Oracle-Efficient Pessimism: Offline Policy Optimization In Contextual Bandits

Lequn Wang[*]
Netflix
lequnw@netflix.com

Akshay Krishnamurthy
Microsoft Research NYC
akshaykr@microsoft.com

Aleksandrs Slivkins
Microsoft Research NYC
slivkins@microsoft.com

Abstract

We consider offline policy optimization (OPO) in contextual bandits, where one is given a fixed dataset of logged interactions. While pessimistic regularizers are typically used to mitigate distribution shift, prior implementations thereof are either specialized or computationally inefficient. We present the first *general oracle-efficient* algorithm for pessimistic OPO: it reduces to supervised learning, leading to broad applicability. We obtain statistical guarantees analogous to those for prior pessimistic approaches. We instantiate our approach for both discrete and continuous actions and perform experiments in both settings, showing advantage over unregularized OPO across a wide range of configurations.

1 INTRODUCTION

Offline Policy Optimization (OPO) is a fundamental variant of reinforcement learning (RL) where one optimizes a decision-making policy using previously collected data. (It is also called *Offline RL.*) OPO is particularly useful when new experimentation via online RL is costly, dangerous, or would take too long. A central challenge in OPO is *distribution shift*, when the logging policy and the learned policy induce different data distributions, possibly leading to high uncertainty on the learned policy and poor overall performance. This challenge is typically mitigated via *pessimism*: optimizing a regularized objective that evaluates each policy via a "pessimistic" confidence bound on its loss, thereby penalizing policies with high empirical variance. While this approach is well-understood from statistical perspective, computationally efficient implementations remain elusive. In this paper, we address this issue for *contextual bandits*, a practically important special case of RL and a research area in its own right.

Thus, we study pessimistic OPO in contextual bandits.[1] We develop a new algorithm for this problem. Our algorithm is *oracle-efficient*, making a single call to an (arbitrary) computational oracle for supervised learning. The algorithm efficiently forms an artificial problem instance of supervised learning which incorporates pessimism and is passed to the oracle. This reduction to supervised learning allows us to handle any "oracle-supported" policy class (*i.e.,* a policy class that an oracle can optimize over), and offers flexibility to employ various oracle implementations developed in prior work. We obtain similar statistical guarantees as prior (computationally inefficient) implementations of pessimism. On a high level, we obtain the first computationally efficient algorithm with statistical guarantees for an arbitrary oracle-supported policy class.

Our approach carries over to contextual bandits with continuous actions, an important, well-studied scenario motivated by optimizing prices and continuous system parameters. Distribution shift and computational tractability are particularly challenging in this scenario. This is due to the complexity of the action space and the large number of hyper-parameters, respectively.

We conduct an extensive empirical study for both discrete and continuous actions. We instantiate our approach across a range of configurations, both for the experimental environment and for the algorithm itself. We find that our approach is broadly superior to the vanilla policy optimization, while it is much more widely applicable—due to oracle efficiency—than the

[*] Most work done while LW was a graduate student at Cornell and an intern at Microsoft Research NYC.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

[1] *I.e.,* OPO with pessimism, as described above. Without further mention, we focus on regularizers that penalize empirical variance, rather than policy complexity. While the latter regularizers may also be construed as "pessimistic", they target overfitting rather than distribution shift.

implementations of pessimism using prior techniques.

Due to space limitations, all proofs and details of the experiments are, resp., in Appendix A and Appendix B.

1.1 Related Work

OPO is extensively studied in contextual bandits (starting from, *e.g.,* Beygelzimer and Langford, 2009; Bottou et al., 2013; Dudík et al., 2014; Athey and Imbens, 2016), and in RL more generally (Levine et al., 2020).

OPO methods typically build on estimators for *offline policy evaluation* (Langford et al., 2008; Dudík et al., 2014; Farajtabar et al., 2018), with inverse probability weighting (IPW) estimator (Horvitz and Thompson, 1952) as the canonical example. It is well known that IPW can have large variance, and a number of variations of this estimator, such as clipping (Strehl et al., 2010; Wang et al., 2017; Su et al., 2019), self-normalization (Swaminathan and Joachims, 2015b), and shrinkage (Su et al., 2020), have been developed to mitigate variance in exchange for introducing bias. These estimators do not alleviate distribution shift and are complementary to our approach. As such, we focus on vanilla IPW for our theoretical treatment.

