LOG-SUM-EXPONENTIAL ESTIMATOR FOR OFF-POLICY EVALUATION AND LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-policy learning and evaluation scenarios leverage logged bandit feedback datasets, which contain context, action, propensity score, and feedback for each data point. These scenarios face significant challenges due to high variance and poor performance with low-quality propensity scores and heavy-tailed reward distributions. We address these issues by introducing a novel estimator based on the log-sum-exponential (LSE) operator, which outperforms traditional inverse propensity score estimators. our LSE estimator demonstrates variance reduction and robustness under heavy-tailed conditions. For off-policy evaluation, we derive upper bounds on the estimator's bias and variance. In the off-policy learning scenario, we establish bounds on the regret—the performance gap between our LSE estimator and the optimal policy—assuming bounded $(1+\epsilon)$-th moment of weighted reward. Notably, we achieve a convergence rate of $O(n^{-\epsilon/(1+\epsilon)})$, where n is the number of training samples for the regret bounds and $\epsilon \in [0,1]$. Theoretical analysis is complemented by comprehensive empirical evaluations in both off-policy learning and evaluation scenarios, confirming the practical advantages of our approach.

1 INTRODUCTION

Off-policy learning and evaluation from logged data are important problems in Reinforcement Learning (RL) theory and practice. The logged bandit feedback (LBF) 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 (reward). It is used in many real applications, e.g., recommendation systems (Aggarwal, 2016; Li et al., 2011), personalized medical treatments (Kosorok & Laber, 2019; Bertsimas et al., 2017), and personalized advertising campaigns (Tang et al., 2013; Bottou et al., 2013). The literature has considered this setting from two perspectives, off-policy evaluation (OPE) and off-policy learning (OPL). In *off-policy evaluation*, we utilize the LBF dataset from a logging (behavioural) policy and an estimator, e.g., Inverse Propensity Score (IPS), to evaluate (or estimate) the performance of a different target policy. In *off-policy learning* we leverage the estimator and LBF dataset to learn an improved policy with respect to logging policy.

In both scenarios, OPL and OPE, the IPS estimator is proposed (Thomas et al., 2015; Swaminathan & Joachims, 2015a). However, this estimator suffers from significant variance in many cases (Rosenbaum & Rubin, 1983). To address this, some improved importance sampling estimators have been proposed, such as the IPS estimator with the truncated ratio of policy and logging policy (Ionides, 2008b), IPS estimator with truncated propensity score (Strehl et al., 2010), self-normalizing estimator (Swaminathan & Joachims, 2015b), exponential smoothing (ES) estimator (Aouali et al., 2023), implicit exploration (IX) estimator (Gabbianelli et al., 2023) 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: estimated propensity scores and heavy-tailed behaviour of weighted reward due to noise or outliers. Previous works such as Swaminathan & Joachims (2015a), Metelli et al. (2021), and Aouali et al. (2023) have made assumptions when dealing with LBF datasets. Specifically, these works assume that rewards are not subject to perturbation (noise) and that true propensity scores are available. However, these assumptions may not hold in real-world scenarios.

Noisy or heavy-tailed reward: three primary sources of noise in reward of LBF datasets can be identified as (Wang et al., 2020): (1) *inherent noise*, arising from physical conditions during feedback collection; (2) *application noise*, stemming from uncertainty in human feedback; and (3) *adversarial noise*, resulting from adversarial perturbations in the feedback process. Furthermore, In addition to noisy (perturbed) reward, a heavy-tailed reward can be observed in many real life applications, e.g., financial markets (Cont & Bouchaud, 2000) and web advertising (Park et al., 2013), the rewards do not behave bounded and follows heavy-tailed distributions.[1]

Noisy (estimated) propensity scores: The access to the exact values of the propensity scores may not be possible, for example, when human agents annotate the LBF dataset. In this situation, one may settle for training a model to estimate the propensity scores. Then, the propensity score stored in the LBF dataset can be considered a noisy version of the true propensity score.

Therefore, there is a need for an estimator that can effectively manage the heavy-tailed condition and noisy rewards or propensity scores in the LBF dataset.

1.1 Contributions

In this work, we propose a novel estimator for off-policy learning and evaluation from the LBF dataset that outperforms existing estimators when dealing with estimated propensity scores and heavy-tailed or noisy weighted rewards. The contribution of our work is three-fold.

First, we propose a novel non-linear estimator based on the Log-Sum-Exponential (LSE) operator which can be applied to both OPE and OPL scenarios. This LSE estimator effectively reduces variance and is applicable to, noisy propensity scores, heavy-tailed reward and noisy reward scenarios.

Second, we provide comprehensive theoretical guarantees for the LSE estimator's performance in OPE and OPL setup. In particular, we first provide bounds on the regret, i.e. the difference between the LSE estimator performance and the true average reward, under mild assumptions. In particular, our theoretical results hold under the heavy-tailed assumption on weighted reward. Furthermore, we studied the convergence rate of regret under heavy-tailed assumption which also holds for unbounded reward. Then, we studied bias and variance analysis for the LSE estimator and the robustness of the LSE estimator under noisy and heavy-tailed reward scenarios.

Finally, we conducted a set of experiments on different datasets to show the performance of the LSE in scenarios with true, estimated propensity scores and noisy reward in comparison with other estimators. We observed an improvement in the performance of learning policy using LSE in comparison with other state-of-the-art algorithms under different scenarios.

1.2 Preliminaries

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}. The cardinality of set $\mathcal{Z}$ is denoted by $|\mathcal{Z}|$. 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., $\log(x) := \log_e(x)$. For two probability measures P and Q defined on the space $\mathcal{Z}$, The *total variation distance* between two densities P and Q, is defined as $\mathbb{TV}(P, Q) := \int_{\mathcal{X}} |P - Q|(\mathrm{d}z)$.

2 Log-Sum-Exponential Estimator

Main Idea: Inspired by the log-sum-exponential operator with applications in multinomial linear regression, naive Bayes classifiers and tilted empirical risk(Calafiore et al., 2019; Murphy, 2012;

[1] A heavy-tailed random variable has a tail distribution heavier than the exponential distribution. For some heavy-tailed random variables, the variance is not well defined.

Williams & Barber, 1998; Li et al., 2023), we define the LSE estimator with parameter $\lambda < 0$,

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

where $\mathbf{Z} = \{z_i\}_{i=1}^n$ are samples from the positive random variable Z. The key property of the LSE operator is its robustness to noisy samples in a limited number of data samples. Here a noisy sample, by intuition, is a point with abnormally large positive z_i. Such points vanish in the exponential sum as $\lim_{z_i \to +\infty} e^{\lambda z_i} = 0$ for $\lambda < 0$. Therefore the LSE operator ignores terms with large values for negative λ. The robustness of LSE has also been explored in the context of supervised learning by Li et al. (2023) from practical perspective. Furthermore, in Appendix (App) C, we discuss the connection between the LSE and entropy regularization.

Motivating example: We provide a toy example to investigate the behaviour of LSE as a general estimator and its difference from the Monte-Carlo estimator (a.k.a. simple average) for *mean estimation*. Suppose that Z is distributed as a Pareto distribution[2] with scale x_m and shape ζ. Let $\zeta = 1.5$ and $x_m = \frac{1}{3}$, then we have $\mathbb{E}[Z] = \frac{\zeta x_m}{\zeta - 1} = 1$. The objective is to estimate $\mathbb{E}[Z]$ with n independent samples drawn from the Pareto distribution. We set $n \in \{10, 50, 100, 1000, 10000\}$ and compute the Monte-Carlo (a.k.a. simple average) and LSE estimation of the expectation of Z. Table 1 shows that LSE (with $\lambda = -0.1$) effectively keeps the variance and MSE, (Bishop & Nasrabadi, 2006), low without significant side-effects on bias. We also observe that the LSE estimator works well under heavy tail distributions.

Table 1: Bias, variance, and MSE of LSE (with $\lambda = -0.1$) and Monte-Carlo estimators. We run the experiment 10000 times and report the variance, bias, and MSE of the estimations.

	Estimator	$n = 10$	$n = 50$	$n = 100$	$n = 1000$	$n = 10000$
Bias	Monte-Carlo	0.0154	0.0155	0.0083	0.0061	0.0044
	LSE	0.1576	0.1606	0.1616	0.1624	0.1629
Variance	Monte-Carlo	1.5406	1.5289	1.3229	1.0203	0.8384
	LSE	0.1038	0.0616	0.0443	0.0335	0.0268
MSE	Monte-Carlo	1.5409	1.5292	1.3229	1.0203	0.8384
	LSE	0.1287	0.0874	0.0704	0.0598	0.0534

3 Related Works

We categorize the estimators based on their approach to reward estimation. Estimators that incorporate reward estimation techniques are classified as model-based estimators. In contrast, those that work without reward estimation are termed model-free estimators. Below, we review model-based estimators, and model-free estimators. Furthermore, we study the estimators which are designed for unbounded reward (heavy-tailed) scenarios in general RL scenarios.

Model-free Estimators: In model-free estimators, e.g., IPS estimators, we have many challenges, including, high variance and heavy-tailed scenarios. Recently, many model-free estimators have been proposed for high variance problems in model-free estimators (Strehl et al., 2010; Ionides, 2008b; Swaminathan & Joachims, 2015b; Aouali et al., 2023; Metelli et al., 2021; Neu, 2015; Aouali et al., 2023; Metelli et al., 2021; Sakhi et al., 2024). However, under heavy-tailed or unbounded reward scenario, the performance of these estimators degrade. In this work, our proposed LSE estimator demonstrates robust performance even under heavy-tailed assumptions, backed by theoretical guarantees.

Model-based Estimators: The direct method for off-policy learning from the LBF datasets is based on the estimation of the reward function, followed by the application of a supervised learning algorithm to the problem. However, this approach does not generalize well, as shown by Beygelzimer & Langford (2009). A different approach where the direct method and the IPS estimator are combined, i.e., doubly-robust, is introduced by Dudík et al. (2014). A different approach based on policy optimization and boosted base learner is proposed to improve the performance in direct

[2]For $Z \sim \mathrm{Pareto}(x_m, \zeta)$ as a heavy-tailed distribution, we have $f_Z(z) = \frac{\zeta x_m^\zeta}{z^{\zeta+1}}$

methods (London et al., 2023). Our approach differs from this area, as we do not estimate the reward function in the LSE estimator. A combination of the LSE estimator with direct method is presented in App. G.3 . Furthermore, the optimistic shrinkage (Su et al., 2020) and Dr-Switch (Wang et al., 2017) as other model-based estimators. In this work, we focus on model-free approach.

Unbounded Reward: Unbounded rewards (or returns) have been observed in various domains, including finance (Lu & Rong, 2018) and robotics (Bohez et al., 2019). In the context of multi-arm bandit problems, unbounded rewards can emerge as a result of adversarial attacks on reward distributions (Guan et al., 2020). Within the broader field of Reinforcement Learning (RL), researchers have investigated poisoning attacks on rewards and the manipulation of observed rewards (Rakhsha et al., 2020; 2021; Rangi et al., 2022). These studies highlight the importance of considering unbounded reward scenarios in RL and bandits algorithms. In particular, in our work, we focus on off-policy learning and evaluation under heavy-tailed (unbounded reward) assumption, employing a bounded $(1+\epsilon)$-th moment of weighted-reward assumption for $\epsilon \in [0,1]$.

4 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 $\pi_\theta \in \Pi_\theta$, where Π_Θ is defined as the set of all policies (policy set) which are parameterized by $\theta \in \Theta$, where Θ is the set of parameters, e.g., the parameters of a neural network. Furthermore, the $\pi_\theta(a|x)$ is defined as the conditional probability of of choosing an action given context x under the policy π_θ.[3]

A reward function [4] $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$, which is unknown, defines the expected reward (feedback) of each observed pair of context and action. In particular, $r(x,a) = \mathbb{E}_{P_{R|X=x,A=a}}[R]$ where $R \in \mathbb{R}^+$ is random reward and $P_{R|X=x,A=a}$ is the conditional distribution of reward R given the pair of context and action, (x,a). Note that, in the LBF setting, we only observe the reward (feedback) for the chosen action a in a given context x, under the known logging policy $\pi_0(a|x)$. We have access to the LBF dataset $S = (x_i, a_i, p_i, r_i)_{i=1}^n$ with n i.i.d. data points where each 'data point' (x_i, a_i, p_i, r_i) contains the context x_i which is sampled from unknown distribution P_X, the action a_i which is sampled from the known logging policy $\pi_0(\cdot|x_i)$, the propensity score $p_i \triangleq \pi_0(a_i|x_i)$, and the observed feedback (reward) r_i as a sample from distribution $P_{R|X=x_i,A=a_i}$ under logging policy $\pi_0(a_i|x_i)$.

We define the expected reward of a learning policy, $\pi_\theta \in \Pi_\theta$, which is called the *value function* evaluated at the learning policy, as

$$V(\pi_\theta) = \mathbb{E}_{P_X}[\mathbb{E}_{\pi_\theta(A|X)}[r(A,X)|X]] = \mathbb{E}_{P_X}[\mathbb{E}_{\pi_\theta(A|X)}[\mathbb{E}_{P_{R|X,A}}[R]]]. \tag{2}$$

We denote the importance weighted reward as $w_\theta(A,X)R$, where $w_\theta(A,X)$ is the weight,

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

As discussed by Swaminathan & Joachims (2015b), the IPS estimator is applied over the LBF dataset S (Rosenbaum & Rubin, 1983) to get an unbiased estimator of the value function by considering the weighted reward as,

$$\widehat{V}(\pi_\theta, S) = \frac{1}{n}\sum_{i=1}^{n} r_i w_\theta(a_i, x_i), \tag{3}$$

where $w_\theta(a_i,x_i) = \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}$.

The IPS estimator as an unbiased estimator has 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). Otherwise, it suffers from a large variance.

[3] In more details, consider an action space $\mathcal{A}$ with a σ-algebra and a σ-finite measure μ. For any policy π and context x, let $\pi(.|x)$ be a probability measure on $\mathcal{A}$ that is absolutely continuous with respect to μ, with density $\pi(.|x) = \frac{\mathrm{d}\pi_c(a|x)}{\mathrm{d}\mu}$ where $\pi_c(a|x)$ is absolute continuous with respect to μ.

[4] The reward can be viewed as the opposite (negative) of the cost. Consequently, a low cost (equivalent to maximum reward) signifies user (context) satisfaction with the given action, and conversely. For the cost function, we have $c(x,a) = -r(x,a)$ as discussed in (Swaminathan & Joachims, 2015a).

LSE in OPE and OPL scenarios: The LSE estimator is defined as

$$\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) := \mathrm{LSE}_\lambda(S) = \frac{1}{\lambda}\log\Big(\frac{1}{n}\sum_{i=1}^{n} e^{\lambda r_i w_\theta(a_i,x_i)}\Big), \tag{4}$$

where $\lambda < 0$ is a tunable parameter which helps us to recover the IPS estimator for $\lambda \to 0$. Furthermore, the LSE estimator is an increasing function with respect to λ.

OPE scenario: One of the evaluation metrics for an estimator in OPE scenarios is the mean squared error (MSE) which is decomposed into squared bias and the variance of the estimator. In particular, for the LSE estimator, we consider the following MSE decomposition in terms of bias and variance,

$$\begin{aligned}\mathrm{MSE}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)) &= \mathbb{B}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta))^2 + \mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)),\\ \mathbb{B}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)) &= \mathbb{E}[w_\theta(A,X)R] - \mathbb{E}[\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)],\\ \mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)) &= \mathbb{E}[(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) - \mathbb{E}[\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)])^2],\end{aligned} \tag{5}$$

where $\mathbb{B}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta))$ and $\mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta))$ are bias and variance of the LSE estimator, respectively.

OPL scenario: Our objective in OPL scenario is to find an optimal $\pi_{\theta^\star}$, one which maximize $V(\pi_\theta)$, i.e.,

$$\pi_{\theta^\star} = \operatorname*{arg\,max}_{\pi_\theta\in\Pi_\Theta} V(\pi_\theta). \tag{6}$$

We define the *generalization error* (or concentration), as the difference between the value function and the LSE estimator for a given learning policy $\pi_\theta \in \Pi_\theta$, i.e.,

$$\mathrm{gen}_\lambda(\pi_\theta) := V(\pi_\theta) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta). \tag{7}$$

For the OPL scenario, we also define $\pi_{\widehat{\theta}}$ as the maximizer of the LSE estimator for a given dataset S,

$$\pi_{\widehat{\theta}}(S) = \operatorname*{arg\,max}_{\pi_\theta\in\Pi_\Theta} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta). \tag{8}$$

Finally, we define *regret*, as the difference between the value function evaluated at π_{θ^*} and $\pi_{\widehat{\theta}}$,

$$\mathfrak{R}_\lambda(\pi_{\widehat{\theta}},S) := V(\pi_{\theta^*}) - V(\pi_{\widehat{\theta}}(S)). \tag{9}$$

More discussion regarding the LSE properties is provided in App. C.

5 Theoretical Foundations of the LSE estimator

In this section, we study the regret, bias-variance and robustness of the LSE estimator. We compare our LSE estimator with other model-free estimators in Table 2. All the proof details are deferred to App.D.

Non-linearity of LSE: The LSE estimator is a non-linear model-free estimator with respect to the weighted reward or reward, which is different from linear model-free estimators. In particular, most of estimators can be represented as the weighted average of reward (feedback),

$$\widehat{V}(\pi_\theta,S) = \frac{1}{n}\sum_{i=1}^{n} r_i g\big(w_\theta(a_i,x_i)\big), \tag{10}$$

where $g:\mathbb{R}\to\mathbb{R}$ is a transformation of $w_\theta(a_i,x_i)$ and is defined for each model-free estimator. For example, we have $g(y)=y$ in the IPS estimator, $g(y)=\min(y,M)$ in the truncated IPS estimator (Ionides, 2008b), $g(y)=((1-\widehat{\lambda})y^s+\widehat{\lambda})^{1/s}$ in the PM estimator (Metelli et al., 2021), $g(y)=y^\beta$ for $\beta\in(0,1)$ in the ES estimator (Aouali et al., 2023) and $g(y)=\frac{\tau y}{y^2+\tau}$ in the optimistic shrinkage (OS) (Su et al., 2020). For the IX-estimator with parameter η (Gabbianelli et al., 2023), we have $g(y)=\frac{y}{1+\eta/\pi_0}$. Furthermore, recently a logarithmic smoothing (LS) estimator and the linear version of LS (LS-LIN) are proposed by Sakhi et al. (2024). However, the LSE estimator is a non-linear function with respect to weighted reward or reward. Therefore, the previous techniques for regret and bias-variance analysis under linear estimators are not applicable.

Table 2: Comparison of estimators. We consider the bounded reward function, i.e., $R_{\max} := \sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} r(a,x)$ for all estimators except LSE. $\mathbb{B}^{\text{SN}}$ and $\mathbb{V}^{\text{SN}}$ are the Bias and the Efron-Stein estimate of the variance of self-normalized IPS. For the ES-estimator, we have $T^{ES} = \mathbb{B}^{ES} + (1/n)\big(D_{\text{KL}}(\pi_\theta\|\pi_0) + \log(4/\delta)\big)$. where $D_{\text{KL}}(\pi_\theta\|\pi_0) = \int_{\mathcal{A}} \pi_\theta(a|x)\log(\pi_\theta(a|x)/\pi_0(a|x))\mathrm{d}a$. We also define power divergence as $P_\alpha(\pi_\theta\|\pi_0) := \int_{\mathcal{A}} \pi_\theta(a|x)^\alpha \pi_0(a|x)^{(1-\alpha)}\mathrm{d}a$ is the power divergence with order α. For the IX-estimator, $C_\eta(\pi)$ is the smoothed policy coverage ratio. We compare the convergence rate of the generalization error for estimators. B and C are constants. For LS estimator, $\mathcal{S}_{\tilde{\lambda}}(\pi_\theta)$ is the discrepancy between π and π_0.

Estimator	Generalization Error (Concentration)	Convergence Rate	Heavy-tailed	Regret Bound	Noisy Reward	Differentiability	Subgaussian Like Tail
IPS	$R^2_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)}{\delta n}}$	$O(n^{-1/2})$	×	✓	×	✓	×
SN-IPS (Swaminathan & Joachims, 2015b)	$R_{\max}(B^{\text{SN}} + \sqrt{V^{\text{ES}}\log\frac{1}{\delta}})$	-	×	×	×	✓	×
IPS-TR $(M > 0)$ (Ionides, 2008a)	$R_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)\log\frac{1}{\delta}}{n}}$	$O(n^{-1/2})$	×	✓	×	×	✓
IX $(\eta > 0)$ (Gabbianelli et al., 2023)	$R_{\max}(2\eta C_\eta(\pi_\theta) + \frac{\log(2/\delta)}{\eta n})$	$O(n^{-1/2})$	×	✓	×	✓	✓
PM $(\lambda \in [0,1])$ (Metelli et al., 2021)	$R_{\max}\sqrt{\frac{P_2(\pi_\theta\|\pi_0)\log\frac{1}{\delta}}{n}}$	$O(n^{-1/2})$	×	×	×	✓	✓
ES $(\alpha \in [0,1])$ (Aouali et al., 2023)	$R_{\max}\sqrt{\frac{D_{\text{KL}}(\pi_\theta\|\pi_0)+\log(4\sqrt{n}/\delta)}{n}} + T^{ES}$	$O((\log(n)/n)^{1/2})$	×	✓	×	✓	×
OS $(\tau > 0)$ (Su et al., 2020)	$\max_{\beta\in\{2,3\}} \sqrt[\beta]{\frac{P_\beta(\pi_\theta\|\pi_0)\big(\log\frac{1}{\delta}\big)^{\beta-1}}{n^{\beta-1}}}$	$O\big(n^{(1-\beta)/\beta}\big)$	×	×	×	✓	×
LS $(\tilde{\lambda} \geq 0)$ (Sakhi et al., 2024)	$\tilde{\lambda}\mathcal{S}_{\tilde{\lambda}}(\pi_\theta) + \frac{\log(2/\delta)}{\tilde{\lambda} n}$	$O\big(n^{-1/2}\big)$	×	✓	×	✓	✓
LSE $(0 > \lambda > -\infty$ and $\epsilon \in [0,1])$ **(ours)**	$C\Big(\frac{2\log(2\lvert\Pi_\theta\rvert/\delta)}{n}\Big)^{\epsilon/(1+\epsilon)}$	$O(n^{-\epsilon/(1+\epsilon)})$	✓	✓	✓	✓	✓

Theoretical comparison with other estimators: The comparison of our LSE estimator with other estimators, including, IPS, self-normalized IPS (Swaminathan & Joachims, 2015b), truncated IPS with weight truncation parameter M, ES-estimator with parameter α (Aouali et al., 2023), IX-estimator with parameter η, PM-estimator with parameter λ (Metelli et al., 2021), OS-estimator with parameter τ (Su et al., 2020) and LS-estimator with parameter $\tilde{\lambda}$ (Sakhi et al., 2024) is provided in Table 2.
Note that the truncated IPS (IPS-TR) (Ionides, 2008a) employs truncation, resulting in a non-differentiable estimator. This non-differentiability complicates the optimization phase, often necessitating additional care and sometimes leading to computationally intensive discretizations (Papini et al., 2019). Furthermore, tuning the threshold M in IPS-TR is sensitive and can result in matching of the learning policy and logging policy in OPL scenario (Aouali et al., 2023).
In the following sections, we provide more details regarding heavy-tail assumption and theoretical results for the LSE estimator.

5.1 Heavy-tail Assumption

In this section, the following heavy-tail assumption is made in our theoretical results.

Assumption 1 (Heavy-tail weighted reward)**.** The reward distribution $P_{R|X,A}$ and $P_X \otimes \pi_0(A|X)$ are such that for all learning policy $\pi_\theta(A|X) \in \Pi_\theta$ and some $\epsilon \in [0,1]$, the $(1+\epsilon)$-th moment of the weighted reward is bounded,

$$\mathbb{E}_{P_X\otimes\pi_0(A|X)\otimes P_{R|X,A}}\Big[\big(w_\theta(A,X)R\big)^{1+\epsilon}\Big] \leq \nu. \tag{11}$$

We make a few remarks. First, in comparison with the bounded reward function assumption in literature, (Metelli et al., 2021; Aouali et al., 2023), in Assumption 1, the reward function can be unbounded. Moreover, our assumptions are weaker with respect to the uniform overlap assumption [5]. In heavy-tailed bandit learning (Bubeck et al., 2013; Shao et al., 2018; Lu et al., 2019), a similar assumption to Assumption 1 on $(1+\epsilon)$-th moment for some $\epsilon \in [0,1]$ of reward is assumed. In contrast, in Assumption 1, we consider the weighted reward. Note that, under uniform coverage (overlap) assumption, Assumption 1 can be interpreted as a heavy-tailed assumption on reward. Furthermore under a bounded reward, Assumption 1 would be equivalent with the heavy-tailed assumption on the $(1+\epsilon)$-th moment of weight function, $w_\theta(a,x)$. More detailed theoretical comparison is provided in App. D.1.

5.2 Regret Bounds

In this section, we provide an upper bound on the regret of the LSE estimator as discussed in the OPL scenario.

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

We will use the following novel and helpful lemma to prove some results in this section.

Lemma 5.1. *Consider the random variable $Z > 0$. For $\epsilon \in [0,1]$, the following upper bound holds on the variance of $e^{\lambda Z}$ for $\lambda < 0$,*

$$\mathbb{V}\left(e^{\lambda Z}\right) \le |\lambda|^{1+\epsilon}\mathbb{E}[Z^{1+\epsilon}]. \tag{12}$$

In the following Theorem, we provide an upper bound on the regret of learning policy under the LSE estimator.

Theorem 5.2 (Regret bounds)**.** *Given Assumption 1 and assuming $n \ge \frac{(2|\lambda|^{1+\epsilon}\nu + \frac{4}{3}\gamma)\log\frac{|\Pi_\theta|}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$, with probability at least $1-\delta$, then there exists $\gamma \in (0,1)$ such that the following upper bound holds on the regret of the LSE estimator,*

$$0 \le \mathfrak{R}_\lambda(\pi_{\widehat{\theta}}, S) \le \frac{|\lambda|^\epsilon}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\nu^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}},$$

where $\pi_{\widehat{\theta}}$ is defined in equation 8.

Sketch of Proof. First, using Bernstein's Inequality, Boucheron et al., 2013 and Lemma 5.1, we provide lower and upper bounds on generalization error for a fixed learning policy π_θ. Then, we consider the following decomposition of regret,

$$V(\pi_{\theta^*}) - V(\pi_{\widehat{\theta}}) = \underbrace{\mathrm{gen}_\lambda(\pi_{\theta^*})}_{I_1} + \underbrace{\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\theta^*}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}})}_{I_2}\underbrace{-\mathrm{gen}_\lambda(\pi_{\widehat{\theta}})}_{I_3}. \tag{13}$$

Note that, the term I_2 is negative. We can provide upper bounds for terms I_1 and I_3 using derived upper and lower bounds on generalization error (Theorem D.1 and Theorem D.2 in App. D.2), respectively. To obtain the final result, we apply the union bound. □

As the regret bound in Theorem 5.2 depends on λ, we need to select an appropriate λ in order to study the convergence rate of regret bound with respect to n.

Proposition 5.3 (Convergence rate)**.** *Given Assumption 1, for any $0 < \gamma < 1$, assuming $n \ge \frac{(2\nu + \frac{4}{3}\gamma)\log\frac{|\Pi_\theta|}{\delta}}{\gamma^2\exp(2\nu^{1/(1+\epsilon)})}$ and setting $\lambda = -n^{-\zeta}$ for $\zeta \in \mathbb{R}^+$, then the overall convergence rate of the regret upper bound is $\max(O(n^{-1+\zeta}), O(n^{-\epsilon\zeta}), O(n^{(-\zeta\epsilon-1)/2}))$ for a finite policy set.*

Remark 5.4. Using Proposition 5.3, the regret upper bound has the convergence rate of $O(n^{-\epsilon/(1+\epsilon)})$ for $\zeta = \frac{1}{1+\epsilon}$. Note that, if Assumption 1 holds for $\epsilon = 1$, then we have the convergence rate of $O(n^{-1/2})$.

Our theoretical results on regret can be applied to unbounded weighted reward under Assumption 1, compared to other estimators where the bounded reward or weighted reward is needed. Furthermore, the dependency of regret or generalization bound on δ can be polynomial $O((\frac{1}{\delta})^\alpha)$ for $\alpha > 0$, sub-exponential $O\left(\frac{\log(1/\delta)}{n}\right)$ or subgaussian $O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$. We also study achieving subgaussian concentration for LSE estimator in App. D.6.

