)$), the generalization error satisfies,*

$$\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \le 2\sqrt{\frac{2(M+R_\theta^2)\log\frac{2}{\delta}}{n\exp(R_\theta)}}\left(\sqrt{\frac{M}{\exp(R_\theta)(M+R_\theta^2)}} + \frac{1}{\sqrt{3}}\right).$$

The robustness of the LSE estimator with respect to noisy propensity scores is studied in App. G.

[*] In the uniform coverage (overlap) assumption, it is assumed that $\sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} w_\theta(a,x) = U_c < \infty$.

5.3 Comparison with previous works

The comparison of our LSE estimator with other estimators, including, IPS, self-normalized IPS [Swaminathan and Joachims, 2015b], truncated IPS with weight truncation parameter M, ES-estimator with parameter α [Aouali et al., 2023], IX-estimator with parameter η and PM-estimator with parameter λ [Metelli et al., 2021] is provided in Table 2.

Table 2: Comparison of estimators. We consider the bounded cost function, i.e., $R_{\max} := \sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} r(a,x)$. $\mathbb{B}^{\mathrm{SN}}$ and $\mathbb{V}^{\mathrm{SN}}$ are the Bias and the Efron-Stein estimate of the variance of self-normalized IPS. For the ES-estimator, we have $T^{ES} = \mathbb{B}^{ES} + (1/n)\big(D_{\mathrm{KL}}(\pi_\theta\|\pi_0) + \log(4/\delta)\big)$. For the IX-estimator, $C_\eta(\pi)$ is the smoothed policy coverage ratio. We compare the convergence rate of the generalization error for estimators.

Estimator	Max Abs	Variance	Bias	Generalization Error	Convergence Rate Order
IPS	$R_{\max}\operatorname{esssup}\frac{\pi_\theta}{\pi_0}$	$\frac{R_{\max}P_2(\pi_\theta\|\pi_0)}{n}$	0	$R^2_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)}{\delta n}}$	$O(n^{-1/2})$
SN-IPS [Swaminathan and Joachims, 2015b]	$R_{\max}$	$R^2_{\max}V^{\mathrm{SN}}$	$R_{\max}B^{\mathrm{SN}}$	$R_{\max}(B^{\mathrm{SN}} + \sqrt{V^{\mathrm{ES}}\log\frac{1}{\delta}})$	-
IPS-TR $(M>0)$ [Ionides, 2008b]	$R_{\max}M$	$R^2_{\max}\frac{P_2(\pi_\theta\|\pi_0)}{n}$	$R_{\max}\frac{P_2(\pi_\theta\|\pi_0)}{M}$	$R_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)\log\frac{1}{\delta}}{n}}$	$O(n^{-1/2})$
IX (η) [Gabbianelli et al., 2023]	$\frac{R_{\max}}{\eta}$	$R_{\max}C_\eta(\pi_\theta)/n$	$R_{\max}\eta C_\eta(\pi_\theta)$	$R_{\max}(2\eta C_\eta(\pi_\theta) + \frac{\log(2/\delta)}{\eta n})$	$O(n^{-1/2})$
PM $(\lambda\in[0,1])$ [Metelli et al., 2021]	$\frac{R_{\max}}{\lambda}$	$\frac{R^2_{\max}P_2(\pi_\theta\|\pi_0)}{n}$	$R_{\max}\lambda P_2(\pi_\theta\|\pi_0)$	$R_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)\log\frac{1}{6}}{n}}$	$O(n^{-1/2})$
ES $(\alpha\in[0,1])$ [Aouali et al., 2023]	$R_{\max}\operatorname{esssup}\frac{\pi_\theta}{\pi_0^\alpha}$	$R^2_{\max}\frac{\mathbb{E}_{\pi_\theta}[\pi_\theta\cdot\pi_0^{1-2\alpha}]}{n}$	$R_{\max}(1-\mathbb{E}_{\pi_\theta}[\pi_0^{1-\alpha}])$	$R_{\max}\sqrt{\frac{D_{\mathrm{KL}}(\pi_\theta\|\pi_0)+\log(4\sqrt{n}/\delta)}{n}} + T^{ES}$	$O(n^{-1/2})$
LSE $(0<\lambda<\infty)$ **(ours)**	$R_{\max}\operatorname{esssup}\frac{\pi_\theta}{\pi_0}$	$(\frac{1}{n}+\lambda)(M+R^2_\theta)$	$\frac{\lambda}{2}(M+R^2_\theta)$	$\frac{\log\frac{2}{\delta}}{n\lambda} + \sqrt{\frac{M\log\frac{2}{\delta}}{n}}$	$O(n^{-1/2})$

From Table 2, we can observe that the upper bound on the generalization error of the LSE estimator has the convergence rate of $O(n^{-1/2})$. Moreover, theoretical results on generalization error, bias and variance can be applied to unbounded weighted cost function under bounded expectation and variance of the weighted cost function, Assumptions 5.1 and 5.2, compared to other estimators where the bounded cost function or weighted cost function is needed.

6 LSE regularized Via αRényi divergence

In this section, we first provide an upper bound on the generalization error of the LSE estimator in terms of α-Rényi divergence. For this purpose, the following assumptions are made.

Assumption 6.1 (Sub-Gaussianity)**.** The weighted square cost function, i.e., $w_\theta(A,X)c^2(A,X)$, is σ-sub-Gaussian[*] under $P_X\otimes\pi_0(A|X)$ and $P_X\otimes\pi_\theta(A|X)$.

Assumption 6.2 (Finite Action Space)**.** The action space $\mathcal{A}$ is a finite set, i.e. we have a finite set of $|\mathcal{A}| = K$ actions.

Proposition 6.3 (Generalization Error Upper Bound based on α-Rényi Divergence)**.** *Given Assumption 6.1 and assuming that the cost function has bounded range $[-C,0]$, then the generalization error of the LSE estimator satisfies,*

$$\mathfrak{L}_\alpha(\gamma,n,\lambda,R_\theta,\delta) \le \operatorname{gen}_\lambda(\pi_\theta) \le \mathfrak{U}_\alpha(\gamma,n_u,\lambda,R_\theta,\delta),$$

where $U_\alpha = C^2 + \sigma\sqrt{\frac{2D_\alpha}{\min(\alpha,1)}}$, $D_\alpha := D_\alpha(\pi_0(A|X)\otimes P_X\|\pi_\theta\otimes P_X)$, $n_u = n$ *for* $n_u \ge \frac{\left(2\lambda^2 M + \frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda R_\theta)}$*, and,*

$$\begin{aligned}\mathfrak{L}_\alpha(\gamma,n,\lambda,R_\theta,\delta) &:= -\frac{\lambda}{2}U_\alpha - \frac{2}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} - \frac{1}{\exp(\lambda R_\theta)}\sqrt{\frac{2(U_\alpha + R_\theta^2)\log\frac{2}{\delta}}{n}},\\ \mathfrak{U}_\alpha(\gamma,n_u,\lambda,R_\theta,\delta) &:= \frac{4}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} + \frac{2}{\exp(\lambda R_\theta)}\sqrt{\frac{2(U_\alpha + R_\theta^2)\log\frac{2}{\delta}}{n}}.\end{aligned} \tag{13}$$

Inspired by Proposition 6.3, we suggest employing α-Rényi as a regularizer for the LSE estimator, denoted as α-LSE, to reduce the upper bound (or increase the lower bound) on the generalization error.

$$\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta) + \beta D_\alpha(\pi_0(A|X)\otimes P_X\|\pi_\theta\otimes P_X). \tag{14}$$

[*] A random variable, X is σ-sub-Gaussian under distribution P'_X if $\log(\mathbb{E}_{X\sim P'_X}[\exp(X - \mathbb{E}_{X\sim P'_X}[X])]) \le \frac{\eta^2\sigma^2}{2}$ for all $\eta\in\mathbb{R}$.

Table 3: Comparison of different algorithms LSE, PM, ES, IX, α-LSE, PM+SM, and IPS-KL accuracy for EMNIST with different qualities of logging policy ($\tau \in \{1, 10\}$) and clean/ noisy propensity scores with $b \in \{5, 0.01\}$ and imbalance scenarios with $\nu \in \{3, 9, 20\}$. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	ν	α-LSE	LSE	PM	ES	IX	PM + SM	IPS-KL	Logging Policy
EMNIST	1	–	–	$\mathbf{91.72 \pm 0.03}$	88.49 ± 0.04	89.19 ± 0.03	88.61 ± 0.06	88.33 ± 0.13	89.38 ± 0.02	90.42 ± 0.11	88.08
		5	–	$\mathbf{91.31 \pm 0.01}$	89.16 ± 0.03	88.94 ± 0.05	88.48 ± 0.03	88.51 ± 0.23	88.83 ± 0.11	90.78 ± 0.08	88.08
		0.01	–	91.39 ± 0.01	86.07 ± 0.01	85.62 ± 0.10	85.71 ± 0.04	81.39 ± 4.02	74.64 ± 3.67	$\mathbf{91.65 \pm 0.01}$	88.08
		–	3	$\mathbf{91.20 \pm 0.03}$	87.83 ± 0.10	88.81 ± 0.11	63.64 ± 0.53	64.82 ± 7.29	89.20 ± 0.02	91.07 ± 0.06	88.08
		–	9	$\mathbf{91.80 \pm 0.02}$	88.01 ± 0.05	88.29 ± 0.10	56.09 ± 0.03	56.08 ± 0.02	88.93 ± 0.04	90.44 ± 0.06	88.08
		–	20	$\mathbf{91.87 \pm 0.01}$	88.00 ± 0.07	87.88 ± 0.07	56.26 ± 0.02	56.32 ± 0.01	88.61 ± 0.05	87.49 ± 0.11	88.08
	10	–	–	$\mathbf{91.02 \pm 0.03}$	88.59 ± 0.03	88.61 ± 0.04	88.38 ± 0.08	87.43 ± 0.19	89.77 ± 0.01	89.90 ± 0.11	79.43
		5	–	$\mathbf{90.20 \pm 0.09}$	88.42 ± 0.07	88.43 ± 0.07	88.39 ± 0.10	88.39 ± 0.06	89.23 ± 0.03	89.91 ± 0.07	79.43
		0.01	–	$\mathbf{89.68 \pm 0.07}$	82.15 ± 0.21	80.85 ± 0.29	81.07 ± 0.07	77.49 ± 2.77	75.38 ± 0.42	86.62 ± 0.13	79.43
		–	3	89.66 ± 0.04	86.96 ± 0.01	87.30 ± 0.03	61.74 ± 0.07	58.76 ± 3.96	$\mathbf{89.98 \pm 0.07}$	84.84 ± 0.10	79.43
		–	9	$\mathbf{89.15 \pm 0.03}$	86.13 ± 0.04	85.84 ± 0.05	55.99 ± 0.04	57.08 ± 3.72	88.73 ± 0.04	85.00 ± 0.04	79.43
		–	20	$\mathbf{89.12 \pm 0.08}$	80.50 ± 2.47	83.36 ± 0.18	56.29 ± 0.08	56.25 ± 0.02	88.18 ± 0.04	80.74 ± 0.06	79.43

Table 4: MSE of LSE, PM, ES and IPS estimators. We run the experiment 1000 times and find the MSE of the estimations.

μ	Metric	Estimator	$n = 10$	$n = 100$	$n = 1000$	$n = 10000$
1.0	MSE	IPS	0.301	0.059	0.007	**0.001**
		ES	0.150	0.056	0.043	0.042
		PM	0.191	0.052	0.006	0.002
		LSE	**0.119**	**0.023**	**0.003**	**0.001**
1.5	MSE	IPS	10.202	2.440	0.354	0.049
		ES	2.202	0.857	0.375	0.304
		PM	12.149	2.960	0.441	0.081
		LSE	**1.516**	**0.408**	**0.111**	**0.020**
2.0	MSE	IPS	171.761	95.488	122.134	5.124
		ES	34.261	24.194	16.584	5.794
		PM	171.761	95.488	122.134	5.124
		LSE	**28.298**	**12.594**	**4.192**	**0.914**

To estimate the α-Rényi Divergence in (14) using the LBF dataset, we use the propensity scores to estimate α-Rényi Divergence between the logging policy and the learning policy,

$$\hat{D}_{\alpha,n} := \frac{1}{\alpha - 1} \log \left(\sum_{\tilde{a} \in \mathcal{A}} \frac{1}{m[\tilde{a}]} \sum_{\substack{(x_i, a_i, p_i) \in D, \\ a_i = \tilde{a}}} \frac{p_i^{\alpha}}{\pi_\theta(a_i|x_i)^{\alpha - 1}} \right),$$

where $m[\tilde{a}]$ is the number of samples in the LBF dataset that have $\tilde{a}$ as their action. The estimation of α-Rényi Divergence, $\hat{D}_{\alpha,n}$, is asymptotically $(m[\tilde{a}] \to \infty, \forall \tilde{a} \in \mathcal{A})$ unbiased, which is shown in App. E. Therefore, the α-LSE objective would be,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) + \beta \hat{D}_{\alpha,n}. \tag{15}$$

We compared the α-Rényi regularization with the KL-regularization problem in App.E.1.

7 Experiments

We briefly present our experiments for synthetic and supervised-to-bandit datasets [Beygelzimer and Langford, 2009]. An experiment on Open Bandit Dataset [Saito et al., 2020a] as a real-world dataset is provided in App. J. More details can be found in App.H.

7.1 Synthetic dataset

Suppose that our bandit dataset has only a single context (state), denoted as x_0, and $\pi_0(\cdot|x_0) \sim \mathcal{N}(2, 1)$ and $\pi_\theta(\cdot|x_0) \sim \mathcal{N}(\mu, 1)$. Also let the cost function be an exponential function $c(a, x_0) = -\exp(\frac{1}{4}a^2)$. The objective is to estimate the expectation of the cost function under π_θ, with independent samples drawn from $\pi_0(\cdot|x_0)$. We set $n = 10, 100, 1000, 10,000$, $\mu = 1.0, 1.5, 2.0$, and estimate the mean square error of IPS, PM estimator [Metelli et al., 2021], ES estimator [Aouali et al., 2023] and LSE estimators of $\mathbb{E}[w_\theta(A, x_0)c(A, x_0)]$. Note that for $\mu = 2$, the learning policy is fitted perfectly to the logging policy. However, due to the unbounded cost function, our LSE estimator outperforms compared to other estimators. Table 4 shows that the LSE estimator achieves a lower MSE compared to other estimators. More details and the results for variance and bias of the estimators are provided in the App. K.

7.2 Supervised-to-bandit datasets

Baselines: For all of our experiments, we consider truncated IPS estimator [Swaminathan and Joachims, 2015a], PM estimator [Metelli et al., 2021] and ES estimator [Aouali et al., 2023] IX

estimator [Neu, 2015] as non-regularized baselines. Furthermore, we also compare α-LSE as regularized LSE with KL-regularized IPS [Aminian et al., 2024], and PM estimator with Second Moment regularization [Metelli et al., 2021].
Datasets: We apply the standard supervised to bandit transformation [Beygelzimer and Langford, 2009] on a classification dataset: Extended-MNIST (EMNIST) [Xiao et al., 2017]. We also run on more datasets, including **CIFAR-10**, **FMNIST** and **LETTER** in App.I. This transformation assumes that each of the classes in the datasets corresponds to an action. Then, a logging policy stochastically selects an action for every sample in the dataset. For each data sample x, action a is sampled by logging policy. For the selected action, propensity score p is determined by the softmax value of that action. If the selected action matches the actual label assigned to the sample, then we have $c = -1$, and $c = 0$ otherwise. So, the 4-tuple (x, a, p, c) makes up the LBF dataset. To generate a noisy propensity score, we use 2 types of noise, log-gamma distribution noise and log-normal distribution noise. The noisy LBF dataset is obtained by multiplying the propensity score, p, of each sample, (x, a, p, c), by the noise term sampled from the specified noise distribution.
Noisy LBF Dataset: We consider noisy propensity scores. For this purpose, motivated by Halliwell [2018] and the discussion in App.G.2, we assume a multiplicative inverse Gamma noise on π_0 for $b \in \mathbb{R}^+$, $\hat{\pi}_0 = \frac{1}{U}\pi_0$, where $\hat{\pi}(a|x)$ is the noisy propensity scores and $U \sim \mathrm{Gamma}(b, b)$[*].
Imbalance LBF dataset: Let $m[a]$ be the number of samples of each action in the dataset. Let $\bar{m} = \frac{n}{k}$ be the supposed number of actions if the dataset were balanced. Hence if the number of samples with each action is $\bar{m}$, the dataset is considered balanced. To create an imbalanced dataset, for each class, we will generate a random variable $Z_i \sim \mathcal{N}(\bar{m}, v^2)$ and set $\hat{m}[i] = \min(\max(Z_i, m_l), m_u)$ where v is the parameter that determines the degree of imbalance of the dataset, $m_l = 0.4\bar{m}$, $m_u = 10\bar{m}$ are the upper bound and lower bounds for the number of samples for each action.
Logging policy: To have logging policies with different performances, given inverse temperature[*] $\tau \in \{1, 10\}$, first, we train a linear softmax logging policy on the fully-labeled dataset. Then, when we apply standard supervised-to-bandit transformation on the dataset, the results obtained from the linear logging policy which are weights of each action according to the input, will be multiplied by the inverse temperature τ and then passed to a softmax layer. Thus, as the inverse temperature τ increases, we will have more uniform and less accurate logging policies.
Metric: We evaluate different algorithms based on the accuracy of the trained model. Inspired by London and Sandler [2019], we calculate the accuracy for a deterministic policy where the accuracy of the model based on the arg max of the softmax layer output for a given context is computed.