The oracle-efficiency framework has been prominent in contextual bandits since Langford and Zhang (2007), with much of this work focusing on supervised learning oracles (e.g., Langford and Zhang, 2007; Dudik et al., 2011; Agarwal et al., 2014). It has been adopted across a range of other problems including structured prediction (Daumé et al., 2009; Ross et al., 2011), active learning (Dasgupta et al., 2007), and online learning (Haghtalab et al., 2022; Block et al., 2022) and in many cases has lead to highly practical algorithms.

Contextual bandits with continuous actions have been studied since Lu et al. (2010); Slivkins (2014), amidst many related papers on *non*-contextual bandits, usually under Lipschitz assumptions (c.f., Slivkins, 2019, Ch 4.4). Our work builds on the "smoothing" approach from Krishnamurthy et al. (2020), see also (Majzoubi et al., 2020; Zhu and Mineiro, 2022; Zhu et al., 2021). OPO with continuous actions has also been studied in Kallus and Zhou (2018); Chernozhukov et al. (2019), but focusing on issues other than pessimism.

Pessimistic OPO in contextual bandits was introduced in (Swaminathan and Joachims, 2015a) via the Empirical Bernstein (EB) regularizer. This approach allows for arbitrary policy classes, but is computationally inefficient, limiting applicability.

The vast follow-up work is either computationally inefficient (*e.g.,* Jin et al., 2022), or lacks statistical guarantees (*e.g.,* Fujimoto et al., 2019; Kumar et al., 2020; Yu et al., 2020; Trabucco et al., 2021), or is substantially restricted in scope. The latter work posits *realizability* – a particular loss model, *e.g.,* linear (*e.g.,* Liu et al., 2020; Li et al., 2022; Rashidinejad et al., 2022; Uehara and Sun, 2022), or focuses on a "tabular" problem[2] (*e.g.,* Kidambi et al., 2020; Rashidinejad et al., 2021), or only applies to specific policy classes (*e.g.,* London and Sandler, 2019; Nguyen-Tang et al., 2022; Sakhi et al., 2023; Aouali et al., 2023).

In contrast to all this follow-up work, our method is computationally efficient and general in scope, not relying on realizability or small number of contexts and handling any "oracle-supported" policy class. Further, none of this follow-up work handles continuous actions.

In simultaneous work,[3] Aouali et al. (2023) use a regularizer similar to ours, but with a specialized scope (mixtures of linear mixed-logit policies) and a different (PAC-Bayesian) perspective in the analysis.

2 PRELIMINARIES

Offline Policy Optimization (OPO). In contextual bandits, an agent interacts with an environment with a context space $\mathcal{X}$ and an action space $\mathcal{A}$. In each round i, the agent observes a context $x_i \in \mathcal{X}$, chooses an action $a_i \in \mathcal{A}$, and observes a loss $\ell_i(a_i) \in [0,1]$ (and nothing else). The pair (x_i, ℓ_i), where ℓ_i is a *loss function* $\mathcal{A} \to [0,1]$, is drawn independently from some fixed (but unknown) distribution $\mathscr{D}$.

As a form of inductive bias, a policy class Π is given, where each policy $\pi \in \Pi$ is a randomized mapping from contexts to actions. $\pi(\cdot \mid x)$ specifies the distribution over actions given context x. We define the risk for policy π and the optimal policy $\pi^\star$, respectively, as

$$R(\pi) := \mathbb{E}_{(x,\ell)\sim\mathscr{D},\; a\sim\pi(\cdot \mid x)}\left[\ell(a)\right],$$
$$\pi^\star \in \operatorname{argmax}_{\pi\in\Pi} R(\pi).$$

In OPO, one seeks a policy $\pi \in \Pi$ with low *excess risk* $R(\pi) - R(\pi^\star)$. The input is a dataset $\mathcal{S} = \{(x_i, a_i, \ell_i(a_i))\}_{i\in[N]}$ collected over N rounds of contextual bandits by a known *logging policy* $\mu : \mathcal{X} \to \Delta(\mathcal{A})$. Here, each action a_i is drawn independently from distribution $\mu(\cdot \mid x_i)$ specified by the logging policy.