Finite policy set: The theorems in this section assumed that the policy set, Π_θ, is finite; this is for example the case in off-policy learning problems with a finite number of policies. If this assumption is violated, we can apply the growth function technique which is bounded by VC-dimension (Vapnik, 2013) or Natarajan dimension (Holden & Niranjan, 1995) as discussed in (Jin et al., 2021). Furthermore, we can apply PAC-Bayesian analysis (Gabbianelli et al., 2023) for the LSE estimator. More discussion regarding the PAC-Bayesian approach is provided in App. D.5.

5.3 Bias and Variance

In this section, we provide an analysis of bias and variance for the LSE estimator.

Proposition 5.5 (Bias bound)**.** *Given Assumption 1, the following lower and upper bounds hold on the bias of the LSE estimator with $\lambda < 0$,*

$$\frac{(n-1)}{2n|\lambda|}\mathbb{V}(e^{\lambda w_\theta(A,X)R}) \le \mathbb{B}(\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \le \frac{1}{1+\epsilon}|\lambda|^\epsilon\nu + \frac{1}{2n\lambda}\mathbb{V}(e^{\lambda w_\theta(A,X)R}). \tag{14}$$

Remark 5.6 (Asymptotically Unbiased). By selecting λ as a function of n, which tends to zero as $n \to \infty$, e.g. $\lambda(n) = -n^{-\varsigma}$ for some $\varsigma > 0$, the bounds in Proposition 5.5 becomes asymptotically zero. The overall convergence rate for upper bound is $O(n^{-\epsilon/(1+\epsilon)})$ by choosing $\varsigma = \frac{1}{1+\epsilon}$. For example, if Assumption 1 holds for $\epsilon = 1$, then by choosing $\varsigma = 1/2$, we have the convergence rate of $O(n^{-1/2})$ for the bias of the LSE estimator. Consequently, the LSE estimator is asymptotically unbiased.

For the variance of the LSE estimator, we provide the following upper bound.

Proposition 5.7 (Variance Bound)**.** *Assume that* $\mathbb{E}[(w_\theta(A,X)R)^2] \le \nu_2$ [6] *holds. Then the variance of the LSE estimator with* $\lambda < 0$*, satisfies,*

$$\mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)) \le \frac{1}{n}\mathbb{V}(w_\theta(A,X)R) \le \frac{1}{n}\nu_2. \tag{15}$$

Remark 5.8 (Variance Reduction). We can observe that the variance of the LSE is less than the variance of IPS estimator for all $\lambda < 0$.

Combining the results in Proposition 5.5 and Proposition 5.7, we can derive an upper bound on MSE of the LSE estimator using equation 5. An upper bound on the moment of the LSE estimator is provided in App. D.3. The bias and variance comparison of different estimators is provided in App. D.1.1.

5.4 Robustness of the LSE Estimator: Noisy Reward

In this section section, we study the robustness of the LSE estimator under noisy reward. We also investigate the performance of the LSE estimator under noisy (estimated) propensity scores in the App. E.

Suppose that due to an outlier or noise in receiving the feedback (reward), the underlying distribution of the reward given a pair of actions and contexts, $P_{R|X,A}$ is shifted via the distribution of noise or outlier, denoted as $\widetilde{P}_{R|X,A}$. We model the distributional shift of reward via distribution $\widetilde{P}_{R|X,A}$ due to inspiration by the notion of influence function (Marceau & Rioux, 2001; Christmann & Steinwart, 2004). Furthermore, we define the noisy reward LBF dataset as $\widetilde{S}$ with n data samples. For our result in this section, the following assumption is made.

Assumption 2 (Heavy-tailed Weighted Noisy Reward)**.** The noisy reward distribution $\widetilde{P}_{R|X,A}$ and $P_X \otimes \pi_0(A|X)$ are such that for all learning policy $\pi_\theta(A|X) \in \Pi_\theta$ and some $\epsilon \in [0,1]$, the $(1+\epsilon)$-th moment of the weighted reward is bounded,

$$\mathbb{E}_{P_X\otimes\pi_0(A|X)\otimes\widetilde{P}_{R|X,A}}\left[\left(w_\theta(A,X)R\right)^{1+\epsilon}\right] \le \widetilde{\nu}. \tag{16}$$

Under noisy reward LBF dataset, we derive the following learning policy,

$$\pi_{\widehat{\theta}}(\widetilde{S}) = \arg\max_{\pi_\theta \Pi_\theta} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\pi_\theta, \widetilde{S})\,. \tag{17}$$

In the following theorem, we provide an upper bound on the regret of $\pi_{\widehat{\theta}}(\widetilde{S})$ as the learning policy under noisy reward LBF dataset.

Theorem 5.9. *Given Assumption 1, Assumption 2 and assuming* $n \ge \frac{(2|\lambda|^{1+\epsilon}\nu + \frac{4}{3}\gamma)\log\frac{|\Pi_\theta|}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$*, with probability at least* $1-\delta$*, then there exists* $\gamma \in (0,1)$ *such that the following upper bound holds on the regret of the LSE estimator under noisy reward logged data,*

$$\begin{aligned}0 \le \mathfrak{R}_\lambda(\pi_{\widehat{\theta}}(\widetilde{S}), \widetilde{S}) \le \frac{|\lambda|^\epsilon}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\widetilde{\nu}\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\widetilde{\nu}^{1/(1+\epsilon)})}} \\ + \mathbb{TV}(P_{R|X,A}, \widetilde{P}_{R|X,A})\Big(\frac{1}{|\lambda|\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})} + \frac{1}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})}\Big),\end{aligned} \tag{18}$$

where $\pi_{\widehat{\theta}}(\widetilde{S})$ *is defined in equation 17.*

[6] Assumption 1 for $\epsilon = 1$.

Table 3: Bias, variance, and MSE of LSE, ES, PM, IX, and IPS-TR estimators. The experiment is run 10000 times with 1000 samples. The variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

	$\alpha = 1.1$			$\alpha = 1.4$		
Estimator	Bias	Variance	MSE	Bias	Variance	MSE
PM	0.004	0.063	0.063	−0.301	164.951	165.041
ES	−0.001	0.054	0.054	1.959	0.396	4.232
LSE	0.052	0.006	**0.009**	0.615	0.292	**0.670**
IPS-TR	0.020	0.052	0.052	0.053	133.688	133.691
IX	0.237	0.002	0.058	1.340	0.048	1.842
SNIPS	−0.005	0.059	0.059	−0.029	133.520	133.521
LS-LIN	0.284	0.001	0.082	2.164	0.005	4.687
LS	0.082	0.007	0.013	0.564	0.458	0.776
OS	0.521	0.020	0.292	0.623	23.589	23.977

Discussion: This term in equation 18, $\mathbb{TV}(P_{R|X,A}, \widetilde{P}_{R|X,A})\left(\frac{1}{|\lambda|\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})} + \frac{1}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})}\right)$, can be interpreted as the cost of noise associated with noisy reward. This cost can be reduced by increasing $|\lambda|$. However, increasing $|\lambda|$ also amplifies the term $\frac{|\lambda|^\epsilon}{1+\epsilon}\nu$ in the upper bound on regret. Therefore, there is a trade-off between robustness and regret, particularly for $\lambda < 0$ in the LSE estimator.

6 Experiments

We present our experiments for OPE and OPL. Our aim is to demonstrate that our proposed estimators not only possess desirable theoretical properties but also compete with baseline estimators in practical scenarios. More details can be found in App.F. Furthermore, an experiment on a real-world dataset, KUAIREC (Gao et al., 2022), is provided in App. G.4.

6.1 Off-policy Evaluation

We conduct synthetic experiments to evaluate our proposed LSE estimator performance in OPE setting. For this purpose, we consider an LBF dataset which has only a single context (state), denoted as x_0. We consider the learning and logging policies as Gaussian distributions, $\pi_\theta(\cdot|x_0) \sim \mathcal{N}(\mu_1, \sigma^2)$ and $\pi_0(\cdot|x_0) \sim \mathcal{N}(\mu_2, \sigma^2)$. The reward function is a positive exponential function $e^{\alpha x^2}$ which is unbounded. We also set our parameters to observe different tail distributions. We fix $\mu_1 = 0.5, \mu_2 = 1, \sigma^2 = 0.25$ and change the value of α which controls the tail of the weighted reward variable, $\alpha \in \{1.4, 1.6\}$. We also examine different values of α and the effect of number of samples for a fixed α in App. G.1. Moreover, we conduct a similar experiment when logging and learning policies are Lomax distributions[7] in App. G.1.

Baselines: For our experiments in OPE setting, we consider truncated IPS estimator (Swaminathan & Joachims, 2015a), PM estimator (Metelli et al., 2021), ES estimator (Aouali et al., 2023), IX estimator (Gabbianelli et al., 2023), SNIPS (Swaminathan & Joachims, 2015b), LS-LIN and LS estimators (Sakhi et al., 2024), and OS (shrinkage) (Su et al., 2020) estimator as baselines.

Metrics: We calculate the Bias, Variance, and MSE of estimators by running the experiments for $10K$ times each one over 1000 samples.

Discussion: The results presented in Table 3 demonstrate that the LSE estimator has better performance in terms of both MSE and variance when compared to other baselines. There is a close performance comparison between LSE and LS. More experiments are provided in App. G.10.

6.2 Off-policy Learning

In off-policy learning scenario, we apply the standard supervised to bandit transformation (Beygelzimer & Langford, 2009) on a classification dataset: Extended-MNIST (EMNIST) (Xiao et al., 2017) to generate the LBF dataset. We also run on FMNIST in App.G.2 . 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

[7]The Lomax distribution is a Pareto Type II distribution which is a heavy-tailed distribution.

Table 4: Comparison of different estimators LSE, PM, ES, IX, BanditNet, LS-LIN and OS accuracy for EMNIST with different qualities of logging policy ($\tau \in \{1, 10\}$) and true / noisy (estimated) propensity scores with $b \in \{5, 0.01\}$ and noisy reward with $P_f \in \{0.1, 0.5\}$. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	P_f	**LSE**	**PM**	**ES**	**IX**	**BanditNet**	**LS-LIN**	**OS**	**Logging Policy**
EMNIST	1	–	–	88.49 ± 0.04	$\mathbf{89.19 \pm 0.03}$	88.61 ± 0.06	88.33 ± 0.13	66.58 ± 6.39	88.70 ± 0.02	88.71 ± 0.26	88.08
		5	–	$\mathbf{89.16 \pm 0.03}$	88.94 ± 0.05	88.48 ± 0.03	88.51 ± 0.23	65.10 ± 0.69	88.38 ± 0.18	88.70 ± 0.15	88.08
		0.01	–	$\mathbf{86.07 \pm 0.01}$	85.62 ± 0.10	85.71 ± 0.04	81.39 ± 4.02	66.55 ± 3.11	84.64 ± 0.17	84.59 ± 0.09	88.08
		–	0.1	$\mathbf{89.29 \pm 0.04}$	89.08 ± 0.05	88.45 ± 0.09	88.14 ± 0.14	59.90 ± 3.78	88.30 ± 0.12	88.74 ± 0.09	88.08
		–	0.5	88.72 ± 0.08	$\mathbf{88.78 \pm 0.03}$	87.27 ± 0.10	87.08 ± 0.14	56.95 ± 3.06	87.20 ± 0.32	88.06 ± 0.09	88.08
	10	–	–	88.59 ± 0.03	$\mathbf{88.61 \pm 0.04}$	88.38 ± 0.08	87.43 ± 0.19	85.48 ± 3.13	88.58 ± 0.08	86.88 ± 0.34	79.43
		5	–	88.42 ± 0.07	$\mathbf{88.43 \pm 0.07}$	88.39 ± 0.10	88.39 ± 0.06	84.90 ± 3.10	88.23 ± 0.27	86.00 ± 0.37	79.43
		0.01	–	$\mathbf{82.15 \pm 0.21}$	80.85 ± 0.29	81.07 ± 0.07	77.49 ± 2.77	27.02 ± 1.92	78.43 ± 3.13	21.70 ± 4.11	79.43
		–	0.1	$\mathbf{88.29 \pm 0.06}$	88.22 ± 0.02	88.19 ± 0.08	87.93 ± 0.35	84.89 ± 3.21	87.50 ± 0.17	87.68 ± 0.16	79.43
		–	0.5	88.71 ± 0.16	88.52 ± 0.07	84.42 ± 0.34	83.25 ± 3.45	63.35 ± 13.39	85.75 ± 0.04	$\mathbf{89.09 \pm 0.05}$	79.43

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 $r = 1$, and $r = 0$ otherwise. So, the 4-tuple (x, a, p, r) makes up the LBF dataset.

Baselines: For all of our experiments in OPL, we compare our LSE estimator against several non-regularized baseline estimators, including, truncated IPS (Swaminathan & Joachims, 2015a), PM (Metelli et al., 2021), ES (Aouali et al., 2023), IX (Gabbianelli et al., 2023), BanditNet (Joachims et al., 2018), LS-LIN (Sakhi et al., 2024) and OS estimator (Su et al., 2020).

Noisy (Estimated) propensity score: For noisy propensity score, motivated by Halliwell (2018) and the discussion in App.E.1, we assume a multiplicative inverse Gamma noise on π_0 for $b \in \mathbb{R}^+$, $\widehat{\pi}_0 = \frac{1}{U}\pi_0$, where $\widehat{\pi}(a|x)$ is the estimated propensity scores and $U \sim \mathrm{Gamma}(b, b)$. [8].

Noisy reward: Inspired by Metelli et al. (2021), we also consider noise in reward samples. In particular, we model noisy reward by a reward-switching probability $P_f \in [0, 1]$ to simulate noise in the reward samples. For example, a reward sample of $r = 1$ may switch to $r = 0$ with probability P_f.

Logging policy: To have logging policies with different performances, given inverse temperature[9] $\tau \in \{1, 10\}$, 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 the performance of the different estimators based on the accuracy of the trained model. Inspired by London & Sandler (2019), we calculate the accuracy for a deterministic policy where the accuracy of the model based on the argmax 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. The deterministic accuracies of LSE, PM, ES, IX, BanditNet, OS and LS-LIN for $\tau \in \{1, 10\}$ are presented in Table 4.

Discussion: The results presented in Table 4 demonstrate that the LSE estimator achieves maximum accuracy (with less variance) in most scenarios compared to all baselines. More discussion and experiments are provided in App. G.

7 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 regret under heavy-tailed assumption. Furthermore, we explore the performance of our estimator in scenarios involving estimated propensity scores or heavy-tailed weighted reward. Results from our experimental evaluation demonstrate that our estimator, guided by our theoretical framework, performs competitively compared to most of baseline estimators in off-policy learning and evaluation. In future work, we plan to study the effect of regularization on the LSE estimator from both theoretical and practical perspectives. Moreover, we envision extending the application of our estimator to more challenging RL setups, (Chen & Jiang, 2022; Zanette et al., 2021; Xie et al., 2019a). Inspired by the application of LSE operator in supervised learning for positive tilt (Li et al., 2023), we can explore the performance of the LSE estimator for positive λ as future work.

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

[9]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 $a_i \in \mathcal{A}$.

REFERENCES

Pranav Agarwal, Pierre de Beaucorps, and Raoul de Charette. Goal-constrained sparse reinforcement learning for end-to-end driving. *arXiv preprint arXiv:2103.09189*, 2021.

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

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Online bandit learning with offline preference data. *arXiv preprint arXiv:2406.09574*, 2024.

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.

Pierre Alquier and Benjamin Guedj. Simpler pac-bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Alexander Amini, Igor Gilitschenski, Jacob Phillips, Julia Moseyko, Rohan Banerjee, Sertac Karaman, and Daniela Rus. Learning robust control policies for end-to-end autonomous driving from data-driven simulation. *IEEE Robotics and Automation Letters*, 5(2):1143–1150, 2020. doi: 10.1109/LRA.2020.2966414.

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.

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

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*, pp. 129–138, 2009.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Steven Bohez, Abbas Abdolmaleki, Michael Neunert, Jonas Buchli, Nicolas Heess, and Raia Hadsell. Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*, 2019.

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.

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

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

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

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.

Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. *The master equation and the convergence problem in mean field games*. Princeton University Press, 2019.

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

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.

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.

Rama Cont and Jean-Philipe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2):170–196, 2000.

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.

Murad Dawood, Nils Dengler, Jorge de Heuvel, and Maren Bennewitz. Handling sparse rewards in reinforcement learning using model predictive control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 879–885. IEEE, 2023.

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*, pp. 248–255. Ieee, 2009.

Abhimanyu Dubey et al. Cooperative multi-agent bandits with heavy tails. In *International conference on machine learning*, pp. 2730–2739. PMLR, 2020.

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

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

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

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 540–550, 2022. doi: 10.1145/3511808.3557220. URL `https://doi.org/10.1145/3511808.3557220`.

Panayiotis G. Georgiou, Panagiotis Tsakalides, and Chris Kyriakakis. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE transactions on multimedia*, 1(3):291–301, 1999.

Ziwei Guan, Kaiyi Ji, Donald J Bucci Jr, Timothy Y Hu, Joseph Palombo, Michael Liston, and Yingbin Liang. Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pp. 4036–4043, 2020.

Maxime Haddouche and Benjamin Guedj. Pac-bayes generalisation bounds for heavy-tailed losses through supermartingales. *arXiv preprint arXiv:2210.00928*, 2022.

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

A Ben Hamza and Hamid Krim. Image denoising: A nonlinear robust statistical approach. *IEEE transactions on signal processing*, 49(12):3045–3054, 2001.

Sean B Holden and Mahesan Niranjan. On the practical applicability of vc dimension bounds. *Neural Computation*, 7(6):1265–1288, 1995.

Matthew Holland. Pac-bayes under potentially heavy tails. *Advances in Neural Information Processing Systems*, 32, 2019.

Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 120, 2018.

Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. *Advances in Neural Information Processing Systems*, 36, 2024.

Yulong Huang and Yonggang Zhang. A new process uncertainty robust student'st based kalman filter for sins/gps integration. *IEEE Access*, 5:14391–14404, 2017.

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

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

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

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.

Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

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

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

Yue Kang, Cho-Jui Hsieh, and Thomas Lee. Low-rank matrix bandits with heavy-tailed rewards. *arXiv preprint arXiv:2404.17709*, 2024.

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

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

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

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities. arXiv:1909.01931, 2019.

Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. *arXiv preprint arXiv:1903.09132*, 2019.

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

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.

Kyungjae Lee and Sungbin Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5280–5294, 2022.

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*, pp. 297–306, 2011.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

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

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

Hengheng Lu and Kang Rong. Unbounded returns and the possibility of credit rationing: A note on the stiglitz–weiss and arnold–riley models. *Journal of Mathematical Economics*, 75:67–70, 2018.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *international conference on machine learning*, pp. 4154–4163. PMLR, 2019.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. 2021.

Gábor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pp. 3524–3546. PMLR, 2022.

Étienne Marceau and Jacques Rioux. On robustness in risk theory. *Insurance: Mathematics and Economics*, 29(2):167–185, 2001.

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.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling, 2018.

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.

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

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

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

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

Jaimie Yejean Park, Kyoung-Won Lee, Sang Yeon Kim, and Chin-Wan Chung. Ads by whom? ads about what? exploring user influence and contents in social advertising. In *Proceedings of the first ACM conference on Online social networks*, pp. 155–164, 2013.

Minjae Park, Seok Young Lee, Jin Seok Hong, and Nam Kyu Kwon. Deep deterministic policy gradient-based autonomous driving for mobile robots in sparse reward environments. *Sensors*, 22 (24):9574, 2022.

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*, pp. 1220–1230, 2023.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, pp. 7974–7984. PMLR, 2020.

Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments. *arXiv preprint arXiv:2102.08492*, 2021.

Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the limits of poisoning attacks in episodic reinforcement learning. *arXiv preprint arXiv:2208.13663*, 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.

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.

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

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.

Laura Ruotsalainen, Martti Kirkko-Jaakkola, Jesperi Rantanen, and Maija Mäkelä. Error modelling for multi-sensor measurements in infrastructure-free indoor navigation. *Sensors*, 18(2):590, 2018.

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.

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*, pp. 29734–29759. PMLR, 2023.

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. Pac-bayesian offline contextual bandits with guarantees. In *International Conference on Machine Learning*, pp. 29777–29799. PMLR, 2023.

Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. *arXiv preprint arXiv:2405.14335*, 2024.

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.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

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

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

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

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020.

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.

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.

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

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*, pp. 1587–1594, 2013.

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

Ilya O Tolstikhin and Yevgeny Seldin. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Sattar Vakili, Keqin Liu, and Qing Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5): 759–767, 2013.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6202–6209, 2020.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

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.

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

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

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, 2019a.

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, 2019b.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

Xiaotian Yu, Han Shao, Michael R Lyu, and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *UAI*, pp. 937–946, 2018.

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.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Xiaoying Zhang, Junpu Chen, Hongning Wang, Hong Xie, and Hang Li. Uncertainty-aware off-policy learning. *arXiv preprint arXiv:2303.06389*, 2023a.

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*, 2023b. URL `https://openreview.net/forum?id=1pWNhmbllE`.

Han Zhong, Jiayi Huang, Lin Yang, and Liwei Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 34:15710–15720, 2021.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline policy evaluation and optimization with heavy-tailed rewards. *arXiv preprint arXiv:2310.18715*, 2023.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 541–549. PMLR, 2024.

Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 3385–3393. PMLR, 2021.

Appendix

Table of Contents

A Other Related Works

In this section, we provide other related works.

Other methods: A balance-based weighting approach, which outperforms traditional estimators, was proposed by Kallus (2018). Other extensions of batch learning as a scenario for off-policy learning 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 value function based on the entropy of the target policy.

Pessimism Method and Off-policy RL: The pessimism concept originally, introduced in offline RL(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 & 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 non-linear estimator. Note that for our theoretical analysis, we consider heavy-tailed assumption for $(1+\epsilon)$-th moment for some $\epsilon \in [0,1]$. However, (Jin et al., 2022) also considers 3rd and 4th 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 & 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. We consider this combination as future work.

Individualized Treatment Effects: The individual treatment effect aims to estimate the expected values of the squared difference between outcomes (reward 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 & van der Schaar, 2017; Athey et al., 2019; Shi et al., 2019; Kennedy, 2020; Nie & Wager, 2021). Our work differs from this line of works by considering multiple action scenario and assuming the access to propensity scores in the LBF dataset.

Noisy/Corrupted Rewards: Agnihotri et al. (2024) utilized offline data with noisy preference feedback as a warm-up step for online bandit learning. In linear bandits, Kveton et al. (2019) estimated a set of pseudo-rewards for each perturbed reward in the history and used it for reward parameter estimation. Lee & Lim (2022) assumes a heavy-tailed noise variable on the observed rewards and proposes two exploration strategies that provide minimax regret guarantees for the multi-arm bandit problem under the heavy-tailed reward noise. In the linear bandits, Kang et al. (2024) Huang et al. (2024) tackles the issue of heavy-tailed noise on cost function by modifying the reward parameter estimation objective. The former one uses Huber loss for reward function parameter estimation and the latter one truncates the rewards. Zhong et al. (2021) and Xue et al. (2024) propose the median of means and truncation to handle the heavy-tailed noise in the observed rewards. In this work, we study the performance of our proposed estimator, the LSE estimator, under noisy and heavy-tailed reward.

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., 2019b) 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 true and estimated propensity scores, where the estimated propensity scores are modeled via Gamma noise. Our work differs from the line of works on the estimation methods of propensity scores.

Bandit and Reinforcement Learning under Heavy-tailed Distributions: Some works discussed the heavy-tailed reward in bandit learning (Medina & Yang, 2016; Bubeck et al., 2013; Shao et al., 2018; Lu et al., 2019; Zhong et al., 2021). Furthermore, some works also discussed the heavy-tailed rewards in RL (Zhuang & Sui, 2021; Zhu et al., 2024). However, off-policy learning with LBF dataset under a heavy-tailed distribution of weighted reward is overlooked.

Mean-estimation under Heavy-tailed Distributions: In (Lugosi & Mendelson, 2019; 2021; Hopkins, 2018), the performance of median-of-means and trimmed mean estimators have been studied and the sub-Gaussian behavior of these estimators are studied. However, median-of-means estimator presents practical challenges in implementation: it requires additional computational resources for data partitioning and mean calculations, while also introducing discontinuities that can prevent gradient-based optimization methods.

Generalization Error under Heavy-tailed Assumption: There are also some works studied the generalization error of supervised learning under unbounded loss functions, in particular, under heavy-tailed assumption via the PAC-Bayesian approach. Losses with heavier tails are studied by Alquier & Guedj (2018) where probability bounds (non-high probability) are developed. Using a different estimator than empirical risk, PAC-Bayes bounds for losses with bounded second and third moments are developed by Holland (2019). Notably, their bounds include a term that can increase with the number of samples n. Kuzborskij & Szepesvári (2019) and Haddouche & Guedj (2022) also provide bounds for losses with a bounded second moment. The bounds in (Haddouche & Guedj, 2022) rely on a parameter that must be selected before the training data is drawn. Information-theoretic bounds based on the second moment of loss function $\sup_{h\in\mathcal{H}} |\ell(h, Z) - \mathbb{E}[\ell(h, \tilde{Z})]|$ are also derived in (Lugosi & Neu, 2022). Furthermore, in (Lugosi & Mendelson, 2019, Section 4), the uniform bound via Rademacher complexity analysis over the L_2 bounded function space is studied for median-of-means estimator. In our work, we focus on generalization error and regret analysis of the LSE estimator as a non-linear estimator in OPL and OPE scenarios.

Heavy-tailed rewards in Bandits: Bandit learning with heavy-tailed reward distributions has been extensively studied. Bubeck et al. (2013) proposed Robust UCB, and Vakili et al. (2013) introduced DSEE as bandit algorithms with theoretical regret guarantees. Yu et al. (2018) proposed a bandit algorithm based on pure exploration with heavy-tailed reward distributions. Heavy-tailed reward distributions are also studied in the context of linear bandits (Shao et al., 2018; Medina & Yang, 2016). Dubey et al. (2020) proposed a decentralized algorithm for cooperative multi-agent bandits when the reward distribution is heavy-tailed. Our work differs from this line of works by considering heavy-tailed assumption on weighted reward.

Heavy-tailed rewards in RL: The challenge of heavy-tailed distributions in decision making has been studied for more than two decades (Georgiou et al., 1999; Hamza & Krim, 2001; Huang & Zhang, 2017; Ruotsalainen et al., 2018). There is a significant amount of study in RL dealing with heavy tailed reward distribution (Zhu et al., 2023; Zhuang & Sui, 2021; Huang et al., 2024). Moreover, big sparse rewards are a prominent issue in reinforcement learning (Park et al., 2022; Agarwal et al., 2021; Dawood et al., 2023). In such scenarios, there is a far-reaching goal, possibly accompanied by sparse failure states in which the agent attains big positive and negative rewards respectively. For example in safe autonomous driving, accidents are so costly and, hence are assigned large negative rewards. They are also delayed and sparse, which means that they are observed after many steps with a lot of exploration in the environment (Kiran et al., 2021; Amini et al., 2020). This hinders the training and leads to an infeasible slow learning curve. A common approach to tackle this issue is reward shaping which inserts new engineered reward functions alongside the agent's trajectory to provide guidelines for the agent (Kiran et al., 2021). This strategy may fail because it biases the model into the strategy hinted by the new rewards, which may not be the optimal solution for the original problem. Moreover, the method of reward shaping will not necessarily avoid the low-probability high-value rewards, because the imputed rewards are mostly small and high-value rewards still happen with low probability. Therefore, handling low-probability large reward is one of the challenges in this field, which can be modeled by heavy-tailed distributions as discussed with more details in App. G.12.

B PRELIMINARIES

B.1 NOTATIONS AND DIAGRAM

All notations are summarized in Table 5. An overview of our main theoretical results is provided in Fig. 1.