For each value of τ, we apply the LSE estimator and observe the accuracy over three runs on EMNIST. Table 3 shows the deterministic accuracy of LSE, PM, ES, IX, α-Rényi-regularized LSE (α-LSE), PM+KL, and IPS-KL for $\tau \in \{1, 10\}$. Moreover, the experiments for noisy LBF dataset (with $b \in \{5, 0.01\}$) and imbalance LBF dataset (with $\nu \in \{3, 9, 20\}$) are provided in Table 3.

Discussion: We observe that α-LSE achieves maximum accuracy(with less variance) in most clean, noisy and imbalance scenarios compared to all baselines. More discussion is provided in App. I.1.

8 Conclusion and future works

In this work, inspired by the log-sum-exponential operator, we proposed a novel estimator for off-policy learning application. Subsequently, we conduct a comprehensive theoretical analysis of the LSE estimator, including a study of bias and variance, along with an upper bound on generalization error. Building on our theoretical insights, we advocate for a regularization approach for our log-sum-exponential estimator based on α-Rényi divergence. Furthermore, we explore the performance of our estimator in scenarios involving noisy and imbalanced logged bandit feedback datasets. Results from our experimental evaluation demonstrate that our estimator, guided by our theoretical framework, performs competitively compared to baseline methods.
In future work, we plan to combine our estimator with doubly robust estimators or augmented-inverse-propensity-weighted estimators [Bang and Robins, 2005, Robins et al., 1994]. Moreover, we envision extending the application of our estimator to more challenging reinforcement learning setups, [Chen and Jiang, 2022, Zanette et al., 2021, Xie et al., 2019b].

[*] If $Z \sim \mathrm{Gamma}(\alpha, \beta)$, then we have $f_Z(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}$.

[*] The inverse temperature τ is defined as $\pi_0(a_i|x) = \frac{\exp(h(x,a_i)/\tau)}{\sum_{j=1}^{k} \exp(h(x,a_j)/\tau)}$ where $h(x, a_i)$ is the i-th input to the softmax layer for context $x \in \mathcal{X}$ and action $a_i \in \mathcal{A}$.

Acknowledgments and Disclosure of Funding

Aminian acknowledges the support of the UKRI Prosperity Partnership Scheme (FAIR) under EPSRC Grant EP/V056883/1 and the Alan Turing Institute. Shi's research is partially supported by the EPSRC Grant EP/W014971/1. We would like to express our sincere gratitude to Mohammad T. Teimuri for providing the GPU used in our experiments.

References

Charu C Aggarwal. *Recommender Systems*. Springer, 2016.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 297–306, 2011.

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–286, 2019.

Dimitris Bertsimas, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, 40(2):210–217, 2017.

Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 1587–1594, 2013.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015a.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008a.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. *Advances in Neural Information Processing Systems*, 23, 2010.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems*, 28, 2015b.

Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In *40th International Conference on Machine Learning (ICML)*, 2023.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in neural information processing systems*, 34:8119–8132, 2021.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, pages 528–535, 2008.

Giuseppe C Calafiore, Stephane Gaubert, and Corrado Possieri. Log-sum-exp neural networks and posynomial models for convex and log-log-convex data. *IEEE transactions on neural networks and learning systems*, 31(3):827–838, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351, 1998.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138, 2009.

Ben London, Levi Lu, Ted Sandler, and Thorsten Joachims. Boosted off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5614–5640. PMLR, 2023.

Ralph B D'Agostino Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281, 1998.

Sherry Weitzen, Kate L Lapane, Alicia Y Toledano, Anne L Hume, and Vincent Mor. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13(12):841–853, 2004.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9 (4):403, 2004.

Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555, 2008.

Yuying Xie, Yeying Zhu, Cecilia A Cotton, and Pan Wu. A model averaging approach for estimating propensity scores by optimizing balance. *Statistical methods in medical research*, 28(1):84–101, 2019a.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Chengchun Shi, Rui Song, and Wenbin Lu. Robust learning for optimal treatment decision with np-dimensionality. *Electronic journal of statistics*, 10:2894, 2016.

Qi Dai, Jian-wei Liu, and Jiapeng Yang. Multi-armed bandit heterogeneous ensemble learning for imbalanced data. *Computational Intelligence*, 39(2):344–368, April 2023. ISSN 0824-7935, 1467-8640. doi: 10.1111/coin.12566. URL `https://onlinelibrary.wiley.com/doi/10.1111/coin.12566`.

Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36:4985–5009, 2023.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 541–549. PMLR, 02–04 May 2024. URL `https://proceedings.mlr.press/v238/zhu24a.html`.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008b. ISSN 10618600. URL `http://www.jstor.org/stable/27594308`.

Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. *arXiv preprint arXiv:2309.15771*, 2023.

Yuta Saito, Aihara Shunsuke, Matsutani Megumi, and Narita Yusuke. Large-scale open dataset, pipeline, and benchmark for bandit algorithms. *arXiv preprint arXiv:2008.07146*, 2020a.

Gholamali Aminian, Armin Behnamnia, Roberto Vega, Laura Toni, Chengchun Shi, Hamid R. Rabiee, Omar Rivasplata, and Miguel R. D. Rodrigues. Semi-supervised batch learning from logged data, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Leigh J Halliwell. The log-gamma distribution and non-normal error. *Variance (Accepted for Publication)*, 2018.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019. Preprint version arXiv:1806.11500.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pages 378–388. PMLR, 2022.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019b.

Nathan Kallus. Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pages 4989–4999. PMLR, 2019.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20, 2007.

Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. *Advances in Neural Information Processing Systems*, 32, 2019.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*, 2020.

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

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 34:4065–4078, 2021.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 2023.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning" without"overlap: Pessimism and generalized empirical Bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. *arXiv preprint arXiv:2202.06317*, 2022.

Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *international conference on Machine learning*, pages 29734–29759. PMLR, 2023.

Jie Peng, Hao Zou, Jiashuo Liu, Shaoming Li, Yibao Jiang, Jian Pei, and Peng Cui. Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *Proceedings of the ACM Web Conference 2023*, pages 1220–1230, 2023.

Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. Off-policy evaluation for large action spaces via policy convolution. *arXiv preprint arXiv:2310.15433*, 2023.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR, 2016.

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

M Wainwright. Basic tail and concentration bounds. `http://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`, 2015.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences, 2021.

Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. *arXiv preprint arXiv:2306.07923*, 2023.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Elvezio M Ronchetti and Peter J Huber. *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA, 2009.

Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *The Journal of Machine Learning Research*, 5:1007–1034, 2004.

Xiaoying Zhang, Junpu Chen, Hongning Wang, Hong Xie, Yang Liu, John C.S. Lui, and Hang Li. Uncertainty-aware instance reweighting for off-policy learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=1pWNhmbllE`.

David Slate. Letter Recognition. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5ZP40.

Yuta Saito, Aihara Shunsuke, Matsutani Megumi, and Narita Yusuke. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

A Other related works

Other methods: A balance-based weighting approach, which outperforms traditional estimators, was proposed by Kallus [2018]. Other extensions of batch learning scenarios have been studied, Papini et al. [2019] consider samples from different policies and Sugiyama et al. [2007] propose Direct Importance Estimation, which estimates weights directly by sampling from contexts and actions. Chen et al. [2019] introduced a convex surrogate for the regularized true risk based on the entropy of the target policy.

Pessimism method and off-policy reinforcement learning: The pessimism concept originally, introduced in offline reinforcement learning [Buckman et al., 2020, Jin et al., 2021], aims to derive an optimal policy within Markov decision processes (MDPs) by utilizing pre-existing datasets [Rashidinejad et al., 2022, 2021, Yin and Wang, 2021, Yan et al., 2023]. This concept has also been adapted to contextual bandits, viewed as a specific MDP instance. Recently, a 'design-based' version of the pessimism principle was proposed by Jin et al. [2022] who propose a data-dependent and policy-dependent regularization inspired by a lower confidence bound (LCB) on the estimation uncertainty of the augmented-inverse-propensity-weighted (AIPW)-type estimators which also includes IPS estimators. Our work differs from that of Jin et al. [2022] as our estimator is a non-linear estimator. Note that for our theoretical analysis, we consider bounded second-moment of weights. However, [Jin et al., 2022] also considers third and fourth moments of weights bounded.

Action embedding and clustering: Due to the extreme bias and variance of IPS and doubly-robust (DR) estimators in large action spaces, Saito and Joachims [2022] proposed using action embeddings to leverage marginalized importance weights and address these issues. Subsequent studies, including [Saito et al., 2023, Peng et al., 2023, Sachdeva et al., 2023], have introduced alternative methods to tackle the challenge of large action spaces. Our work can be integrated with these approaches to further mitigate the effects associated with large action spaces.

Individualized treatment effects: The individual treatment effect aims to estimate the expected values of the squared difference between outcomes (cost or feedback) for control and treated contexts [Shalit et al., 2017]. In the individual treatment effect scenario, the actions are limited to two actions (treated/not treated) and the propensity scores are unknown [Shalit et al., 2017, Johansson et al., 2016, Alaa and van der Schaar, 2017, Athey et al., 2019, Shi et al., 2019, Kennedy, 2020, Nie and Wager, 2021]. Our work differs from this line of works by considering larger action spaces and assuming the access to propensity scores in the LBF dataset.

B Preliminaries

B.1 Definitions

We define the softmax function

$$\begin{aligned}\operatorname{softmax}(x_1, x_2, \cdots, x_n) &= (s_1, s_2, \cdots, s_n),\\ s_i &= \frac{e^{x_i}}{\sum_{j=1}^{n} x^{x_j}}, \quad 1 \leq i \leq n.\end{aligned}$$

The diag function, $\operatorname{diag}(a_1, a_2, \cdots, a_n) \in \mathbb{R}^{n \times n}$, defines a diagonal matrix with $a_1, a_2, \cdots, a_n$ as elements on its diagonal.

Definition B.1 (Asymptotic Value)**.** We define the symbol $O(f(n))$ (order) and $\Omega(f(n))$, for an arbitrary function $f : \mathbb{R} \to \mathbb{R}^+$ as the upper and lower asymptotic behaviors, respectively. For a function $g : \mathbb{R} \to \mathbb{R}$, we call g to be of order f, or $g(n) = O(f(n))$, if there exists C_0, N_0, such that for any $n \geq N_0$, we have,

$$g(n) \leq C_0 f(n).$$

Similarly, we state that $g(n) = \Omega(f(n))$, if there exists C_0', N_0' such that for any $n \geq N_0'$,

$$g(n) \geq C_0' f(n).$$

B.2 Theoretical tools

In this section, we provide the main lemmas which are used in our theoretical proofs.
Here are some popular lemmas that are used in our theoretical analysis.

Lemma B.2 (Hoeffding Inequality, Boucheron et al., 2013)**.** *Suppose that X_i are sub-Gaussian independent random variables, with means μ_i and sub-Gaussian parameter σ_i^2, then we have:*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_1-\mu_i)\geq t\right)\leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right) \tag{16}$$

Lemma B.3 (One-sided Bernstein's inequality [Wainwright, 2015])**.** *Let $X\leq b_1$ and $\mathbb{E}[X]<\infty$, then the following holds for all $\lambda\in[0,3/b_1)$,*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]\leq \exp\left(\frac{(\lambda^2/2)\mathbb{E}[X^2]}{1-\frac{b_1\lambda}{3}}\right).$$

Lemma B.4 (Variational representation of α-Rényi Divergence [Birrell et al., 2021])**.** *Suppose that P and Q are two probability measure over the set $\mathcal{X}$, then the following variational representation for α-Rényi holds,*

$$D_\alpha(P\|Q):=\sup_{g\in\mathcal{G}}\frac{\alpha}{\alpha-1}\log\left[\mathbb{E}_P[\exp((\alpha-1)g(X))]\right]-\log\left[\mathbb{E}_Q[\exp(\alpha g(X))]\right], \tag{17}$$

where $\mathcal{G}=\{g:\mathcal{X}\to\mathbb{R}|\mathbb{E}_Q[\exp(\alpha g(X))]<\infty,\quad \mathbb{E}_P[\exp((\alpha-1)g(X))]<\infty\}$.

The rest of the lemmas are provided with proofs.

Lemma B.5 (Change of variables)**.** *Assume that the following equation holds,*

$$\epsilon=\exp\left\{-\frac{A\delta^2}{B+C\delta}\right\},$$

for $A,B,C,\epsilon\geq 0$ and $0\leq\delta\leq 1$. Then, we have,

$$\delta\leq\frac{C\log\frac{1}{\epsilon}}{A}+\sqrt{\frac{B\log\frac{1}{\epsilon}}{A}}.$$

Also, for some $D>0$, if

$$A\geq\frac{B\log\frac{1}{\epsilon}+2DC\log\frac{1}{\epsilon}}{D^2}$$

we have,

$$\delta\leq D$$

Proof. We have,

$$\epsilon=\exp\left\{-\frac{A\delta^2}{B+C\delta}\right\}\leftrightarrow A\delta^2-C\log\frac{1}{\epsilon}\delta-B\log\frac{1}{\epsilon}=0$$

Given $\delta>0$ and solving the quadratic equation, we have,

$$\begin{aligned}\delta=\frac{1}{2A}\left(C\log\frac{1}{\epsilon}+\sqrt{C^2\log^2\frac{1}{\epsilon}+4AB\log\frac{1}{\epsilon}}\right)&=\frac{C}{2}\sqrt{\frac{\log\frac{1}{\epsilon}}{A}}\left(\sqrt{\frac{\log\frac{1}{\epsilon}}{A}}+\sqrt{\frac{\log\frac{1}{\epsilon}}{A}+4\frac{B}{C^2}}\right)\\&\leq C\sqrt{\frac{\log\frac{1}{\epsilon}}{A}}\left(\sqrt{\frac{\log\frac{1}{\epsilon}}{A}}+\sqrt{\frac{B}{C^2}}\right)\\&=\frac{C\log\frac{1}{\epsilon}}{A}+\sqrt{\frac{B\log\frac{1}{\epsilon}}{A}},\end{aligned}$$

where the inequality is derived from $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$.
For the second part, similar argument works for $a=\sqrt{A}$ as the variable ,

$$\frac{C\log\frac{1}{\epsilon}}{A}+\sqrt{\frac{B\log\frac{1}{\epsilon}}{A}}\leq D\leftrightarrow Da^2-\sqrt{B\log\frac{1}{\epsilon}}a-C\log\frac{1}{\epsilon}\geq 0$$

which is satisfied if a is greater than the bigger root,

$$a \geq \frac{\sqrt{B \log \frac{1}{\epsilon}} + \sqrt{B \log \frac{1}{\epsilon} + 4DC \log \frac{1}{\epsilon}}}{2D}$$