We posit access to a *computational oracle*: an algorithm for some hard but well-studied problem. Indeed, OPO tends to be NP-hard even with full feedback: given datapoints (x_i, a_i, ℓ_i), $i \in [N]$, where $\ell_i : \mathcal{A} \to [0,1]$ is the entire loss function. However, this is precisely

[2] *I.e.,* optimizes over the class of all policies, implicitly assuming a small number of contexts.

[3] According to resp. publication dates on `arxiv.org`.

cost-sensitive classification (CSC), a classical and well-studied problem in supervised learning. Therefore, we posit an oracle which exactly solves CSC for a particular action set $\mathcal{A}$ and policy class Π.[4] We allow only a small number of oracle calls; such algorithms are called *oracle-efficient* (and our algorithms only call the oracle once). This is a standard approach in prior work on contextual bandits and OPO (Dudík et al., 2014; Swaminathan and Joachims, 2015a).

Naive Solution: IPW. The prevailing approach to OPO from the statistical perspective is to construct an estimator $\widehat{R}:\Pi\to\mathbb{R}_+$ for the policy risk $R(\cdot)$ based on the dataset $\mathcal{S}$ and minimize $\widehat{R}(\pi)$ over the policy class Π. The standard estimator is the inverse probability weighting (IPW) estimator. Define:

$$\widehat{\ell}_i(\pi) := \frac{\pi(a_i \mid x_i)}{\mu(a_i \mid x_i)}\ell_i(a_i), \quad \widehat{R}_{\mathrm{IPW}}(\pi) := \frac{1}{N}\sum_{i=1}^{N}\widehat{\ell}_i(\pi),$$

and set $\widehat{\pi}_{\mathrm{IPW}} \in \operatorname{argmin}_{\pi\in\Pi}\widehat{R}_{\mathrm{IPW}}(\pi)$. The IPW estimator is unbiased and (for finitely many actions) asymptotically consistent whenever the support of the logging policy μ is the entire action space (for any context). To formalize this:

Assumption 2.1. *$\mu(a \mid x) > 0$ for any context $x\in\mathcal{X}$ and action $a\in\mathcal{A}$.*

Further, the IPW-based approach is oracle-efficient: optimizing $\widehat{R}_{\mathrm{IPW}}(\cdot)$ is equivalent to calling the oracle with the loss vectors $a\mapsto \ell_i(a_i)/\mu(a_i \mid x_i)\cdot\mathbf{1}\{a=a_i\}$.

The *variance* of the IPW estimator, specifically $V(\pi) := \mathrm{Var}\left[\widehat{\ell}_i(\pi)\right]$, will be crucial in what follows. We call it the *IPW-variance* of policy π. Denote $V(\Pi) := \sup_{\pi\in\Pi}V(\pi)$. Also important is the worst-case density ratios, $\delta_{\sup}(\pi,\mu) := \sup_{x\in\mathcal{X},a\in\mathcal{A}}\pi(a \mid x)/\mu(a \mid x)$ for a particular policy π and $\delta_{\sup}(\Pi,\mu) := \sup_{\pi\in\Pi}\delta_{\sup}(\pi,\mu)$.