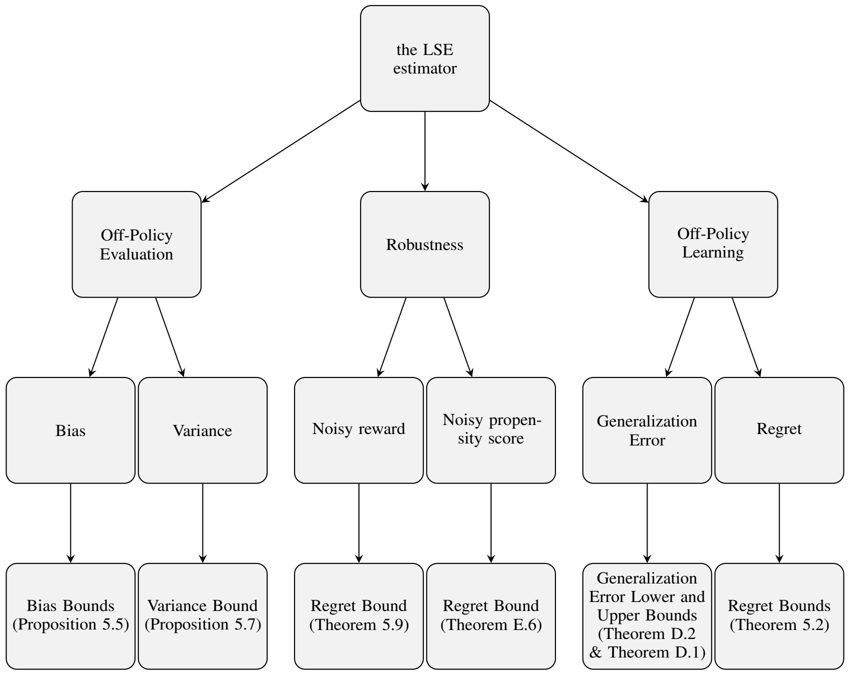


Figure 1: Overview of the main results

Table 5: Summary of notations in the paper

Notation	Definition	Notation	Definition
X	Context	A	Action
$r(X, A)$	Reward function	R	Reward
n	The number of logged data samples	P_X	Distribution over context set
S	LBF dataset	p_i	Propensity score $(\pi_0(a_i\|x_i))$
π_θ	Learning policy	$w_\theta(A\|X)$	weight $(\pi_\theta(A\|X)/\pi_0(A\|X))$
$\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)$	the LSE estimator	$V(\pi_\theta)$	Value of learning policy π_θ
ν	Upper bound on $(1+\epsilon)$-th moment of weighted reward (Assumption 1)	$\pi_0(a\|X)$	Logging policy
$\mathrm{gen}_\lambda(\pi_\theta)$	Generalization error of the LSE estimator	$\mathfrak{R}_\lambda(\pi_{\widehat{\theta}}, S)$	Regret of the LSE estimator
$\mathbb{B}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))$	Bias of the LSE estimator	$\mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta))$	Variance of the LSE estimator

B.2 Definitions

We define the softmax function

$$\mathrm{softmax}(x_1, x_2, \cdots, x_n) = (s_1, s_2, \cdots, s_n),$$
$$s_i = \frac{e^{x_i}}{\sum_{j=1}^{n} x^{x_j}}, \quad 1 \le i \le n.$$

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

Definition B.1. (Cardaliaguet et al., 2019) A functional $U : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ admits a *functional linear derivative* if there is a map $\frac{\delta U}{\delta m} : \mathcal{P}(\mathbb{R}^n) \times \mathbb{R}^n \to \mathbb{R}$ which is continuous on $\mathcal{P}(\mathbb{R}^n)$, such that for all $m, m' \in \mathcal{P}(\mathbb{R}^n)$, it holds that

$$U(m') - U(m) = \int_0^1 \int_{\mathbb{R}^n} \frac{\delta U}{\delta m}(m_\lambda, a)\,(m' - m)(da)\,\mathrm{d}\lambda,$$

where $m_\lambda = m + \lambda(m' - m)$.

B.3 Theoretical Tools

In this section, we provide the main lemmas which are used in our theoretical proofs.

Lemma B.2 (Kantorovich-Rubenstein duality of total variation distance, see (Polyanskiy & Wu, 2022))**.** *The Kantorovich-Rubenstein duality (variational representation) of the total variation distance is as follows:*

$$\mathbb{TV}(m_1, m_2) = \frac{1}{2L} \sup_{g \in \mathcal{G}_L} \left\{ \mathbb{E}_{Z\sim m_1}[g(Z)] - \mathbb{E}_{Z\sim m_2}[g(Z)] \right\}, \tag{19}$$

where $\mathcal{G}_L = \{g : \mathcal{Z} \to \mathbb{R}, ||g||_\infty \le L\}$.

Lemma B.3 (Hoeffding Inequality, Boucheron et al., 2013)**.** *Suppose that Z_i are sub-Gaussian independent random variables, with means μ_i and sub-Gaussian parameter σ_i^2, then we have:*

$$\mathbb{P}\left(\sum_{i=1}^{n} (Z_i - \mu_i) \ge t \right) \le \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} \sigma_i^2} \right) \tag{20}$$

Lemma B.4 (Bernstein's Inequality, Boucheron et al., 2013)**.** *Suppose that $S = \{Z_i\}_{i=1}^n$ are i.i.d. random variable such that $|Z_i - \mathbb{E}[Z]| \le R$ almost surely for all i, and $\mathbb{V}(Z) = \sigma^2$. Then the following inequality holds with probability at least $(1-\delta)$ under P_S,*

$$\left| \mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i \right| \le \sqrt{\frac{4\sigma^2 \log(2/\delta)}{n}} + \frac{4R\log(2/\delta)}{3n}. \tag{21}$$

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 some positive parameters $A, B, C, \epsilon \ge 0$ and $0 \le \delta \le 1$. Then, we have,

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

Also, for some $D > 0$, if $A \ge \frac{B\log\frac{1}{\epsilon} + 2DC\log\frac{1}{\epsilon}}{D^2}$, then we have $\delta \le D$.

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,

$$Ax + \sqrt{Bx} \leq C \tag{22}$$

Proof. Given $Ax \leq C$, equation equation 22 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 22,

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

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 equation 22 to hold. □

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

Proof. Taking the second derivative gives us the result, $\frac{d^2}{dy^2}\left(\log(y) + \beta y^2\right) = -\frac{1}{y^2} + 2\beta$.

□

Lemma B.8. *We have the following inequality for* $y < 0$ *and* $\epsilon \in [0, 1]$*,*

$$e^y \leq 1 + y + \frac{|y|^{1+\epsilon}}{1+\epsilon}. \tag{23}$$

Proof. For $y = 0$, equality holds. If suffices to prove that the derivative of LHS of equation 23 is more than the derivative of RHS $\forall y < 0$, i.e.,

$$e^y - 1 + |y|^\epsilon \geq 0.$$

Note that for $y \leq -1$, $|y|^\epsilon \geq 1$ and the inequality trivially holds. For $y > -1$, $|y|^\epsilon$ is minimized at $\epsilon = 1$, so it is sufficient to prove the inequality only for $\epsilon = 1$, which is,

$$e^y - 1 - y \geq 0 \leftrightarrow e^y \geq y + 1$$

and holds $\forall y \leq 0$. □

Lemma B.9. *For a positive random variable,* $Z > 0$*, suppose* $\mathbb{E}[Z^{1+\epsilon}] < \nu_z$ *for some* $\epsilon \in [0, 1]$*. Then, the following inequality holds,*

$$\mathbb{E}[Z] \leq \nu_z^{1/(1+\epsilon)}$$

Proof. Due to Jensen's inequality, we have,

$$\mathbb{E}[Z] = \mathbb{E}\big[(Z^{1+\epsilon})^{1/(1+\epsilon)}\big] \leq \mathbb{E}\big[Z^{1+\epsilon}\big]^{1/(1+\epsilon)} \leq \nu_z^{1/(1+\epsilon)}.$$

□

Lemma B.10. *For a positive random variable,* $Z > 0$*, suppose* $\mathbb{E}[Z^{1+\epsilon}] < \infty$ *for some* $\epsilon \in [0, 1]$*. Then, following inequality for* $\lambda < 0$ *holds,*

$$\mathbb{E}[Z] \geq \frac{1}{\lambda}\log \mathbb{E}[e^{\lambda Z}] \geq \mathbb{E}[Z] - \frac{1}{1+\epsilon}|\lambda|^\epsilon \mathbb{E}[Z^{1+\epsilon}].$$

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

$$1 + z \leq e^z \leq 1 + z + \frac{1}{1+\epsilon}|z|^{1+\epsilon}.$$

Therefore, we have,

$$\begin{aligned}
\frac{1}{\lambda}\log \mathbb{E}[e^{\lambda Z}] &\geq \frac{1}{\lambda}\log \mathbb{E}[1 + \lambda Z + \frac{1}{1+\epsilon}|\lambda|^{1+\epsilon} Z^{1+\epsilon}] \\
&= \frac{1}{\lambda}\log\left(1 + \lambda\mathbb{E}[Z] + \frac{1}{1+\epsilon}|\lambda|^{1+\epsilon}\mathbb{E}\big[Z^{1+\epsilon}\big]\right) \\
&\geq \frac{1}{\lambda}\left(\lambda\mathbb{E}[Z] + \frac{1}{1+\epsilon}|\lambda|^{1+\epsilon}\mathbb{E}[Z^{1+\epsilon}]\right) \\
&= \mathbb{E}[Z] - \frac{1}{1+\epsilon}|\lambda|^\epsilon \mathbb{E}[Z^{1+\epsilon}].
\end{aligned}$$

□

C Other Properties of the LSE estimator

Proposition C.1 (LSE Asymptotic Properties)**.** *The following asymptotic properties of LSE with respect to λ holds,*

$$\lim_{\lambda\to 0} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S) = \frac{1}{n}\left(\sum_{i=1}^{n} r_i w_\theta(a_i, x_i)\right),$$
$$\lim_{\lambda\to -\infty} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S) = \min_i r_i w_\theta(a_i, x_i),$$
$$\lim_{\lambda\to \infty} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S) = \max_i r_i w_\theta(a_i, x_i).$$

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

$$\begin{aligned}
\lim_{\lambda\to 0} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S) &= \lim_{\lambda\to 0} \frac{\log\left(\frac{\sum_{i=1}^{n} e^{\lambda r_i w_\theta(a_i,x_i)}}{n}\right)}{\lambda} \\
&= \lim_{\lambda\to 0} \frac{\left(\frac{\sum_{i=1}^{n} r_i w_\theta(a_i,x_i) e^{\lambda r_i w_\theta(a_i,x_i)}}{\sum_{i=1}^{n} e^{\lambda r_i w_\theta(a_i,x_i)}}\right)}{1} \\
&= \frac{\sum_{i=1}^{n} r_i w_\theta(a_i, x_i)}{n}.
\end{aligned}$$

For the second limit for $\lambda \to -\infty$ we have:

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

As both lower and upper tends to $\min_i r_i w_\theta(a_i, x_i)$ we conclude that:

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

A similar argument proves the third limit ($\lambda \to \infty$). □

Remark C.2. As shown in (Zhang, 2006, Proposition 1.1), the LSE function is an increasing function with respect to λ.

Derivative of the LSE estimator: The derivative of the LSE estimator can be represented as,

$$\nabla_\theta \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i e^{\lambda(r_i w_\theta(a_i,x_i) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta))} \nabla_\theta w_\theta(a_i, x_i). \tag{24}$$

Note that, in equation 24, we have a weighted average of the gradient of the weighted reward samples. In contrast to the linear estimators for which the gradient is a uniform mean of reward samples, in the LSE estimator, the gradient for large values of $r_i w_\theta(a_i, x_i)$, $\forall i \in [n]$ (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 $r_i w_\theta(a_i, x_i)$ (i.e. high $w_\theta(a, x)$), $\forall i \in [n]$.

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,*

$$\nabla \widehat{\mathrm{V}}^{\lambda}_{\text{LSE}}(S, \pi_\theta) = \operatorname{softmax}(\lambda r_1 w_\theta(a_1, x_1), \cdots, \lambda r_n w_\theta(a_n, x_n)), \tag{25}$$

$$\nabla^2 \widehat{\mathrm{V}}^{\lambda}_{\text{LSE}}(S) = \lambda \operatorname{diag}(S_n) - \lambda S_n S_n^T, \tag{26}$$

where $S_n = \operatorname{softmax}(\lambda r_1 w_\theta(a_1, x_1), \cdots, \lambda r_n w_\theta(a_n, x_n))$. *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 $\widehat{\mathrm{V}}^{\lambda}_{\text{LSE}}$, we prove that for $\lambda \geq 0$ the Hessian matrix is positive semi-definite. The proof for concavity for $\lambda < 0$ is similar.

$$\begin{aligned}\mathbf{z}^T \nabla^2 \widehat{\mathrm{V}}^{\lambda}_{\text{LSE}} \mathbf{z} &= \lambda \left(\mathbf{z}^T \operatorname{diag}(S_n)\mathbf{z} - \mathbf{z}^T S_n S_n^T \mathbf{z}\right) = \lambda \left(\sum_{i=1}^n S_n(i) z_i^2 - \left(\sum_{i=1}^n S_n(i) z_i \right)^2 \right) \\ &= \lambda \left(\left(\sum_{i=1}^n S_n(i) z_i^2 \right) \left(\sum_{i=1}^n S_n(i) \right) - \left(\sum_{i=1}^n S_n(i) z_i \right)^2 \right) \geq 0.\end{aligned}$$

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

Using Lemma C.3, we can show that $\widehat{\mathrm{V}}^{\lambda}_{\text{LSE}}$ is convex for $\lambda \geq 0$ and concave for $\lambda < 0$. Applying Lemma C.3, we can prove that the derivative of the LSE estimator is positive and less than one, i.e.,

$$0 \leq \nabla \widehat{\mathrm{V}}^{\lambda}_{\text{LSE}}(S, \pi_\theta) \leq 1. \tag{27}$$

Furthermore, we prove equation 24 by applying Lemma C.3.

C.1 LSE ESTIMATOR AND KL REGULARIZATION

In this section, we will discuss the connection between the LSE estimator,

$$\text{LSE}_\lambda(\mathbf{Z}) = \frac{1}{\lambda} \log \Big(\frac{1}{n} \sum_{i=1}^n e^{\lambda z_i} \Big), \tag{28}$$

and the KL regularization problem.

Consider the following KL-regularized expected minimization for $\lambda < 0$,

$$\min_{\mathbf{P} \in \Delta^{n-1}} \sum_{i=1}^n p_i z_i - \frac{1}{\lambda} D_{\text{KL}}(\mathbf{P} \| \text{Uni}(n)), \tag{29}$$

where Δ^{n-1} denotes the probability simplex and $\text{Uni}(n)$ in the discrete uniform distribution over n mass points. Note that $\lambda < 0$, and the KL divergence is strictly convex with respect to $\mathbf{P}$. Therefore, the objective function in equation 29 is convex. Then, the solution of regularized problem in equation 29, is the Gibbs distribution as follows,

$$p_i^\star = \frac{\exp(\lambda z_i)}{\sum_{i=1}^n \exp(\lambda z_i)}, \quad \forall i \in [n], \tag{30}$$

Using equation 30 in equation 29, we have,

$$\begin{aligned}&\sum_{i=1}^n \frac{\exp(\lambda z_i) z_i}{\sum_{j=1}^n \exp(\lambda z_j)} - \frac{1}{\lambda} \sum_{i=1}^n \frac{\exp(\lambda z_i)}{\sum_{j=1}^n \exp(\lambda z_j)} \Big(\lambda z_i - \log \Big(\frac{1}{n} \sum_{i=1}^n \exp(\lambda z_i) \Big) \Big) \\ &= \frac{1}{\lambda} \log \Big(\frac{1}{n} \sum_{i=1}^n \exp(\lambda z_i) \Big).\end{aligned} \tag{31}$$

Therefore, the final value of KL-regularized minimization problem is the LSE estimator with $\lambda < 0$. Therefore, the LSE estimator with negative parameter can be interpreted as KL-regularized expected minimization problem.

D PROOFS AND DETAILS OF SECTION 5

D.1 DETAILS OF THEORETICAL COMPARISON

In this section, we compare our estimator with PM, ES, IX, LS and OS from a theoretical perspective in more details.

D.1.1 BIAS AND VARIANCE COMPARISON

In this section we present the bias and variance comparison of different estimators in Table 6. We define power divergence as $P_\alpha(\pi_\theta \| \pi_0) := \int_a \pi_\theta(a|x)^\alpha \pi_0(a|x)^{(1-\alpha)} \mathrm{d}a$ is the power divergence with order α. For a fair comparison, we consider the bounded reward function, i.e., $R_{\max} := \sup_{(a,x)\in\mathcal{A}\times\mathcal{X}} r(a,x)$. Therefore, we have $\nu \le R_{\max}^{1+\epsilon} P_{1+\epsilon}(\pi_\theta\|\pi_0)$ and $\nu_2 \le R_{\max}^2 P_2(\pi_\theta\|\pi_0)$. We can observe that LSE has the same behavior in comparison with other estimators.

Table 6: Comparison of bias and variance of estimators. $\mathbb{B}^{\mathrm{SN}}$ 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 $T^{ES} = \mathbb{B}^{ES} + (1/n)\big(D_{\mathrm{KL}}(\pi_\theta\|\pi_0) + \log(4/\delta)\big)$. where $D_{\mathrm{KL}}(\pi_\theta\|\pi_0) = \int_a \pi_\theta(a|x)\log(\pi_\theta(a|x)/\pi_0(a|x))\mathrm{d}a$. For the IX-estimator, $C_\eta(\pi)$ is the smoothed policy coverage ratio. We compare the convergence rate of the generalization error for estimators. B and C are constants. For LS estimator, $\mathcal{S}_{\tilde{\lambda}}(\pi_\theta)$ is the discrepancy between π and π_0.

Estimator	Variance	Bias
IPS	$\frac{R_{\max}^2 P_2(\pi_\theta\|\pi_0)}{n}$	0
SN-IPS (Swaminathan & Joachims, 2015b)	$R_{\max}^2 V^{\mathrm{SN}}$	$R_{\max} B^{\mathrm{SN}}$
IPS-TR ($M > 0$) (Ionides, 2008a)	$R_{\max}^2 \frac{P_2(\pi_\theta\|\pi_0)}{n}$	$R_{\max} \frac{P_2(\pi_\theta\|\pi_0)}{M}$
IX ($\eta > 0$) (Gabbianelli et al., 2023)	$R_{\max} C_\eta(\pi_\theta)/n$	$R_{\max}\eta C_\eta(\pi_\theta)$
PM ($\lambda \in [0,1]$) (Metelli et al., 2021)	$\frac{R_{\max}^2 P_2(\pi_\theta\|\pi_0)}{n}$	$R_{\max}\lambda P_2(\pi_\theta\|\pi_0)$
ES ($\alpha \in [0,1]$) (Aouali et al., 2023)	$R_{\max}^2 \frac{\mathbb{E}_{\pi_\theta}[\pi_\theta \cdot \pi_0^{1-2\alpha}]}{n}$	$R_{\max}(1 - \mathbb{E}_{\pi_\theta}[\pi_0^{1-\alpha}])$
OS ($\tau > 0$) (Su et al., 2020)	$\frac{R_{\max}^2 P_2(\pi_\theta\|\pi_0)}{n}$	$R_{\max} \frac{P_3(\pi_\theta\|\pi_0)}{\tau}$
LS ($\tilde{\lambda} \ge 0$) (Sakhi et al., 2024)	$\frac{\mathcal{S}_{\tilde{\lambda}}(\pi_\theta)}{n}$	$\tilde{\lambda}\mathcal{S}_{\tilde{\lambda}}(\pi_\theta)$
LSE ($0 > \lambda > -\infty$ and $\epsilon \in [0,1]$) **(ours)**	$\frac{R_{\max}^2 P_2(\pi_\theta\|\pi_0)}{n}$	$\frac{1}{1+\epsilon}\lvert\lambda\rvert^\epsilon R_{\max}^{1+\epsilon} P_{1+\epsilon}(\pi_\theta\|\pi_0) - \frac{B}{2n\lvert\lambda\rvert}$

Note that in variance comparison between IPS and LSE, the LSE variance is less than IPS. However in Table 6, we use a looser upper bound to compare bounds in terms of the same parameter $R_{\max}$.

Bias and Variance Trade-off: Observe that for the bias and variance of the LSE estimator, there is a trade-off with respect to $\lambda < 0$. Specifically, reducing λ increases the bias of the LSE estimator,

$$\mathbb{B}(\widehat{\mathrm{V}}_{\mathrm{LSE}}^{\lambda}(S,\pi_\theta)) = \mathbb{E}[w_\theta(A,X)R] - \mathbb{E}[\widehat{\mathrm{V}}_{\mathrm{LSE}}^{\lambda}(S,\pi_\theta)]. \tag{32}$$

This is a consequence of the increasing property of the LSE with respect to λ (see Remark C.2).

Additionally, for the variance, we have the following bound,

$$\mathrm{Var}(\widehat{\mathrm{V}}_{\mathrm{LSE}}^{\lambda}(S,\pi_\theta)) \le \mathbb{E}[(\widehat{\mathrm{V}}_{\mathrm{LSE}}^{\lambda}(S,\pi_\theta))^2]. \tag{33}$$

It is important to note that decreasing λ reduces the upper bound on the variance of the LSE estimator.

Therefore, by decreasing $\lambda < 0$, the bias increases and the variance decreases.

D.1.2 Comparison with PM Estimator

In (Metelli et al., 2018), the authors proposed the following PM estimator for two hyper-parameter (λ_p, s),

$$\widehat{V}_{\mathrm{PM}}(S,\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n}((1-\lambda_p)w_\theta(a_i,x_i)^s + \lambda_p)^{\frac{1}{s}}\, r_i.$$

An upper bound on generalization error of PM estimator for $(\lambda_p, s=-1)$, is provided in (Metelli et al., 2018, Theorem 5.1),

$$\mathrm{gen}_{\mathrm{PM}}(S,\pi_\theta) \le \|R\|_\infty (2+\sqrt{3})\left(\frac{2P_\alpha(\pi_\theta\|\pi_0)^{\frac{1}{\alpha-1}}\log\frac{1}{\delta}}{3(\alpha-1)^2 n}\right)^{1-\frac{1}{\alpha}}, \tag{34}$$

where $\mathrm{gen}_{\mathrm{PM}}(S,\pi_\theta) = V(\pi_\theta) - \widehat{V}_{\mathrm{PM}}(S,\pi_\theta)$ and $\alpha \in (1,2]$. In contrast to the bound presented in equation 34, which necessitates a bounded reward, exhibits a dependence on $\log(1/\delta)^{\frac{\epsilon}{1+\epsilon}}$ and two hyper-parameter (s,λ_p), our work offers several advancements. We derive both upper and lower bounds on generalization error, as detailed in Theorem D.2 and Theorem D.1, respectively. These bounds help us for our subsequent derivation of an upper bound on regret. Notably, our bounds demonstrate a more favorable dependence of $\log(1/\delta)^{1/2}$. This improvement not only eliminates the requirement for bounded rewards but also provides a tighter concentration. Furthermore, we provide theoretical analysis for robustness with respect to both noisy reward and noisy propensity scores, and we just have one hyperparameter. Note that the assumption on $P_\alpha(\pi_\theta\|\pi_0)$ for $\alpha = 1+\epsilon$ in (Metelli et al., 2018) is similar to bounded $(1+\epsilon)$-th moment of weight function, $w_\theta(a,x)$ for a bounded reward function.

D.1.3 Comparison with ES Estimator

The ES estimator (Aouali et al., 2023)is represented as,

$$\widehat{V}^{\alpha}_{\mathrm{ES}}(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_|x_i)^\alpha}, \quad \alpha \in [0,1]. \tag{35}$$

In (Aouali et al., 2023, Theorem 4.1), an upper bound on generalization error is derived via PAC-Bayesian approach for $\alpha \in [0,1]$,

$$\begin{aligned}|V(\pi_{\mathbb{Q}}) - \widehat{V}^{\alpha}_{\mathrm{ES}}(\pi_{\mathbb{Q}})| \le \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}})}{n\lambda} \\ + \frac{\lambda}{2}\bar{V}_n^\alpha(\pi_{\mathbb{Q}}).\end{aligned}$$

$$\begin{aligned}
&\text{where } \mathrm{KL}_1(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4\sqrt{n}}{\delta}, \text{ and} \\
&\mathrm{KL}_2(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta}, \\
&B_n^\alpha(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\pi_0^{1-\alpha}(a|x_i)\right], \\
&\bar{V}_n^\alpha(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)\|R\|_\infty^2}{\pi_0(a_i|x_i)^{2\alpha}},
\end{aligned} \tag{36}$$

where $\mathbb{Q}$ and $\mathbb{P}$ are posterior and prior distributions over the set of hypothesis, $\widehat{R}_n^\alpha(\pi_{\mathbb{Q}})$ is ES estimator and $R(\pi_{\mathbb{Q}})$ is true risk. The ES estimator's bound exhibits several limitations. Primarily, it requires a bounded reward. Moreover, the upper bound on the generalization error of the ES estimator converges at a rate of $O(\log(n)n^{-1/2})$, which is suboptimal. A notable drawback is the presence of the term $B_n^\alpha(\pi_{\mathbb{Q}})$, which remains constant for $\alpha > 1$ and does not decrease with increasing sample size n. In contrast, we derive an upper bound on the Regret with a convergence rate of $O(n^{-1/2})$ under the condition of bounded second moment ($\epsilon = 1$) and can be extended for heavy-tailed scenarios under bounded reward. This improved rate not only eliminates the logarithmic factor but also demonstrates a tighter concentration. Furthermore, we have a theoretical analysis for robustness with respect to both noisy reward and noisy propensity scores. Finally, the noisy reward scenario is not studied under the ES estimator.

D.1.4 Comparison with IX Estimator

The IX estimator (Gabbianelli et al., 2023) is defined as for $\eta > 0$,

$$\widehat{V}_{\mathrm{ES}}^{\eta}(S,\pi_\theta) := \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_\theta(a_i|x_i)}{\pi_\theta(a_i|x_i)+\eta}r_i.$$

The following upper bound on regret of IX estimator is derived in (Gabbianelli et al., 2023, Theorem 1),

$$\mathfrak{R}(\pi_{\theta^*}) \leq \sqrt{\frac{\log(2|\Pi_\theta|/\delta)}{n}}(2\eta C_\eta(\pi_{\theta^*})+1), \tag{37}$$

where

$$C_\eta(\pi_\theta) = \mathbb{E}\left[\sum_a \frac{\pi_\theta(a|X)}{\pi_0(a|X)+\eta}\cdot r(X,a)\right]. \tag{38}$$

In equation 37, it is assumed that reward is bounded. The term $C_\eta(\pi_\theta)$ can be large if η is small. While a small η is desirable for reducing bias, it can simultaneously increase $C_\eta(\pi_\theta)$, potentially compromising the tightness of the bound. The bounded reward in $[0,1]$ is needed for the proof of regret bound as $R^2 \leq R$ for $R \in [0,1]$. Moreover, the process of tuning η in the IX estimator is particularly sensitive.

D.1.5 Comparison with Logarithmic Smoothing

We provide theoretical comparison with the Logarithmic Smoothing (LS) estimator (Sakhi et al., 2024).

The LS estimator is,

$$\hat{V}_n^{\tilde{\lambda}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{\lambda}}\log(1+\tilde{\lambda}w_\theta(x_i,a_i)r_i),$$

for $\tilde{\lambda} > 0$. As mentioned in (Sakhi et al., 2024), a Taylor expansion of LS estimator around $\tilde{\lambda} = 0$ yields,

$$\hat{V}_n^{\tilde{\lambda}}(\pi) = \hat{V}_n(\pi) + \sum_{\ell=2}^{\infty}\frac{(-1)^\ell\tilde{\lambda}^{\ell-1}}{\ell}\left(\frac{1}{n}\sum_{i=1}^{n}(w_\theta(x_i,a_i)r_i)^\ell\right).$$

Furthermore, the authors introduced,

$$\mathcal{S}_{\tilde{\lambda}}(\pi) = \mathbb{E}\left[\frac{(w_\theta(X,A)r)^2}{(1+\tilde{\lambda}w_\pi(X,A)r)}\right],$$

where in (Sakhi et al., 2024, Proposition 7), a bounded second moment is needed to derive the generalization error bound. Furthermore, for PAC-Bayesian analysis, the author proposed a linearized version,

$$\hat{V}_n^{\tilde{\lambda}\text{-LIN}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\tilde{\lambda}}\log\left(1+\frac{\tilde{\lambda}r_i}{\pi_0(a_i|x_i)}\right),$$

Note that, the linearized version of LS estimator is bounded by IPS estimator due to $\log(1+x) \leq x$ inequality. Then, for LS-LIN estimator the PAC-Bayesian upper bound on the Regret of LS-LIN estimator is derived in (Sakhi et al., 2024, Proposition 11) as follows,

$$0 \leq V(\hat{\pi}_n) - V(\pi_Q^*) \leq \tilde{\lambda} S_{\tilde{\lambda}}^{\text{LIN}}(\pi_Q^*) + \frac{2(\text{KL}(Q||P) + \ln(2/\delta))}{\tilde{\lambda} n},$$

where $S_{\tilde{\lambda}}^{\text{LIN}}(\pi) = \mathbb{E}\left[\frac{\pi(a|x)r^2}{\pi_0(a|x)+\tilde{\lambda}\pi_0(a|x)r}\right]$.