So,

$$A \geq \frac{B \log \frac{1}{\epsilon} + 2DC \log \frac{1}{\epsilon}}{D^2} \geq \left(\frac{\sqrt{B \log \frac{1}{\epsilon}} + \sqrt{B \log \frac{1}{\epsilon} + 4DC \log \frac{1}{\epsilon}}}{2D} \right)^2$$

where the last inequality comes from $\frac{a^2+b^2}{2} \geq \left(\frac{a+b}{2}\right)^2$. Hence if $A \geq \frac{B \log \frac{1}{\epsilon} + 2DC \log \frac{1}{\epsilon}}{D^2}$, a is bigger than the largest root and the proposed inequality holds. □

Lemma B.6. *Assume $A, B, C \in \mathbb{R}^+$. For any $x \in \mathbb{R}^+$ such that,*

$$x \leq \frac{C^2}{2AC + B},$$

we have,

$$Ax + \sqrt{Bx} \leq C \tag{18}$$

Proof. Given $Ax \leq C$, equation (18) is equivalent to the following quadratic form.

$$A^2x^2 - (B + 2AC)x + C^2 \geq 0$$

Let $0 < r_1 < r_2$ be the roots of the abovementioned quadratic form. If $X < r_1$, $Ax \leq C$ holds and the quadratic form is positive. So we have the following condition on x to satisfy Equation 18,

$$x \leq \frac{B + 2AC - \sqrt{(B + 2AC)^2 - 4A^2C^2}}{2A^2} = \frac{2C^2}{B + 2AC + \sqrt{(B + 2AC)^2 - 4A^2C^2}}.$$

Since,

$$\frac{C^2}{2AC + B} \leq \frac{2C^2}{B + 2AC + \sqrt{(B + 2AC)^2 - 4A^2C^2}},$$

the condition in the lemma is sufficient for (18) to hold. □

Lemma B.7. *Let us consider the functions $h_b(x) = \log(x) + \frac{1}{2b^2}x^2$ and $h_a(x) = \log(x) + \frac{1}{2a^2}x^2$ for $a < x < b$. Then $h_b(x)$ and $h_a(x)$ are concave and convex, respectively.*

Proof. Taking the second derivative gives us the result:

$$\frac{d^2}{dx^2}\left(\log(x) + \beta x^2\right) = -\frac{1}{x^2} + 2\beta.$$

□

Lemma B.8. *Suppose $\mathbb{E}[X^2] < \infty$. Then, following inequality for $\lambda > 0$ and $X < 0$ holds,*

$$\mathbb{E}[X] \leq \frac{1}{\lambda} \log \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[X] + \frac{\lambda}{2}\mathbb{E}[X^2].$$

Proof. The left side inequality follows from Jensen's inequality on $f(x) = \log(x)$. For the right side, we have for $x < 0$,

$$1 + x \leq e^x \leq 1 + x + \frac{1}{2}x^2.$$

Therefore, we have,

$$\begin{aligned}
\frac{1}{\lambda} \log \mathbb{E}[e^{\lambda X}] &\leq \frac{1}{\lambda} \log \mathbb{E}[1 + \lambda X + \frac{1}{2}\lambda^2 X^2] \\
&= \frac{1}{\lambda} \log \left(1 + \lambda \mathbb{E}[X] + \frac{1}{2}\lambda^2 \mathbb{E}[X^2]\right) \\
&\leq \frac{1}{\lambda} \left(\lambda \mathbb{E}[X] + \frac{1}{2}\lambda^2 \mathbb{E}[X^2]\right) \\
&= \mathbb{E}[X] + \frac{\lambda}{2}\mathbb{E}[X^2].
\end{aligned}$$

□

Lemma B.9. *We have,*

$$\frac{1}{2n\lambda}\mathbb{V}(e^{\lambda w_\theta(A,X)c(A,X)}) \leq R_\lambda(\pi_\theta) - \mathbb{E}\left[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)\right],$$

Proof. For the sake of simplicity of notation, we consider $y_\theta(a_i,x_i) = c(a_i,x_i)w_\theta(a_i,x_i)$. Note that, an upper bound 1 on $\frac{\sum_{i=1}^n e^{\lambda c_i w_\theta(a_i,x_i)}}{n}$ holds. Now, we have,

$$\begin{aligned}
&\mathbb{E}[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)] = \frac{1}{\lambda}\mathbb{E}\left[\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)\right] \\
&= \frac{1}{\lambda}\mathbb{E}\left[\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right) + \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2 - \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2\right] \\
&\leq \frac{1}{\lambda}\left(\log\left(\mathbb{E}\left[\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right]\right) + \frac{1}{2}\mathbb{E}\left[\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right]^2 - \frac{1}{2}\mathbb{E}\left[\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2\right]\right) \\
&= \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \frac{1}{2\lambda}\mathbb{V}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right) \\
&= \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right),
\end{aligned}$$

where the third line is derived by applying Jensen inequality on function

$$\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right) + \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2,$$

which is concave based on Lemma B.7 for $b = 1$. Finally, we have,

$$\frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda w_\theta(A,X)c(A,X)}\right]\right) - \mathbb{E}[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)] \geq \frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda w_\theta(A,X)c(A,X)}\right).$$

□

Lemma B.10. *We have,*

$$\mathbb{E}\left[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)\right] - R(\pi_\theta) \geq \frac{n-1}{n}\mathbb{V}\left(e^{\lambda w_\theta(A,X)c(A,X)}\right)$$

Proof. Setting $y_\theta(a_i,x_i) = c(a_i,x_i)w_\theta(a_i,x_i)$, according to Lemma B.7 for b = 1, $f(x) = \log(x) + \frac{1}{2}x^2$ is concave. So we have,

$$\begin{aligned}
\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right) + \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2 &\geq \frac{1}{n}\left(\sum_{i=1}^n \log\left(e^{\lambda y_\theta(a_i,x_i)}\right) + \frac{1}{2}e^{2\lambda y_\theta(a_i,x_i)}\right) \\
&= \frac{\lambda}{n}\sum_{i=1}^n y_\theta(a_i,x_i) + \frac{1}{2n}\sum_{i=1}^n e^{2\lambda y_\theta(a_i,x_i)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)\right] \\
&\geq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n y_\theta(a_i,x_i) + \frac{1}{2n\lambda}\sum_{i=1}^n e^{2\lambda y_\theta(a_i,x_i)} - \frac{1}{2\lambda}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2\right] \\
&= \mathbb{E}[Y_\theta(A,X)] + \frac{n-1}{2n\lambda}\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right).
\end{aligned}$$

□

C Details of section 3

Proposition C.1 (LSE asymptotic properties)**.** *The following asymptotic properties of LSE holds (w.r.t λ):*

$$\begin{aligned}\lim_{\lambda\to 0}\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) &= \frac{1}{n}\left(\sum_{i=1}^{n} c_i w_\theta(a_i,x_i)\right),\\ \lim_{\lambda\to -\infty}\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) &= \min_i c_i w_\theta(a_i,x_i),\\ \lim_{\lambda\to \infty}\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) &= \max_i c_i w_\theta(a_i,x_i).\end{aligned}$$

Proof. For the first limit, we use L'Hopital's rule:

$$\lim_{\lambda\to 0}\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) = \lim_{\lambda\to 0}\frac{\log\left(\frac{\sum_{i=1}^{n} e^{\lambda x_i}}{n}\right)}{\lambda} = \lim_{\lambda\to 0}\frac{\left(\frac{\sum_{i=1}^{n} x_i e^{\lambda x_i}}{\sum_{i=1}^{n} e^{\lambda x_i}}\right)}{1} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

For the second limit for $\lambda < 0$ we have:

$$\begin{aligned}\min_i x_i = \frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^{n} e^{\lambda \min_i x_i}}{n}\right) &\le \frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^{n} e^{\lambda x_i}}{n}\right)\\ &\le \frac{1}{\lambda}\log\left(\frac{e^{\lambda \min_i x_i}}{n}\right)\\ &= \min_i x_i - \frac{1}{\lambda}\log n.\end{aligned}$$

As both lower and upper tends to $\min_i x_i$ we conclude that:

$$\lim_{\lambda\to -\infty}\frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^{n} e^{\lambda x_i}}{n}\right) = \min_i x_i.$$

A similar argument proves the third limit. □

Remark C.2. For $\lambda \to \infty$, the LSE estimator only minimizes the sample with the maximum value. In fact, due to $w_\theta(An, X)c(A, X) < 0$, the LSE estimator minimizes the sample with minimum absolute value. In particular, it considers the sample with a very small importance weight, ignoring all samples with high discrepancy between logging and learning policy, including the ones contributing to the high variance of the IPS estimator. It can also be interpreted as the robustness of noisy samples, where the variance of the IPS estimator.

It is interesting to study the sensitivity of the LSE estimator with respect to its values.

Lemma C.3. *The gradient and hessian of the LSE estimator with respect to its values are as follows,*

$$\nabla\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) = \mathrm{softmax}(\lambda c_1 w_\theta(a_1,x_1),\cdots,\lambda c_n w_\theta(a_n,x_n)), \tag{19}$$

$$\nabla^2\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) = \lambda\mathrm{diag}(S_m) - \lambda S_m S_m^T, \tag{20}$$

where $S_m = \mathrm{softmax}(\lambda c_1 w_\theta(a_1,x_1),\cdots,\lambda c_n w_\theta(a_n,x_n))$*. Also, LSE is convex when* $\lambda > 0$ *and concave otherwise.*

Proof. The two equations can be derived with simple calculations. About the convexity and concavity of $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}$, we prove that for $\lambda \ge 0$ Hessian matrix is positive semi-definite. The proof for concavity for $\lambda < 0$ is similar.

$$\begin{aligned}\mathbf{x}^T\nabla^2\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}\mathbf{x} = \lambda\left(\mathbf{x}^T\mathrm{diag}(S_m)\mathbf{x} - \mathbf{x}^T S_m S_m^T\mathbf{x}\right) &= \lambda\left(\sum_{i=1}^{n} S_m(i)x_i^2 - \left(\sum_{i=1}^{n} S_m(i)x_i\right)^2\right)\\ = \lambda\left(\left(\sum_{i=1}^{n} S_m(i)x_i^2\right)\left(\sum_{i=1}^{n} S_m(i)\right) - \left(\sum_{i=1}^{n} S_m(i)x_i\right)^2\right) &\ge 0.\end{aligned}$$

Where the last inequality is derived from the Cauchy–Schwarz inequality. □

Using Lemma C.3, it can be shown that $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}$ is convex for $\lambda \geq 0$ and concave for $\lambda < 0$. Applying Lemma C.3, it can be shown that the derivative of the LSE estimator is positive and less than one, i.e.,

$$0 \leq \nabla \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) \leq 1. \tag{21}$$

We can also prove Equation. (7) by applying Lemma C.3.

D Proofs and details of section 5

D.1 Proofs and details of bias-variance

Lemma 5.3 (restated). *Suppose that $X < 0$ is a random variable. The following upper bound holds on the variance of $e^{\lambda X}$ for $\lambda > 0$,*

$$\mathbb{V}\left(e^{\lambda X}\right) \leq \lambda \min_{C_1} \mathbb{E}[|X - C_1|] \leq \lambda \mathbb{E}[-X], \tag{22}$$

In addition, by assuming $\mathbb{V}(X) < \infty$, we have,

$$\mathbb{V}\left(e^{\lambda X}\right) \leq \lambda^2 \mathbb{V}(X). \tag{23}$$

Proof. We have,

$$|e^{\lambda X} - e^{\lambda C_1}| = \left| \int_{\lambda C_1}^{\lambda x} e^y dy \right| \leq |\lambda(x - C_1)| e^{\max(\lambda x, \lambda C_1)} \leq \lambda |x - C_1|.$$

Then it holds that

$$\mathbb{V}(e^{\lambda X}) = \min_{C_1} \mathbb{E}\left[(e^{\lambda X} - e^{\lambda C_1})^2\right] = \min_{C_1} \mathbb{E}\left[|e^{\lambda X} - e^{\lambda C_1}| \lambda |X - C_1|\right] \leq \min_{C_1} \mathbb{E}\left[\lambda |X - C_1|\right].$$

Replacing C_1 with 0 gives the first inequality. Note that,

$$\mathbb{V}(e^{\lambda X}) = \min_{C_1} \mathbb{E}\left[(e^{\lambda X} - e^{\lambda C_1})^2\right] \leq \min_{C_1} \mathbb{E}\left[\lambda^2 (X - C_1)^2\right] = \lambda^2 \mathbb{V}(X),$$

where proves the second inequality. □

Proposition 5.4 (restated). *Under Assumption 5.2, we have the following bound on the bias,*

$$\frac{n-1}{2n\lambda} \mathbb{V}(e^{\lambda w_\theta(A,X) c(A,X)}) \leq \mathbb{B} \leq \frac{\lambda}{2}(M + R_\theta^2) - \frac{1}{2n\lambda} \mathbb{V}(e^{\lambda w_\theta(A,X) c(A,X)}).$$

Proof. The lower bound is directly derived from Lemma B.10. For the upper bound, we combine Lemma B.9 and the upper bound in Lemma B.8.

$$\begin{aligned}
\mathbb{E}[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)] - \frac{1}{\lambda} \log\left(\mathbb{E}\left[e^{\lambda w_\theta(A,X) c(A,X)}\right]\right) &\leq -\frac{1}{2n\lambda} \mathbb{V}\left(e^{\lambda w_\theta(A,X) c(A,X)}\right) \\
\frac{1}{\lambda} \log\left(\mathbb{E}\left[e^{\lambda w_\theta(A,X) c(A,X)}\right]\right) &\leq \mathbb{E}[w_\theta(A,X) c(A,X)] + \frac{\lambda}{2} \mathbb{E}[c^2(A,X) w_\theta^2(A,X)] \\
\to \mathbb{E}[\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)] &\leq \mathbb{E}[w_\theta(A,X) c(A,X)] + \frac{\lambda}{2} \mathbb{E}[c^2(A,X) w_\theta^2(A,X)] \\
&\quad - \frac{1}{2n\lambda} \mathbb{V}\left(e^{\lambda w_\theta(A,X) c(A,X)}\right).
\end{aligned}$$

Which gives the proposed upper bound. □

Proposition 5.6 (restated). *Suppose Assumption 5.2 holds. Let $\mu_i = \mathbb{E}[c^i(A,X) w_\theta^i(A,X)]$ for $i \in \{1, 2\}$. Also, suppose that $\lambda < -2\frac{\mu_1}{\mu_2}$ We have the following bound on the variance of the LSE estimator,*

$$\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \leq \frac{M}{n} + \frac{\lambda}{2}\left(-2R_\theta - \frac{\lambda}{2}(M + R_\theta^2)\right)(M + R_\theta^2).$$

Proof. Suppose that $Z = \frac{\sum_{i=1}^{n} e^{\lambda c_i w_\theta(a_i,x_i)}}{n}$. Also set $y_{i,\theta}(a_i,x_i) = c_i(a_i,x_i)w_\theta(a_i,x_i)$. According to Lemma B.7 for $b = 1$, we have $\log(x) + \frac{1}{2}x^2$ is concave for $x < 1$. Using Jensen's inequality we have,

$$\begin{aligned}
\frac{1}{\lambda}\log Z &= \hat{\mathrm{R}}^{\lambda}_{\text{LSE}}(S,\pi_\theta) \\
&= \frac{1}{\lambda}\log Z + \frac{1}{2\lambda}Z^2 - \frac{1}{2\lambda}Z^2 \\
&\geq \frac{1}{\lambda}\log\left(e^{\frac{\sum_{i=1}^{n}\lambda c_i w_\theta(a_i,x_i)}{n}}\right) \\
&= \frac{\sum_{i=1}^{n} c_i w_\theta(a_i,x_i)}{n}.
\end{aligned}$$

Since $\log Z < 0$ a.s., We have,

$$\mathbb{E}\left[\frac{1}{\lambda^2}\log^2 Z\right] \leq \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} c_i w_\theta(a_i,x_i)}{n}\right)^2\right].$$