Known Issue: Distribution Shift. OPO with the IPW estimator can have poor finite-sample behavior, particularly when the support of the logging policy μ is highly non-uniform across contexts (e.g., see Jin et al. (2021) and references therein). Indeed, the standard (and essentially best) bound for IPW is that for any $\alpha\in(0,1)$, with probability at least $1-\alpha$,

$$R(\widehat{\pi}_{\mathrm{IPW}}) - R(\pi^\star) \lesssim \sqrt{V(\Pi)\cdot\tfrac{1}{N}\cdot\ln(|\Pi|/\alpha)} + \delta_{\sup}(\Pi,\mu)\cdot\tfrac{1}{N}\cdot\ln\left(|\Pi|/\alpha\right), \quad (2.1)$$

where $\lesssim$ ignores constant factors. The undesirable behavior of IPW manifests in the dependence on the worst-case IPW-variance $V(\Pi)$. In more detail, if $\pi^\star$ (or, more generally, any high-quality policy) has good coverage under the logging policy μ, its IPW-variance $V(\pi^\star)$ would be small, and we would hope that the excess risk of $\widehat{\pi}_{\mathrm{IPW}}$ would be correspondingly small. Unfortunately, this is not the case for IPW-based policy optimization; a low quality policy with large variance can significantly degrade the finite sample performance.

Known Fix: Pessimism. *Pessimism* mitigates the effects of distribution shift in OPO. One now minimizes an *upper confidence bound* (UCB) on policy risk, penalizing policies with high uncertainty. Formally, one minimizes a pessimistic estimator $\widehat{R}:\Pi\times[0,1]\to\mathbb{R}_+$ which satisfies $\Pr\left[\forall\pi\in\Pi : \widehat{R}(\pi,\alpha)\geq R(\pi)\right]\geq 1-\alpha$, where $\alpha\in(0,1)$ is a parameter.[5] This yields policy $\widehat{\pi}\in\Pi$ which w.h.p. satisfies

$$R(\widehat{\pi}) \leq \widehat{R}(\widehat{\pi},\alpha) \leq \min_{\pi\in\Pi}\widehat{R}(\pi,\alpha). \quad (2.2)$$

Thus, $R(\widehat{\pi})$ is compared to the best policy risk guaranteed by the data, under this pessimistic estimator. Here, we interpret $\widehat{R}(\pi,\alpha)$ as the "guaranteed policy risk", which tends to be smaller for policies of similar risk but better coverage in the data. Guarantees of this form are sometimes called *best-effort guarantees* (e.g., Xie et al., 2021; Jin et al., 2021).

Remark 2.2. *The key advantage of "pessimistic" guarantee* (2.2) *is that the estimator only needs to be "sharp" on* some *good policy π, regardless of how well it can estimate other policies. This suffices to guarantee that the learned policy $\widehat{\pi}$ has low risk. In particular, if the logging policy μ has good support for $\pi^\star$, the best policy, then we can expect $\widehat{\pi}$ to perform well, even if other policies are poorly supported.*

To characterize the *quality* of best-effort guarantees, one provides a data-independent upper bound for $\widehat{R}(\pi,\alpha) - R(\pi)$, and therefore for $R(\widehat{\pi}) - R(\pi)$.

Prior Implementations of Pessimism in OPO. Swaminathan and Joachims (2015a) obtain an upper confidence bound (UCB) on the policy risk using the empirical Bernstein (EB) inequality (Maurer and Pontil, 2009). Letting $\widehat{V}(\pi) := \frac{1}{N(N-1)}\sum_{1\leq i<j\leq N}\left(\widehat{\ell}_i(\pi)-\widehat{\ell}_j(\pi)\right)^2$ be the sample variance, they minimize the UCB on the policy risk,

$$\widehat{R}_{\mathrm{IPW}}(\pi) + \sqrt{\widehat{V}(\pi)\cdot\tfrac{1}{N}\cdot\ln(|\Pi|/\alpha)}. \quad (2.3)$$

This is advantageous as per Remark 2.2. The following data-independent guarantee holds: letting $\widehat{\pi}_{\mathrm{IPW+EB}}$ be

[4] Sometimes the oracle needs to handle losses that range on $\mathbb{R}_+$, so we allow this without further mention. For continuous actions, we use a standard CSC oracle that handles a finite action space, see Section 5.