Theoretical Comparison: The key distinction between the LS estimator and our LSE estimator is that we explicitly assume the heavy-tailed weighted reward and can drive the better convergence rate.

In (Sakhi et al., 2024, Proposition 7), the authors demonstrate that under the assumption of *a bounded second moment of the weighted reward*, the convergence rate is $O(1/\sqrt{n})$.

However, if the second moment is not bounded, from (Sakhi et al., 2024) we only know that:

$$S_{\tilde{\lambda}}(\pi) = \mathbb{E}\left[\frac{(w(X,A)r)^2}{1+\tilde{\lambda}w(X,A)r}\right] \leq \min\left(\frac{1}{\tilde{\lambda}}\mathbb{E}\left[w(X,A)r\right], \mathbb{E}\left[(w(X,A)r)^2\right]\right).$$

If we replace $S_{\tilde{\lambda}}(\pi)$ with $\frac{1}{\tilde{\lambda}}\mathbb{E}\left[w(X,A)r\right]$ in (Sakhi et al., 2024, Proposition 7), we get $O(1)$ as convergence rate. In contrast, our analysis yields a convergence rate of

$$O(n^{-\epsilon/(1+\epsilon)}),$$

for bounded $(1+\epsilon)$-th moment.

This result demonstrates that our assumption is both precise and necessary to achieve the optimal convergence rate for regret under the heavy-tailed assumption.

D.1.6 Comparison with Optimistic Shrinkage

The OS estimator (Su et al., 2020) is represented as for $\tau \geq 0$.

$$\widehat{V}_{\text{OS}}(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\tau w_\theta(a_i,x_i)}{w_\theta^2(a_i,x_i)+\tau} r_i. \tag{39}$$

In (Metelli et al., 2021, Theorem E.1), an upper bound for the right tail of the concentration inequality for the OS estimator is established, which depends on $P_3(\pi_\theta \| \pi_0)$. Consequently, this estimator fails to ensure reliable performance under heavy-tailed assumptions, even when the reward is bounded. Furthermore, due to applying the Bernstein inequality in the proof, theoretical results can not be extended to unbounded reward.

D.1.7 Comparison under Bounded Reward Assumption

In this section, we compare different estimators by assuming bounded reward. Note that, under bounded reward assumption, $R \in [0, R_{\max}]$, our Assumption 1, would be simplified as follows,

Assumption 3. The $P_X \otimes \pi_0(A|X)$ are such that for all learning policy $\pi_\theta(A|X) \in \Pi_\theta$ and some $\epsilon \in [0,1]$, the $(1+\epsilon)$-th moment of the weight function is bounded,

$$\mathbb{E}_{P_X \otimes \pi_0(A|X)}\left[\big(w_\theta(A,X)\big)^{1+\epsilon}\right] \leq \nu_w. \tag{40}$$

Note that, under Assumption 3, our theoretical results hold by replacing ν with $\nu_w R_{\max}^{1+\epsilon}$. In the following, we compare main estimators, PM, ES, IX, LS and OS with LSE under Assumption 3,

- The PM estimator provides an upper bound on concentration inequality under Assumption 3. However, a lower bound on generalization error (concentration inequality) is not provided. Furthermore, for $\epsilon = 0$, we can have a bounded upper bound on generalization error. However, (Metelli et al., 2021, Theorem 5.1) is infinite for $\epsilon = 0$.[10]

[10] Note that in Metelli et al. (2021), the authors consider $\alpha \in (1,2]$ where $\alpha = \epsilon + 1$ and $\epsilon \in (0,1]$.

- The ES estimator, does not support Assumption 3 and an assumption on bounded $\frac{\pi_\theta}{\pi_0^{2\alpha}}$ for $\alpha \in (0,1)$ is needed. Furthermore, the convergence rate of generalization bound on ES estimator is worse than ours in $\epsilon = 1$.
- For OS estimator, the bounded assumption on third moment of weight function is needed. Therefore, it does not support Assumption 3.
- The theoretical results for LS estimator do not need bounded $(1+\epsilon)$-th moment of weight function, Assumption 3. However, under Assumption 3, we can not derive the optimal rate of regret, $O(n^{\frac{-\epsilon}{1+\epsilon}})$ for $\epsilon \in [0,1]$ under LS estimator.
- For IX estimator, using the upper bound on regret in (Gabbianelli et al., 2023, Theorem 7), requires bounded $C_0(\pi_{\theta^*})$, which can impose a stronger condition than Assumption 3.

D.1.8 Detailed Comparison with Tilted Empirical Risk

Inspired by the log-sum-exponential function, the authors in (Li et al., 2023) proposed a non-linear form known as tilted empirical risk. They established connections between tilted empirical risk and other risk measures, particularly demonstrating that tilted empirical risk acts as a risk regularization via the KL divergence between uniform and weighted distributions. Furthermore, they explored the connection between tilted empirical risk and conditional value at risk. However, the generalization error and excess risk analysis of tilted empirical risk remained unexplored. Since our LSE estimator is also based on the log-sum-exponential function, we believe our current analysis of generalization error and regret in OPL/OPE could be extended to analyze tilted empirical risk under heavy-tailed assumptions and improve the understanding of tilted empirical risk under heavy-tailed scenario in supervised learning scenario.

D.1.9 Comparison with the Assumption 1 in Switch Estimator

The switch estimator introduced in (Wang et al., 2017) adaptively chooses between model-free estimation and an estimated reward function based on importance weights. While (Wang et al., 2017) requires the existence of finite $(2+\tilde{\epsilon})$-th moments (for $\epsilon > 0$) in their Assumption 1, our work operates under a weaker condition. We only require bounded $(1+\epsilon)$-th moments for some $\epsilon \in [0,1]$. This distinction is significant—our assumption (Assumption 1) encompasses cases where the second moment and $(2+\tilde{\epsilon})$-th moment for $\tilde{\epsilon} > 0$ do not exist. In contrast, (Wang et al., 2017, Assumption 1), which requires the finiteness of the $(2+\tilde{\epsilon})$-th moments, imposes a strictly stronger condition on the underlying distribution. Therefore, we can not apply the approach in (Wang et al., 2017) in our case.

D.2 Proofs and Details of Regret Bounds

Lemma 5.1 (Restated). *Consider the random variable $Z > 0$. For $\epsilon \in [0,1]$, the following upper bound holds on the variance of $e^{\lambda Z}$ for $\lambda < 0$,*

$$\mathbb{V}\left(e^{\lambda Z}\right) \le |\lambda|^{1+\epsilon}\mathbb{E}[Z^{1+\epsilon}]. \tag{41}$$

Proof. We have,

$$|e^{\lambda Z} - e^{\lambda C_1}| = \left|\int_{\lambda C_1}^{\lambda z} e^y dy\right| \le |\lambda(z - C_1)|e^{\max(\lambda z, \lambda C_1)} \le |\lambda||z - C_1|.$$

Then it holds that

$$\begin{aligned}\mathbb{V}(e^{\lambda Z}) = \min_{C_1 \in \mathbb{R}^+} \mathbb{E}\left[(e^{\lambda Z} - e^{\lambda C_1})^2\right] &= \min_{C_1 \in \mathbb{R}^+} \mathbb{E}\left[|e^{\lambda Z} - e^{\lambda C_1}|^{1-\epsilon}|e^{\lambda Z} - e^{\lambda C_1}|^{1+\epsilon}\right] \\ &= \min_{C_1 \in \mathbb{R}^+} \mathbb{E}\left[|e^{\lambda Z} - e^{\lambda C_1}|^{1-\epsilon}|\lambda|^{1+\epsilon}|Z - C_1|^{1+\epsilon}\right] \\ &\le \min_{C_1 \in \mathbb{R}^+} \mathbb{E}\left[|\lambda|^{1+\epsilon}|Z - C_1|^{1+\epsilon}\right] \le |\lambda|^{1+\epsilon}\mathbb{E}[Z^{1+\epsilon}],\end{aligned}$$

where the last inequality holds due to the fact that $|e^{\lambda Z} - e^{\lambda C_1}|^{1-\epsilon} \le 1$. □

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

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

where $0 < \delta < 1$ and n is the number of samples in LBF dataset. We first provide an upper bound on generalization error.

Theorem D.1. *Given Assumption 1, with probability at least $1 - \delta$, then the following upper bound holds on the generalization error of the LSE for a learning policy $\pi_\theta \in \Pi_\theta$*

$$\text{gen}_\lambda(\pi_\theta) \leq \frac{1}{1+\epsilon}|\lambda|^\epsilon \nu - \frac{1}{\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n \exp(2\lambda \nu^{1/(1+\epsilon)})}} - \frac{4\log(2/\delta)}{3\lambda \exp(\lambda \nu^{1/(1+\epsilon)})n}.$$

Proof. To ease the notation, we consider $Y_\theta(A, X) = w_\theta(A, X)R$. Using Bernstein's inequality (Lemma B.4), with probability $(1 - \delta)$, we have,

$$\mathbb{E}[\exp(\lambda Y_\theta(A, X))] - \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda Y_\theta(a_i, x_i)) \geq -\sqrt{\frac{4\mathbb{V}(\exp(\lambda Y_\theta(A, X)))\log(2/\delta)}{n}} - \frac{4\log(2/\delta)}{3n}.$$

Using Lemma 5.1, $\mathbb{V}(\exp(\lambda Y_\theta(A, X))) \leq |\lambda|^{1+\epsilon}\nu$, we have,

$$\mathbb{E}[\exp(\lambda Y_\theta(A, X))] - \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda Y_\theta(a_i, x_i)) \geq -\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} - \frac{4\log(2/\delta)}{3n}.$$

As the log function is an increasing function, the following holds with probability at least $1 - \delta$,

$$\widehat{\mathrm{V}}^\lambda_{\text{LSE}}(S, \pi_\theta) \geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3n}\right).$$

where recall that $\widehat{\mathrm{V}}^\lambda_{\text{LSE}}(S, \pi_\theta) = \frac{1}{\lambda}\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda y_\theta(a_i, x_i))\right)$. With probability at least $1 - \delta$, using the inequality $\log(x + y) \leq \log(x) + y/x$ for $x > 0$,

$$\begin{aligned}
\widehat{\mathrm{V}}^\lambda_{\text{LSE}}(S, \pi_\theta) &\geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3n}\right) \\
&\geq \frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) + \frac{1}{\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]n}.
\end{aligned}$$

Using Lemma B.10, we have with probability at least $1 - \delta$,

$$\begin{aligned}
\widehat{\mathrm{V}}^\lambda_{\text{LSE}}(S, \pi_\theta) &\geq \mathbb{E}[Y_\theta(A, X)] - \frac{1}{1+\epsilon}|\lambda|^\epsilon \mathbb{E}[Y_\theta(A, X)^{1+\epsilon}] \\
&\quad + \frac{1}{\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]n} \\
&\geq \mathbb{E}[Y_\theta(A, X)] - \frac{1}{1+\epsilon}|\lambda|^\epsilon \mathbb{E}[Y_\theta(A, X)^{1+\epsilon}] \\
&\quad + \frac{1}{\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3\lambda \mathbb{E}[e^{\lambda Y_\theta(A,X)}]n} \\
&\geq \mathbb{E}[Y_\theta(A, X)] - \frac{1}{1+\epsilon}|\lambda|^\epsilon \nu + \frac{1}{\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log(2/\delta)}{n \exp(2\lambda \nu^{1/(1+\epsilon)})}} + \frac{4\log(2/\delta)}{3\lambda \exp(\lambda \nu^{1/(1+\epsilon)})n}.
\end{aligned}$$

The final result holds by by applying Lemma B.9 to $\mathbb{E}[e^{\lambda Y_\theta(A,X)}] \geq \exp(\lambda \nu^{1/(1+\epsilon)})$. □

Next, we provide a lower bound on generalization error.

Theorem D.2. *Given Assumption 1, and assuming* $n \geq \frac{\left(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$, *then there exists* $\gamma\in(0,1)$ *such that with probability at least* $1-\delta$, *the following lower bound on generalization error of the LSE for a learning policy* $\pi_\theta\in\Pi_\theta$ *holds*

$$\mathrm{gen}_\lambda(\pi_\theta)\geq\frac{1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}}+\frac{4\log(2/\delta)}{3(1-\gamma)\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}$$

Proof. To ease the notation, we consider $Y_\theta(A,X)=Rw_\theta(A,X)$. Using Bernstein's inequality (Lemma B.4), with probability $(1-\delta)$, we have,

$$\mathbb{E}[\exp(\lambda Y_\theta(A,X))]-\frac{1}{n}\sum_{i=1}^n\exp(\lambda Y_\theta(a_i,x_i))\leq\sqrt{\frac{4\mathbb{V}(\exp(\lambda Y_\theta(A,X)))\log(2/\delta)}{n}}+\frac{4\log(2/\delta)}{3n}.$$

Using Lemma 5.1, $\mathbb{V}(\exp(\lambda Y_\theta(A,X)))\leq|\lambda|^{1+\epsilon}\nu$, we have,

$$\mathbb{E}[\exp(\lambda Y_\theta(A,X))]-\frac{1}{n}\sum_{i=1}^n\exp(\lambda Y_\theta(a_i,x_i))\leq\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}+\frac{4\log(2/\delta)}{3n}.$$

As the log function is an increasing function, the following holds with probability at least $1-\delta$,

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)\leq\frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]-\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}-\frac{4\log(2/\delta)}{3n}\right).$$

where recall that $\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)=\frac{1}{\lambda}\log\left(\frac{1}{n}\sum_{i=1}^n\exp(\lambda y_\theta(a_i,x_i))\right)$. Without loss of generality, we can assume that,

$$\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}+\frac{4\log(2/\delta)}{3n}\leq\gamma\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\tag{42}$$

for some $\gamma\in(0,1)$. Using the inequality $\log(z-y)\geq\log(z)-\frac{y}{z-y}$ for $z>y>0$, and assuming $z=\mathbb{E}[e^{\lambda Y_\theta(A,X)}]$ and $y=\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}+\frac{4\log(2/\delta)}{3n}$ and combining with equation 42, then with probability $(1-\delta)$, we have,

$$\begin{aligned}\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)&\leq\frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]-\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}-\frac{4\log(2/\delta)}{3n}\right)\\&\leq\frac{1}{\lambda}\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right)-\frac{1}{\lambda(1-\gamma)\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}}\\&\quad-\frac{4\log(2/\delta)}{(1-\gamma)\lambda3\mathbb{E}[e^{\lambda Y_\theta(A,X)}]n}.\end{aligned}$$

Equation 42 can be considered as quadratic equation in terms of $\frac{1}{\sqrt{n}}$. Then, using lemma B.6, we have,

$$\frac{\left(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma\right)\log(2/\delta)}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}\leq n.\tag{43}$$

Using Lemma B.10, with probability at least $(1-\delta)$ we have

$$\begin{aligned}\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) &\leq \mathbb{E}[Y_\theta(A,X)] - \frac{1}{\lambda(1-\gamma)\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}} - \frac{4\log(2/\delta)}{3(1-\gamma)\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]n} \\ &\leq \mathbb{E}[Y_\theta(A,X)] - \frac{1}{(1-\gamma)\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n}} - \frac{4\log(2/\delta)}{3(1-\gamma)\lambda\mathbb{E}[e^{\lambda Y_\theta(A,X)}]n} \\ &\leq \mathbb{E}[Y_\theta(A,X)] - \frac{1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2/\delta)}{3(1-\gamma)\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}\,.\end{aligned}$$

The final result holds by applying Lemma B.9 to $\mathbb{E}[e^{\lambda Y_\theta(A,X)}] \geq \exp(\lambda\nu^{1/(1+\epsilon)})$. □

Using the previous upper and lower bounds on generalization error, we can provide an upper bound on the regret of the LSE estimator.

Theorem 5.2 (Restated). *Given Assumption 1 and assuming $n \geq \frac{(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$, with probability at least $1-\delta$, then there exists $\gamma\in(0,1)$ such that the following upper bound holds on the regret of the LSE estimator,*

$$0 \leq \mathfrak{R}_\lambda(\pi_{\widehat{\theta}},S) \leq \frac{|\lambda|^\epsilon}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\nu^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}}.$$

Proof. We have,

$$V(\pi_{\theta^*}) - V(\pi_{\widehat{\theta}}) = \underbrace{V(\pi_{\theta^*}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\theta^*})}_{I_1} + \underbrace{\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\theta^*}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}})}_{I_2} + \underbrace{\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}}) - V(\pi_{\widehat{\theta}})}_{I_3}. \tag{44}$$

Using upper bound on generalization error, Theorem D.1, and union bound (Shalev-Shwartz & Ben-David, 2014), with probability at least $1-\delta$, the following upper bound holds on term I_1,

$$\begin{aligned}V(\pi_{\theta^*}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\theta^*}) &\leq \sup_{\pi_\theta\in\Pi_\theta} V(\pi_\theta) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) \\ &\leq \frac{1}{1+\epsilon}|\lambda|^\epsilon\nu - \frac{1}{\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}.\end{aligned} \tag{45}$$

Using lower bound on generalization error, Theorem D.2, and union bound (Shalev-Shwartz & Ben-David, 2014), with probability at least $1-\delta$, the following upper bound holds on term I_3,

$$\begin{aligned}\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}}) - V(\pi_{\widehat{\theta}}) &\leq \sup_{\pi_\theta\in\Pi_\theta} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) - V(\pi_\theta) \\ &\leq \frac{-1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3(1-\gamma)\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}.\end{aligned} \tag{46}$$

Note that the term I_2 is negative as the $\pi_{\widehat{\theta}}$ is the maximizer of the LSE estimator over Π_θ. Combining equation 45 and equation 46 with equation 44, and applying the union bound, completes the proof.

□

Proposition 5.3 (Restated). *Given Assumption 1, for any $0<\gamma<1$, assuming $n \geq \frac{(2\nu+\frac{4}{3}\gamma)\log\frac{1}{\delta}}{\gamma^2\exp(2\nu^{1/(1+\epsilon)})}$ and setting $\lambda=-n^{-\zeta}$ for $\zeta\in\mathbb{R}^+$, then the overall convergence rate of the regret upper bound is $\max(O(n^{-1+\zeta}),O(n^{-\epsilon\zeta}),O(n^{(-\zeta\epsilon-1)/2}))$ for finite policy set.*

Proof. Without loss of generality, we can assume that $\lambda \geq -1$. Therefore, we have $|\lambda|^{1+\epsilon} \leq 1$ and $\nu^{1/(1+\epsilon)} \geq 0$, which results in $n \geq \frac{\left(2\nu+\frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2\exp(-2\nu^{1/(1+\epsilon)})} \geq \frac{\left(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma\right)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$. Using Theorem 5.2, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\mathfrak{R}_\lambda(\pi_{\widehat{\theta}}, S)\\
&\leq \frac{|\lambda|^\epsilon}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\nu^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} \qquad (47)\\
&\leq \frac{|\lambda|^\epsilon}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\nu^{1/(1+\epsilon)})} + \frac{(2-\gamma)}{(1-\gamma)\exp(\lambda\nu^{1/(1+\epsilon)})}\sqrt{\frac{4|\lambda|^{\epsilon}\nu\log\frac{4|\Pi_\theta|}{\delta}}{n}}. \qquad (48)
\end{aligned}$$

Since $\lambda \geq -1$, we have $\exp(\lambda\nu^{1/(1+\epsilon)}) \geq \exp(-\nu^{1/(1+\epsilon)})$ (note that $\nu^{1/(1+\epsilon)} \geq 0$ and $-1 < \lambda < 0$). Replacing λ with $\lambda^\star = -n^{-\zeta}$ and $\exp\left(\lambda\nu^{1/(1+\epsilon)}\right)$ with $\exp\left(-\nu^{1/(1+\epsilon)}\right)$, then we have the overall convergence rate of $\max(O(n^{-\epsilon\zeta}), O(n^{-1+\zeta}), O(n^{(-\zeta\epsilon-1)/2}))$. □

D.3 Proofs and Details of Bias and Variance

Proposition 5.5 (Restated). *Given Assumption 1, the following lower and upper bounds hold on the bias of the LSE estimator,*

$$\frac{(n-1)}{2n|\lambda|}\mathbb{V}(e^{\lambda w_\theta(A,X)R}) \leq \mathbb{B}(\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \leq \frac{1}{1+\epsilon}|\lambda|^\epsilon\nu + \frac{1}{2n\lambda}\mathbb{V}(e^{\lambda w_\theta(A,X)R}).$$

Proof. In the proof, for the sake of simplicity of notation, we consider $Y_\theta(A,X) = w_\theta(A,X)R$. For lower bound we need to prove the following,

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

Setting $y_\theta(a_i,x_i) = r_i w_\theta(a_i,x_i)$, 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,

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

Note that $\mathbb{E}[Y_\theta(A,X)] = V(\pi_\theta)$. It completes the proof for lower bound.

For upper bound, we need to prove the following

$$\frac{1}{2n\lambda}\mathbb{V}(e^{\lambda w_\theta(A,X)R}) \geq \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \mathbb{E}\left[\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)\right]. \tag{49}$$

Note that, an upper bound 1 on $\frac{\sum_{i=1}^n e^{\lambda r_i w_\theta(a_i,x_i)}}{n}$ holds. Now, we have,

$$\mathbb{E}[\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)] = \frac{1}{\lambda}\mathbb{E}\left[\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)\right]$$
$$= \frac{1}{\lambda}\mathbb{E}\left[\log\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right) + \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2 - \frac{1}{2}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2\right]$$
$$\geq \frac{1}{\lambda}\left(\log\left(\mathbb{E}\left[\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right]\right) + \frac{1}{2}\mathbb{E}\left[\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right]^2 - \frac{1}{2}\mathbb{E}\left[\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)^2\right]\right)$$
$$= \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \frac{1}{2\lambda}\mathbb{V}\left(\frac{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}{n}\right)$$
$$= \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right),$$

where the first inequality 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 $b=1$. Then, we have,

$$\frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) - \mathbb{E}[\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)] \leq \frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right).$$

Finally, we combine the upper bound in equation 49 .

$$\mathbb{E}[\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)] - \frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) \geq -\frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right),$$

and the upper bound in Lemma B.10,

$$\frac{1}{\lambda}\log\left(\mathbb{E}\left[e^{\lambda Y_\theta(A,X)}\right]\right) \geq \mathbb{E}[Y_\theta(A,X)] - \frac{1}{1+\epsilon}|\lambda|^\epsilon\mathbb{E}[|Y_\theta(A,X)|^{1+\epsilon}].$$

Therefore, we have,

$$\begin{aligned}\mathbb{E}[\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)] \geq \mathbb{E}[Y_\theta(A,X)] &- \frac{1}{1+\epsilon}|\lambda|^\epsilon\mathbb{E}[|Y_\theta(A,X)|^{1+\epsilon}] \\ &- \frac{1}{2n\lambda}\mathbb{V}\left(e^{\lambda w_\theta(A,X)R}\right).\end{aligned} \tag{50}$$

It completes the proof. □

Proposition 5.7. *Assume that $\mathbb{E}[(w_\theta(A,X)R)^2] \leq \nu_2$ (Assumption 1 for $\epsilon = 1$) holds. Then the variance of the LSE estimator with $\lambda < 0$, satisfies,*

$$\mathbb{V}(\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)) \leq \frac{1}{n}\mathbb{V}(w_\theta(A,X)R) \leq \frac{1}{n}\nu_2. \tag{51}$$

Proof. Let $Y_\theta(A,X) = w_\theta(A,X)R$ and $Y_\theta^{(c)} = Y_\theta(A,X) - \mathbb{E}[Y_\theta(A,X)]$ be the centered $Y_\theta(A,X)$. We have,

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta) = \frac{1}{\lambda}\ln\left(\frac{\sum_{i=1}^n e^{\lambda y_{i,\theta}}}{n}\right) = \frac{1}{\lambda}\ln\left(\frac{\sum_{i=1}^n e^{\lambda(y_{i,\theta}^{(c)} - m_\theta)}}{n}\right) = \frac{1}{\lambda}\ln\left(\frac{\sum_{i=1}^n e^{\lambda y_{i,\theta}^{(c)}}}{n}\right) + m_\theta$$

where $m_\theta = \mathbb{E}[Y_\theta(A, X)]$. Note that, we also have $\mathbb{V}(Y_\theta^{(c)}) = \mathbb{V}(Y_\theta)$.

Now, setting $Z = \frac{\sum_{i=1}^n e^{\lambda y_{i,\theta}^{(c)}}}{n}$, we have,

$$\mathbb{V}(\widehat{\mathrm{V}}_{\mathrm{LSE}}^\lambda(S, \pi_\theta)) = \mathbb{V}\left(\frac{1}{\lambda}\log Z\right)$$

Furthermore, using Jensen's inequality for $\lambda < 0$, we have,

$$\frac{1}{\lambda}\log Z \le \frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}.$$

Hence we have,

$$\begin{aligned}
\mathbb{V}(\widehat{\mathrm{V}}_{\mathrm{LSE}}^\lambda(S, \pi_\theta)) &= \mathbb{E}\left[\frac{1}{\lambda^2}\log^2 Z\right] - \left(\mathbb{E}\left[\frac{1}{\lambda}\log Z\right]\right)^2 \\
&\le \mathbb{E}\left[\left(\frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}\right)^2\right] \\
&= \mathbb{V}\left(\frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}\right) + \mathbb{E}\left[\frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}\right]^2 \\
&= \mathbb{V}\left(\frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}\right) + \mathbb{E}\left[Y_\theta^{(c)}\right]^2 \\
&= \mathbb{V}\left(\frac{\sum_{i=1}^n y_{i,\theta}^{(c)}}{n}\right) + 0 \\
&= \frac{1}{n}\mathbb{V}(Y_\theta^{(c)}) = \frac{1}{n}\mathbb{V}(Y_\theta).
\end{aligned}$$

It completes the proof. □

For the moment of the LSE estimator, we provide the following upper bound.

Proposition D.3 (Moment bound)**.** *Given Assumption 1, the following upper bound hold on the moment of the LSE estimator,*

$$\mathbb{E}\left[\left|\frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^n e^{\lambda w_\theta(a_i,x_i) r_i}}{n}\right)\right|^{1+\epsilon}\right] \le \nu. \tag{52}$$

Proof. Suppose that $Z = \frac{\sum_{i=1}^n e^{\lambda r_i w_\theta(a_i,x_i)}}{n}$. Also set $y_{i,\theta}(a_i, x_i) = r_i(a_i, x_i) w_\theta(a_i, x_i)$. For negative $\lambda < 0$ and $Z > 0$, we have,

$$\begin{aligned}
\widehat{\mathrm{V}}_{\mathrm{LSE}}^\lambda(S, \pi_\theta) &= \frac{1}{\lambda}\log(Z) \\
&\le \frac{\sum_{i=1}^n r_i w_\theta(a_i, x_i)}{n}.
\end{aligned}$$

Since $\log Z < 0$ for $0 < Z < 1$, we have,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{\lambda}\log Z\right|^{1+\epsilon}\right] &\le \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n w_\theta(a_i, x_i) r_i\right|^{1+\epsilon}\right] \\
&\le \mathbb{E}[|w_\theta(A, X)R|^{1+\epsilon}] \\
&\le \nu,
\end{aligned}$$

where the second inequality holds due to Jensen inequality. □

D.4 Proof and Details of Robustness of the LSE estimator: Noisy Reward

Using the functional derivative (Cardaliaguet et al., 2019), we can provide the following results.

Proposition D.4. *Given Assumption 1, then the following holds,*

$$\frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_\theta(A,X)R)]) - \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_\theta(A,X)R)]) \leq \frac{\mathbb{TV}(P_{R|X,A},\widetilde{P}_{R|X,A})}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})}, \tag{53}$$

where $P_1 = P_X \otimes \pi_0(A|X) \otimes P_{R|X,A}$ *and* $P_2 = P_X \otimes \pi_0(A|X) \otimes \widetilde{P}_{R|X,A}$.

Proof. We have that

$$\begin{aligned}
&\frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_\theta(A,X)R)]) - \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_\theta(A,X)R)]) \\
&\overset{(a)}{=} \int_{\mathbb{R}\times\mathcal{X}\times\mathcal{A}} \frac{\exp(\lambda w_\theta(A,X)R)}{|\lambda|\mathbb{E}[\exp(\lambda w_\theta(a,x)r)]} P_X \otimes \pi_0(A|X)(\widetilde{P}_{R|X,A} - \widetilde{P}_{R|X,A})(\mathrm{d}a\mathrm{d}x\mathrm{d}r) \\
&\overset{(b)}{\leq} \frac{\mathbb{TV}(P_{R|X,A},\widetilde{P}_{R|X,A})}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})}.
\end{aligned} \tag{54}$$

where (a) and (b) follow from the functional derivative and Lemma B.2. □

Combining Proposition D.4 with generalization error bounds, Theorem D.2 and Theorem D.1, we derive the upper bound on the regret under noisy reward scenario.