Also due to the concavity of $f(x) = \log(x)$ we have,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\lambda}\log Z\right] \leq \frac{1}{\lambda}\log\mathbb{E}\left[Z\right] &= \frac{1}{\lambda}\log\mathbb{E}[e^{\lambda c(A,X)w_\theta(A,X)}] \\
&\leq \mathbb{E}[c(A,X)w_\theta(A,X)] + \frac{\lambda}{2}\mathbb{E}[c^2(A,X)w_\theta^2(A,X)].
\end{aligned}$$

where the last inequality is proposed by Lemma B.8. Hence we have,

$$\begin{aligned}
\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\text{LSE}}(S,\pi_\theta)) &= \mathbb{E}\left[\frac{1}{\lambda^2}\log^2 Z\right] - \left(\mathbb{E}\left[\frac{1}{\lambda}\log Z\right]\right)^2 \\
&\leq \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} c_i w_\theta(a_i,x_i)}{n}\right)^2\right] \\
&\quad - \left(\mathbb{E}[c(A,X)w_\theta(A,X)] + \frac{\lambda}{2}\mathbb{E}[c^2(A,X)w_\theta^2(A,X)]\right)^2.
\end{aligned}$$

given $\lambda \leq \frac{-2\mathbb{E}[c(A,X)w_\theta(A,X)]}{\mathbb{E}[c^2(A,X)w_\theta^2(A,X)]}$. Simplifying the right-hand side we have,

$$\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\text{LSE}}(S,\pi_\theta)) \leq \frac{M}{n} + \frac{\lambda}{2}\left(-2R_\theta - \frac{\lambda}{2}(M + R_\theta^2)\right)(M + R_\theta^2).$$

□

D.2 Proofs and details of generalization error bounds

Theorem 5.8 (restated). *Given Assumptions 5.1 and 5.2, and assuming* $0 < \gamma < 1$*, then with probability at least* $1 - \delta$ *we have,*

$$\mathfrak{L}(\gamma, n, \lambda, R_\theta, M) \leq \mathrm{gen}_\lambda(\pi_\theta) \leq \mathfrak{U}(\gamma, n_u, \lambda, R_\theta, M),$$

where $n_u = n$ *for* $n_u \geq \frac{\left(2\lambda^2 M + \frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda R_\theta)}$*, and,*

$$\begin{aligned}
\mathfrak{L}(\gamma, n, \lambda, R_\theta, M, \delta) &:= \frac{-\lambda}{2}(M + R_\theta^2) - \frac{2}{3}\frac{\log\frac{1}{\delta}}{n\lambda\exp(\lambda R_\theta)} - \sqrt{\frac{2M\log\frac{1}{\delta}}{n\exp(2\lambda R_\theta)}}, \\
\mathfrak{U}(\gamma, n_u, \lambda, R_\theta, M, \delta) &:= \frac{2}{3(1-\gamma)}\frac{\log\frac{1}{\delta}}{n_u\lambda\exp(\lambda R_\theta)} + \frac{1}{1-\gamma}\sqrt{\frac{2M\log\frac{1}{\delta}}{n_u\exp(2\lambda R_\theta)}}.
\end{aligned} \tag{24}$$

Proof. To ease the notation, we consider $y_\theta(a_i,x_i) = c(a_i,x_i)w_\theta(a_i,x_i)$. Using Bernstein's inequality we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda y_\theta(a_i,x_i)) - \mathbb{E}[\exp(\lambda Y_\theta(A,X))] \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}nt^2}{\mathbb{V}(\exp(\lambda Y_\theta(A,X))) + \frac{1}{3}t}\right),$$

So we have,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda y_\theta(a_i,x_i)) - \mathbb{E}[\exp(\lambda Y_\theta(A,X))] \leq t\right) \geq 1-\exp\left(-\frac{\frac{1}{2}nt^2}{\mathbb{V}(\exp(\lambda Y_\theta(A,X))) + \frac{1}{3}t}\right).$$

As the log function is an increasing function, we have,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) \leq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + t\right).$$

where recall that $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) = \frac{1}{\lambda}\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda y_\theta(a_i,x_i))\right)$. Therefore, we have with probability at least $1-\exp\left(-\frac{\frac{1}{2}nt^2}{\mathbb{V}(e^{\lambda Y_\theta(A,X)})+\frac{1}{3}t}\right)$,

$$\begin{aligned}
\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) &\leq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + t\right) \\
&\leq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) + \frac{t}{\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]} \\
&\leq \mathbb{E}[Y_\theta(A,X)] + \frac{\lambda}{2}\mathbb{E}[c^2(A,X)w_\theta^2(A,X)] + \frac{t}{\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}.
\end{aligned}$$

Using Lemma B.5, we have with probability at least $1-\delta$,

$$\begin{aligned}
\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) &\leq \mathbb{E}[Y_\theta(A,X)] + \frac{\lambda}{2}\mathbb{E}[c^2(A,X)w_\theta^2(A,X)] + \frac{\frac{\frac{1}{3}\log\frac{1}{\delta}}{\frac{n}{2}} + \sqrt{\frac{\mathbb{V}(e^{\lambda Y_\theta(A,X)})\log\frac{1}{\delta}}{\frac{n}{2}}}}{\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]} \\
&= \mathbb{E}[Y_\theta(A,X)] + \frac{\lambda}{2}\mathbb{E}[c^2(A,X)w_\theta^2(A,X)] \\
&\quad + \frac{2}{3}\frac{\log\frac{1}{\delta}}{n\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]} + \sqrt{\frac{2\mathbb{V}(e^{\lambda Y_\theta(A,X)})\log\frac{1}{\delta}}{n\lambda^2\mathbb{E}[e^{\lambda Y_\theta(A,X)}]^2}}.
\end{aligned}$$

Similarly, we have,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) \geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - t\right),$$

and,

$$\begin{aligned}
\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) &\geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - t\right) \\
&\geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) - \frac{t}{\lambda(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - t)} \\
&\geq \mathbb{E}[Y_\theta(A,X)] - \frac{t}{\lambda(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - t)}.
\end{aligned}$$

So for $t \leq \gamma\mathbb{E}[e^{\lambda Y_\theta(A,X)}]$, with probability at least $1-\delta$,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) \geq \mathbb{E}[Y_\theta(A,X)] - \frac{2}{3(1-\gamma)}\frac{\log\frac{1}{\delta}}{n\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]} - \frac{1}{(1-\gamma)}\sqrt{\frac{2\mathbb{V}(e^{\lambda Y_\theta(A,X)})\log\frac{1}{\delta}}{n\lambda^2\mathbb{E}[e^{\lambda Y_\theta(A,X)}]^2}}.$$

We can replace $\mathbb{V}(e^{\lambda Y_\theta(A,X)})$ with $\lambda^2 M$ according to Lemma 5.3 and Assumption 5.2, which gives the proposed upper and lower bounds on the generalization.
In order for t to be less than $\gamma\mathbb{E}[e^{\lambda Y_\theta(A,X)}]$, according to the second part of Lemma B.5, it is sufficient to have,

$$n \geq \frac{\left(2\mathbb{V}(e^{\lambda Y_\theta(A,X)}) + \frac{4}{3}\gamma\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right)\log\frac{2}{\delta}}{\gamma^2\mathbb{E}[e^{\lambda Y_\theta(A,X)}]^2}.$$

The final result holds by using $\exp(\lambda R_\theta) \leq \mathbb{E}[e^{\lambda Y_\theta(A,X)}] \leq 1$ and $\mathbb{V}(e^{\lambda Y_\theta(A,X)}) \leq \lambda^2 M$ □

Proposition 5.10 (restated). *Given Assumptions 5.1 and 5.2 and assuming* $n \geq \max\left(\frac{8\log\frac{2}{\delta}}{3(M+R_\theta^2)\exp(R_\theta)}, \frac{\left(8M+\frac{8}{3}\right)\log\frac{2}{\delta}}{\exp(2R_\theta)}\right)$ *and setting*

$$\lambda = \sqrt{\frac{8\log\frac{2}{\delta}}{3n(M+R_\theta^2)\exp(R_\theta)}},$$

then with a probability of at least $1-\delta$ *(*$\delta \in (0,1)$*), the generalization error satisfies,*

$$\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \leq 2\sqrt{\frac{2(M+R_\theta^2)\log\frac{2}{\delta}}{n\exp(R_\theta)}}\left(\sqrt{\frac{M}{\exp(R_\theta)(M+R_\theta^2)}} + \frac{1}{\sqrt{3}}\right).$$

Proof. $n \geq \frac{8\log\frac{2}{\delta}}{3(M+R_\theta^2)\exp(R_\theta)}$, concludes that $\lambda \leq 1$. Hence, $n \geq \frac{\left(8M+\frac{8}{3}\right)\log\frac{2}{\delta}}{\exp(2R_\theta)} \geq \frac{\left(8\lambda^2 M+\frac{8}{3}\right)\log\frac{2}{\delta}}{\exp(2\lambda R_\theta)}$ and using Theorem 5.8 with $\gamma = \frac{1}{2}$ and applying union bound, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \\
&\leq \max\left(\frac{\lambda}{2}(M+R_\theta^2) + \frac{2}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} + \frac{1}{\exp(\lambda R_\theta)}\sqrt{\frac{2\mathbb{V}(e^{\lambda w_\theta(A,X)c(A,X)})\log\frac{2}{\delta}}{n\lambda^2}}\right. \\
&\quad \left., \frac{4}{3}\frac{\log\frac{2}{\delta}}{n\lambda\mu^{(\lambda)}} + \frac{2}{\exp(\lambda R_\theta)}\sqrt{\frac{2\mathbb{V}(e^{\lambda w_\theta(A,X)c(A,X)})\log\frac{2}{\delta}}{n\lambda^2}}\right) \\
&\leq \frac{\lambda}{2}(M+R_\theta^2) + \frac{4}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} + \frac{2}{\exp(\lambda R_\theta)}\sqrt{\frac{2\mathbb{V}(e^{\lambda w_\theta(A,X)c(A,X)})\log\frac{2}{\delta}}{n\lambda^2}}.
\end{aligned}$$

According to Lemma 5.3, we have

$$\mathbb{V}(e^{\lambda w_\theta(A,X)c(A,X)}) \leq \lambda^2\mathbb{V}(w_\theta(A,X)c(A,X)) \leq \lambda^2 M.$$

Hence we have,

$$\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \leq \frac{\lambda}{2}(M+R_\theta^2) + \frac{4}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} + \frac{2}{\exp(\lambda R_\theta)}\sqrt{\frac{2M\log\frac{2}{\delta}}{n}}. \tag{25}$$

Since $\lambda \leq 1$ and $\exp(\lambda R_\theta) \geq \exp(R_\theta)$. Replacing λ in RHS of (25),

$$\begin{aligned}
|\mathrm{LSE}_{\lambda^\star}(S,\pi_\theta) - \mathbb{E}[w_\theta(A,X)c(A,X)]| &\leq 2\sqrt{\frac{2}{3}\frac{(M+R_\theta^2)\log\frac{2}{\delta}}{n\exp(R_\theta)}} + \frac{2}{\exp(R_\theta)}\sqrt{\frac{2M\log\frac{2}{\delta}}{n}} \\
&= 2\sqrt{\frac{2(M+R_\theta^2)\log\frac{2}{\delta}}{n\exp(R_\theta)}}\left(\sqrt{\frac{M}{\exp(R_\theta)(M+R_\theta^2)}} + \frac{1}{\sqrt{3}}\right).
\end{aligned}$$

with a probability of at least $1-\delta$. □

E Proofs and details of αLSE

Proposition E.1. *Suppose that for each action, the number of samples of that action in the LBF dataset goes to infinity,* $(m[\tilde{a}] \to \infty, \forall \tilde{a} \in \mathcal{A})$. *Then we have,*

$$\lim_{\substack{m[\tilde{a}]\to\infty,\\ \forall\tilde{a}\in\mathcal{A}}} \hat{D}_{\alpha,n} = D_\alpha.$$

Proof. Set $Z_a = \frac{\pi_\theta^\alpha(a|X)}{\pi_0^{\alpha-1}(a|X)}$ and $Z_a^{(i)} = \frac{\pi_\theta^\alpha(a|X=x_i)}{\pi_0^{\alpha-1}(a|X=x_i)}$, where $X \sim P_X$, and $a \in \mathcal{A}$ is an arbitrary action. According to the strong law of large numbers,

$$\frac{1}{m[\tilde{a}]}\sum_{\substack{(x_i,a_i,p_i)\\ a_i=a}} \frac{p_i^\alpha}{\pi_\theta^{\alpha-1}(a|x_i)} = \frac{1}{m[\tilde{a}]}\sum_{i=1}^{m[\tilde{a}]} Z_a^{(i)} \to \mathbb{E}[Z_a] = \mathbb{E}_X\left[\frac{\pi_0^\alpha(a|X)}{\pi_\theta^{\alpha-1}(a|X)}\right],$$

with probability one. This statement is true for all $a \in \mathcal{A}$. Therefore, for all actions it holds. Hence with probability one,

$$\sum_{a\in\mathcal{A}} \frac{1}{m[\tilde{a}]} \sum_{\substack{(x_i,a_i,p_i)\\ a_i=a}} \frac{p_i^{\alpha}}{\pi_\theta^{\alpha-1}(a|x_i)} \sum_{a\in\mathcal{A}} \frac{1}{m[\tilde{a}]} \sum_{i=1}^{m[\tilde{a}]} Z_a^{(i)} \to \sum_{a\in\mathcal{A}} \mathbb{E}[Z_a] = \sum_{a\in\mathcal{A}} \mathbb{E}_X\left[\frac{\pi_0^{\alpha}(a|X)}{\pi_\theta^{\alpha-1}(a|X)}\right].$$

The proof is complete by the continuity of $\frac{1}{\alpha-1}\log(x)$ function. □

Proposition E.2 (Upper bound via α-Rényi Divergence for $\alpha \in (0,1)$)**.** *Given Assumption 6.1 and assuming that the cost function has bounded range* $[-C, 0]$*, then the following upper bound holds on the second moment of the weighted cost function,*

$$\mathbb{E}[w_\theta^2(A,X)c^2(A,X)] \le C^2 + \sqrt{\frac{2\sigma^2}{\alpha} D_\alpha(\pi_0(A|X)\otimes P_X \| \pi_\theta(A|X)\otimes P_X)}.$$

Proof. Note that,

$$\mathbb{E}_{\pi_0\otimes P_X}[w_\theta^2(A,X)c^2(A,X)] = \mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X)]. \tag{26}$$

Using the sub-Gaussian assumption, for all $\lambda \in \mathbb{R}$, we have,

$$\log\left(\mathbb{E}_{\pi_\theta\otimes P_X}[\exp(\alpha\lambda w_\theta(A,X)c^2(A,X))]\right) \le \alpha\lambda\mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X))] + \frac{\lambda^2\alpha^2\sigma^2}{2}. \tag{27}$$

Similarly, we have,

$$\begin{aligned}&\log\left(\mathbb{E}_{\pi_0\otimes P_X}[\exp((\alpha-1)\lambda w_\theta(A,X)c^2(A,X))]\right)\\ &\le -(1-\alpha)\lambda\mathbb{E}_{\pi_0\otimes P_X}[w_\theta(A,X)c^2(A,X))] + \frac{\lambda^2(1-\alpha)^2\sigma^2}{2}.\end{aligned} \tag{28}$$

Combining (27), (28) with Lemma B.4 for $\alpha \in (0,1)$, we have,

$$\begin{aligned}&D_\alpha(\pi_0\otimes P_X\|\pi_\theta\otimes P_X)\\ &\ge \frac{\alpha}{\alpha-1}\log\left[\mathbb{E}_{\pi_0\otimes P_X}[\exp(\lambda(\alpha-1)w_\theta(A,X)c^2(A,X))]\right]\\ &\quad -\log\left[\mathbb{E}_{\pi_\theta\otimes P_X}[\exp(\lambda\alpha w_\theta(A,X)c^2(A,X))]\right]\\ &\ge \alpha\lambda\Big(\mathbb{E}_{\pi_0\otimes P_X}[w_\theta(A,X)c^2(A,X)] - \mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X)]\Big) - \frac{\lambda^2\alpha\sigma^2}{2},\end{aligned} \tag{29}$$

Note that, we have a non-negative parabola in λ,

$$D_\alpha(\pi_0\otimes P_X\|\pi_\theta\otimes P_X) - \alpha\lambda A + \frac{\lambda^2\alpha\sigma^2}{2} \ge 0, \tag{30}$$

where $A = \Big(\mathbb{E}_{\pi_0\otimes P_X}[w_\theta(A,X)c^2(A,X)] - \mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X)]\Big)$ and the parabola's discriminant has to be non-positive, and we have,

$$\Big|\mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X)] - \mathbb{E}_{\pi_0\otimes P_X}[w_\theta(A,X)c^2(A,X)]\Big| \le \sqrt{\frac{2\sigma^2 D_\alpha(\pi_0\otimes P_X\|\pi_\theta\otimes P_X)}{\alpha}}. \tag{31}$$

Therefore, we have,

$$\begin{aligned}\mathbb{E}_{\pi_\theta\otimes P_X}[w_\theta(A,X)c^2(A,X)] &\le \mathbb{E}_{\pi_\theta\otimes P_X}[c^2(A,X)] + \sqrt{\frac{2\sigma^2 D_\alpha(\pi_0\otimes P_X\|\pi_\theta\otimes P_X)}{\alpha}}\\ &\le C^2 + \sqrt{\frac{2\sigma^2 D_\alpha(\pi_0\otimes P_X\|\pi_\theta\otimes P_X)}{\alpha}}.\end{aligned} \tag{32}$$

□

We can also discuss the connection of sub-Gaussian assumption with uniform coverage assumption [Wang et al., 2023, Gabbianelli et al., 2023].

The result in Proposition E.2 holds for $\alpha \in (0,1)$. For $\alpha > 1$, we provide the following proposition.

Proposition E.3 (Upper bound via α-Rényi Divergence for $\alpha \geq 1$)**.** *Under the same Assumptions in Proposition E.2, for $\alpha \geq 1$, we have,*

$$\mathbb{E}[w_\theta^2(A,X)c^2(A,X)] \leq C^2 + \sqrt{2\sigma^2 D_\alpha(\pi_0(A|X) \otimes P_X \| \pi_\theta(A|X) \otimes P_X)}.$$

Proof. The proof is similar to Proposition E.2 by replacing (28) with the following inequality, derived by applying Jensen inequality,

$$\log\left(\mathbb{E}_{\pi_0 \otimes P_X}[\exp((\alpha-1)\lambda w_\theta(A,X)c^2(A,X))]\right) \geq -(1-\alpha)\lambda \mathbb{E}_{\pi_0 \otimes P_X}[w_\theta(A,X)c^2(A,X))]. \tag{33}$$

□

Proposition 6.3 (restated). *Given Assumption 6.1 and assuming that the cost function has bounded range $[-C,0]$, then the generalization error of the LSE estimator satisfies,*

$$\mathfrak{L}_\alpha(\gamma, n, \lambda, R_\theta, \delta) \leq \text{gen}_\lambda(\pi_\theta) \leq \mathfrak{U}_\alpha(\gamma, n_u, \lambda, R_\theta, \delta),$$

where $U_\alpha = 1 + \sigma\sqrt{\frac{2D_\alpha}{\min(\alpha,1)}}$, $D_\alpha := D_\alpha(\pi_0(A|X) \otimes P_X \| \pi_\theta \otimes P_X)$, $n_u = n$ *for* $n_u \geq \frac{\left(2\lambda^2 M + \frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2 \exp(2\lambda R_\theta)}$*, and,*

$$\begin{aligned}
\mathfrak{L}_\alpha(\gamma, n, \lambda, R_\theta, \delta) &:= -\frac{\lambda}{2}U_\alpha - \frac{2}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} - \frac{1}{\exp(\lambda R_\theta)}\sqrt{\frac{2(U_\alpha + R_\theta^2)\log\frac{2}{\delta}}{n}}, \\
\mathfrak{U}_\alpha(\gamma, n_u, \lambda, R_\theta, \delta) &:= \frac{4}{3}\frac{\log\frac{2}{\delta}}{n\lambda\exp(\lambda R_\theta)} + \frac{2}{\exp(\lambda R_\theta)}\sqrt{\frac{2(U_\alpha + R_\theta^2)\log\frac{2}{\delta}}{n}}.
\end{aligned} \tag{34}$$

Proof. The proofs follows directly from combining Proposition E.2 with Theorem 5.8. □

Remark E.4 (Uniform Coverage Assumption). In the uniform coverage (overlap) assumption, it is assumed that

$$\sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} \frac{\pi_\theta(a|x)}{\pi_0(a|x)} = U_c < \infty. \tag{35}$$

In Proposition 6.3, we assume that the importance weighted of the squared cost function, i.e., $w_\theta(A,X)c^2(A,X)$, is σ-sub-Gaussian under $P_X \otimes \pi_0(A|X)$ and $P_X \otimes \pi_\theta(A|X)$. Given the constraint of a bounded cost function, the uniform coverage assumption (35) implies $\sigma = \frac{U_c}{2}$, leading to the validity of the result in Proposition E.2. It's important to highlight that the sub-Gaussian assumption is a weaker assumption compared to the uniform coverage assumption.

E.1 Comparison with KL-regularized risk

For KL divergence $\text{KL}(\pi_\theta(A|X) \otimes P_X \| \pi_0(A|X) \otimes P_X)$, it is required that $\pi_\theta(A|X) \otimes P_X$ would be absolute continuous [*] with respect to $\pi_0(A|X) \otimes P_X$. However, in α-Rényi divergence for $\alpha \in (0,1)$, it is required that $\pi_\theta(A|X) \otimes P_X$ would not be mutually singular[*] with respect to $\pi_0(A|X) \otimes P_X$.

Hence, the conditions imposed by α-Rényi divergence on the logging policy and the learning policy are comparatively more weaker than those dictated by KL divergence. This distinction can also be interpreted as an implicit exploration effect [Kocák et al., 2014, Strehl et al., 2010]. Consequently, in KL-regularized risks [Aminian et al., 2024, London and Sandler, 2019], the assumptions regarding $\pi_0(A|X) \otimes P_X$ and $\pi_\theta(A|X) \otimes P_X$ are more rigorous than those associated with α-Rényi regularization.

[*] $p(x)$ is absolutely continuous with respect to $q(x)$, if $p(A) = 0$ whenever $q(A) = 0$, for measurable $A \subset \mathcal{X}$.

[*] $p(x)$ is mutually singular with respect to $q(x)$, if there is an measurable event $A \subset \mathcal{X}$ such that $q(A) = 0$ and $p(\mathcal{X} \backslash A) = 0$.

F Regret bound

In this section, we provide an upper bound on the regret under LSE estimator.

Proposition F.1 (Regret Bound)**.** *Let θ^* be the optimal parameter that minimizes the true risk,*

$$\pi_{\theta^*} = \arg\inf_{\pi_\theta \in \Pi_\theta} R(\pi_\theta),$$

and $\hat{\theta}$ be the minimizer of the LSE estimator for a given dataset S,

$$\pi_{\hat{\theta}} = \arg\min_{\pi_\theta \in \Pi_\Theta} \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta).$$

Given the assumptions of Theorem 5.8, we have the following regret bound with a probability of at least $1-\delta$*,*

$$0 \le R(\pi_{\hat{\theta}}) - R(\pi_{\theta^*}) \le \mathfrak{U} - \mathfrak{L},$$

where $\mathfrak{U}$ and $\mathfrak{L}$ are defined in (12)*.*

Proof. We have,

$$R(\pi_{\hat{\theta}}) - R(\pi_{\theta^*}) = \underbrace{R(\pi_{\hat{\theta}}) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\hat{\theta}})}_{\le \mathfrak{U}_{2\delta}} + \underbrace{\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\hat{\theta}}) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\theta^*})}_{\le 0} + \underbrace{\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\theta^*}) - R(\pi_{\theta^*})}_{\le -\mathfrak{L}_{2\delta}}$$

where $\mathfrak{U}_{2\delta} = \mathfrak{U}(\gamma, n_u, \lambda, R_\theta, M, 2\delta), \mathfrak{L}_{2\delta} = \mathfrak{L}(\gamma, n_u, \lambda, R_\theta, M, 2\delta)$. The first and third bounds each are true with a probability of at least $1-\delta$, according to the proof of the 5.8 before applying the final union bound. Also the second inequality is a.s. true because $\pi_{\hat{\theta}}$ is the minimizer of $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)$. Putting together, using union bound, we have with a probability of at least $1-2\delta$,

$$R(\pi_{\hat{\theta}}) - R(\pi_{\theta^*}) \le \mathfrak{U}_{2\delta} - \mathfrak{L}_{2\delta},$$

where replacing 2δ with δ proves the proposition. □

Remark F.2. Setting $\lambda = \lambda(n) = O(\frac{1}{\sqrt{n}})$, the abovementioned proposition provides a regret bound of $O(\frac{1}{\sqrt{n}})$.

G Noisy propensity scores

To model the noisy propensity scores, we consider $\hat{\pi}_0(a|x)$ as the noisy version of the logging policy $\pi_0(a|x)$. Similarly, we define $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta)$ for the LSER on the noisy data samples $\hat{S}$, with noisy propensity scores.

Definition G.1. The log-sum error of the noisy (or estimated) propensity score $\hat{\pi}_0(a|x)$ is defined as

$$\Delta(\hat{\pi}_0, \pi_0) = \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda \hat{w}_\theta(A, X) c(A, X))] - \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda w_\theta(A, X) c(A, X))]. \tag{36}$$

where $\hat{w}_\theta(A, X) = \frac{\pi_\theta(A|X)}{\hat{\pi}_0(A|X)}$.

Definition G.1 captures a notion of bias in the noise that is applied to the propensity score. It indicates the change in the population form of the LSE estimator. Analogously, for the Monte Carlo estimator, the change in the expected value shows the bias of the noise, and for additive noise, the zero-mean assumption ensures that in expectation, the noisy value is the same as the original value. For the LSE estimator instead, we require the exponential forms to be close to each other. It is also inspired by influence function definition and robust statistic [Ronchetti and Huber, 2009, Christmann and Steinwart, 2004].

Let us assume the following assumption for noisy propensity scores.

Assumption G.2 (Bounded true risk under noise)**.** The learning policy $\pi_\theta(A|X)$, cost function $c(A, X)$ and P_X are such that the expected true risk under noise is bounded, $\mathbb{E}_{P_X \otimes \pi_0(A|X)}[\hat{w}_\theta(A, X) c(A, X)] \le R_{\hat{\theta}}$.

In the following proposition, we study the robustness of the LSE estimator with respect to noisy propensity scores.

Proposition G.3. *Given Assumption 5.1 and Assumption G.2, and assuming* $n > \frac{\frac{4}{3}\mu_{\min}+4}{\mu_{\min}^2}\log\frac{4}{\delta}$ *where* $\mu_{\min} = \min\left(e^{\lambda R_\theta}, e^{\lambda R_{\hat\theta}}\right)$*, then with probability at least* $1-\delta$*, it holds that,*

$$\left|\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S},\pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) - \Delta(\hat\pi_0,\pi_0)\right| \le \frac{2\epsilon}{\lambda}\Big(\frac{1}{e^{\lambda R_{\hat\theta}}} + \frac{1}{e^{\lambda R_\theta}}\Big),$$

where, $\epsilon = \frac{\log\frac{4}{\delta}}{3n} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}$.

Proof. Set $Y_\theta(A,X) = w_\theta(A,X)c(A,X)$, $\hat{Y}_\theta(A,X) = \hat{w}_\theta(A,X)c(A,X)$, $u_i = \frac{1}{\lambda}\left(e^{\hat{y}_i} - e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mu\right)$ and $v_i = \frac{1}{\lambda}(e^{y_\theta(a_i,x_i)} - \mu)$, where $\mu = \mathbb{E}[e^{\lambda Y_\theta(A,X)}]$. We have $-\frac{\mu}{\lambda} \le v_i \le \frac{1}{\lambda} - \frac{\mu}{\lambda}$ and $-\frac{e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mu}{\lambda} \le u_i \le \frac{1}{\lambda} - \frac{e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mu}{\lambda}$. Then, using the one-sided Bernstein's inequality (Lemma B.3), and changing variables (Lemma B.5), we have:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda y_\theta(a_i,x_i)} - \mathbb{E}[e^{\lambda Y_\theta(A,X)}] > \frac{(1-\mu)\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \le \delta,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda y_\theta(a_i,x_i)} - \mathbb{E}[e^{\lambda Y_\theta(A,X)}] < -\frac{\mu\log\frac{1}{\delta}}{3n} - \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \le \delta,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda\hat{y}_i} - e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] > \frac{(1-e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mu)\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda\hat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \le \delta,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda\hat{y}_i} - e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] < -\frac{e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mu\log\frac{1}{\delta}}{3n} - \sqrt{\frac{\mathbb{V}\left(e^{\lambda\hat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \le \delta.$$

Therefore, with probability at least $1-2\delta$, for $\epsilon_2 < \frac{1}{2}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]$, we have,

$$\begin{aligned}
&\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S},\pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)\\
&= \frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^{n} e^{\lambda\hat{y}_i}}{\sum_{i=1}^{n} e^{\lambda y_\theta(a_i,x_i)}}\right)\\
&\le \frac{1}{\lambda}\log\left(\frac{e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \epsilon_1}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \epsilon_2}\right)\\
&= \frac{1}{\lambda}\left(\log\left(e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \epsilon_1\right) - \log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \epsilon_2\right)\right)\\
&\le \frac{1}{\lambda}\Bigg(\log\left(e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) + \frac{\epsilon_1}{e^{\lambda\Delta(\hat\pi_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\\
&\qquad - \left(\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) - \frac{\epsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \epsilon_2}\right)\Bigg)\\
&\le \Delta(\hat\pi_0,\pi_0) + \frac{1}{\lambda}\left(\frac{\epsilon_1}{\mathbb{E}[e^{\lambda\hat{Y}_\theta(A,X)}]} + \frac{2\epsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right)\\
&\le \Delta(\hat\pi_0,\pi_0) + \frac{2}{\lambda}\left(\frac{\epsilon_1}{\mathbb{E}[e^{\lambda\hat{Y}_\theta(A,X)}]} + \frac{\epsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right).
\end{aligned}$$

where

$$\epsilon_1 = \frac{(1-\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}])\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda \hat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}},$$
$$\epsilon_2 = \frac{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}.$$

Similarly, with probability at least $1-2\delta$ we have, given $\epsilon_3 < \frac{1}{2}\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]$,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) \geq \Delta(\hat{\pi}_0, \pi_0) - \frac{2}{\lambda}\left(\frac{\epsilon_3}{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]} + \frac{\epsilon_4}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right),$$

where,

$$\epsilon_3 = \frac{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda \hat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}},$$
$$\epsilon_4 = \frac{(1-\mathbb{E}[e^{\lambda Y_\theta(A,X)}))]\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}.$$

Therefore, with probability at least $1-4\delta$ we have,

$$\Delta(\hat{\pi}_0, \pi_0) - \frac{2}{\lambda}\left(\frac{\epsilon_3}{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]} + \frac{\epsilon_4}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right) \leq \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)$$
$$\leq \Delta(\hat{\pi}_0, \pi_0) + \frac{2}{\lambda}\left(\frac{\epsilon_1}{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]} + \frac{\epsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right).$$

We have for $i \in [4]$,

$$\epsilon_i \leq \frac{\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}.$$

So, replacing δ with $\delta/4$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) - \Delta(\hat{\pi}_0, \pi_0)\right| \\
&\leq \frac{2}{\lambda}\left(\frac{\log\frac{4}{\delta}}{3n} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)\left(\frac{1}{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]} + \frac{1}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right) \\
&\leq \frac{2}{\lambda}\left(\frac{\log\frac{4}{\delta}}{3n} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)\frac{2\epsilon}{\lambda}\Big(\frac{1}{e^{\lambda R_{\hat{\theta}}}} + \frac{1}{e^{\lambda R_\theta}}\Big),
\end{aligned}$$

which is true given $\frac{\log\frac{4}{\delta}}{3n} + \sqrt{\frac{\log\frac{4}{\delta}}{n}} < \frac{1}{2}\min\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}], \mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]\right)$. According to Lemma B.6, this is satisfied when,

$$n > \frac{\frac{4}{3}\mu_{\min} + 4}{\mu_{\min}^2}\log\frac{4}{\delta}.$$

□

According to Proposition G.3, we can derive an upper bound on the range of LSER when the logging policy is noisy, with probability $1-\delta, \delta \in (0,1)$.