Theorem 5.9. *Given Assumption 1, Assumption 2 and assuming* $n \geq \frac{\left(2|\lambda|^{1+\epsilon}\nu + \frac{4}{3}\gamma\right)\log\frac{|\Pi_\theta|}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}$, *with probability at least* $1-\delta$, *then there exists* $\gamma \in (0,1)$ *such that the following upper bound holds on the regret of the LSE estimator under noisy reward logged data,*

$$\begin{aligned}
0 \leq \mathfrak{R}_\lambda(\pi_{\widehat{\theta}}(\widetilde{S}),\widetilde{S}) \leq{}& \frac{|\lambda|^\epsilon}{1+\epsilon}\nu \\
&- \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\widetilde{\nu}\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\widetilde{\nu}^{1/(1+\epsilon)})}} \\
&+ \mathbb{TV}(P_{R|X,A},\widetilde{P}_{R|X,A})\Big(\frac{1}{|\lambda|\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})} + \frac{1}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})}\Big),
\end{aligned}$$

where $\pi_{\widehat{\theta}}(\widetilde{S}) = \arg\max_{\pi_\theta \Pi_\theta} \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\pi_\theta,\widetilde{S})$.

Proof. We have,

$$\begin{aligned}
V(\pi_{\theta^*}) - V(\pi_{\widehat{\theta}}(\widetilde{S})) ={}& \underbrace{V(\pi_{\theta^*}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widetilde{S},\pi_{\theta^*})}_{I_1} + \underbrace{\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widetilde{S},\pi_{\theta^*}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widetilde{S},\pi_{\widehat{\theta}}(\widetilde{S}))}_{I_2} \\
&+ \underbrace{\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widetilde{S},\pi_{\widehat{\theta}}(\widetilde{S})) - V\big(\pi_{\widehat{\theta}}(\widetilde{S})\big)}_{I_3}.
\end{aligned} \tag{55}$$

Using upper bound on generalization error, Theorem D.1, and union bound (Shalev-Shwartz & Ben-David, 2014), with probability at least $1-\delta$, the following upper bound holds on term I_1,

$$\begin{aligned}
&V(\pi_{\theta^*}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widetilde{S},\pi_{\theta^*}) \\
&= V(\pi_{\theta^*}) - \frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_\theta(A,X)R)]) \\
&\quad + \frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_\theta(A,X)R)]) - \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_\theta(A,X)R)]) \\
&\quad + \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_\theta(A,X)R)]) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widetilde{S},\pi_{\theta^*}) \\
&\leq \frac{1}{1+\epsilon}|\lambda|^{\epsilon}\nu \\
&\quad + \frac{\mathbb{TV}(P_{R|X,A},\widetilde{P}_{R|X,A})}{|\lambda|\exp(\lambda\nu^{1/(1+\epsilon)})} \\
&\quad - \frac{1}{\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\widetilde{\nu}\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\widetilde{\nu}^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}.
\end{aligned} \tag{56}$$

Using lower bound on generalization error, Theorem D.2, and union bound (Shalev-Shwartz & Ben-David, 2014), with probability at least $1-\delta$, the following upper bound holds on term I_3,

$$\begin{aligned}
&\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widetilde{S},\pi_{\widehat{\theta}}(\widetilde{S})) - V(\pi_{\widehat{\theta}}(\widetilde{S})) \\
&= \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widetilde{S},\pi_{\widehat{\theta}}(\widetilde{S})) - \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_{\widehat{\theta}}(A,X)R)]) \\
&\quad + \frac{1}{\lambda}\log(\mathbb{E}_{P_2}[\exp(\lambda w_{\widehat{\theta}}(A,X)R)]) - \frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_{\widehat{\theta}}(A,X)R)]) \\
&\quad + \frac{1}{\lambda}\log(\mathbb{E}_{P_1}[\exp(\lambda w_{\widehat{\theta}}(A,X)R)]) - V(\pi_{\widehat{\theta}}(\widetilde{S})) \\
&\leq \frac{-1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\widetilde{\nu}\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\widetilde{\nu}^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3(1-\gamma)\lambda\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})n} \\
&\quad + \frac{\mathbb{TV}(P_{R|X,A},\widetilde{P}_{R|X,A})}{|\lambda|\exp(\lambda\widetilde{\nu}^{1/(1+\epsilon)})}.
\end{aligned} \tag{57}$$

Note that the term I_2 is negative as the $\pi_{\widehat{\theta}}(\widetilde{S})$ is the maximizer of the LSE estimator over Π_θ. Combining equation 56 and equation 57 with equation 55, and applying the union bound, completes the proof. □

D.5 PAC-Bayesian Discussion

In this section, we explore the PAC-Bayesian approach and its application in extending our previous results. Given that the methodology for deriving these results closely resembles our earlier approach, we will outline the key steps in the derivation process rather than providing a full detailed analysis.

For this purpose, we introduce several additional definitions inspired by Gabbianelli et al. (2023). For PAC-Bayesian approach, we focus on randomized algorithms that output a distribution $\widehat{Q}_n \in \mathcal{P}(\Pi_\theta)$ over policies. Our interest lies in performance guarantees that satisfy two conditions: (1) they hold in expectation with respect to the random selection of $\widehat{\pi}_n \sim \widehat{Q}_n$, and (2) they maintain high probability with respect to the realization of the LBF dataset. For this purpose, we define the following integral forms of our previous formulation,

$$
\begin{aligned}
V(Q) &= \int V(\pi_\theta)\mathrm{d}Q(\pi_\theta), \\
\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,Q) &= \int \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi)\mathrm{d}Q(\pi), \\
\mathfrak{R}(Q,S) &= \int \mathfrak{R}(\pi,S)\mathrm{d}Q(\pi).
\end{aligned} \tag{58}
$$

These expressions capture relevant quantities evaluated in expectation under the distribution $\mathbb{Q} \in \mathcal{P}(\Pi_\theta)$ where $\mathcal{P}(\Pi_\theta)$ is the set of distributions over policy set. Let $\mathbb{P} \in \mathcal{P}(\Pi_\theta)$ a prior distribution over policy class.

We can relax the uniform assumption on $(1+\epsilon)$-th moment Assumption 1, as follows,

Assumption 4. The reward distribution $P_{R|X,A}$ and $P_X \otimes \pi_0(A|X)$ are such that for a posterior distribution Q over the set of policies Π_θ and some $\epsilon \in (0,1]$, the $(1+\epsilon)$-th moment of the weighted reward is bounded,

$$
\mathbb{E}_{\pi_\theta\sim\mathbb{Q}}\mathbb{E}_{P_X\otimes\pi_0(A|X)\otimes P_{R|X,A}}\left[\big(w_\theta(A,X)R\big)^{1+\epsilon}\right] \le \nu_q. \tag{59}
$$

In order to derive the upper bound on regret, we need to derive the upper and lower PAC-Bayesian bound on generalization error. For this purpose, we can apply the following bound from (Tolstikhin & Seldin, 2013, Theorem 2) which holds with probability $1-\delta$ and for a fixed $c_1 > 1$,

$$
\begin{aligned}
&\Big|\int_{\pi_\theta\sim\mathbb{Q}} \mathbb{E}[\exp(\lambda Y_\theta(A,X))] - \int_{\pi_\theta\sim\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n}\exp(\lambda Y_\theta(a_i,x_i))\Big| \\
&\le (1+c_1)\sqrt{\frac{(e-2)\mathbb{E}_Q[\mathbb{V}(\exp(\lambda Y_\theta(A,X)))]\left(\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{\nu_1}{\delta}\right)}{n}},
\end{aligned} \tag{60}
$$

where $Y_\theta(a_i,x_i) = w_\theta(a_i,x_i)r_i$ and

$$
\nu_1 = \left\lceil \frac{1}{\ln c_1}\ln\left(\sqrt{\frac{(e-2)n}{4\ln(1/\delta)}}\right)\right\rceil + 1. \tag{61}
$$

Similar to Theorem D.2 and Theorem D.1, we can replace $\mathbb{E}_Q[\mathbb{V}(\exp(\lambda Y_\theta(A,X)))]$ with $|\lambda|^{1+\epsilon}\mathbb{E}[Y_\theta(A,X)^{1+\epsilon}]$. Given Assumption 4, the following upper bounds holds on generalization error,

$$
\mathbb{E}_{\pi_\theta\sim\mathbb{Q}}[\mathrm{gen}_\lambda(\pi_\theta)] \le \frac{1}{1+\epsilon}|\lambda|^{\epsilon}\nu_q - \frac{(1+c_1)}{\lambda}\sqrt{\frac{(e-2)|\lambda|^{1+\epsilon}\nu_q\left(\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\nu_1}{\delta}\right)}{\exp(2\lambda\nu_q^{1/(1+\epsilon)})n}}. \tag{62}
$$

For lower bound, given Assumption 4, there exists n_0 such that for $n \ge n_0$ and $\gamma_q \in (0,1)$ the following holds with probability $(1-\delta)$,

$$
\mathbb{E}_{\pi_\theta\sim\mathbb{Q}}[\mathrm{gen}_\lambda(\pi_\theta)] \ge \frac{(1+c_1)}{\lambda(1-\gamma_q)}\sqrt{\frac{(e-2)|\lambda|^{1+\epsilon}\nu_q\left(\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\nu_1}{\delta}\right)}{\exp(2\lambda\nu_q^{1/(1+\epsilon)})n}}. \tag{63}
$$

Combining equation 63 and equation 62, we can derive an upper bound on $\mathfrak{R}(\widehat{Q},S)$ in a similar approach to Theorem 5.2 under Assumption 4 and assuming $\widehat{Q}_n := \arg\max_{Q\in\mathcal{P}(\Pi_\theta)} \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,Q)$.

$$
\mathfrak{R}(\widehat{Q}_n,S) \le \frac{1}{1+\epsilon}|\lambda|^{\epsilon}\nu_q - \frac{(1+c_1)(2-\gamma_q)}{(1-\gamma_q)\lambda}\sqrt{\frac{(e-2)|\lambda|^{1+\epsilon}\nu_q\left(\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2\nu_1}{\delta}\right)}{\exp(2\lambda\nu_q^{1/(1+\epsilon)})n}}. \tag{64}
$$

Note that, the PAC-Bayesian approach in (London & Sandler, 2019; Sakhi et al., 2023; 2024; Aouali et al., 2023) is different. However, their PAC-Bayesian model can also be applied to our LSE estimator.

D.6 Sub-Gaussian Discussion

In this section, we investigate the sub-Gaussianity concentration inequality (generalization error) under LSE estimator.

We first present the following general result.

Proposition D.5. *Given Assumption 1, for any $0 < \gamma < 1$, assuming $n \geq \max\left(\frac{(2\nu+\frac{4}{3}\gamma)\log\frac{1}{\delta}}{\gamma^2\exp(2\nu^{1/(1+\epsilon)})}, \frac{\log\frac{2}{\delta}}{\nu}\right)$ and setting*

$$\lambda = -\left(\frac{\log\frac{2}{\delta}}{\nu n}\right)^{\frac{1}{1+\epsilon}},$$

then with a probability at least $1-\delta$ for $\delta \in (0,1)$, the absolute of generalization error of the LSE estimator satisfies for a fixed $\pi_\theta \in \Pi_\theta$,

$$\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \leq \left(\frac{1}{1+\epsilon} + \frac{4}{(1-\gamma)\exp(\nu^{1/(1+\epsilon)})}\right)\nu^{\frac{1}{1+\epsilon}}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

Proof. Choosing $n \geq \frac{2\log\frac{2}{\delta}}{\nu}$, we have $\lambda \geq -1$, $|\lambda|^{1+\epsilon} \leq 1$ and $\nu \geq 0$, which results in $n \geq \frac{(2\nu+\frac{4}{3}\gamma)\log\frac{1}{\delta}}{\gamma^2\exp(2\nu^{1+\epsilon})} \geq \frac{(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma)\log\frac{1}{\delta}}{\gamma^2\exp(2\lambda\nu^{1+\epsilon})}$. Using Theorem D.1 and Theorem D.2, we have with probability at least $1-\delta$,

$$\begin{aligned}&\left|\mathrm{gen}_\lambda(\pi_\theta)\right| \\ &\leq \frac{1}{1+\epsilon}|\lambda|^\epsilon\nu - \frac{1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2/\delta)}{3(1-\gamma)\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}\end{aligned} \tag{65}$$

Since $\lambda \geq -1$, we have $\exp(\lambda\nu^{1+\epsilon}) \geq \exp(-\nu^{1+\epsilon})$ (note that $\nu \geq 0$). Replacing λ with $\lambda^\star = -\left(\frac{\log\frac{2}{\delta}}{\nu n}\right)^{\frac{1}{1+\epsilon}}$ and $\exp\left(\lambda\nu^{1+\epsilon}\right)$ with $\exp\left(\nu^{1+\epsilon}\right)$, we have,

$$\begin{aligned}\left|\mathrm{gen}_\lambda(\pi_\theta)\right| &\leq \frac{\nu^{\frac{1}{1+\epsilon}}}{1+\epsilon}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{\epsilon}{1+\epsilon}} + \frac{4\nu^{\frac{1}{1+\epsilon}}}{3(1-\gamma)\exp(\nu^{1/(1+\epsilon)})}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \\ &\quad + \frac{2\nu^{\frac{1}{1+\epsilon}}}{(1-\gamma)\exp(\nu^{1/(1+\epsilon)})}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \\ &\leq \left(\frac{1}{1+\epsilon} + \frac{4}{(1-\gamma)\exp(\nu^{1/(1+\epsilon)})}\right)\nu^{\frac{1}{1+\epsilon}}\left(\frac{\log\frac{2}{\delta}}{n}\right)^{\frac{\epsilon}{1+\epsilon}}\end{aligned}$$

with a probability at least $1-\delta$. As the upper bound on absolute value of the generalization error holds. □

Remark D.6. Suppose that the second moment of weighted reward is bounded which is equal to Assumption 1 with $\epsilon = 1$. As a result, using Proposition D.5 for $\epsilon = 1$, we can establish a concentration inequality (generalization bound) for the LSE even in cases where the rewards are unbounded.

D.7 Implicit Shrinkage

Su et al. (2020) proposed the optimistic shrinkage where the weights are less than the main weights of IPS estimator. Other transformation of weights in other estimators are also lower bound to main weights of IPS estimators. For example, in TR-IPS, we have $\min(M, w_\theta(a,x))$ which is a lower bound to $w_\theta(a,x)$. Our LSE estimator is also a lower bound to IPS estimator,

$$\frac{1}{\lambda}\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda w_\theta(a_i,x_i)r_i)\right) \leq \frac{1}{n}\sum_{i=1}^{n}w_\theta(a_i,x_i)r_i, \tag{66}$$

which can be interpreted as implicit shrinkage. Furthermore, note that the LSE is not separable with respect to the samples, so instead of per sample shrinkage, we investigate LSE's shrinkage effect on the entire output, which is the estimated average reward. In can be derived by simple calculation that for $\lambda < 0$,

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^{n} e^{\lambda y_i}}{n}\right) = \frac{1}{|\lambda|} D_{\mathrm{KL}}\left(\frac{1}{n}\mathbb{1}_n, \mathrm{softmax}(\lambda y_i)\right),$$

where $\mathbb{1}_n$ is all-one vector with size n. Hereby we see that LSE shrinks the Monte-Carlo average by the KL-divergence between the uniform vector and softmax of the samples (with temperature $1/\lambda$). This way, when outlier values or large values are out of the normal range of the data are observed, the amount of shrinkage increases. Also when the variance is high or we have heavy-tailed distributions, the softmax of λy_i goes further from the uniform vector and more shrinkage is applied.

E ROBUSTNESS OF THE LSE ESTIMATOR: ESTIMATED PROPENSITY SCORES

In this section, we study the robustness of the LSE estimator with respect to estimated (noisy) propensity scores.

To model the estimated propensity scores, we consider $\widehat{\pi}_0(a|x)$ as the noisy version of the logging policy $\pi_0(a|x)$. Similarly, we define $\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S}, \pi_\theta)$ for the LSE estimator on the noisy data samples $\widehat{S}$, with estimated propensity scores. In this section, we made the following definitions.

Definition E.1 (Discrepancy metric)**.** We define the general discrepancy metric between $\widehat{w}_\theta(A, X)R$ and $w_\theta(A, X)R$ with bounded $1+\epsilon$-th moment as,

$$d_{\pi_0}(\widehat{w}_\theta(A, X)R, w_\theta(A, X)R) := \mathbb{E}\big[\big(\widehat{w}_\theta(A, X) - w_\theta(A, X)\big)R\big]. \tag{67}$$

Definition E.2. The log-sum error of the noisy (or estimated) propensity score $\widehat{\pi}_0(a|x)$ is defined as

$$\Delta_{\pi_\theta}(\widehat{\pi}_0, \pi_0) = \frac{1}{\lambda}\log \mathbb{E}_{P_1}[\exp(\lambda \widehat{w}_\theta(A, X)R)] - \frac{1}{\lambda}\log \mathbb{E}_{P_1}[\exp(\lambda w_\theta(A, X)R)]. \tag{68}$$

where $\widehat{w}_\theta(A, X) = \frac{\pi_\theta(A|X)}{\widehat{\pi}_0(A|X)}$ and where $P_1 = P_X \otimes \pi_0(A|X) \otimes P_{R|X,A}$.

Definition E.2 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. Similarly, 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 & Huber, 2009; Christmann & Steinwart, 2004).

We made the following assumption on estimated propensity scores.

Assumption 5 (Bounded moment under noise)**.** The reward function $r(A, X)$ and P_X are such that for all learning policy $\pi_\theta(A|X) \in \Pi_\theta$, the moment of weighted reward is bounded under estimated propensity score scenario, $\mathbb{E}_{P_X \otimes \pi_0(A|X) \otimes P_{R|X,A}}[(\widehat{w}_\theta(A, X)R)^{1+\epsilon}] \leq \widehat{\nu}$.

Remark E.3. Under Assumption 5 and Assumption 1 and using Lemma B.9, it can be shown that the discrepancy metric in Definition E.1 is bounded,

$$-\nu^{1/(1+\epsilon)} \leq d_{\pi_0}(\widehat{w}_\theta(A, X)R, w_\theta(A, X)R) \leq \widehat{\nu}^{1/(1+\epsilon)}. \tag{69}$$

We define the achieved policy under the estimated propensity scores as

$$\pi_{\widehat{\theta}}(S) := \underset{\pi_\theta \in \Pi_\Theta}{\arg\max}\, \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S}, \pi_\theta).$$

In order to derive an upper bound on the regret under noisy propensity score, the following results are needed.

Proposition E.4. *Given Assumption 1 and Assumption 5, the following upper and lower bound hold on $\Delta_{\pi_\theta}(\widehat{\pi}_0, \pi_0)$,*

$$d_{\pi_0}(w_\theta(A, X)R, \widehat{w}_\theta(A, X)R) - \frac{|\lambda|^\epsilon \widehat{\nu}}{1+\epsilon} \leq \Delta_{\pi_\theta}(\widehat{\pi}_0, \pi_0),$$
$$\textit{and, } \Delta_{\pi_\theta}(\widehat{\pi}_0, \pi_0) \leq \frac{|\lambda|^\epsilon \nu}{1+\epsilon} + d_{\pi_0}(\widehat{w}_\theta(A, X)R, w_\theta(A, X)R).$$

Proof. It follows directly from applying Lemma B.10 to $\frac{1}{\lambda}\log \mathbb{E}_{P_1}[\exp(\lambda \widehat{w}_\theta(A, X)R)]$ and $\frac{1}{\lambda}\log \mathbb{E}_{P_1}[\exp(\lambda w_\theta(A, X)R)]$ and combining the lower and upper bounds. Then, we have,

$$\mathbb{E}\big[\big(w_\theta(A, X) - \widehat{w}_\theta(A, X)\big)R\big] - \frac{|\lambda|^\epsilon \widehat{\nu}}{1+\epsilon} \leq \Delta_{\pi_\theta}(\widehat{\pi}_0, \pi_0) \leq \frac{|\lambda|^\epsilon \nu}{1+\epsilon} + \mathbb{E}\big[\big(\widehat{w}_\theta(A, X) - w_\theta(A, X)\big)R\big].$$

□

Proposition E.5. *Given Assumption 5, and assuming* $n > \frac{\frac{4}{3}\mu_{\min}+4}{\mu_{\min}^2}\log\frac{4}{\delta}$ *where* $\mu_{\min} = \min\left(e^{\lambda\nu^{1/(1+\epsilon)}}, e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}\right)$*, then with probability at least* $(1-\delta)$ *for a fixed* $\pi_\theta \in \Pi_\theta$*, we have,*

$$\left|\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S},\pi_\theta) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta) - \Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)\right| \leq \frac{2\upsilon(\delta)}{\lambda}\Big(\frac{1}{e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}} + \frac{1}{e^{\lambda\nu^{1/(1+\epsilon)}}}\Big),$$

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

Proof. Set $Y_\theta(A,X) = w_\theta(A,X)R$, $\widehat{Y}_\theta(A,X) = \widehat{w}_\theta(A,X)r(A,X)$, $u_i = \frac{1}{\lambda}\left(e^{\widehat{y}_i} - e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mu\right)$ 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 $-\frac{\mu}{\lambda} \leq v_i \leq \frac{1}{\lambda} - \frac{\mu}{\lambda}$ and $-\frac{e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mu}{\lambda} \leq u_i \leq \frac{1}{\lambda} - \frac{e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mu}{\lambda}$. Then, using the one-sided Bernstein's inequality (Lemma B.4), and changing variables (Lemma B.5), we have:

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

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

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n e^{\lambda\widehat{y}_i} - e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] > \frac{(1-e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mu)\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda\widehat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \leq \delta,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n e^{\lambda\widehat{y}_i} - e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] < -\frac{e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mu\log\frac{1}{\delta}}{3n} - \sqrt{\frac{\mathbb{V}\left(e^{\lambda\widehat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}\right) \leq \delta.$$

Therefore, with probability at least $1-2\delta$, for $\upsilon_2 < \frac{1}{2}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]$, we have,

$$\begin{aligned}
&\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S},\pi_\theta) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)\\
&= \frac{1}{\lambda}\log\left(\frac{\sum_{i=1}^n e^{\lambda\widehat{y}_i}}{\sum_{i=1}^n e^{\lambda y_\theta(a_i,x_i)}}\right)\\
&\leq \frac{1}{\lambda}\log\left(\frac{e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \upsilon_1}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \upsilon_2}\right)\\
&= \frac{1}{\lambda}\left(\log\left(e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}] + \upsilon_1\right) - \log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \upsilon_2\right)\right)\\
&\leq \frac{1}{\lambda}\Bigg(\log\left(e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) + \frac{\upsilon_1}{e^{\lambda\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)}\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\\
&\qquad - \left(\log\left(\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\right) - \frac{\upsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}] - \upsilon_2}\right)\Bigg)\\
&\leq \Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0) + \frac{1}{\lambda}\left(\frac{\upsilon_1}{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]} + \frac{2\upsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right)\\
&\leq \Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0) + \frac{2}{\lambda}\left(\frac{\upsilon_1}{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]} + \frac{\upsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right).
\end{aligned}$$

where

$$\upsilon_1 = \frac{(1-\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}])\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda\widehat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}},$$
$$\upsilon_2 = \frac{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}.$$

Similarly, with probability at least $1-2\delta$ we have, given $\upsilon_3 < \frac{1}{2}\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]$,

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_\theta) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta) \geq \Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0) - \frac{2}{\lambda}\left(\frac{\upsilon_3}{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]} + \frac{\upsilon_4}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right),$$

where,

$$\upsilon_3 = \frac{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda\widehat{Y}_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}},$$
$$\upsilon_4 = \frac{(1-\mathbb{E}[e^{\lambda Y_\theta(A,X)}])\log\frac{1}{\delta}}{3n} + \sqrt{\frac{\mathbb{V}\left(e^{\lambda Y_\theta(A,X)}\right)\log\frac{1}{\delta}}{n}}.$$

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

$$\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0) - \frac{2}{\lambda}\left(\frac{\upsilon_3}{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]} + \frac{\upsilon_4}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right) \leq \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\hat{S},\pi_\theta) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta)$$
$$\leq \Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0) + \frac{2}{\lambda}\left(\frac{\upsilon_1}{\mathbb{E}[e^{\lambda\widehat{Y}_\theta(A,X)}]} + \frac{\upsilon_2}{\mathbb{E}[e^{\lambda Y_\theta(A,X)}]}\right).$$

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

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

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

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

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

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

□

In the following theorem, we study the regret of the LSE estimator under $\pi_{\widetilde{\theta}}(S)$ policy.

Theorem E.6. *Suppose that,*

$$\pi_{\widetilde{\theta}}(\widehat{S}) = \arg\max_{\pi_\theta\in\Pi_\Theta} \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_\theta),$$

where $\widehat{S}$ is the data with noisy propensity scores. Given Assumption 1, and 5, and assuming that $n \geq \max\left(\frac{\frac{4}{3}\mu_{\min}+4}{\mu_{\min}^2}\log\frac{4|\Pi_\theta|}{\delta}, \frac{\left(2|\lambda|^{1+\epsilon}\nu+\frac{4}{3}\gamma\right)\log\frac{4|\Pi_\theta|}{\delta}}{\gamma^2\exp(2\lambda\nu^{1/(1+\epsilon)})}\right)$ where $\mu_{\min} = \min\left(e^{\lambda\nu^{1/(1+\epsilon)}}, e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}\right)$, then there exists $\gamma\in(0,1)$ such that the following upper bound holds on the regret of the LSE estimator under $\pi_{\widetilde{\theta}}(S)$ with probability at least $1-\delta$ for $\delta\in(0,1)$,

$$\begin{aligned}\mathfrak{R}_\lambda(\pi_{\widetilde{\theta}},S) \leq &\frac{2|\lambda|^\epsilon}{1+\epsilon}\nu + \frac{|\lambda|^\epsilon}{1+\epsilon}\widehat{\nu}\\ &- \frac{4(2-\gamma)}{3(1-\gamma)}\frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda\exp(\lambda\nu^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}}\\ &+ d_{\pi_0}(\widehat{w}_{\widehat{\theta}}(A,X)R, w_{\widehat{\theta}}(A,X)R) + d_{\pi_0}(\widehat{w}_{\widetilde{\theta}}(A,X)R, w_{\widetilde{\theta}}(A,X)R)\\ &+ \frac{4\upsilon(\frac{\delta}{4|\Pi_\theta|})}{\lambda}\Big(\frac{1}{e^{\lambda\nu^{1/(1+\epsilon)}}} + \frac{1}{e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}}\Big),\end{aligned} \tag{70}$$

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

Proof. Let $\widehat{\theta}$ be,

$$\pi_{\widehat{\theta}}(S) = \arg\max_{\pi_\theta\in\Pi_\Theta} \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_\theta).$$

We decompose the regret as follows,

$$\begin{aligned}&\mathfrak{R}_\lambda(\pi_{\widetilde{\theta}},S)\\ &= V(\pi_{\theta^*}) - V(\pi_{\widetilde{\theta}})\\ &= \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widetilde{\theta}}) - V(\pi_{\widetilde{\theta}})\\ &\quad - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widetilde{\theta}}) + \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widetilde{\theta}})\\ &\quad - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widetilde{\theta}}) + \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widehat{\theta}})\\ &\quad - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widehat{\theta}}) + \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}})\\ &\quad - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}}) + \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\theta^*})\\ &\quad - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\theta^*}) + V(\pi_{\theta^*}).\end{aligned}$$

Using the generalization error bounds at Theorem D.2 and Theorem D.1 and using the union bound, with probability $(1-\delta)$ we have,

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widetilde{\theta}}) - V(\pi_{\widetilde{\theta}}) \leq -\frac{1}{\lambda(1-\gamma)}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3(1-\gamma)\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}, \tag{71}$$

$$V(\pi_{\theta^*}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\theta^*}) \leq \frac{1}{1+\epsilon}|\lambda|^\epsilon\nu - \frac{1}{\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu\log(2|\Pi_\theta|/\delta)}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}} - \frac{4\log(2|\Pi_\theta|/\delta)}{3\lambda\exp(\lambda\nu^{1/(1+\epsilon)})n}. \tag{72}$$

In addition, using Proposition E.5, we have,

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widetilde{\theta}}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widetilde{\theta}}) \leq \Delta_{\pi_{\widetilde{\theta}}}(\widehat{\pi}_0,\pi_0) + \frac{2\upsilon(\delta/|\Pi_\theta|)}{\lambda}\left(\frac{1}{e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}} + \frac{1}{e^{\lambda\nu^{1/(1+\epsilon)}}}\right), \tag{73}$$

$$\widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(S,\pi_{\widehat{\theta}}) - \widehat{\mathrm{V}}^\lambda_{\mathrm{LSE}}(\widehat{S},\pi_{\widehat{\theta}}) \leq \Delta_{\pi_{\widehat{\theta}}}(\widehat{\pi}_0,\pi_0) + \frac{2\upsilon(\delta/|\Pi_\theta|)}{\lambda}\left(\frac{1}{e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}} + \frac{1}{e^{\lambda\nu^{1/(1+\epsilon)}}}\right). \tag{74}$$

As $\pi_{\widetilde{\theta}}$ is the maximizer of $\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S}, \pi_\theta)$, we have,

$$\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S}, \pi_{\widehat{\theta}}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S}, \pi_{\widetilde{\theta}}) \leq 0, \tag{75}$$

and as $\pi_{\widehat{\theta}}$ is the maximizer of $\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_\theta)$ we have,

$$\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\theta^*}) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S, \pi_{\widehat{\theta}}) \leq 0. \tag{76}$$

So putting all together, using the union bound we have with probability at least $1-\delta$,

$$\begin{aligned} V(\pi_{\widetilde{\theta}}) - V(\pi_{\theta^*}) \leq{}& \frac{|\lambda|^{\epsilon}}{1+\epsilon}\nu - \frac{4(2-\gamma)}{3(1-\gamma)} \frac{\log\frac{4|\Pi_\theta|}{\delta}}{n\lambda \exp(\lambda \nu^{1/(1+\epsilon)})} - \frac{(2-\gamma)}{(1-\gamma)\lambda}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log\frac{4|\Pi_\theta|}{\delta}}{n\exp(2\lambda\nu^{1/(1+\epsilon)})}}. \\ &+ \Delta_{\pi_{\widehat{\theta}}}(\widehat{\pi}_0, \pi_0) - \Delta_{\pi_{\widetilde{\theta}}}(\widehat{\pi}_0, \pi_0) \\ &+ \frac{2\upsilon(\frac{\delta}{4|\Pi_\theta|})}{\lambda}\Big(\frac{1}{e^{\lambda\nu^{1/(1+\epsilon)}}} + \frac{1}{e^{\lambda\widehat{\nu}^{1/(1+\epsilon)}}}\Big), \end{aligned}$$

where $\upsilon\big(\frac{\delta}{4|\Pi_\theta|}\big) = \frac{\log\left(\frac{16\Pi_\theta}{\delta}\right)}{3n} + \sqrt{\frac{\log\left(\frac{16\Pi_\theta}{\delta}\right)}{n}}$. The final result holds by applying Proposition E.4 to $\Delta_{\pi_{\widehat{\theta}}}(\widehat{\pi}_0, \pi_0) - \Delta_{\pi_{\widetilde{\theta}}}(\widehat{\pi}_0, \pi_0)$. □

Discussion: The term $d_{\pi_0}(\widehat{w}_{\widehat{\theta}}(A,X)R, w_{\widehat{\theta}}(A,X)R) + d_{\pi_0}(\widehat{w}_{\widetilde{\theta}}(A,X)R, w_{\widetilde{\theta}}(A,X)R)$ in equation 70 can be interpreted as the cost of estimated propensity scores which is independent from n. Similar to Remark 5.4, we have the convergence rate of $O(n^{-\epsilon/(1+\epsilon)})$ for all remaining terms in equation 70.