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) + \Delta(\hat{\pi}_0, \pi_0) - B\epsilon \leq \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) \leq \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) + \Delta(\hat{\pi}_0, \pi_0) + B\epsilon, \qquad (37)$$

where $\epsilon = \frac{2}{\lambda}\left(\frac{\log\frac{4}{\delta}}{3n} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)$, and $B = \frac{1}{\mathbb{E}[e^{\lambda \hat{Y}_\theta(A,X)}]} + \frac{1}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}$. Hence, the range of change of LSER is an interval of size $2B\epsilon$ which is $O(\frac{1}{\sqrt{n}})$. Hence as the number of samples increases, independent of the type and intensity of the noise, the variation of LSER goes to zero with a rate of $n^{-1/2}$. This can be interpreted as the robustness of LSER to noise.

As discussed in the following Corollary, the small range of variation of the noise gives an upper bound on the variance of the noisy version of the LSER.

Corollary G.4. *Under the same assumptions in Proposition G.3, then the following upper bound holds on the variance of the LSE estimator under noisy propensity scores with probability at least* $1-\delta, \delta \in (0,1)$*,*

$$\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta)) \leq 2\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) + 2B^2\epsilon^2,$$

where $\epsilon = \frac{2}{\lambda}\left(\frac{\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$*, and* $B = \left(\frac{1}{e^{\lambda R_{\hat{\theta}}}} + \frac{1}{e^{\lambda R_\theta}}\right)$*.*

Proof. As $\Delta(\hat{\pi}_0, \pi_0)$ is a constant with respect to $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta)$ and $\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)$, then we have,

$$\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \leq \left(\frac{2B\epsilon}{2}\right)^2 = B^2\epsilon^2.$$

Therefore,

$$\begin{aligned}
\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta)) &= \mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) + \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \\
&= \mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) + \mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \\
&\quad + 2\mathrm{Cov}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta), \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \\
&\leq \mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) + \mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) \\
&\quad + 2\sqrt{\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))} \\
&= \left(\sqrt{\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))} + \sqrt{\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))}\right)^2 \\
&\leq \left(\sqrt{\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))} + B\epsilon\right)^2 \leq 2\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)) + 2B^2\epsilon^2.
\end{aligned}$$

□

From Corollary G.4, we have an upper bound on the variance of the LSER under noisy propensity scores, in terms of the variance of the LSER under clean propensity scores. Therefore, if $\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))$ is bounded, then we can expect bounded $\mathbb{V}(\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(\hat{S}, \pi_\theta))$.

G.1 Robustness to outliers

In this section, we provide a lower bound on LSE under outlier samples, where the absolute value of the importance-weighted cost is large.

Proposition G.5. *Let* p *proportion of the importance-weighted costs be set to a large negative value* $-C$*. Set the newly obtained dataset as* S_p*. Under Assumption 5.2 and Assumption 5.1, for any* $t < \exp(\lambda R_\theta)$ *with a probability at least* $1 - \exp\left(\frac{-n(\exp(\lambda R_\theta) - t)^2}{M + \frac{1}{3}(\exp(\lambda R_\theta) - t)}\right)$ *we have,*

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S_p) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) \geq -\frac{p}{\lambda(1-p)t}.$$

Proof. Let $Z' = \sum_{i=1}^{np} e^{\lambda y_{i,\theta}}$, $Y_\theta(A,X) = c(A,X)w_\theta(A,X)$, and $Z = \frac{1}{n(1-p)}\sum_{i=np+1}^{n} e^{\lambda y_{i,\theta}}$, where $y_{i,\theta} = c(a_i, x_i)w_\theta(a_i, x_i)$.

$$\begin{aligned}
\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S_p) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) &= \frac{1}{\lambda}\log\left(\frac{np\exp(-\lambda C) + n(1-p)Z}{Z' + n(1-p)Z}\right) \\
&\geq \frac{1}{\lambda}\log\left(\frac{np\exp(-\lambda C) + n(1-p)Z}{np + n(1-p)Z}\right).
\end{aligned}$$

where the last inequality holds because $Z' \leq np$. Using Bernstein's inequality, for any $t' < \mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}]$, if we set $\epsilon = \mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}] - t'$, we have with probability at least $1 - \exp\left(\frac{-\frac{n}{2}\epsilon^2}{\mathbb{V}+\frac{1}{3}\epsilon}\right)$,

$$Z - \mathbb{E}[Z] \geq -\epsilon \leftrightarrow Z \geq \mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}] - \epsilon = t'.$$

where $\mathbb{V} = \mathbb{V}(e^{\lambda Y_\theta(A,X)})$. Hence we have,

$$\begin{aligned}
\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S_p) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) &\geq \frac{1}{\lambda} \log\left(\frac{np\exp(-\lambda C) + n(1-p)Z}{np + n(1-p)Z}\right) \\
&\geq \frac{1}{\lambda} \log\left(\frac{np\exp(-\lambda C) + n(1-p)t'}{np + n(1-p)t'}\right) \\
&\geq \frac{1}{\lambda} \log\left(\frac{(1-p)t'}{p + (1-p)t'}\right) \\
&= -\frac{1}{\lambda} \log\left(\frac{p}{(1-p)t'} + 1\right) \geq -\frac{p}{\lambda(1-p)t'}.
\end{aligned}$$

Now if $t < \exp(\lambda R_\theta) \leq \mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}]$, we have with probability at least $1 - \exp\left(\frac{-n(\exp(\lambda R_\theta)-t)^2}{M+\frac{1}{3}(\exp(\lambda R_\theta)-t)}\right) \leq 1 - \exp\left(\frac{-n(\exp(\lambda R_\theta)-t)^2}{\mathbb{V}+\frac{1}{3}(\exp(\lambda R_\theta)-t)}\right) \leq 1 - \exp\left(\frac{-n(\mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}]-t)^2}{M+\frac{1}{3}(\mathbb{E}[e^{\lambda c(A,X) w_\theta(A,X)}]-t)}\right)$,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S_p) - \hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S) \geq -\frac{p}{\lambda(1-p)t}.$$

□

The Proposition G.5 provides a lower bound on the LSE estimator when some proportion of the data is corrupted by being set to a large value. Although the LSE is not a bounded estimator, the provided bound is independent of C, the scale of the corrupted value.

G.2 Gamma noise discussion

For statistical modeling of the noisy propensity scores, as discussed in [Zhang et al., 2023], suppose that the logging policy is a softmax policy with respect to a.

$$\pi_0(a|x) = \mathrm{softmax}(f_{\phi^*}(x,a)). \tag{38}$$

Where f_ϕ is a function parameterized by ϕ that indicates the policy's function output before softmax operation and ϕ^* is the parameter of this function for the true logging policy.

We have an estimation of the function $f_{\phi^*}(x,a)$, as $f_{\hat{\phi}}(x,a)$ and we can model the error in the estimation of $f_{\phi^*}(x,a)$ as a random variable Z.

$$f_{\hat{\phi}}(X,A) = f_{\phi^*}(X,A) + Z(X,A).$$

Then we have,

$$\hat{\pi}_0 = \mathrm{softmax}(f_{\phi^*} + Z) \propto e^{Z}\pi_0.$$

One straightforward choice for the distribution of Z is normal distribution, which results in a log-normal noise for propensity scores. Motivated by Halliwell [2018], we use a negative log-gamma distribution for Z, which results in an inverse Gamma multiplicative noise on the propensity scores. Negative log-gamma distribution is skewed towards negative values, resulting in inverse gamma noise on the logging policy which is skewed towards values less than one. This pushes the propensity scores $\frac{\pi_\theta}{\pi_0}$ towards the higher variance, i.e., the logging policy is near zero and the importance weight becomes large.

G.3 Robustness

To study our method for noise robustness, we define two different scenarios based on the definition in G.2.

G.3.1 Model-based propensity scores

First, we consider a model-based setting in which the noise is modeled with an inverse Gamma distribution. We use inverse gamma distribution $1/U$ as a multiplicative noise, so we have,

$$\hat{\pi}_0 = \frac{1}{U}\pi_0 \rightarrow \hat{w}_\theta(A, X) = U w_\theta(A, X).$$

which results in a multiplicative gamma noise on the importance-weighted cost. We choose $U \sim \mathrm{Gamma}(b, b)$, so $\mathbb{E}[U] = 1$. Hence, the expected value of the noisy version is the same as the original noiseless variable.

$$\mathbb{E}[U w_\theta(A, X) c(A, X)] = \mathbb{E}[U]\mathbb{E}[w_\theta(A, X) c(A, X)] = \mathbb{E}[w_\theta(A, X) c(A, X)].$$

Note that we have

$$\mathbb{E}\left[e^{\lambda w_\theta(A,X)c(A,X)U}\right] = \mathbb{E}\left[\left(\frac{1}{1 - \lambda w_\theta(A, X)c(A, X)/b}\right)^b\right],$$

Therefore, $\mathbb{E}[e^{\lambda U w_\theta(A,X)c(A,X)}]$ converges to $\mathbb{E}[e^{\lambda w_\theta(A,X)c(A,X)}]$ for $b \rightarrow \infty$. Furthermore, we can assume that for a large value b, $\Delta(\hat{\pi}_0, \pi_0) \approx 0$ and using Proposition G.3, with a probability of at least $1 - \delta$, we have,

$$\left|\hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(\hat{S}, \pi_\theta) - \hat{\mathrm{R}}^\lambda_{\mathrm{LSE}}(S, \pi_\theta)\right| \leq \epsilon\Big(\frac{1}{\mathbb{E}[e^{\lambda \hat{w}_\theta(A,X)c(A,X)}]} + \frac{1}{\mathbb{E}[e^{\lambda w_\theta(A,X)c(A,X)}]}\Big). \quad (39)$$

Therefore, if the domain of the inverse Gamma noise is small enough, then the amount of change in the LSE is small. In addition, we can arbitrarily decrease the deviation from the original noiseless LSER by increasing the size of the LBF dataset.

G.3.2 Estimated propensity scores

We also consider another setting in which we empirically test the robustness of our estimator in a more realistic and applicable setting. Suppose that the propensity scores are completely missing from the dataset. In this situation, one trivial solution is to estimate the propensity as a function of state and action. The estimated propensity score is then a noisy version of the true propensity score. Moreover, complying with the setting mentioned in G.2, if we model the error Z of the estimated logit values as an additive negative Log-Gamma distribution, we would have the same model as in the model-based setting.

In order to estimate the propensity scores computed by the logging policy, we consider the same model as the logging policy,

$$\tilde{\pi}(A|X) = \mathrm{softmax}(f_{\tilde{\phi}}(X, A)).$$

This model is trained to act similarly to the logging policy, hence it is trained with cross-entropy loss, $\mathrm{CE}(., .)$, with respect to the actions that are logged in the datasets by the logging policy.

$$L(\tilde{\phi}, S) = \sum_{i=1}^{n} \mathrm{CE}(\tilde{\pi}(\cdot|x_i), a_i).$$

where, $\mathrm{CE}(\hat{\mathbf{a}}, a) = -\log(\hat{\mathbf{a}}_a)$.

The intuition with such a training procedure becomes more clear with a simple example. Suppose that we have a single state x_0, and two different actions, a_1, a_2. Now suppose that the logging policy selects actions with probabilities $p, (1 - p)$ respectively. Hence, p proportion of the dataset is created with action a_1. Hence the final objective for the propensity estimator would become,

$$L(\tilde{\phi}, S) = -p\log(\tilde{\pi}(\cdot|x_0)_1) - (1-p)\log(\tilde{\pi}(\cdot|x_0)_2) = -(pf_1 + (1-p)f_2) + \log(\exp(f_1) + \exp(f_2))$$

where $f_1 = f_{\tilde{\phi}}(s, a_1)$, $f_2 = f_{\tilde{\phi}}(s, a_2)$. The abovementioned objective is minimized when,

$$\frac{e^{f_1}}{e^{f_1} + e^{f_2}} = p,$$

which means that the propensity estimator should give exactly the same scores as the logging policy.

Table 5: Statistics of the datasets used in our experiments. For image datasets the 2048-dimensional features from pretrained ResNet-50 are used.

Data set	training samples	test samples	number of actions	Dimension
FMNIST	$60,000$	10000	10	2048
EMNIST	$60,000$	10000	10	2048
CIFAR-10	$50,000$	10000	10	2048
Letter	$20,000$	–	26	16
Open Bandit Dataset	$12,357,200$	-	80	27

H Experiment details

Datasets: In addition to two datasets EMNIST and CIFAR-10, we also run our estimator over Fashion-MNIST (FMNIST) [Xiao et al., 2017], UCI Letter Recognition [Slate, 1991], and Open Bandit Dataset [Saito et al., 2020b]. In CIFAR-10, EMNIST, and FMNIST datasets we have separate validation and test sets. For Letter and Open Bandit Dataset (OPD) we split them into training, validation, and test sets with a ratio of 0.8, 0.1, 0.1, respectively. The details of these datasets are demonstrated in Table 5. More details on the OPD dataset is discussed in section J.

Setup details: We use mini-batch SGD as an optimizer for all experiments. The learning learning used for EMNIST and FMNIST datasets is 0.001, for CIFAR-10 and OPD it is 0.01 and for Letter it's 1.0 with cosine annealing scheduler. Also, we use early stopping in our training phase and the maximum number of epochs is 300. For the image datasets, EMNIST, FMNIST, and CIFAR-10, we use the last layer features from ResNet-50 model pretrained on the ImageNet dataset [Deng et al., 2009].

Imbalance LBF dataset details: As discussed in Section 7, we aim to create an imbalanced dataset with $\hat{m}[i]$ samples for action i, for each $1 \leq i \leq k$. Hence, we will (over-)sample each data sample such that the expected number of samples of i-th action would be $\hat{m}[i]$. To achieve this, let $p[i] = \frac{\hat{m}[i]}{m[i]} = l_i + \alpha_i, l_i \in \mathbb{Z}, \alpha_i \in [0,1)$. We repeat each sample l_i times and with probability α_i, add another instance of the sample.

H.1 Hyper-parameter tuning

In order to find the value for each hyper-parameter, we put aside a part of the training dataset as a validation set and find the parameter that results in the highest accuracy on the validation set, and then we report the method's performance on the test set.

There are two main hyper-parameters in our setup. First, λ in the LSE estimator, and second β as regularization multiplier for α-Rényi regularizer. Note that, the parameter α for α-Rényi belongs to $(0,1)$. Therefore, for simplifying the hyper-parameter tuning process, we set $\alpha = \frac{1}{1+\lambda}$, and the final objective for regularized LSE estimator via α-Rényi divergence is,

$$\hat{\mathrm{R}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) + \beta \hat{D}_{\frac{1}{1+\lambda}, n}. \tag{40}$$

In order to tune λ we use grid search over the values in $\{0.01, 0.1, 1, 10, 100\}$ and to tune β parameter, we use Optuna, a hyper-parameter optimization Python-based library, over the range $[0.01, 10]$ with 3 trials and 3 runs for each trial. The reason for using Optuna is to reduce the number of trials and find reasonable values for hyper-parameters more efficiently.

Hyper-parameter tuning for PM, ES, IPS-KL, and IX estimators: For the PM, ES, and IX estimators, grid search will be used for hyper-parameter tuning. To tune the PM parameter λ, we will use $\lambda \in \{0, 0.1, 0.3, 0.5, 0.8\}$ values. For the ES estimator, the parameter α will be varied across $\alpha \in \{0.1, 0.4, 0.7, 1\}$. For the IX estimator, the γ parameter will be tested with values in the set $\gamma \in \{0.01, 0.1, 1, 10, 100\}$.

To tune the β parameter of the IPS-KL estimator (the regularization coefficient), we use Optuna over the range $[0.01, 5]$ with 6 trials and 3 runs for each trial.

α-Rényi regularization parameter: For the α parameter in the α-Rényi regularization component, we establish the α value as $\frac{1}{1+\lambda}$. This approach ensures that α remains within the range $0 < \alpha < 1$, making the regularization term $D_\alpha(P|Q)$ convex with respect to both P and Q.

As a result, if we set λ to 0, the α-LSE estimator will have the following modifications: the LSE component of the estimator will transform into an IPS estimator, and the α-Rényi component will convert to KL Divergence. Therefore, the IPS-KL estimator can be considered a special case of the α-LSE estimator.

Approximation for $\boldsymbol{\lambda}$ value: Although we use grid search to tune the λ in our algorithm, inspired by Proposition 5.10, we can select the following value,

$$\lambda^* = \frac{1}{n^{1/2}}, \tag{41}$$

where n is the batch size. For example, we have tested λ^* on EMNIST dataset with $n \in \{512, 256, 128, 64, 16\}$ with corresponding values $\lambda^* \in \{0.044, 0.0625, 0.088, 0.125, 0.25\}$ which its results are presented in the Table 6. We can observe that the suggested value of $\lambda^* = \frac{1}{\sqrt{n}}$ does not only have a theoretical generalization bound of $O(\frac{1}{\sqrt{n}})$ (according to 5.10), but also achieves reasonable performance in experiments. Hence, we can choose λ^* to avoid the overhead of grid search.

Table 6: Performance of λ^* with $n \in \{16, 64, 128, 256, 512\}$ on EMNIST dataset and $\tau = 1$.

n λ	16	64	128	256	512
0.01	92.83 ± 0.10	91.52 ± 0.01	90.26 ± 0.02	88.71 ± 0.26	85.43 ± 0.44
0.1	$\mathbf{92.83 \pm 0.01}$	91.45 ± 0.01	90.37 ± 0.02	88.93 ± 0.10	85.50 ± 0.58
1	92.66 ± 0.01	$\mathbf{91.66 \pm 0.02}$	$\mathbf{90.76 \pm 0.02}$	$\mathbf{89.54 \pm 0.01}$	$\mathbf{87.79 \pm 0.01}$
10	91.33 ± 0.01	89.48 ± 0.09	88.86 ± 0.05	88.03 ± 0.03	86.73 ± 0.03
$\boldsymbol{\lambda^*}$	92.78 ± 0.01	91.52 ± 0.05	90.38 ± 0.05	88.83 ± 0.02	85.09 ± 0.51

H.2 Code

The code for this study is written in Python. We use Pytorch for the training of our model. The supplementary material includes a zip file named rl_without_reward.zip with the following files:

- **preprocess_raw_dataset_from_model.py**: The code to generate the base pre-processed version of the datasets with raw input values.
- **preprocess_feature_dataset_from_model.py**: The code to generate the base pre-processed version of the datasets with pre-trained features.
- The **data** folder consists of any potentially generated bandit dataset (which can be generated by running the scripts in code).
- The **code** folder contains the scripts and codes written for the experiments.
  - **requirements.txt** contains the Python libraries required to reproduce our results.
  - **readme.md** includes the syntax of different commands in the code.
  - **accs**: A folder containing the result reports of different experiments.
  - **saved_logs**: Training log for different experiments.
  - **data.py** code to load data for image datasets.
  - **eval.py & eval_rec2.py** code to evaluate estimators for image datasets and open bandit dataset.
  - **config**: Contains different configuration files for different setups.
  - **runs**: Folder containing different batch running scripts.
  - **loss.py**: Script of our loss functions including LSE and α-LSE.
  - **train_logging_policy.py**: Script to train the logging policy.
  - **create_bandit_dataset.py**: Code for the generation of the bandit dataset using the logging policy.

  - **main_semi_ot.py**: Main training code which implements different methods proposed by our paper.
  - **main_semi_rec2.py**: Main training code for Open Bandit dataset.
- The **prepare_real_data** folder contains the scripts and codes written for open bandit dataset.
  - **create.ipynb**: The notebook for preparing open bandit dataset.
  - **data**: A folder containing open bandit dataest data.

To use this code, the user needs to download and store the dataset using *preprocess_raw_dataset_from_model.py* script. All downloaded data will be stored in *data* directory. Then, to train the logging policy, the *code/train_logging_policy.py* should be run. Then, by using *code/create_bandit_dataset.py*, the LBF dataset corresponding to the experiment setup, will be created. Finally, to train the desired estimator, the user should use *code/main_semi_ot.py* script. To prepare and utilize the Open Bandit dataset, users should follow the same procedure as outlined, using scripts that have a 'rec2' suffix.

Computational resources: We have taken all our experiments using 3 servers, one with a nvidia 1080 Ti and one with two nvidia GTX 4090, and one with three nvidia 2070-Super GPUs.

I Additional experiments

We present the results of our experiments for EMNIST and FMNIST in Table 7, for CIFAR-10 and LETTER in Table 9, and for Open Bandit Dataset in Table 12.

I.1 Experiment discussion

As we can observe in the results for different scenarios and datasets, our estimator, equipped with our proposed regularization, shows dominant performance among other baselines. The details of the number of best-performing estimator is provided in Table 11.

In the noisy scenario, where noise robustness is critical, increasing the noise on the propensity scores by reducing the b value results in a marked decrease in the performance of all estimators, with the notable exception of α-LSE, which exhibits superior noise robustness.

One of the key advantages of the α-LSE estimator is its consistent performance across various scenarios. Even in the rare instances where it does not outperform other methods, the performance gap is minimal. In contrast, other estimators either maintain lower performance in most settings (e.g., PM and ES) or demonstrate significantly poor performance in specific settings and datasets (e.g., PM+SM).

In all four datasets, without noise, increasing τ has a negligible impact on the estimators. However, in noisy scenarios, a higher τ leads to decreased performance. This happens because as τ increases, the logging policy distribution approaches a uniform distribution, making it easier for noise to affect the argmax value, thereby reducing the estimators' performance. Notably, the α-LSE estimator demonstrates better robustness compared to other estimators, consistently showing superior performance in all noisy setups when $b = 0.01$. (Note that IPS-KL is a special case of α-LSE.)

Moreover, the impact of α-R'enyi regularization is clearly evident in the observed outcomes. For the EMNIST, FMNIST, and Letter datasets, this regularization significantly enhances the overall accuracy and generalization capabilities of the α-LSE estimator. In contrast, for the CIFAR-10 dataset, the effectiveness of regularization is limited, primarily because the linear model has reached its maximum capacity. As a result, regularization has a minimal effect on improving accuracy. However, it greatly strengthens the α-LSE estimator's resilience to noise in propensity score true values, showing a marked improvement in noise robustness. This same observation holds true for scenarios with data imbalance.

Table 7: Comparison of different base algorithms LSE, PM, ES, IX accuracy for EMNIST and FMNIST with different qualities of logging policy ($\tau \in \{1, 2, 5, 10, 20\}$) and clean/ noisy propensity scores with $b \in \{5, 0.01\}$ and imbalance scenario ($\nu \in \{3, 9, 20\}$). The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	ν	**LSE**	**PM**	**ES**	**IX**	**Logging Policy**
EMNIST	1	–	–	88.49 ± 0.04	**89.19 ± 0.03**	88.61 ± 0.06	88.33 ± 0.13	88.08
		5	–	**89.16 ± 0.03**	88.94 ± 0.05	88.48 ± 0.03	88.51 ± 0.23	88.08
		0.01	–	**86.07 ± 0.01**	85.62 ± 0.10	85.71 ± 0.04	81.39 ± 4.02	88.08
		–	3	87.83 ± 0.10	**88.81 ± 0.11**	63.64 ± 0.53	64.82 ± 7.29	88.08
		–	9	88.01 ± 0.05	**88.29 ± 0.10**	56.09 ± 0.03	56.08 ± 0.02	88.08
		–	20	**88.00 ± 0.07**	87.88 ± 0.07	56.26 ± 0.02	56.32 ± 0.01	88.08
	2	–	–	**89.00 ± 0.01**	88.75 ± 0.03	88.28 ± 0.11	88.20 ± 0.05	74.33
		5	–	**89.01 ± 0.03**	88.82 ± 0.07	88.43 ± 0.20	88.50 ± 0.07	74.33
		0.01	–	**86.83 ± 0.06**	83.60 ± 0.12	76.74 ± 3.52	73.73 ± 7.40	74.33
	5	–	–	**88.70 ± 0.01**	88.69 ± 0.05	88.60 ± 0.15	87.58 ± 0.24	39.16
		5	–	**88.63 ± 0.04**	88.43 ± 0.03	90.51 ± 0.13	88.30 ± 0.07	39.16
		0.01	–	**84.08 ± 0.02**	83.00 ± 0.06	83.09 ± 0.03	81.11 ± 3.17	39.16
	10	–	–	88.59 ± 0.03	**88.61 ± 0.04**	88.38 ± 0.08	87.43 ± 0.19	79.43
		5	–	88.42 ± 0.07	**88.43 ± 0.07**	88.39 ± 0.10	88.39 ± 0.06	79.43
		0.01	–	**82.15 ± 0.21**	80.85 ± 0.29	81.07 ± 0.07	77.49 ± 2.77	79.43
		–	3	86.96 ± 0.01	**87.30 ± 0.03**	61.74 ± 0.07	58.76 ± 3.96	79.43
		–	9	**86.13 ± 0.04**	85.84 ± 0.05	55.99 ± 0.04	57.08 ± 3.72	79.43
		–	20	80.50 ± 2.47	**83.36 ± 0.18**	56.29 ± 0.08	56.25 ± 0.02	79.43
FMNIST	1	–	–	76.38 ± 0.03	**78.54 ± 0.01**	72.90 ± 2.35	69.12 ± 0.26	78.38
		5	–	73.20 ± 2.43	**78.43 ± 0.03**	70.38 ± 2.59	70.80 ± 2.38	78.38
		0.01	–	**74.08 ± 1.64**	70.74 ± 0.16	57.93 ± 2.66	63.34 ± 3.64	78.38
		–	3	76.52 ± 0.15	**78.56 ± 0.09**	64.89 ± 2.72	61.55 ± 0.96	78.38
		–	9	76.73 ± 0.20	**78.71 ± 0.07**	51.23 ± 1.59	51.35 ± 1.35	78.38
		–	20	**76.71 ± 0.16**	73.37 ± 3.92	45.27 ± 0.04	45.26 ± 0.01	78.38
	2	–	–	**78.55 ± 0.17**	78.40 ± 0.09	69.33 ± 2.40	70.62 ± 2.40	66.94
		5	–	77.97 ± 0.09	**78.20 ± 0.04**	70.49 ± 2.43	69.25 ± 0.11	66.94
		0.01	–	**73.17 ± 1.93**	69.26 ± 0.28	60.87 ± 2.48	60.18 ± 3.52	66.94
	5	–	–	77.32 ± 0.03	**77.40 ± 0.06**	69.19 ± 0.32	69.08 ± 0.15	37.76
		5	–	77.16 ± 0.09	**77.29 ± 0.04**	69.54 ± 3.70	68.80 ± 0.46	37.76
		0.01	–	**71.22 ± 0.07**	65.83 ± 0.09	56.49 ± 2.08	54.76 ± 7.66	37.76
	10	–	–	76.14 ± 0.11	**76.80 ± 0.27**	69.25 ± 0.10	70.69 ± 2.39	21.43
		5	–	75.42 ± 0.16	**76.73 ± 0.06**	71.42 ± 2.53	69.21 ± 0.25	21.43
		0.01	–	**74.04 ± 0.15**	65.90 ± 0.14	53.69 ± 1.37	63.57 ± 3.91	21.43
		–	3	76.83 ± 0.14	**77.56 ± 0.03**	58.93 ± 4.80	61.16 ± 5.11	21.43
		–	9	**76.35 ± 0.04**	76.06 ± 0.11	49.22 ± 4.22	50.12 ± 0.02	21.43
		–	20	70.86 ± 2.13	**73.82 ± 0.07**	44.90 ± 0.20	42.70 ± 0.04	21.43

Table 8: Comparison of regularized algorithms α-LSE, PM + SM, and IPS-KL accuracy for EMNIST and FMNIST with different qualities of logging policy $\tau \in \{1, 2, 5, 10, 20\}$ and clean/ noisy propensity scores with $b \in \{5, 0.01\}$ and imbalance scenario ($\nu \in \{3, 9, 20\}$). The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	ν	α-LSE	PM + SM	IPS-KL	Logging Policy
EMNIST	1	–	–	**91.72 ± 0.03**	89.38 ± 0.02	90.42 ± 0.11	88.08
		5	–	**91.31 ± 0.01**	88.83 ± 0.11	90.78 ± 0.08	88.08
		0.01	–	91.39 ± 0.01	74.64 ± 3.67	**91.65 ± 0.01**	88.08
		–	3	**91.20 ± 0.03**	89.20 ± 0.02	91.07 ± 0.06	88.08
		–	9	**91.80 ± 0.02**	88.93 ± 0.04	90.44 ± 0.06	88.08
		–	20	**91.87 ± 0.01**	88.61 ± 0.05	87.49 ± 0.11	88.08
	2	–	–	90.76 ± 0.03	89.16 ± 0.05	**90.90 ± 0.04**	74.33
		5	–	**91.43 ± 0.04**	88.89 ± 0.07	90.94 ± 0.10	74.33
		0.01	–	**91.13 ± 0.02**	66.75 ± 3.20	90.80 ± 0.01	74.33
	5	–	–	**91.47 ± 0.02**	89.51 ± 0.04	90.04 ± 0.06	39.16
		5	–	**91.13 ± 0.11**	90.25 ± 0.06	90.51 ± 0.13	39.16
		0.01	–	**91.05 ± 0.02**	83.40 ± 0.06	89.08 ± 0.04	39.16
	10	–	–	**91.02 ± 0.03**	89.77 ± 0.01	89.90 ± 0.11	79.43
		5	–	**90.20 ± 0.09**	89.23 ± 0.03	89.91 ± 0.07	79.43
		0.01	–	**89.68 ± 0.07**	75.38 ± 0.42	86.62 ± 0.13	79.43
		–	3	89.66 ± 0.04	**89.98 ± 0.07**	84.84 ± 0.10	79.43
		–	9	**89.15 ± 0.03**	88.73 ± 0.04	85.00 ± 0.04	79.43
		–	20	**89.12 ± 0.08**	88.18 ± 0.04	80.74 ± 0.06	79.43
FMNIST	1	–	–	81.05 ± 0.01	78.69 ± 0.11	**81.31 ± 0.13**	78.38
		5	–	**82.05 ± 0.04**	49.64 ± 3.16	80.49 ± 0.05	78.38
		0.01	–	81.89 ± 0.05	10.00 ± 0.01	**82.26 ± 0.02**	78.38
		–	3	**81.74 ± 0.13**	78.79 ± 0.05	79.97 ± 0.09	78.38
		–	9	**81.99 ± 0.04**	78.91 ± 0.04	79.91 ± 0.18	78.38
		–	20	**82.16 ± 0.02**	77.46 ± 0.06	79.62 ± 0.16	78.38
	2	–	–	**80.85 ± 0.04**	78.61 ± 0.09	80.17 ± 0.11	66.94
		5	–	**81.79 ± 0.02**	75.23 ± 4.10	80.37 ± 0.04	66.94
		0.01	–	79.85 ± 0.07	59.13 ± 4.16	**81.54 ± 0.01**	66.94
	5	–	–	**81.56 ± 0.03**	78.57 ± 0.14	80.41 ± 0.16	37.76
		5	–	**82.12 ± 0.10**	79.98 ± 0.06	81.02 ± 0.07	37.76
		0.01	–	**82.14 ± 0.05**	33.42 ± 3.84	78.12 ± 2.54	37.76
	10	–	–	**81.63 ± 0.06**	79.73 ± 0.17	79.81 ± 0.10	21.43
		5	–	**81.64 ± 0.10**	80.04 ± 0.12	80.10 ± 0.07	21.43
		0.01	–	**81.02 ± 0.04**	73.10 ± 1.94	77.92 ± 0.08	21.43
		–	3	**80.93 ± 0.01**	80.18 ± 0.07	78.66 ± 0.16	21.43
		–	9	**80.77 ± 0.03**	78.34 ± 0.12	77.43 ± 0.06	21.43
		–	20	80.14 ± 0.05	**81.18 ± 0.08**	76.12 ± 0.02	21.43