In the following Corollary, we discuss that the small range of variation of the noise gives an upper bound on the variance of the LSE estimator under estimated propensity score.

Corollary E.7. *Under the same assumptions in Proposition E.5, then the following upper bound holds on the variance of the LSE estimator under estimated propensity scores with probability at least* $(1-\delta)$*,*

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

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

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

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

Therefore,

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

□

From Corollary E.7, we have an upper bound on the variance of the LSE estimator under estimated propensity scores, in terms of the variance of the LSE estimator under true propensity scores. Therefore, if $\mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta))$ is bounded, then we expect bounded $\mathbb{V}(\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S},\pi_\theta))$.

E.1 Gamma Noise Discussion

For statistical modeling of the estimated propensity scores, as discussed in (Zhang et al., 2023b), suppose that the logging policy is a softmax policy with respect to a.

$$\pi_0(A|X) = \mathrm{softmax}(f_{\theta^*}(X,A)), \tag{77}$$

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

We have an estimation of the function $f_{\theta^*}(X,A)$, as $f_{\widehat{\theta}}(X,A)$ and we model the error in the estimation of $f_{\theta^*}(X,A)$ as a random variable Z which is a function of X and A,

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

Then we have,

$$\begin{aligned}\widehat{\pi}_0 &= \mathrm{softmax}(f_{\widehat{\theta}}(X,A))\\ &= \mathrm{softmax}(f_{\theta^*} + Z)\\ &\propto e^{Z}\pi_0.\end{aligned}$$

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.

In particular, 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,

$$\widehat{\pi}_0 = \frac{1}{U}\pi_0 \to \widehat{w}_\theta(A,X) = Uw_\theta(A,X).$$

which results in a multiplicative gamma noise on the importance-weighted reward. 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.

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

Note that we have

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

Therefore, $\mathbb{E}[e^{\lambda U w_\theta(A,X)R}]$ converges to $\mathbb{E}[e^{\lambda w_\theta(A,X)R}]$ for $b\to\infty$. Furthermore, we assume that for a large value b, $\Delta_{\pi_\theta}(\widehat{\pi}_0,\pi_0)\approx 0$ and using Proposition E.5, with a probability at least $1-\delta$, we have,

$$\left|\widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(\widehat{S},\pi_\theta) - \widehat{\mathrm{V}}^{\lambda}_{\mathrm{LSE}}(S,\pi_\theta)\right| \le \epsilon\Big(\frac{1}{\mathbb{E}[e^{\lambda\widehat{w}_\theta(A,X)R}]} + \frac{1}{\mathbb{E}[e^{\lambda w_\theta(A,X)R}]}\Big). \tag{78}$$

The impact of inverse Gamma noise on the LSE estimator is constrained when the noise's domain is sufficiently small. This property ensures that the LSE remains relatively stable under certain noise conditions. Furthermore, we can reduce the deviation from the original noiseless LSE by increasing the size of the Logged Bandit Feedback (LBF) dataset. This relationship demonstrates the estimator's robustness and scalability in practical applications.

Table 7: Statistics of the datasets used in our experiments. For image datasets the 2048-dimensional features from pretrained ResNet-50 are used.

DATA SET	IPS-TRAINING SAMPLES	TEST SAMPLES	NUMBER OF ACTIONS	DIMENSION
FMNIST	$60,000$	10000	10	2048
EMNIST	$60,000$	10000	10	2048
KUAIREC	12,530,806	4,676,570	10,728	1555

F EXPERIMENT DETAILS

Datasets: In addition to dataset EMNIST, we also run our estimator over Fashion-MNIST (FMNIST) (Xiao et al., 2017).

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. Furthermore, we use early stopping in our training phase and the maximum number of epochs is 300. For the image datasets, EMNIST and FMNIST, we use the last layer features from ResNet-50 model pretrained on the ImageNet dataset (Deng et al., 2009).

F.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.

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, 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 data-driven approach proposed in (Metelli et al., 2021). 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 $\gamma \in \{0.01, 0.1, 1, 10, 100\}$.

F.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.
- 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.
  - **data.py** code to load data for image datasets.
  - **eval.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.
  - **train_logging_policy.py**: Script to train the logging policy.

  - **train_reward_estimator.py**: Script to train the reward estimator for DM and DR methods.
  - **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.
  - **synthetic_experiment_v3.py**: Code for synthetic experiments.
  - **motivation.ipynb**: Code for motivating example.
  - **OPE_classification**: The codes for the OPE experiments on real-world datasets from UCI repository.
    * **train_on_uci.ipynb**: Main code running experiments on UCI datasets.
    * **faulty_policy.py**: The code for the faulty policy model for the logging and training polices.
    * **UCI**: The folder containing UCI datasets used in the experiments.
- The **real_world** folder contains the scripts and codes written for Kuai-Rec dataset.
  - **preprocess_data.ipynb**: The code that preprocess the KuaiRec dataset and makes it ready for training.
  - **run_kuairec_experiments.py**: The main code for real dataset experiments. It contains the training of the logging policy as well as the learning policy
  - **eval.py**: Code containing the implementation of the evaluation metrics.

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. For OPE synthetic experiments, the code *synthetic_experiment_v3.py* should be run. For real-world OPL experiments, the Kuairec (version 2) dataset should be downloaded and put in *real_world/KuaiRec 2.0/* folder and first *real_world/preprocess_data.ipynb* notebook should be run and then *real_world/run_kuairec_experiments.py* code will train the estimators on Kuairec dataset. The code itself trains and stores a logging policy before the main training phase. For OPE real-world experiments, the notebook *OPE_classification/train_on_uci.ipynb* would train the estimators on the UCI datasets in the folder *OPE_classification/UCI*.

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.

G Additional Experiments

This section presents supplementary experiments to further validate our LSE approach in off-policy learning and evaluation. We extend our experiments as follows:

1. Comparison with the Model-based estimators: We conduct a series of experiments to assess the performance of model-based estimators in comparison with our LSE estimator.
2. Combined method: We investigate the efficacy of combining the LSE estimator with the Doubly Robust (DR) estimator, exploring potential synergies between these methods.
3. Real-world application: To demonstrate the practical relevance of our approach, we apply our methods to a real-world dataset, providing insights into their performance under real world datasets in off-policy learning scenarios.
4. λ Effect: We study the effect of λ in different scenarios.
5. Sample number effect: We study the performance of LSE estimator with different number of samples n.
6. Off-policy evaluation: We conduct more off-policy evaluation using Lomax distribution.
7. Off-policy learning: We run more experiments for off-policy learning scenario under FM-NIST dataset.
8. Selection of λ: Different methods of the selection of λ, data-driven selection of λ and sensitivity of λ are explored.
9. Distributional properties: In OPE scenario under heavy-tailed assumption, the distributional properties of LSE are studied.
10. Comparison with LS estimator: More Comparison with LS estimator in OPE setting based on choosing λ is provided.

These additional experiments aim to provide a comprehensive evaluation of our proposed LSE estimator.

G.1 Off-policy evaluation experiment

We conduct synthetic experiments to test our model's performance and behavior compared to other models and the effectiveness of our approach in the case of heavy-tailed rewards. We have two different settings. Gaussian setting in which the distributions are Gaussian random variables, having exponential tails, and Lomax setting in which the distributions are Lomax random variables, with polynomial tails. In all experiments we run 10K trials to estimate the bias, variance and MSE of each method, given MSE as the main criteria to compare the performance of different approaches. We conduct experiments on our method (LSE), power-mean estimator (PM) (Metelli et al., 2021), exponential smoothing (ES) (Aouali et al., 2023), IX estimator (Gabbianelli et al., 2023), truncated IPS (IPS-TR) (Ionides, 2008b), self-normalized IPS (SNIPS) (Swaminathan & Joachims, 2015b), OS estimator (Su et al., 2020) and LS estimator (Sakhi et al., 2024). The number of samples changes in different settings. In each setting, we grid search the hyperparameter of each method with 5 different values and select the one that leads to the least estimated MSE value. Note that the hyperparameter for each method is selected independently in each setting, but the candidate values are fixed throughout all settings.

Gaussian: In this setting, as explained in section 6, we have $\pi_\theta(\cdot|x_0) \sim \mathcal{N}(\mu_1, \sigma^2), \pi_0(\cdot|x_0) \sim \mathcal{N}(\mu_2, \sigma^2)$ and $r(x_0, u) = -e^{\alpha u^2}$. Given $2\alpha\sigma^2 < 1$, with simple calculations we have,

$$\mathbb{E}_{\pi_\theta}[r] = -\frac{1}{\sqrt{1-2\alpha\sigma^2}} \exp\left(\frac{\alpha\mu_1^2}{1-2\alpha\sigma^2}\right) \tag{79}$$

$$\mathbb{E}_{\pi_0}\left[\left|\frac{\pi_\theta}{\pi_0} r\right|^{1+\epsilon}\right] = |\mathbb{E}_{\pi_\theta}[r]| \exp\left(\frac{\epsilon(\mu_1-\mu_2)((1+\epsilon+2\alpha\sigma^2)\mu_1 - (1+\epsilon-2\alpha\sigma^2)\mu_2)}{2\sigma^2(1-2\alpha\sigma^2)}\right) \tag{80}$$

We fix $\mu_1 = 0.5, \mu_2 = 1, \sigma^2 = 0.25$, but we change α as it increases the $1+\epsilon$-moment of the weighted reward variable as it tends to $\frac{1}{2\sigma^2}$ and (given $\mu_1 > 0, \epsilon \leq \frac{\mu_1}{|\mu_1-\mu_2|}$ or $\mu_1 > \mu_2$) leads to

unbounded $1+\epsilon$-moment for $\alpha = \frac{1}{2\sigma^2}$. We report the experiment results in Tables 8 and 9 As we can observe that LSE effectively keeps the variance low without significant side-effects on bias, leading to an overall low MSE, making it a viable choice with general unbounded reward functions.
We also try different values for the number of samples, and observe the methods capability to work well on small number of samples and their performance growth with the number of samples. For $\alpha = 1.4$, the results of different methods for $n = 100, 1K, 10K, 100K$ are illustrated in Table 10.

Discussion: We observe that either in small sample size or large sample size, LSE beats other methods with significant gap. Inspecting the bias of LSE though different sample sizes, the bias becomes fixed and doesn't decrease as the number of samples in the LBF dataset goes beyond 1K. This is due to the fixed candidate set for the parameter λ in LSE and the presence of λ in our derived bias upper bound in Proposition 5.5. This shows that the dependence of the bias on λ that appears in the bias upper bound is tight and with a fixed λ, the bias doesn't vanish, no matter how much data we have and for large number of samples it is critical to select λ as a function of n. Furthermore, we can see that the variance of LSE effectively decreases as the number of samples increase. Here we can observe the decrease rate of $1/n$ in the variance, as it is proved in Proposition 5.7 under bounded second moment assumption. We also observe that as α increases and the reward function's growth becomes bigger PM, IPS-TR, SNIPS, and OS suffer from a very large variance, while ES, LSE, IX, and LS-LIN manage to keep the variance relatively low. Among these low-variance methods, LSE achieves the lowest bias, indicating a better bias-variance trade-off. Also, LS-LIN achieves the lowest variance among all methods. We hypothesize that is is do to the fact that LS-LIN, along LSE, is the only method that is not linear w.r.t. reward and compresses the reward besides the importance weight.

Lomax: In the Lomax setting, we use Lomax distributions with scale 1 for the learning and logging policies, $\pi_\theta(u|x_0) \sim \frac{\alpha}{(u+1)^{\alpha+1}}, \pi_0(u|x_0) \sim \frac{\alpha}{(u+1)^{\alpha'+1}}, \alpha, \alpha' > 0$. We use a polynomial function for the reward, $r(u) = (1+u)^\beta, \beta > 0$. The main difference in this setting compared to Gaussian setting is that here the tails of the distributions are polynomial, in contrast to the Gaussian setting in which the tails are exponential. In this setting, for $\alpha > \beta$, we have,

$$\mathbb{E}_{\pi_\theta}[r] = \frac{\alpha}{\alpha - \beta}$$
$$\mathbb{E}_{\pi_0}\left[\left|\frac{\pi_\theta}{\pi_0} r\right|^{1+\epsilon}\right] = \left(\frac{\alpha}{\alpha-\beta}\right)^{1+\epsilon} k^{-\epsilon}(1+\epsilon(1-k))^{-1}$$

where $k = \frac{\alpha'}{\alpha-\beta}$ and for the second inequality to hold we should have $1+\epsilon(1-k) > 0$. The condition $\alpha > \beta$ is sufficient for the weighted reward function to be ϵ-heavy-tailed for some $\epsilon > 0$ (either $k < 1$ or $\epsilon < \frac{1}{|1-k|}$. We change the value of β to $0.5, 1, 2$. We also fix $\alpha - \beta = 0.5$, to keep the value function in an appropriate range. We change k to get different values for $\alpha' = k(\alpha - \beta)$ which determines the tail of the weighted reward variable. We set $k = 2, 3, 4$. The results are shown in Tables 11 and 12. We observe the superior performance of LSE compared to other methods.

Discussion: In Lomax experiments the LSE estimator has the best performance in most of settings. In two settings, i.e., $\beta = 0.5$ and $\alpha' \in \{1.5, 2.0\}$, IPS-TR does better than LSE with a very small margin, yet LSE is the second best model in these two settings. Similar to the Gaussian setting, we also run the experiments for different numbers of samples to inspect the effect of the number of samples on the performance of the models. We fix $\alpha = 2.5$, $\beta = 2$ and $\alpha' = 1.5$ in this scenario. Table 13 reports the performance of LSE across different number of samples. The same conclusions as the Gaussian setting are also observable in the Lomax setting. We can observe that LSE has better performance for $n = 100, 10K, 100K$.

Table 8: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Gaussian distributions for $\alpha = 1.0, 1.1, 1.2, 1.3$. The experiment was run 10000 times and the variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

α	Estimator	Bias	Variance	MSE
1.0	PM	0.037	0.004	0.006
	ES	−0.001	0.006	0.006
	LSE	0.021	0.003	**0.003**
	IPS-TR	0.019	0.004	0.004
	IX	0.168	0.001	0.029
	SNIPS	−0.003	0.008	0.008
	LS-LIN	0.151	0.001	0.024
	LS	0.006	0.005	0.005
	OS	0.505	0.005	0.260
1.1	PM	0.004	0.063	0.063
	ES	−0.001	0.054	0.054
	LSE	0.052	0.006	**0.009**
	IPS-TR	0.020	0.052	0.052
	IX	0.237	0.002	0.058
	SNIPS	−0.005	0.059	0.059
	LS-LIN	0.284	0.001	0.082
	LS	0.082	0.007	0.0135
	OS	0.521	0.020	0.292
1.2	PM	−0.043	0.435	0.437
	ES	0.000	0.357	0.357
	LSE	0.152	0.014	**0.037**
	IPS-TR	0.024	0.353	0.354
	IX	0.373	0.005	0.144
	SNIPS	−0.003	0.366	0.366
	LS-LIN	0.545	0.002	0.299
	LS	0.183	0.016	0.050
	OS	0.541	0.116	0.409
1.3	PM	−0.121	1.731	1.746
	ES	1.162	0.026	1.377
	LSE	0.158	0.124	**0.148**
	IPS-TR	0.030	1.404	1.405
	IX	0.662	0.016	0.453
	SNIPS	−0.000	1.491	1.491
	LS-LIN	1.069	0.003	1.145
	LS	0.155	0164	0.188
	OS	0.463	56.581	56.796

Table 9: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Gaussian distributions for $\alpha = 1.4, 1.5, 1.6, 1.7$. The experiment was run 10000 times and the variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

α	Estimator	Bias	Variance	MSE
1.4	PM	-0.301	164.951	165.041
	ES	1.959	0.396	4.232
	LSE	0.615	0.292	**0.670**
	IPS-TR	0.053	133.688	133.691
	IX	1.340	0.048	1.842
	SNIPS	-0.029	133.520	133.521
	LS-LIN	2.164	0.005	4.687
	LS	0.564	0.458	0.776
	OS	0.623	23.589	23.977
1.5	PM	-0.205	222.003	222.045
	ES	3.850	1.505	16.324
	LSE	2.132	0.645	5.190
	IPS-TR	0.349	179.990	180.112
	IX	3.116	0.153	9.865
	SNIPS	0.315	194.830	194.929
	LS-LIN	4.682	0.009	21.927
	LS	1.968	1.156	**5.028**
	OS	1.096	504.001	505.205
1.6	PM	0.726	5095.725	5096.252
	ES	9.420	22.685	111.416
	LSE	7.541	1.233	58.105
	IPS-TR	1.903	4131.016	4134.636
	IX	8.665	0.502	75.589
	SNIPS	1.860	4426.166	4429.625
	LS-LIN	11.547	0.015	133.349
	LS	7.148	2.595	**53.689**
	OS	3.669	1303.684	1317.146
1.7	PM	9.943	125126.550	125225.418
	ES	38.531	0.301	1484.959
	LSE	32.107	2.244	1033.093
	IPS-TR	12.880	101427.776	101593.680
	IX	32.923	1.802	1085.753
	SNIPS	12.704	102027.853	102189.250
	LS-LIN	38.112	0.024	1452.556
	LS	31.227	5.267	**980.41**
	OS	29.171	17767.954	18618.899

Table 10: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Gaussian distributions setup. The experiment was run 10000 times fixing $\alpha = 1.4$ and different number of samples $n \in \{100, 1000, 10000, 100000\}$. The variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

n	Estimator	Bias	Variance	MSE
100	PM	−0.1288	203.5015	203.5181
	ES	1.9769	1.7696	5.6775
	LSE	1.2210	0.5015	**1.9925**
	IPS-TR	0.1617	164.9972	165.0234
	IX	1.3459	0.4783	2.2897
	SNIPS	0.0074	196.8881	196.8881
	LS-LIN	2.1683	0.0568	4.7585
	LS	1.1817	0.8115	2.2079
	OS	0.7661	10.2588	10.8458
1000	PM	−0.1963	18.3363	18.3749
	ES	1.9587	0.1694	4.0058
	LSE	0.6030	0.2999	**0.6635**
	IPS-TR	0.1007	14.8696	14.8798
	IX	1.3375	0.0486	1.8376
	SNIPS	0.0594	15.0741	15.0776
	LS-LIN	2.1646	0.0056	4.6910
	LS	0.5640	0.4580	0.7761
	OS	0.6432	8.7698	9.1835
10000	PM	−0.2282	10.4458	10.4979
	ES	1.9625	0.0285	3.8800
	LSE	0.6159	0.0296	0.4089
	IPS-TR	0.0464	8.4660	8.4681
	IX	1.3410	0.0048	1.8031
	SNIPS	0.0435	8.5986	8.6005
	LS-LIN	2.1644	0.0005	4.6852
	LS	0.5606	0.0466	**0.3609**
	OS	0.5564	4.8936	5.2032
100000	PM	−0.2505	1.8148	1.8775
	ES	0.0246	1.4707	1.4713
	LSE	0.6160	0.0029	0.3823
	IPS-TR	0.0250	1.4706	1.4712
	IX	1.3408	0.0005	1.7982
	SNIPS	0.0246	1.4757	1.4763
	LS-LIN	2.1629	5.6014	4.6783
	LS	0.5584	0.0049	**0.3167**
	OS	0.5823	0.8251	1.1643

Table 11: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Lomax distributions setup for $\beta = 1.0, 1.5$. The experiment was run 10000 times with different values of α, α' and β. The variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

β	α	α'	Method	Bias	Variance	MSE
0.5	1.0	1.0	PM	−0.0004	0.0197	0.0197
			ES	−0.0004	0.0197	0.0197
			LSE	0.0361	0.0047	0.0060
			IPS-TR	−0.0004	0.0197	0.0197
			IX	0.6958	0.0001	0.4842
			SNIPS	−0.0004	0.0197	0.0197
			LS-LIN	0.4475	0.0002	0.2005
			LS	0.0266	0.0046	**0.0053**
			OS	0.3332	0.0094	0.1204
		1.5	PM	0.2191	0.0154	0.0634
			ES	0.0145	0.2011	0.2013
			LSE	0.1702	0.0117	0.0407
			IPS-TR	0.1341	0.0146	0.0326
			IX	0.7815	0.0003	0.6111
			SNIPS	0.0181	0.1668	0.1671
			LS-LIN	0.5303	0.0011	0.2822
			LS	0.0697	0.0346	**0.0395**
			OS	0.7636	0.0007	0.5838
		2.0	PM	0.4784	0.0084	0.2372
			ES	0.9554	0.0020	0.9147
			LSE	0.1586	0.0801	0.1052
			IPS-TR	0.2965	0.0171	**0.1050**
			IX	0.8641	0.0006	0.7472
			SNIPS	0.0580	1.1500	1.1533
			LS-LIN	0.6106	0.0023	0.3751
			LS	0.3086	0.0238	0.1190
			OS	1.0176	0.0003	1.0358
1	1.5	1.0	PM	−0.0823	0.0440	0.0508
			ES	0.0006	0.0357	0.0357
			LSE	0.0731	0.0092	0.0146
			IPS-TR	0.0006	0.0357	0.0357
			IX	1.0438	0.0002	1.0897
			SNIPS	−0.0003	0.0418	0.0418
			LS-LIN	0.8513	0.0004	0.7252
			LS	0.0429	0.0104	**0.0122**
			OS	0.3566	0.0364	0.1635
		1.5	PM	0.0167	0.7885	0.7888
			ES	0.0167	0.7885	0.7888
			LSE	0.1122	0.0820	0.0946
			IPS-TR	0.0167	0.7885	0.7888
			IX	1.1723	0.0006	1.3749
			SNIPS	0.0167	0.7885	0.7888
			LS-LIN	0.9551	0.0014	0.9136
			LS	0.1183	0.0717	**0.0857**
			OS	0.5122	0.6815	0.9439
		2.0	PM	0.3839	0.3198	0.4672
			ES	1.4337	0.0035	2.0589
			LSE	0.2731	0.1353	**0.2099**
			IPS-TR	0.2280	0.2424	0.2944
			IX	1.2957	0.0013	1.6801
			SNIPS	0.0614	2.3202	2.3239
			LS-LIN	1.0580	0.0030	1.1223
			LS	0.2548	0.1785	0.2434
			OS	1.2544	0.0059	1.5793

Table 12: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Lomax distributions setup for $\beta = 2$. The experiment was run 10000 times with different values of α, α' and β. The variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

β	α	α'	Method	Bias	Variance	MSE
2	2.5	1.0	PM	-0.2267	0.1913	0.2427
			ES	-0.0049	0.1540	0.1540
			LSE	0.0304	0.0461	0.0471
			IPS-TR	-0.0049	0.1540	0.1540
			IX	1.7392	0.0007	3.0256
			SNIPS	-0.0100	0.1858	0.1859
			LS-LIN	1.9231	0.0011	3.6995
			LS	0.0819	0.0281	**0.0348**
			OS	0.5571	0.0849	0.3953
		1.5	PM	-0.2510	17.7398	17.8028
			ES	2.2891	0.0024	5.2425
			LSE	0.2266	0.1688	**0.2201**
			IPS-TR	-0.0042	14.3693	14.3694
			IX	1.9546	0.0018	3.8224
			SNIPS	-0.0062	14.4548	14.4549
			LS-LIN	2.0374	0.0016	4.1529
			LS	0.2330	0.1699	0.2242
			OS	0.3995	13.5957	13.7553
		2.0	PM	-0.2114	27.6307	27.6754
			ES	2.3886	0.0113	5.7167
			LSE	0.5334	0.2729	**0.5574**
			IPS-TR	-0.0086	22.5415	22.5416
			IX	2.1606	0.0035	4.6717
			SNIPS	-0.0107	22.6954	22.6955
			LS-LIN	2.1601	0.0034	4.6694
			LS	0.4946	0.3696	0.61424
			OS	0.5158	7.4515	7.7175

Table 13: Bias, variance, and MSE of LSE, ES, PM, IX, IPS-TR, SNIPS, LS-LIN, and OS estimators with Lomax distributions setup. The experiment is conducted for 10000 times and different number of samples $n \in \{100, 1000, 10000, 100000\}$. The variance, bias, and MSE of the estimations are reported. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

n	Estimator	Bias	Variance	MSE
100	PM	-0.2486	75.480	75.542
	ES	2.2895	0.0244	5.2663
	LSE	0.6217	0.4035	**0.7900**
	IPS-TR	0.0021	61.140	61.140
	IX	1.9546	0.0182	3.8388
	SNIPS	-0.0331	67.583	67.583
	LS-LIN	2.0369	0.0168	4.1660
	LS	0.6339	0.5402	0.9421
	OS	0.4287	61.159	61.343
1000	PM	-0.2421	10.960	11.019
	ES	2.2889	0.0024	5.2415
	LSE	0.2245	0.1702	0.2206
	IPS-TR	0.0037	8.8781	8.8780
	IX	1.9540	0.0018	3.8198
	SNIPS	0.0010	9.0742	9.0742
	LS-LIN	2.0375	0.0016	4.1531
	LS	0.2330	0.1699	**0.2242**
	OS	0.4345	8.8799	9.0687
10000	PM	-0.2317	0.6596	0.7132
	ES	0.0131	0.5343	0.5345
	LSE	0.2253	0.0171	**0.0679**
	IPS-TR	0.0131	0.5342	0.5345
	IX	1.9539	0.0002	3.8180
	SNIPS	0.0133	0.5364	0.5366
	LS-LIN	2.0375	0.0002	4.1517
	LS	0.2338	0.0171	0.0717
	OS	0.4438	0.5345	0.7315
100000	PM	-0.2619	0.6546	0.7232
	ES	-0.0140	0.5302	0.5304
	LSE	0.2267	0.0019	**0.0533**
	IPS-TR	-0.0140	0.5302	0.5304
	IX	1.9538	1.6977	3.8175
	SNIPS	-0.0137	0.5284	0.5286
	LS-LIN	2.0374	1.6805	4.1509
	LS	0.2351	0.0019	0.0572
	OS	0.4166	0.5302	0.7038

G.2 OFF-POLICY LEARNING EXPERIMENT

We present the results of our experiments for EMNIST and FMNIST in Table 15.

As we can observe in the results for different scenarios and datasets, our estimator, shows dominant performance among other baselines. The details of the number of best-performing and second rank estimator is provided in Table 14. We observe that in 21 out of 30 experiments, the LSE estimator outperforms other estimators. Additionally, it ranks second in 7 of the remaining 9 experiments.

Table 14: Comparison of different estimators in terms of the number of best|second rank performances of all true propensity score/ reward , estimated (noisy) propensity scores and noisy reward experiment setups in OPL scenario.

Estimator	True PS & Reward	Noisy PS	Noisy Reward	Total
LSE	3|2	10|1	8|4	21|7
OS	1|2	1|0	3|3	5|5
PM	2|1	1|7	1|5	4|13
ES	0|0	0|3	0|0	0|4
LS-LIN	0|1	0|0	0|0	0|1
IX	0|0	0|1	0|0	0|1

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 LSE, which exhibits superior noise robustness.

In all two 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 LSE estimator demonstrates better robustness compared to other estimators, consistently showing superior performance in all noisy setups when $b = 0.01$.

Table 15: Comparison of different estimators LSE, PM, ES, IX, BanditNet, LS-LIN and OS accuracy for EMNIST and FMNIST with different qualities of logging policy ($\tau \in \{1, 10, 20\}$) and true / estimated propensity scores with $b \in \{5, 0.01\}$ and noisy reward with $P_f \in \{0.1, 0.5\}$. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	P_f	LSE	PM	ES	IX	BanditNet	LS-LIN	OS	Logging Policy
EMNIST	1	–	–	88.49 ± 0.04	**89.19 ± 0.03**	88.61 ± 0.06	88.33 ± 0.13	66.58 ± 6.39	88.70 ± 0.02	88.71 ± 0.26	88.08
		5	–	**89.16 ± 0.03**	88.94 ± 0.05	88.48 ± 0.03	88.51 ± 0.23	65.10 ± 0.69	88.38 ± 0.18	88.70 ± 0.15	88.08
		0.01	–	**86.07 ± 0.01**	85.62 ± 0.10	85.71 ± 0.04	81.39 ± 4.02	66.55 ± 3.11	84.64 ± 0.17	84.59 ± 0.09	88.08
		–	0.1	**89.29 ± 0.04**	89.08 ± 0.05	88.45 ± 0.09	88.14 ± 0.14	59.90 ± 3.78	88.30 ± 0.12	88.74 ± 0.09	88.08
		–	0.5	88.72 ± 0.08	**88.78 ± 0.03**	87.27 ± 0.10	87.08 ± 0.14	56.95 ± 3.06	87.20 ± 0.32	88.06 ± 0.09	88.08
	10	–	–	88.59 ± 0.03	**88.61 ± 0.04**	88.38 ± 0.08	87.43 ± 0.19	85.48 ± 3.13	88.58 ± 0.08	86.88 ± 0.34	79.43
		5	–	88.42 ± 0.07	**88.43 ± 0.07**	88.39 ± 0.10	88.39 ± 0.06	84.90 ± 3.10	88.23 ± 0.27	86.00 ± 0.37	79.43
		0.01	–	**82.15 ± 0.21**	80.85 ± 0.29	81.07 ± 0.07	77.49 ± 2.77	27.02 ± 1.92	78.43 ± 3.13	21.70 ± 4.11	79.43
		–	0.1	**88.29 ± 0.06**	88.22 ± 0.02	88.19 ± 0.08	87.93 ± 0.35	84.89 ± 3.21	87.50 ± 0.17	87.68 ± 0.16	79.43
		–	0.5	88.71 ± 0.16	88.52 ± 0.07	84.42 ± 0.34	83.25 ± 3.45	63.35 ± 13.39	85.75 ± 0.04	**89.09 ± 0.05**	79.43
	20	–	–	**88.28 ± 0.05**	88.20 ± 0.08	87.96 ± 0.34	86.82 ± 1.30	83.69 ± 3.32	88.21 ± 0.06	80.64 ± 0.25	14.86
		5	–	**88.42 ± 0.12**	87.98 ± 0.05	88.27 ± 0.33	88.27 ± 0.07	86.82 ± 0.17	88.19 ± 0.11	79.31 ± 0.61	14.86
		0.01	–	**81.36 ± 0.14**	75.53 ± 2.61	73.45 ± 2.78	72.31 ± 1.46	26.92 ± 2.51	72.33 ± 0.35	11.12 ± 0.39	14.86
		–	0.1	**88.10 ± 0.05**	87.93 ± 0.16	87.69 ± 0.22	87.67 ± 0.18	81.73 ± 3.09	87.08 ± 0.14	82.95 ± 0.31	14.86
		–	0.5	**86.83 ± 0.10**	86.67 ± 0.19	84.01 ± 0.32	80.79 ± 3.06	75.20 ± 3.01	83.05 ± 0.75	86.03 ± 0.48	14.86
FMNIST	1	–	–	76.45 ± 0.12	73.33 ± 2.67	72.90 ± 2.35	69.12 ± 0.26	60.66 ± 2.16	69.29 ± 0.19	**77.77 ± 0.09**	78.38
		5	–	73.20 ± 2.43	75.07 ± 0.27	70.38 ± 2.59	70.80 ± 2.38	22.41 ± 4.50	69.33 ± 0.20	**77.57 ± 0.10**	78.38
		0.01	–	**74.08 ± 1.64**	70.35 ± 0.12	57.93 ± 2.66	63.34 ± 3.64	30.20 ± 8.17	63.86 ± 3.40	37.57 ± 3.16	78.38
		–	0.1	76.07 ± 0.02	74.54 ± 0.02	70.42 ± 2.53	70.58 ± 2.47	50.37 ± 5.43	70.41 ± 2.20	**77.71 ± 0.22**	78.38
		–	0.5	76.96 ± 0.23	74.03 ± 0.30	66.32 ± 0.44	66.66 ± 1.41	54.53 ± 1.32	66.57 ± 2.76	**77.46 ± 0.11**	78.38
	10	–	–	**76.14 ± 0.11**	74.42 ± 0.17	69.25 ± 0.10	70.69 ± 2.39	65.70 ± 3.78	69.31 ± 0.24	74.89 ± 0.96	21.43
		5	–	**75.42 ± 0.16**	74.79 ± 0.15	71.42 ± 2.53	69.21 ± 0.25	69.53 ± 0.29	70.15 ± 2.53	72.87 ± 0.47	21.43
		0.01	–	**74.04 ± 0.15**	60.77 ± 0.09	53.69 ± 1.37	63.57 ± 3.91	26.96 ± 1.87	60.65 ± 3.83	13.22 ± 0.91	21.43
		–	0.1	**76.78 ± 0.23**	73.91 ± 0.13	68.58 ± 0.09	68.07 ± 0.18	64.05 ± 2.34	68.10 ± 0.58	76.24 ± 0.29	21.43
		–	0.5	**77.66 ± 0.17**	74.02 ± 0.05	61.46 ± 4.72	62.60 ± 0.16	43.33 ± 2.83	61.35 ± 1.83	77.52 ± 0.26	21.43
	20	–	–	**75.12 ± 0.03**	74.32 ± 0.12	69.26 ± 0.09	72.46 ± 2.14	64.92 ± 3.82	72.86 ± 2.32	65.78 ± 1.10	14.84
		5	–	**75.13 ± 0.09**	74.17 ± 0.15	69.23 ± 0.46	68.72 ± 0.30	62.41 ± 4.24	69.06 ± 0.11	63.53 ± 1.70	14.84
		0.01	–	**69.16 ± 0.22**	55.20 ± 1.14	60.91 ± 2.75	61.11 ± 4.92	28.23 ± 2.18	61.46 ± 1.96	13.04 ± 4.76	14.84
		–	0.1	**75.48 ± 0.09**	71.84 ± 2.47	65.41 ± 4.23	67.91 ± 0.16	65.21 ± 2.93	68.03 ± 0.46	70.90 ± 0.26	14.84
		–	0.5	**75.96 ± 0.05**	73.12 ± 0.25	61.79 ± 3.13	60.19 ± 3.13	55.13 ± 0.15	60.51 ± 3.28	73.32 ± 0.81	14.84

Table 16: Comparison of different model-based estimators DR, DR-OS, MRDR, SWITCH-DR, SWITCH-DR-LSE, DM and DR-LSE with LSE for EMNIST and FMNIST under a logging policy with $\tau = 10$, true / estimated propensity scores with $b \in \{5, 0.01\}$ and noisy reward with $P_f \in \{0.1, 0.5\}$. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	τ	b	P_f	DR-LSE	DR	DR-OS	MRDR	DR-Switch	DR-Switch-LSE	LSE	DM	Logging Policy
EMNIST	10	–	–	**88.79 ± 0.03**	88.71 ± 0.07	87.79 ± 0.36	80.57 ± 4.00	79.40 ± 5.21	87.73 ± 0.31	88.59 ± 0.03	76.52 ± 2.68	79.43
		5	–	**88.67 ± 0.04**	88.49 ± 0.13	87.83 ± 0.17	80.08 ± 4.62	79.28 ± 0.65	85.80 ± 3.40	88.42 ± 0.07	76.73 ± 4.95	79.43
		0.01	–	**83.30 ± 3.13**	78.24 ± 0.57	80.53 ± 0.32	10.00 ± 0.01	74.81 ± 0.57	41.11 ± 2.87	82.15 ± 0.21	75.65 ± 0.29	79.43
		–	0.1	**88.51 ± 0.02**	88.32 ± 0.16	87.50 ± 0.28	45.49 ± 9.14	75.28 ± 0.09	79.86 ± 0.64	88.29 ± 0.06	78.85 ± 2.69	79.43
		–	0.5	85.88 ± 0.13	83.53 ± 0.54	85.46 ± 0.73	7.04 ± 4.18	72.76 ± 0.56	81.73 ± 0.23	**88.71 ± 0.16**	75.26 ± 2.39	79.43
FMNIST	10	–	–	**80.15 ± 0.09**	68.70 ± 5.12	63.66 ± 0.39	58.61 ± 3.89	54.20 ± 6.27	34.47 ± 0.02	76.14 ± 0.11	51.24 ± 4.16	79.43
		5	–	**79.64 ± 0.05**	66.67 ± 3.50	64.80 ± 2.36	56.62 ± 1.52	56.61 ± 7.37	29.59 ± 3.83	75.42 ± 0.16	59.65 ± 3.13	79.43
		0.01	–	55.10 ± 0.25	52.19 ± 3.84	60.92 ± 1.81	10.00 ± 0.01	63.35 ± 1.62	41.13 ± 2.84	**74.04 ± 0.15**	58.94 ± 4.18	79.43
		–	0.1	**79.91 ± 0.11**	68.94 ± 0.35	63.19 ± 1.69	10.00 ± 0.01	57.54 ± 3.05	52.79 ± 4.04	76.78 ± 0.23	56.33 ± 7.70	79.43
		–	0.5	**79.14 ± 0.04**	56.47 ± 7.08	56.72 ± 7.19	22.05 ± 4.50	59.54 ± 2.95	75.31 ± 0.55	77.66 ± 0.17	53.70 ± 7.19	79.43

G.3 MODEL-BASED ESTIMATORS

There are some approaches where utilise the estimation of reward. For example, in direct method (DM), the reward is estimated from logged data via regression. In particular, an estimation of reward function, $\widehat{r}(x,a)$, is learning from LBF dataset S using a regression. The objective function for DM can be represented as,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a}\pi_\theta(a|x_i)\widehat{r}(a,x_i). \tag{81}$$

In doubly-robust (DR) approach (Dudík et al., 2014) DM is combined with IPS estimator and has a promising performance in off-policy learning scenario. The object function for doubly robust can be represented,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a}\pi_\theta(a|x_i)\widehat{r}(a,x_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}(r_i - \widehat{r}(a,x_i)). \tag{82}$$

There are also some improvements regarding the DR, including DR based on optimistic Shrinkage (DR-OS) (Su et al., 2020), DR-Switch (Wang et al., 2017) and MRDR (Farajtabar et al., 2018).

As these methods are based estimation of reward, we consider them as model-based methods. Inspired by DR method, we combine the LSE estimator with the DM method (DR-LSE)

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a}\pi_\theta(a|x_i)\widehat{r}(a,x_i) + \frac{1}{\lambda}\log\Big(\frac{1}{n}\sum_{i=1}^{n}\exp\big(\lambda\frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}(r_i - \widehat{r}(a,x_i))\big)\Big). \tag{83}$$

We also combine, LSE with DR-Switch as (DR-Switch-LSE) where the IPS estimator in DR-Switch is replaced with LSE estimator.

In this section, we aim to show that the combination of our LSE estimator with the DR method as a model-based method can improve the performance of these methods. For our experiments, we use the same experiment setup as described in App. F. We compare model-based methods, DM, DR and DR-LSE, DR-Switch, DR-OS, DR-Switch-LSE with our LSE estimator. The results are shown in Table 16. We observed that DR-LSE outperforms the standard DR in many scenarios.

G.4 Real-World Dataset

We applied our method to the Kuairec, a public real-world recommendation system dataset ((Gao et al., 2022)). This dataset is gathered from the recommendation logs of the video-sharing mobile app Kuaishou. In each instance, a user watches an item (video) and the watch duration divided by the entire duration of the video is reported. We use the same procedure as (Zhang et al., 2023a) to prepare the logged bandit dataset. We also use the same architecture for the logging policy and the learning policy, with some modifications in the hidden size and number of layers of the deep models. We use separate models for the logging and learning policies. We first train the logging policy using cross-entropy loss and fix it to use as the propensity score estimator for the training of the OPL models. We report Precision@K, and NDCG@K for K=1, 3, 5, 10. Recall@K is very low for small K values because the number of positive items for each use of much more than K. For each method, we use grid search to find the hyperparameter that maximizes the Precision@1 in the validation dataset. The comparison of different estimators is presented in Table 17. We can observe that in Precision@1, Precision@3, Precision@10, NDCG@1, NDCG@3 and NDCG@10, we have the best performance.

Table 17: Comparison of different estimators LSE, PM, ES, IX, LS-LIN, OS and SNIPS in different metrics. The best-performing result is highlighted in **bold** text, while the second-best result is colored in red for each scenario.

Dataset	Method	Precision@1	Precision@3	Precision@5	Precision@10	NDCG@1	NDCG@3	NDCG@5	NDCG@10
KuaiRec	PM	0.8885	0.5723	0.5201	0.4275	0.8585	0.6551	0.5932	0.4988
	SNIPS	0.0289	0.6177	0.5995	0.6462	0.0289	0.4981	0.5226	0.5830
	IX	0.8794	0.5824	0.6355	0.6586	0.8794	0.6164	0.6410	0.6548
	ES	0.8951	0.7495	**0.7187**	0.6644	0.8951	0.7787	**0.7483**	0.7006
	OS	0.8993	0.3215	0.2015	0.1403	0.8993	0.4381	0.3227	0.2378
	LS-LIN	0.8836	0.6680	0.7159	0.6904	0.8836	0.7159	0.7368	0.7108
	LSE	**0.9257**	**0.7534**	0.6999	**0.7206**	**0.9257**	**0.7917**	0.7441	**0.7431**

G.5 Sample Number Effect

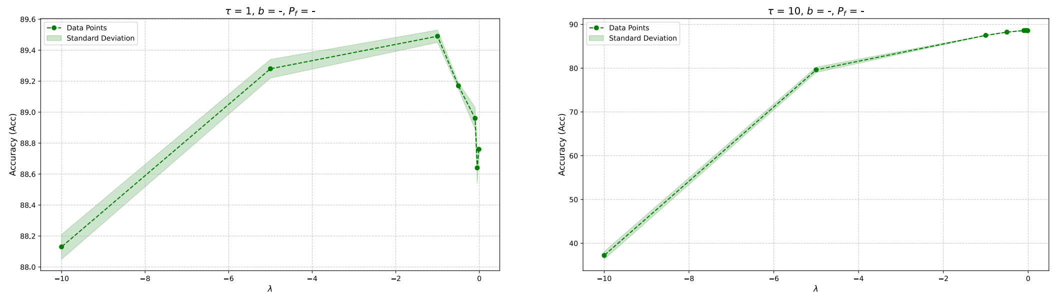
We also conduct experiments on our LSE estimator and PM estimator to examine the effect of limited training samples in the OPL scenario. For this purpose, we considered different ratios of training LBF dataset, $R_n \in \{1, 0.5, 0.2, 0.05\}$. The results are shown in Table 18. We observed that reducing R_n decreased the accuracy for both estimators. However, our LSE estimator demonstrated robust performance under different ratios of training LBF dataset, R_n. Therefore, for small-size LBF datasets, we can apply the LSE estimator for off-policy learning.

G.6 λ Effect

The impact of λ across various scenarios and τ values was investigated using the experimental setup described in Appendix F for the EMNIST dataset. Figure 2 illustrates the accuracy of the LSE estimator for $\tau \in 1, 10$. For $\tau = 1$, corresponding to a logging policy with higher accuracy, an optimal λ value of approximately -1.5 was observed. In contrast, for $\tau = 10$, representing a logging policy with lower accuracy, the optimal λ approached zero. Additionally, in scenarios with noisy rewards Fig.3, both $\tau = 1$ and $\tau = 10$ we observed an optimal λ values larger -2. As for $\tau = 1$, the logging policy has higher accuracy, the effect of noisy reward should be canceled by larger $|\lambda|$. However, for $\lambda = 10$, we need a smaller $|\lambda|$.

Table 18: Comparison of LSE and PM accuracy for EMNIST dataset with different ratio of training LBF dataset ($R_n \in \{1, 0.5, 0.2, 0.05\}$) and true / estimated propensity scores with $b \in \{5, 0.01\}$ and noisy reward with $P_f \in \{0.1, 0.5\}$. The best-performing result is highlighted in **bold** text.

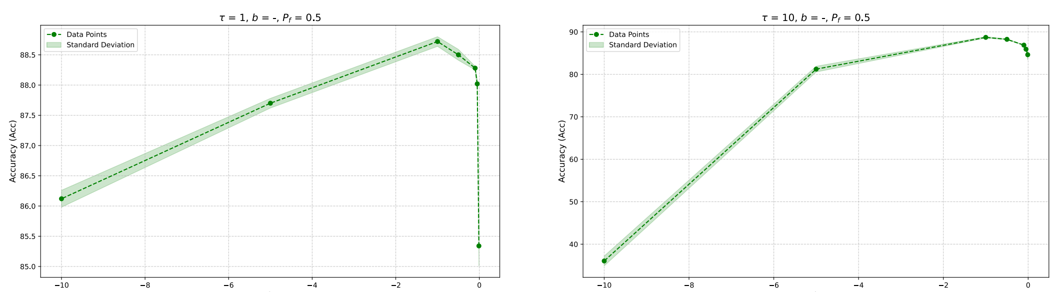
Dataset	τ	R_n	b	P_f	**LSE**	**PM**	**Logging Policy**
EMNIST	1	1	–	–	$\mathbf{88.49 \pm 0.04}$	$\mathbf{89.19 \pm 0.03}$	88.08
			0.01	–	$\mathbf{86.07 \pm 0.01}$	85.62 ± 0.10	88.08
			–	0.5	88.72 ± 0.08	$\mathbf{88.78 \pm 0.03}$	88.08
		0.5	–	–	$\mathbf{87.79 \pm 0.08}$	86.42 ± 0.11	88.08
			0.01	–	$\mathbf{81.13 \pm 0.08}$	48.70 ± 15.46	88.08
			–	0.5	$\mathbf{86.24 \pm 0.07}$	85.17 ± 0.36	88.08
		0.2	–	–	$\mathbf{83.76 \pm 0.25}$	74.57 ± 1.01	88.08
			0.01	–	$\mathbf{67.64 \pm 3.89}$	23.18 ± 5.02	88.08
			–	0.5	$\mathbf{80.39 \pm 0.19}$	69.54 ± 0.65	88.08
		0.05	–	–	$\mathbf{70.16 \pm 2.44}$	53.51 ± 2.77	88.08
			0.01	–	$\mathbf{36.06 \pm 0.62}$	15.56 ± 3.21	88.08
			–	0.5	$\mathbf{50.06 \pm 2.10}$	47.57 ± 5.19	88.08
	10	1	–	–	88.59 ± 0.03	$\mathbf{88.61 \pm 0.04}$	79.43
			0.01	–	$\mathbf{82.15 \pm 0.21}$	80.85 ± 0.29	79.43
			–	0.5	$\mathbf{88.71 \pm 0.16}$	88.52 ± 0.07	79.43
		0.5	–	–	$\mathbf{86.30 \pm 0.04}$	86.02 ± 0.06	79.43
			0.01	–	$\mathbf{75.02 \pm 2.67}$	28.12 ± 1.94	79.43
			–	0.5	$\mathbf{86.61 \pm 0.08}$	83.21 ± 0.10	79.43
		0.2	–	–	80.67 ± 0.35	$\mathbf{80.83 \pm 0.22}$	79.43
			0.01	–	$\mathbf{53.32 \pm 1.47}$	17.03 ± 0.30	79.43
			–	0.5	$\mathbf{80.89 \pm 0.19}$	73.42 ± 1.14	79.43
		0.05	–	–	$\mathbf{48.51 \pm 0.81}$	42.27 ± 1.48	79.43
			0.01	–	$\mathbf{34.15 \pm 0.61}$	14.70 ± 2.20	79.43
			–	0.5	$\mathbf{56.64 \pm 2.40}$	41.75 ± 1.95	79.43



(a) $\tau = 1$

(b) $\tau = 10$

Figure 2: Accuracy of the LSE estimator over different values of λ for true propensity score and reward. (a) $\tau = 1$. (b) $\tau = 10$.



(a) $\tau = 1$

(b) $\tau = 10$

Figure 3: Plots of Accuracy of the LSE estimator over different values of λ for true propensity score and noisy reward with $P_f = 0.5$. (a) $\tau = 1$. (b) $\tau = 10$.

G.7 SELECTION OF λ

Although we use grid search to tune the λ in our algorithm, inspired by Proposition 5.3, we can select the following value,

$$\lambda^* = \frac{1}{n^{1/(\epsilon+1)}}, \tag{84}$$

where n is the number of samples. With such a selection we have a regret rate of $O(n^{-\epsilon/(1+\epsilon)})$. We test and evaluate our selection in OPL and OPE. We examine also a data driven approach for selecting λ in Section G.7.2.

G.7.1 λ SELECTION FOR OPL

We have tested λ^* on EMNIST dataset. In OPL experiments we have truncated the propensity score to 0.001 in order to avoid numerical overflow. Hence, our distributions are effectively heavy-tailed with $\epsilon = 1$, leading to $\lambda^* = \frac{1}{\sqrt{n}}$. We change $n = 512, 256, 128, 64, 16$ with corresponding values $\lambda^* \in \{0.044, 0.0625, 0.088, 0.125, 0.25\}$ which its results are presented in the Table 19. Note that because we use stochastic gradient descent in training, here n is the batch size. 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 Proposition 5.3), but also achieves reasonable performance in experiments.

Table 19: Comparison of accuracy (%) for different λ values and sample sizes n

$\lambda\backslash n$	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
λ^*	92.78 ± 0.01	91.52 ± 0.05	90.38 ± 0.05	88.83 ± 0.02	85.09 ± 0.51

G.7.2 DATA-DRIVEN SELECTION OF λ

In Theorem 5.2, we assume a fixed value of λ. However, it is often important in practical applications to have a method for adjusting λ dynamically based on the data.

Recall the following regret bound proposed by Theorem 5.2,

$$\begin{aligned}\mathfrak{R}_\lambda(\pi_{\widehat{\theta}}, S) \leq \frac{|\lambda|^\epsilon}{1+\epsilon}\nu + \frac{4(2-\gamma)}{3(1-\gamma)} \frac{\log \frac{4|\Pi_\theta|}{\delta} \exp(|\lambda|\nu^{1/(1+\epsilon)})}{n|\lambda|} \\ + \frac{(2-\gamma)}{(1-\gamma)|\lambda|}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log \frac{4|\Pi_\theta|}{\delta} \exp(2|\lambda|\nu^{1/(1+\epsilon)})}{n}}\end{aligned}$$

which is true for any γ. If γ tends to zero, we have,

$$\mathfrak{R}_\lambda(\pi_{\widehat{\theta}}, S) \leq \frac{|\lambda|^\epsilon}{1+\epsilon}\nu + \frac{8}{3}\frac{\exp(|\lambda|\nu^{1/(1+\epsilon)}) \log \frac{4|\Pi_\theta|}{\delta}}{n|\lambda|} + \frac{2}{|\lambda|}\sqrt{\frac{4|\lambda|^{1+\epsilon}\nu \log \frac{4|\Pi_\theta|}{\delta} \exp(2|\lambda|\nu^{1/(1+\epsilon)})}{n}}.$$

Let the upper bound be U_R and $x = \sqrt{\nu|\lambda|^{1+\epsilon}}$. We have,

$$\begin{aligned}U_R &= \frac{x^{\frac{2\epsilon}{1+\epsilon}}}{(1+\epsilon)\nu^{\frac{\epsilon}{1+\epsilon}}}\nu + \frac{8}{3}\frac{\nu^{\frac{1}{1+\epsilon}} \exp(x^{\frac{2}{1+\epsilon}}) \log \frac{4|\Pi_\theta|}{\delta}}{nx^{\frac{2}{1+\epsilon}}} + 2\sqrt{\frac{4\nu \log \frac{4|\Pi_\theta|}{\delta} \exp(2x^{\frac{2}{1+\epsilon}})}{n(x^{\frac{2}{1+\epsilon}}\nu^{\frac{-1}{1+\epsilon}})^{1-\epsilon}}} \\ &= \nu^{\frac{1}{1+\epsilon}}\left(\frac{x^{\frac{2\epsilon}{1+\epsilon}}}{(1+\epsilon)} + \frac{8}{3}\frac{\exp(x^{\frac{2}{1+\epsilon}}) \log \frac{4|\Pi_\theta|}{\delta}}{nx^{\frac{2}{1+\epsilon}}} + 2\sqrt{\frac{4 \log \frac{4|\Pi_\theta|}{\delta} \exp(2x^{\frac{2}{1+\epsilon}})}{nx^{\frac{2(1-\epsilon)}{1+\epsilon}}}}\right).\end{aligned} \tag{85}$$

Finally, we assume that $|\lambda| \leq 1$ and bound and replace the exponential $\exp(x^{\frac{2}{1+\epsilon}})$ by e. Minimizing the upper bound in equation 85, we derive the following optimum λ for the optimization of the upper bound in Theorem 5.2,

$$\lambda^* = \max\left\{ -f(\epsilon) \cdot \left(\frac{\ln\left(\frac{1}{\delta}\right)}{\nu n} \right)^{\frac{1}{1+\epsilon}}, -1 \right\} \tag{86}$$

where $f(\epsilon) = \left(\frac{e(1+\epsilon)}{\epsilon} \left(1 - \epsilon + \sqrt{(1-\epsilon)^2 + \frac{8\epsilon}{3e(1+\epsilon)}} \right) \right)^{\frac{2}{1+\epsilon}}$. Note that, we can compute the empirical value of ν based on the available LBF dataset,

$$\hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} \left(w_\theta(a_i, x_i) r_i \right)^{1+\epsilon}. \tag{87}$$

Using empirical $\hat{\nu}$ in equation 86, we derive the value for data driven λ. Note that, in our experiments, we consider $\epsilon = 1$.