Table 9: Comparison of different base algorithms LSE, PM, ES, and IX accuracy for CIFAR-10, and Letter with different qualities of logging policy $\tau \in \{1, 2, 5, 10, 20\}$ and clean/ noisy propensity scores with $b \in \{5, 0.01\}$ and imbalance scenarios ($\nu \in \{3, 9, 20\}$). The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	ν	**LSE**	**PM**	**ES**	**IX**	**Logging Policy**
CIFAR-10	1	–	–	65.02 ± 0.03	**65.13 ± 0.10**	64.03 ± 1.38	64.58 ± 0.26	64.42
		5	–	65.28 ± 0.12	**65.39 ± 0.08**	64.61 ± 0.15	64.61 ± 0.21	64.42
		0.01	–	59.11 ± 0.07	59.79 ± 0.11	**63.24 ± 0.11**	57.11 ± 2.09	64.42
		–	3	63.74 ± 0.07	**64.56 ± 0.08**	44.54 ± 0.01	44.47 ± 0.11	64.42
		–	9	64.14 ± 0.05	**64.62 ± 0.07**	49.56 ± 0.05	49.28 ± 0.02	64.42
		–	20	64.28 ± 0.20	**64.71 ± 0.12**	46.15 ± 0.04	46.15 ± 0.01	64.42
	2	–	–	65.42 ± 0.14	**65.60 ± 0.16**	64.63 ± 0.13	64.31 ± 0.25	59.95
		5	–	**65.44 ± 0.06**	65.26 ± 0.02	64.87 ± 0.02	63.93 ± 0.21	59.95
		0.01	–	58.96 ± 0.18	59.30 ± 0.04	**60.05 ± 0.02**	58.63 ± 0.08	59.95
	5	–	–	**64.99 ± 0.08**	64.93 ± 0.10	64.04 ± 0.43	63.32 ± 0.09	42.61
		5	–	**65.22 ± 0.08**	65.19 ± 0.34	63.73 ± 0.33	63.45 ± 0.05	42.61
		0.01	–	57.29 ± 0.32	57.72 ± 0.13	**59.20 ± 0.26**	56.18 ± 0.19	42.61
	10	–	–	**64.18 ± 0.05**	64.14 ± 0.08	63.04 ± 0.08	61.54 ± 1.14	27.12
		5	–	63.98 ± 0.24	**63.98 ± 0.13**	61.07 ± 1.42	62.85 ± 0.15	27.12
		0.01	–	53.83 ± 0.13	54.67 ± 0.04	**58.11 ± 0.04**	53.31 ± 0.14	27.12
		–	3	61.97 ± 0.13	**62.12 ± 0.05**	43.37 ± 0.02	43.43 ± 0.04	27.12
		–	9	**62.12 ± 0.18**	61.83 ± 0.06	40.12 ± 1.67	44.24 ± 3.79	27.12
		–	20	62.75 ± 0.06	**62.76 ± 0.12**	45.39 ± 0.06	45.44 ± 0.15	27.12
Letter	1	–	–	42.67 ± 0.01	**43.65 ± 0.03**	29.97 ± 1.56	34.21 ± 4.53	41.46
		5	–	32.85 ± 0.21	**32.98 ± 0.04**	28.46 ± 3.02	26.23 ± 5.68	41.46
		0.01	–	31.88 ± 0.04	**32.08 ± 0.02**	28.83 ± 6.98	33.41 ± 0.51	41.96
		–	3	42.49 ± 0.04	**43.18 ± 0.22**	10.59 ± 2.56	17.37 ± 2.95	41.46
		–	9	42.99 ± 0.05	**43.81 ± 0.01**	23.11 ± 1.51	25.64 ± 4.31	41.46
		–	20	43.11 ± 0.07	**44.33 ± 0.01**	34.31 ± 2.53	29.20 ± 1.47	41.46
	2	–	–	42.29 ± 0.08	**43.53 ± 0.01**	31.22 ± 4.60	29.06 ± 2.91	34.64
		5	–	33.11 ± 0.02	33.22 ± 0.02	33.55 ± 0.41	**33.63 ± 0.52**	34.64
		0.01	–	**32.36 ± 0.02**	31.60 ± 3.55	26.76 ± 5.42	18.73 ± 6.19	34.64
	5	–	–	43.64 ± 0.01	**46.60 ± 0.04**	30.24 ± 0.08	31.09 ± 0.14	15.07
		5	–	**38.01 ± 0.84**	37.16 ± 0.02	32.81 ± 1.68	32.10 ± 3.22	15.07
		0.01	–	23.56 ± 0.02	**26.45 ± 0.26**	13.45 ± 5.64	12.03 ± 10.04	15.07
	10	–	–	49.22 ± 1.81	**57.98 ± 0.12**	31.63 ± 4.90	33.61 ± 2.83	7.91
		5	–	36.36 ± 0.20	**48.10 ± 0.10**	34.11 ± 0.23	29.46 ± 3.48	7.91
		0.01	–	17.55 ± 4.27	**23.55 ± 0.14**	20.28 ± 0.14	16.78 ± 4.56	7.91
		–	3	46.96 ± 0.01	**53.36 ± 0.01**	15.74 ± 1.26	21.59 ± 0.48	7.91
		–	9	47.05 ± 0.03	**57.22 ± 1.40**	27.83 ± 3.07	26.98 ± 1.98	7.91
		–	20	47.99 ± 0.03	**58.55 ± 0.01**	32.03 ± 1.67	31.95 ± 1.29	7.91

Table 10: Comparison of different regularized algorithms α-LSE, PM + SM, IPS-KL for CIFAR-10, and Letter with different qualities of logging policy $\tau \in \{1, 2, 5, 10, 20\}$ and clean/ noisy propensity scores with $b \in \{5, 0.01\}$ and imbalance scenarios ($\nu \in \{3, 9, 20\}$). The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	ν	**α-LSE**	**PM + SM**	**IPS-KL**	**Logging Policy**
CIFAR-10	1	–	–	**65.38 ± 0.05**	64.47 ± 0.17	60.00 ± 0.18	64.42
		5	–	**65.41 ± 0.05**	62.17 ± 4.31	64.70 ± 0.04	64.42
		0.01	–	**65.24 ± 0.01**	59.20 ± 0.18	64.43 ± 0.05	64.42
		–	3	**65.15 ± 0.20**	64.63 ± 0.14	58.10 ± 5.40	64.42
		–	9	**65.22 ± 0.09**	64.76 ± 0.09	58.73 ± 0.14	64.42
		–	20	**65.30 ± 0.01**	64.78 ± 0.16	60.89 ± 2.85	64.42
	2	–	–	**65.64 ± 0.10**	65.60 ± 0.16	64.25 ± 0.42	59.95
		5	–	**65.37 ± 0.03**	65.10 ± 0.24	64.20 ± 0.18	59.95
		0.01	–	**65.34 ± 0.04**	58.08 ± 1.51	64.20 ± 0.18	59.95
	5	–	–	64.89 ± 0.13	**65.23 ± 0.03**	63.57 ± 0.25	42.61
		5	–	**65.24 ± 0.10**	65.34 ± 0.07	63.23 ± 0.43	42.61
		0.01	–	**63.90 ± 0.13**	45.20 ± 6.86	61.16 ± 0.02	42.61
	10	–	–	64.28 ± 0.09	**65.21 ± 0.06**	63.37 ± 0.10	27.12
		5	–	63.77 ± 0.17	**64.01 ± 0.10**	61.90 ± 0.11	27.12
		0.01	–	**62.53 ± 0.03**	54.29 ± 0.12	55.69 ± 2.19	27.12
		–	3	62.33 ± 0.09	**63.49 ± 0.33**	57.04 ± 0.16	27.12
		–	9	62.46 ± 0.02	**62.59 ± 0.22**	56.79 ± 0.13	27.12
		–	20	62.95 ± 0.06	**63.98 ± 0.08**	58.50 ± 0.41	27.12
Letter	1	–	–	**44.15 ± 0.01**	33.41 ± 3.16	40.79 ± 0.56	41.46
		5	–	**33.40 ± 0.01**	11.04 ± 10.74	32.03 ± 0.14	41.46
		0.01	–	**33.45 ± 0.02**	32.00 ± 0.21	30.68 ± 2.81	41.96
		–	3	**43.81 ± 0.01**	29.11 ± 11.31	37.94 ± 0.76	41.46
		–	9	**44.16 ± 0.01**	44.04 ± 0.01	43.05 ± 0.05	41.46
		–	20	**44.60 ± 0.08**	42.66 ± 1.60	41.38 ± 0.18	41.46
	2	–	–	**43.85 ± 0.02**	43.53 ± 0.01	41.04 ± 0.22	34.64
		5	–	**38.51 ± 0.10**	32.25 ± 0.17	32.16 ± 1.07	34.64
		0.01	–	**33.36 ± 0.02**	13.56 ± 0.41	28.90 ± 3.28	34.64
	5	–	–	50.76 ± 0.06	43.79 ± 0.01	**57.60 ± 1.42**	15.07
		5	–	45.41 ± 0.29	37.23 ± 2.06	**60.05 ± 0.10**	15.07
		0.01	–	**33.63 ± 0.06**	7.58 ± 0.04	26.11 ± 5.79	15.07
	10	–	–	61.76 ± 0.07	57.18 ± 0.16	**67.28 ± 0.09**	7.91
		5	–	52.41 ± 0.67	32.53 ± 0.02	**60.83 ± 4.72**	7.91
		0.01	–	**38.15 ± 0.01**	7.61 ± 0.15	13.66 ± 0.37	7.91
		–	3	60.51 ± 0.06	53.62 ± 0.03	**61.85 ± 0.13**	7.91
		–	9	59.11 ± 0.01	53.77 ± 1.42	**67.63 ± 0.02**	7.91
		–	20	**62.85 ± 0.01**	57.58 ± 0.72	34.15 ± 2.75	7.91

Table 11: Comparison of different algorithms in terms of the number of best performance over all clean (normal), noisy and imbalance experiment setups.

Estimator	Clean (Normal)	Noisy	Imbalance	Total
α-LSE	10	25	17	52
IPS-KL	4	5	2	11
PM+SM	2	1	5	8
LSE	0	1	0	1

J Real-world dataset

We applied our method to the Open Bandit Dataset, a public real-world logged bandit dataset, which is provided by a Japanese e-commerce company, ZOZO, Inc([Saito et al., 2020a]). The logging policy, provided by the company, suggests items to users in their e-commerce platform and each user can click on each of the suggested items. If an item is clicked, a cost equal to -1 and otherwise a cost of 0 is assigned. Hence, the context is the user features and actions(arms) are the items and the cost (reward) for this pair of context and action is defined as stated. The propensity score is also reported in the dataset. In order to have logging policy with different performances, for $p = 0.4, 0.6$ we sample zero-cost samples such that the number of samples with the cost of -1 is 40% and 60% of the dataset, respectively. This way we would have LBF datasets of sizes $153,019$ and $102,013$ respectively, both with 61208 samples with -1 cost.

For the evaluation metric, as Open Bandit Dataset is a recommendation system dataset and multiple actions for each context can lead to 0 cost, accuracy is not a suitable evaluation metric. We use IPS [Swaminathan and Joachims, 2015b] as our main evaluation metric. We also report SNIPS, which is a biased, but low variance estimation of the expected cost over the distribution induced by learning policy. Hence it shows the average cost of each method. Lower values of IPS and SNIPS indicate better performance. The comparison of different methods is presented in Table 12.

Table 12: Comparison of different algorithms LSE, α-LSE, PM, PM+SM, ES, and IPS with different performances of logging policy. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	Logging Policy	Metric	α-LSE	LSE	PM+SM	PM	ES	IPS
Open Bandit Dataset	40%	IPS	$\mathbf{-0.99 \pm 0.01}$	-0.82 ± 0.03	-0.60 ± 0.34	-0.81 ± 0.02	-0.69 ± 0.01	-0.69 ± 0.01
		SNIPS	$\mathbf{-0.73 \pm 0.00}$	-0.68 ± 0.02	-0.65 ± 0.05	-0.67 ± 0.10	-0.65 ± 0.05	-0.65 ± 0.05
	60%	IPS	$\mathbf{-0.70 \pm 0.03}$	-0.55 ± 0.00	-0.41 ± 0.00	-0.52 ± 0.02	-0.51 ± 0.03	-0.36 ± 0.07
		SNIPS	$\mathbf{-0.52 \pm 0.01}$	-0.49 ± 0.00	-0.42 ± 0.00	-0.47 ± 0.01	-0.43 ± 0.03	-0.37 ± 0.06

As we observe, in both settings and according to both metrics α-LSE significantly performs better than other methods. Also, we see that the second best estimator is LSE estimator without regularization. Hence, not only LSE outperforms other methods, but adding α-Rényi regularization has a significant effect on its performance.

K Synthetic experiment

The synthetic experiments results for Bias, variance and MSE of LSE, ES, PM, and IPS are shown in Table 13. Here we also report the bias and variance of each estimator in each setting. As we can observe LSE effectively keeps the variance low without significant side-effects on bias, making it a viable choice with general unbounded cost functions. Note that even in the case when the learning policy is perfectly fitted to the logging policy (i.e. when $\mu = 2$), because of the unboundedness of the cost (reward) function, estimation of the average cost is not easy, leading to very high variances in other methods, while LSE can control the variance in this case. We also observe that even with significant data ($n = 10,000$), the unbiased IPS estimator still doesn't outperform LSE, and its performance decays as μ increases. This phenomenon also holds for other methods compared to LSE. In order to find the parameter for each method, we use grid search and find the parameter that archives the highest MSE.

L Limitation

In our theoretical results (Section 5), we assumed that the expected value and variance of the weighted cost function are bounded. Although these assumptions are weaker compared to previous assumptions in the literature, they cannot be applied to heavy-tailed costs where the variance is unbounded. As future work, we plan to provide new theoretical results which hold under heavy-tailed weighted cost functions assumption.

Table 13: Bias, variance and MSE of LSE, ES, PM, and IPS estimators. We run the experiment 1000 times and find the variance, bias, and MSE of the estimations.

μ_1	Metric	Estimator	$n = 10$	$n = 100$	$n = 1000$	$n = 10000$
1.0	Bias	IPS	0.031	-0.002	-0.000	0.001
		ES	0.220	0.202	0.203	0.204
		PM	0.063	0.029	0.038	0.038
		LSE	0.118	0.107	0.021	0.003
	Variance	IPS	0.300	0.059	0.007	0.001
		ES	0.101	0.015	0.002	0.000
		PM	0.187	0.051	0.005	0.000
		LSE	0.105	0.011	0.003	0.000
	MSE	IPS	0.301	0.059	0.007	$\mathbf{0.001}$
		ES	0.150	0.056	0.043	0.042
		PM	0.191	0.052	0.006	0.002
		LSE	$\mathbf{0.119}$	$\mathbf{0.023}$	$\mathbf{0.003}$	$\mathbf{0.001}$
1.5	Bias	IPS	0.095	-0.024	0.025	0.006
		ES	1.237	0.514	0.556	0.544
		PM	-0.050	-0.180	-0.126	-0.147
		LSE	0.988	0.321	0.117	0.100
	Variance	IPS	10.193	2.440	0.353	0.049
		ES	0.672	0.592	0.066	0.008
		PM	12.147	2.927	0.425	0.059
		LSE	0.540	0.306	0.098	0.010
	MSE	IPS	10.202	2.440	0.354	0.049
		ES	2.202	0.857	0.375	0.304
		PM	12.149	2.960	0.441	0.081
		LSE	$\mathbf{1.516}$	$\mathbf{0.408}$	$\mathbf{0.111}$	$\mathbf{0.020}$
2.0	Bias	IPS	1.624	0.645	-0.487	-0.037
		ES	4.835	4.748	2.132	2.280
		PM	1.624	0.645	-0.487	-0.037
		LSE	3.726	1.941	1.759	0.650
	Variance	IPS	169.123	95.072	121.897	5.123
		ES	10.885	1.647	12.036	0.595
		PM	169.123	95.072	121.897	5.123
		LSE	14.413	8.828	1.099	0.491
	MSE	IPS	171.761	95.488	122.134	5.124
		ES	34.261	24.194	16.584	5.794
		PM	171.761	95.488	122.134	5.124
		LSE	$\mathbf{28.298}$	$\mathbf{12.594}$	$\mathbf{4.192}$	$\mathbf{0.914}$
2.5	Bias	IPS	4.173	3.423	-0.842	-0.654
		ES	23.746	20.450	20.197	11.039
		PM	20.555	20.699	18.753	13.605
		LSE	22.093	16.231	9.842	9.934
	Variance	IPS	11540.778	6809.889	6644.712	2814.923
		ES	43.215	50.898	9.144	165.695
		PM	535.469	169.391	128.223	139.309
		LSE	37.125	35.887	36.814	3.100
	MSE	IPS	11558.191	6821.605	6645.421	2815.350
		ES	607.074	469.117	417.063	287.543
		PM	957.980	597.851	479.882	324.417
		LSE	$\mathbf{525.231}$	$\mathbf{299.346}$	$\mathbf{133.688}$	$\mathbf{101.787}$