G.7.3 λ SELECTION FOR OPE

We tested our λ selection in the OPE setting with Lomax distributions. We changed the number of samples and set $n = 100, 500, 1K, 5K, 10K, 50K, 500K$ and tested all estimators as we as LSE with selected $\lambda = \lambda^\star$. The results are illustrated at Tables 20, and 21. The first observation is that in all settings, the selected $\lambda^\star$ outperforms all other estimators, except LS which loses in $n \leq 5000$ experiments with a very small margin and is not significantly worse than the λ found by grid search.

Another critical observation is that as the number of samples increases, the selected λ works better than compared to other methods, even LSE with λ found by grid-search. In $n = 100K$, not only $\lambda^\star$ performs the best, but also the λ found by grid-search falls behind IPS-TR and ES. This shows the significance of selective λ when the number of samples is large.

Third observation is the lower performance of $\lambda^\star$ when we have very small number of samples, e.g. $n = 100$. This also conforms our theoretical results, as upper and lower bounds on generalization and regret bounds in Theorem D.1, Theorem D.2 and Theorem 5.2 requires a minimum number of samples as an assumption.

Table 20: Summary of Bias, Variance, and MSE for Different Estimators for Lomax OPE experiments. We change the number of samples $n = 100, 500, 1K, 10K$ and report the metrics for PM, ES, LSE, LSE(λ^*), LS, LS-LIN, OS, IPS-TR, IX, SNIPS

n	Estimator	Bias	Var	MSE
100	PM	-0.2623	30.6419	30.7106
	ES	2.2894	0.0247	5.2662
	LSE	0.6194	0.3967	**0.7803**
	LSE(λ^*)	0.9144	0.1952	1.0314
	LS	0.6386	0.5336	0.9414
	LS-LIN	2.0377	0.0167	4.1689
	OS	0.4485	22.7449	22.9461
	IPS-TR	-0.0144	24.8212	24.8214
	IX	1.9517	0.0171	3.8264
	SNIPS	-0.0483	25.8348	25.8371
500	PM	-0.2002	3.1605	3.2006
	ES	0.0415	2.5603	2.5620
	LSE	0.2221	0.3375	**0.3869**
	LSE(λ^*)	0.5542	0.0984	0.4055
	LS	0.2309	0.3449	0.3983
	LS-LIN	2.0377	0.0033	4.1557
	OS	0.42724	7.6075	7.7901
	IPS-TR	0.0415	2.5603	2.5620
	IX	1.9536	0.0035	3.8200
	SNIPS	0.0347	2.6865	2.6877
1000	PM	-0.2379	4.8325	4.8891
	ES	0.0076	3.9145	3.9145
	LSE	0.2262	0.1720	**0.2231**
	LSE(λ^*)	0.4335	0.0712	0.2591
	LS	0.2270	0.1751	0.2266
	LS-LIN	2.0368	0.0016	4.1502
	OS	0.4178	4.0558	4.2303
	IPS-TR	0.0076	3.9145	3.9145
	IX	1.9536	0.0018	3.8186
	SNIPS	0.0040	4.0054	4.0054
5000	PM	-0.2428	3.7591	3.8180
	ES	0.0032	3.0449	3.0449
	LSE	0.2277	0.0343	**0.0862**
	LSE(λ^*)	0.2448	0.0319	0.0919
	LS	0.2334	0.0342	0.0887
	LS-LIN	2.0374	0.0003	4.1513
	OS	0.4626	0.4477	0.6617
	IPS-TR	0.0032	3.0449	3.0449
	IX	1.9535	0.0004	3.8166
	SNIPS	0.0025	2.9976	2.9976
10000	PM	-0.2318	0.4702	0.5239
	ES	0.0131	0.3809	0.3811
	LSE	0.2254	0.0171	0.0679
	LSE(λ^*)	0.1867	0.0212	**0.0560**
	LS	0.2341	0.0173	0.0721
	LS-LIN	2.0376	0.0002	4.1518
	OS	0.4336	0.5004	0.6884
	IPS-TR	0.0131	0.3809	0.3811
	IX	1.9536	0.0002	3.8168
	SNIPS	0.0123	0.3830	0.3832

Table 21: Summary of Bias, Variance, and MSE for Different Estimators for Lomax OPE experiments. We change the number of samples $n = 50K, 100K$ and report the metrics for PM, ES, LSE, LSE(λ^*), LS, LS-LIN, OS, IPS-TR, IX, SNIPS

n	Estimator	Bias	Var	MSE
50000	PM	-0.2418	0.2152	0.2736
	ES	0.0040	0.1743	0.1743
	LSE	0.2261	0.0033	0.0544
	LSE(λ^*)	0.1020	0.0085	**0.0189**
	LS	0.2324	0.0035	0.0574
	LS-LIN	2.0374	0.0000	4.1512
	OS	0.3872	5.0487	5.1987
	IPS-TR	0.0040	0.1743	0.1743
	IX	1.9538	0.0000	3.8172
	SNIPS	0.0040	0.1745	0.1746
100000	PM	-0.2347	0.0633	0.1184
	ES	0.0105	0.0513	0.0514
	LSE	0.2267	0.0017	0.0531
	LSE(λ^*)	0.0790	0.0056	**0.0119**
	LS	0.2338	0.0017	0.0564
	LS-LIN	2.0375	0.0000	4.1516
	OS	0.4294	0.2179	0.4021
	IPS-TR	0.0105	0.0513	0.0514
	IX	1.9538	0.0000	3.8172
	SNIPS	0.0105	0.0515	0.0516

G.7.4 SENSITIVITY TO THE SELECTION OF λ

In this section, we investigate the performance of our proposed data-driven λ in App. G.7.2, where can avoid any sort of hyper-parameter tuning. Hence solving the problem of selection of λ and any concerns related to the selection of a "bad" λ.

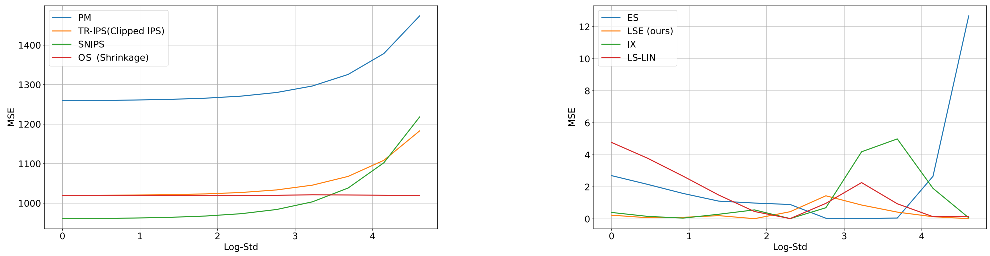
In order to measure the sensitivity of the selection of λ we compare three different methods. First, λ is found by grid search which provides the best MSE. Second, λ^* is found by the data-driven suggest in App.G.7.2. In the third method, we select λ uniformly randomly from [0, 1], $\tilde{\lambda} \sim \mathrm{Uniform}(0, 1)$. This method shows the performance of LSE by choosing random λ as hyperparameter. We test these methods on the Lomax scenario where we have the more challenging heavy-tailed (for $\epsilon \neq 1$) condition. The MSE of each method for the same setting of parameters as in the original OPE experiments and for $n = 1K, 10K, 100K$ is reported in table 22.

Table 22: MSE of LSE with fine-tuned, data-driven and random λ for $\beta = 1.0, 1.5, 2.0$. The experiment was run 100000 times with different values of α, α', and β.

β	α	α'	**Estimator**	$n = 1K$	$n = 10K$	$n = 100K$
0.5	1.0	1.0	LSE	**0.006**	**0.0009**	**0.0001**
			LSE-λ^*	0.049	0.0076	0.0009
			LSE-$\tilde{\lambda}$	0.131	0.131	0.131
		1.5	LSE	**0.041**	**0.0.008**	**0.0039**
			LSE-λ^*	0.463	0.138	0.03
			LSE-$\tilde{\lambda}$	0.449	0.449	0.449
		2.0	LSE	**0.105**	**0.033**	**0.026**
			LSE-λ^*	1.044	0.450	0.148
			LSE-$\tilde{\lambda}$	0.764	0.762	0.760
1.0	1.5	1.0	LSE	**0.014**	**0.002**	**0.0003**
			LSE-λ^*	0.110	0.018	0.002
			LSE-$\tilde{\lambda}$	0.398	0.398	0.394
		1.5	LSE	**0.093**	**0.020**	**0.012**
			LSE-λ^*	1.042	0.311	0.067
			LSE-λ_r	1.227	1.226	1.223
		2.0	LSE	**0.211**	**0.088**	**0.0754**
			LSE-λ^*	3.05	1.013	0.333
			LSE-$\tilde{\lambda}$	1.991	1.99	1.985
2.0	2.5	1.0	LSE	**0.0463**	**0.005**	**0.0014**
			LSE-λ^*	0.3071	0.048	0.0054
			LSE-$\tilde{\lambda}$	1.550	1.548	1.552
		1.5	LSE	**0.222**	**0.058**	**0.052**
			LSE-λ^*	2.894	0.864	0.187
			LSE-$\tilde{\lambda}$	4.242	4.236	4.246
		2.0	LSE	**0.548**	**0.313**	**0.289**
			LSE-λ^*	6.530	2.817	0.928
			LSE-$\tilde{\lambda}$	6.534	6.531	6.535

Our experimental results demonstrate that LSE with grid-searched λ consistently achieves the lowest MSE across all experimental configurations. The data-driven λ selection approach exhibits strong performance, ranking second in scenarios with larger sample sizes ($n = 10K, 100K$). For smaller samples ($n = 1K$), random λ selection occasionally outperforms the data-driven approach. Notably, LSE maintains robust variance control under heavy-tailed distributions even with randomly selected λ values. The performance gap between data-driven and random λ selection widens significantly as the sample size increases, suggesting a clear strategy for parameter selection: while the estimator remains robust to arbitrary λ choices, the data-driven approach becomes increasingly reliable with larger sample sizes.

- If n is small (e.g. $n \approx 1000$), we have fewer computational concerns, and a grid-search based on the performance on a validation set can find an appropriate λ for our problem.



(a) PM, TR-IPS, SNIPS, OS (b) LSE, LS-LIN, IX, ES

Figure 4: MSE of the PM, TR-IPS, SNIPS, OS, LS-LIN, IX, OS, ES, and LSE estimators over different values of $\log \sigma$

- For larger values of n, we can hold to the data-driven proposal of λ which gives a comparable performance with the grid-search method.

Another hint about the selection of λ is that for problems where the variance of the importance weights of the unbounded behavior of the reward function is not an issue, a very small λ (e.g. $\lambda = 0.01$) can be a better option because as $\lambda \to 0$, LSE tends to vanilla IPS. For heavy-tailed problems, selecting bigger λ values around 1 can lead to better performance.

G.8 OPE WITH NOISE

Here we discuss the performance of estimators in OPE when reward noise is available. In all experiments, the number of samples is 1000 and the number of trials is $100K$.

G.8.1 GAUSSIAN SETTING

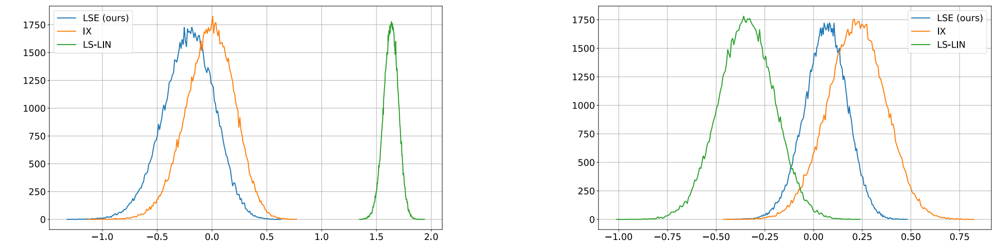
We run the same experiments as mentioned in Section 6.1 by adding noise to the observed reward. We add a positive Gaussian noise,

$$\tilde{R}(S, A) = R(S, A) + |W| : W \sim \mathcal{N}(0, \sigma^2).$$

where $\tilde{R}(S, A)$ is noisy reward function. We increase σ from 1 to 100 and observe the behavior of different estimators under the noise. We report the MSE of different estimators. There is a discrepancy between the performance of different estimators. LSE, LS, LSE-LIN, IX, and ES demonstrated robust performance under high noise conditions, while PM, TR-IPS, SNIPS, and OS exhibited substantially higher MSE values, often differing by several orders of magnitude from the better-performing estimators. We draw the MSE of these two groups against $\log \sigma$ in Figure 4. We observe that ES, LSE, IX, and LS-LIN are better suited for the noisy scenario. Also we observe that ES is more sensitive to the increase of the variance of the noise. We also investigate the distributional form of the estimators with the same levels of noise. Estimators other than LSE, LS-LIN, and IX keep proposing outlier estimations. But these three estimators stay stable in this setting and are compared in Figure 5 for two levels of noise. Among these three estimators, LSE can keep a low bias with almost the same variance in comparison to IX nad LS-LIN, hence leading to the lowest MSE.

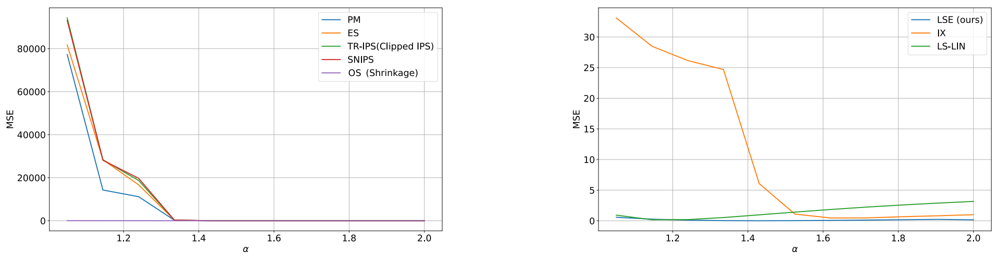
G.8.2 LOMAX SETTING

When we examine the Lomax setting, the estimators' performance deteriorates as we introduce heavier-tailed noise distributions. To test this, we add Pareto-distributed (with parameter α) noise to the reward, varying the parameter α from 1.05 to 2.0. The parameter α controls the tail weight of the distribution, with values closer to 1 producing heavier tails. Our results, shown in Figure 6, reveal a clear split in estimator performance. The estimators - PM, ES, TR-IPS, OS, and SNIPS - struggle significantly with the heavy-tailed noise and show poor performance based on their MSE. In contrast, the more robust estimators - LSE, LS-LIN, and IX - maintain better performance across different noise levels, similar to what we observed in the Gaussian scenario. Note that the IX estimator, despite



(a) small noise, $\sigma^2 \approx 1.82$ (b) large noise, $\sigma^2 = 20$

Figure 5: The error distribution of the LS-LIN, LSE, and IX estimators

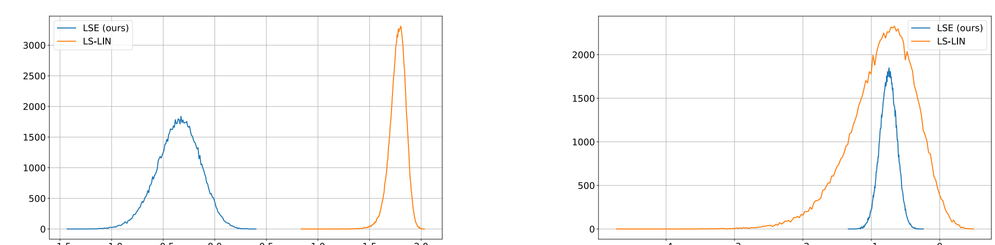


(a) PM, TR-IPS, SNIPS, OS, ES (b) LSE, LS-LIN, IX

Figure 6: MSE of the PM, TR-IPS, SNIPS, OS, LS-LIN, IX, OS, ES, and LSE estimators over different values of α

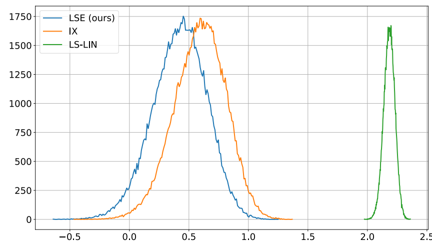
having significantly less error than the poorly performing estimators, compared to LSE and LS-LIN is much more worse in the tail of the noise.

For the distributional behavior of the estimators, we observe that except LSE and LS-LIN, the estimators produce extreme outlier values. Error distribution is the distribution of the difference between the estimated value and the true value. Hence, we plot the error distribution of the LSE and LS-LIN with respect to noise in Figure 7. Here we see that in the small noise scenario LSE despite having more variance, is significantly less bias. Under large noise, LSE keeps the variance lower than LS-LIN, while showing the same bias. Hence, in both cases LSE achieve less MSE than LS-LIN and performs better in both small and large noise scenarios.

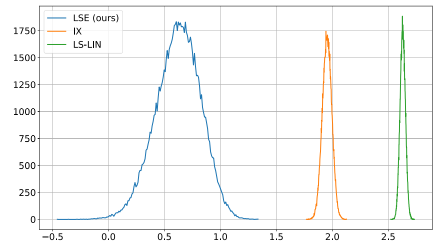


(a) Small noise, $\alpha = 2.0$ (b) Large noise, $\alpha = 1.05$

Figure 7: The error distribution of the LS-LIN and LSE estimators



(a) Gaussian Scenario



(b) Lomax Scenario

Figure 8: The error distribution of the LS-LIN, IX, and LSE estimators

G.9 Distributional properties in OPE

In this section, we investigate the error distribution of different estimators. In both Gaussian and Lomax settings, SNIPS, TR-IPS, OS, ES, and PM show extreme outlier values, but LS-LIN, LSE, and IX avoid outliers. In Figure 8, we show the error distribution of these estimators. We can see the competitive performance of IX and LSE in the Gaussian scenario, while LS-LIN induces a relatively large bias in this setting. In the Lomax setting, LSE has a bigger variance than IX and LS-LIN, while having significantly less bias. LSE has the property that it keeps bias significantly low while trading it for some small variance, leading to less MSE and better performance.

G.10 More Comparison with LS Estimator

We conduct experiments to measure and compare the sensitivity of LSE and LS with respect to the selection of λ. To measure the sensitivity, we choose grid method where we test the followings set of $\lambda \in \{0.001, 0.01, 0.1, 1.0, 5.0\}$ in Table 23, adaptive method where $\lambda_n := \frac{1}{\sqrt{n}}$ is chosen, Table 24, and random method where $\hat{\lambda}$ chosen uniformly random from $[0, 1]$, Table 25. Then we compare these two estimators among these different methods of selecting λ. The results are reported below for Lomax setup.

We can observe that in a close competitions, using the grid search method, LSE outperforms in 4 out of 9 experiments. With the adaptive method, LSE performs better in 7 out of 9 experiments, and when using the random method, LSE outshines in all 9 experiments.

Table 23: MSE of LSE and LS estimators with grid-searched for $\lambda \in 0.001, 0.01, 0.1, 1.0, 5.0$ and $\beta = 1.0, 1.5, 2.0$. The experiment was run 100000 times with different values of α, α', and β.

β	α	α'	**Estimator**	Bias	Variance	MSE
0.5	1.0	1.0	LSE	0.0362	0.0047	0.0060
			LS	0.0266	0.0047	**0.0054**
		1.5	LSE	0.1693	0.0118	0.0404
			LS	0.0697	0.0346	**0.0395**
		2.0	LSE	0.1590	0.0813	**0.1066**
			LS	0.3086	0.0238	0.1190
1.0	1.5	1.0	LSE	0.0728	0.0091	0.0144
			LS	0.0429	0.0104	**0.0122**
		1.5	LSE	0.1065	0.0829	0.0942
			LS	0.1183	0.0717	**0.0857**
		2.0	LSE	0.2726	0.1367	**0.2111**
			LS	0.2548	0.1785	0.2434
2.0	2.5	1.0	LSE	0.0302	0.0452	0.0461
			LS	0.0819	0.0281	**0.0348**
		1.5	LSE	0.2245	0.1702	**0.2206**
			LS	0.2330	0.1699	0.2242
		2.0	LSE	0.5345	0.2645	**0.5502**
			LS	0.4946	0.3696	0.6142

Table 24: MSE of LSE_{λ_n} and LS_{λ_n} estimators with data-driven $\lambda_n = \frac{1}{\sqrt{n}}$ for $\beta = 1.0, 1.5, 2.0$ and $n = 1000$. The experiment was run 100000 times with different values of α, α', and β.

β	α	α'	**Estimator**	Bias	Variance	MSE
0.5	1.0	1.0	LSE_{λ_n}	0.0816	0.0029	**0.0096**
			LS_{λ_n}	0.1314	0.0028	0.0200
		1.5	LSE_{λ_n}	0.2756	0.0054	**0.0814**
			LS_{λ_n}	0.2841	0.0073	0.0880
		2.0	LSE_{λ_n}	0.4651	0.0063	0.2226
			LS_{λ_n}	0.4476	0.0099	**0.2103**
1.0	1.5	1.0	LSE_{λ_n}	0.1596	0.0053	**0.0308**
			LS_{λ_n}	0.2610	0.0052	0.0733
		1.5	LSE_{λ_n}	0.4857	0.0091	**0.2449**
			LS_{λ_n}	0.5129	0.0123	0.2754
		2.0	LSE_{λ_n}	0.7817	0.0100	0.6211
			LS_{λ_n}	0.7645	0.0159	**0.6004**
2.0	2.5	1.0	LSE_{λ_n}	0.3652	0.0111	**0.1445**
			LS_{λ_n}	0.6177	0.0108	0.3924
		1.5	LSE_{λ_n}	0.9792	0.0169	**0.9757**
			LS_{λ_n}	1.0722	0.0227	1.1723
		2.0	LSE_{λ_n}	1.4919	0.0180	**2.2437**
			LS_{λ_n}	1.4952	0.0282	2.2637

Table 25: MSE of $\mathrm{LSE}_{\hat{\lambda}}$ and $\mathrm{LS}_{\hat{\lambda}}$ estimators with random $\hat{\lambda}$ for $\beta = 1.0, 1.5, 2.0$. The experiment was run 100000 times with different values of α, α', and β.

β	α	α'	**Estimator**	Bias	Variance	MSE
0.5	1.0	1.0	$\mathrm{LSE}_{\hat{\lambda}}$	0.3418	0.0139	**0.1308**
			$\mathrm{LS}_{\hat{\lambda}}$	0.6779	0.0640	0.5236
		1.5	$\mathrm{LSE}_{\hat{\lambda}}$	0.6516	0.0247	**0.4493**
			$\mathrm{LS}_{\hat{\lambda}}$	0.8335	0.0581	0.7528
		2.0	$\mathrm{LSE}_{\hat{\lambda}}$	0.8583	0.0262	**0.7629**
			$\mathrm{LS}_{\hat{\lambda}}$	0.9635	0.0491	0.9775
1.0	1.5	1.0	$\mathrm{LSE}_{\hat{\lambda}}$	0.6019	0.0365	**0.3987**
			$\mathrm{LS}_{\hat{\lambda}}$	1.2196	0.1783	1.6656
		1.5	$\mathrm{LSE}_{\hat{\lambda}}$	1.0803	0.0594	**1.2264**
			$\mathrm{LS}_{\hat{\lambda}}$	1.4290	0.1521	2.1941
		2.0	$\mathrm{LSE}_{\hat{\lambda}}$	1.3890	0.0603	**1.9898**
			$\mathrm{LS}_{\hat{\lambda}}$	1.6017	0.1237	2.6890
2.0	2.5	1.0	$\mathrm{LSE}_{\hat{\lambda}}$	1.1942	0.1180	**1.5442**
			$\mathrm{LS}_{\hat{\lambda}}$	2.4803	0.6019	6.7537
		1.5	$\mathrm{LSE}_{\hat{\lambda}}$	2.0218	0.1686	**4.2565**
			$\mathrm{LS}_{\hat{\lambda}}$	2.7676	0.4797	8.1390
		2.0	$\mathrm{LSE}_{\hat{\lambda}}$	2.5258	0.1654	**6.5451**
			$\mathrm{LS}_{\hat{\lambda}}$	3.0074	0.3796	9.4239

G.11 OPE ON REAL-WORLD DATASETS

Table 26: UCI datasets specifications. N is the number of samples, K is the number of actions, and p is the number of features.

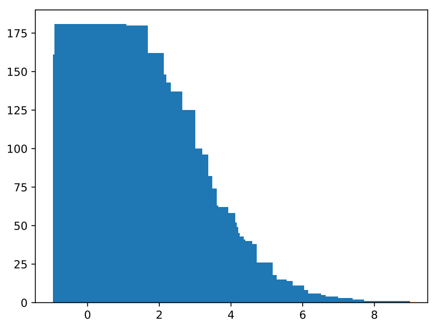
Dataset	N	K	p
Yeast	1,484	10	8
Page-blocks	5,473	5	10
Optdigits	5,620	10	64
Satimage	6,430	6	36
Kropt	28,056	18	6

We evaluate our method's performance in OPE by conducting experiments on 5 UCI classification datasets, as explained in Table 26,

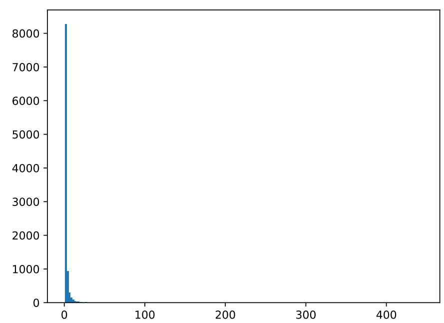
We use the same supervised-to-bandit approach as in OPL experiments. Suggested by Sakhi et al. (2024), we consider a set of softmax policies as the target and logging policy. Consider an ideal policy as a softmax policy peaked on the true label of the sample. Moreover, a faulty policy is an ideal policy that has a set of its actions shifted by 1, hence, doing mostly wrong on the samples from the shifted labels. For the logging policy, we use faulty policies on the first $K/2$ actions with temperatures $\tau_0 = \{0.6, 0.7, 0.8\}$, and faulty policies on the last $K/2$ actions with $\tau = \{0.1, 0.3, 0.5\}$ as target policies, a total of 9 different experiments for each dataset. We create a bandit dataset using the logging policy π_0 and estimate the expected reward of the π_θ which is calculated as below,

$$V(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} \pi_\theta\left(y_i|x_i\right)$$

where y_i is the true label of the data sample x_i. We also add a random uniform noise $\epsilon \sim \mathrm{Uniform}(0,1)$ to the policy logits before softmax. We ran each experiment in each setting 10 times and calculated the average MSE of each estimator over all 90 experiments. For hyperparameter selection, for LS, OS, IPS-TR, PM, and IX, we use their own proposals. For LSE and ES, we use 0.2



(a) Gaussian Distribution, $\mathcal{N}(0, 2)$



(b) Lomax Distribution, $\mathcal{L}(1.5)$

Figure 9: Histogram of 10K samples generated from Gaussian and Lomax distributions (we consider the absolute value of the Gaussian samples to focus on the tail of the distributions)

of the dataset as a validation set to find the hyperparameter with the lowest MSE by grid search and evaluate the method on the remaining 0.8 of the dataset. Table 27 illustrates this on the 5 datasets for different estimators.

Table 27: MSE of LSE, PM, ES, IX, OS, LS, IPS-TR and SNIPS estimators on 5 UCI classification datasets on the OPE task.

Dataset	PM	ES	IX	OS	LS	IPS-TR	SN-IPS	LSE
Yeast	0.237	0.0096	0.0573	0.0131	0.0146	0.0255	0.0088	**0.0077**
Satimage	0.0033	0.0066	0.0057	0.0035	0.0047	0.0043	0.0086	**0.0028**
Kropt	0.0160	0.0041	0.0056	0.0169	0.0208	0.0189	0.0256	**0.0015**
Optdigits	0.0079	0.0066	0.0150	0.0076	0.0083	0.0098	0.0110	**0.0042**
Page-Blocks	0.0440	**0.0002**	0.0236	0.0487	0.0513	0.0445	0.0639	0.0008

G.12 Connection between heavy-tailed distributions and outlier modeling

We illustrate how heavy-tailed distributions can model outlier samples. Consider two sets of observations, the first one from a normal distribution $\mathcal{N}(0, 2)$ which has an exponential tail, and the second from a Lomax distribution $\mathcal{L}(1.5)$, which is heavy-tailed with $\epsilon = 0.5$. Figure 9 depicts the histogram of observed 10K samples from each distribution. We can observe that the Lomax distribution contains large, low-probability values (values around 400), but the total range for Gaussian observations is less than 10. The occurrence of sparse very low probability outlier values is possible by sampling from a heavy-tailed distribution like Lomax distribution. However, it's does not hold for an exponential-tailed distribution like Gaussian. Hence, heavy-tailed distributions seem to be able to model scenarios with sparse large rewards or outliers, which is not possible using an exponential-tailed distribution. In the following, we discuss the heavy-tailed reward scenario in RL applications.