[5] For implementation, it may be convenient to modify $\widehat{R}(\pi)$ so as to drop any additive "constants" that do not depend on π (as this preserves the minimizer).

the learned policy, $\forall \pi \in \Pi$,

$$R(\widehat{\pi}_{\mathrm{IPW+EB}}) - R(\pi) \lesssim \sqrt{V(\pi) \cdot \tfrac{1}{N} \cdot \ln(|\Pi|/\alpha)} + \delta_{\sup}(\Pi, \mu) \cdot \tfrac{1}{N-1} \cdot \ln(|\Pi|/\alpha). \quad (2.4)$$

It is also a best-effort guarantee (since it is obtained by upper-bounding (2.3)). The technical advantage over Eq. (2.1) is that the worst-case IPW-variance $V(\Pi)$ is replaced with policy-specific $V(\pi)$. This method outperforms the vanilla IPW approach in experiments.

However, this approach suffers from **computational inefficiency**. Particularly, the EB-based objective in Eq. (2.3): (a) does not decompose across data points so it is not amenable to streaming or stochastic optimization methods,[6] (b) yields a non-convex landscape with a differentiable policy class, and (c) does not support non-differentiable policy classes except in highly specialized cases (London et al., 2023). Note that non-differentiable policy classes are employed by a variety of methods, *e.g.,* those that train a regression model $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and induce the policy $\pi_f : x \mapsto \operatorname{argmin}_a f(x, a)$.

3 PSEUDO-LOSS

We introduce a new regularizer, dubbed *pseudo-loss* (PL), and show that it provides pessimism-style guarantees for OPO while admitting an oracle-efficient implementation. We focus on discrete actions here, i.e., when $|\mathcal{A}| < \infty$.[7] Pseudo-loss is defined as follows:

Definition 3.1. *Given a policy π, pseudo-loss $\widehat{\mathrm{PL}}(\pi)$ and its expectation are*

$$\widehat{\mathrm{PL}}(\pi) := \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \frac{\pi(a \mid x_i)}{\mu(a \mid x_i)}$$

$$\mathrm{PL}(\pi) := \mathbb{E}\left[\widehat{\mathrm{PL}}(\pi)\right] = \mathbb{E}_{x \sim \mathscr{D}}\left[\sum_{a \in \mathcal{A}} \frac{\pi(a \mid x)}{\mu(a \mid x)}\right].$$

This is well-defined by Assumption 2.1 (which we assume throughout without further mention).

We optimize this objective, parameterized by $\beta > 0$:

$$\widehat{\pi}_{\mathrm{IPW+PL},\beta} \in \operatorname*{argmin}_{\pi \in \Pi} \widehat{R}_{\mathrm{IPW}}(\pi) + \beta \cdot \widehat{\mathrm{PL}}(\pi). \quad (3.1)$$

Remark 3.2. *Our regularizer is inspired by a technique in the EXP3.P algorithm (Auer et al., 2002). This algorithm works in a very different scenario: high-probability regret bounds in online adversarial bandits. In particular, it is not concerned with contexts, offline optimization, pessimism, or oracles. Our analysis is technically different (because of the different scenario), and, in some sense, stronger: e.g., we remove the dependence on the range of β in Eq. (3.1).*

3.1 PL Implements Pessimism

Consider the objective in Eq. (3.1) plus some term Ψ_β that does not depend on policy π (so the optimization stays the same). We prove that the modified objective is an upper confidence bound on the policy risk.

We need some notation to define Ψ_β. Denote the supremum and infimum of the probability mass function (p.m.f.) induced by policy π as, resp., $\delta_{\sup}(\pi) := \sup \pi(a \mid x)$ and $\delta_{\inf}(\pi) := \inf \pi(a \mid x)$, where both extrema are over contexts $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$. Write $\delta_{\sup}(\Pi) = \sup_{\pi \in \Pi} \delta_{\sup}(\pi)$. Denote

$$\Delta(\Pi, \mu) := \max\left(\sqrt{\delta_{\sup}(\Pi)/\delta_{\inf}(\mu)},\ \delta_{\sup}(\Pi, \mu)\right),$$

where $\delta_{\sup}(\Pi, \mu)$ was defined in Section 2. Thus:

Lemma 3.3. *Fix $\alpha \in (0, 1)$. Let*

$$\Psi_\beta = \frac{O(\delta_{\sup}(\Pi)/\beta + \Delta(\Pi, \mu)) \cdot \ln(|\Pi|/\alpha)}{N}. \quad (3.2)$$

With probability at least $1 - \alpha$, the following holds for all policies $\pi \in \Pi$ and $\beta > 0$:

$$R(\pi) \le \widehat{R}_{\mathrm{IPW}}(\pi) + \beta \cdot \widehat{\mathrm{PL}}(\pi) + \Psi_\beta. \quad (3.3)$$

3.2 Oracle-Efficient Implementation

Proposition 3.4. *The optimization in Eq. (3.1) can be solved by a single call to any CSC oracle for policy class Π, with modified loss vectors*

$$a \mapsto \ell_i(a_i)/\mu(a_i \mid x_i) \cdot \mathbf{1}\{a = a_i\} + \beta/\mu(a \mid x_i).$$

In practice, we treat β as a hyper-parameter, following prior implementations of EB in Swaminathan and Joachims (2015a),[8] with the goal of selecting a near-optimal value during a subsequent policy selection step.

3.3 Performance Guarantees

Lemma 3.3 immediately implies a best-effort guarantee via (2.2). Further, we obtain best-effort guarantees

[6] However, Swaminathan and Joachims (2015a) proposed an approach to optimize this objective using stochastic gradient descent by iteratively (across epochs) optimizing an upper bound on the objective.

[7] Given context x, each policy π produces a probability mass function (p.m.f.) $\pi(\cdot \mid x)$ over the actions. We call such policies mass-based.

[8] There, the bound in Eq. (2.3) is not used directly, because it may be too loose or the exact complexity of the policy class Π may be unknown. Instead, a hyper-parameter similar to β is introduced.

that are data-independent.[9] This is advantageous as per Remark 2.2 and similar to EB.

Theorem 3.5. *Fix $\alpha \in (0,1)$. With probability at least $1-\alpha$, for any $\beta > 0$ we have*

$$R(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) \leq \min_{\pi \in \Pi} \left\{ R(\pi) + O(\Phi) \right\}, \tag{3.4}$$

where Φ equals $\beta \cdot \widehat{\mathrm{PL}}(\pi)$ plus Ψ_β from Eq. (3.2).

Further, with probability at least $1-\alpha$, for some $\beta^\star > 0$ Eq. (3.4) holds with $\beta = \beta^*$ and Φ given by

$$\sqrt{\frac{\delta_{\sup}(\Pi) \cdot \mathrm{PL}(\pi) \cdot \ln \frac{|\Pi|}{\alpha}}{N}} + \frac{\Delta(\Pi,\mu) \cdot \ln \frac{|\Pi|}{\alpha}}{N}. \tag{3.5}$$

The key advantage over EB is the oracle-efficient implementation. However, our guarantee is worse than that of EB, since each terms in Eq. (3.5) is lower-bounded by the respective term in Eq. (2.4). (This is because $V(\pi) \leq \delta_{\sup}(\Pi) \cdot \mathrm{PL}(\pi)$ for any policy π, see Prop. A.2.)

4 EMPIRICAL EVALUATION

Scope. We compare pseudo-loss (PL) to other "general" approaches for pessimistic OPO that accommodate an arbitrary oracle: Empirical Bernstein and "no pessimism". We consider two representative oracles, based, resp., on policy gradient and linear regression. The full scope of our experiments is explained below.

To keep our scope manageable, we do not compare PL to the numerous "specialized" approaches for pessimistic OPO (see Related Work). While we believe such comparisons are somewhat unfair to the general method such as ours, we leave open the possibility that some of these specialized approaches are superior for their respective policy classes. Likewise, we did not consider deep learning oracles.[10]

Experimental Setup. We simulate offline contextual bandit instances from full-information classification datasets.[11] This semi-synthetic setup gives us the ground-truth for evaluation and allows to precisely vary experimental conditions. We use four datasets from OpenML, with 1M datapoints, 14-36 real-valued features, and 6-26 classes (see Table 6 in Appendix B). For the experimental environment, we vary the following factors: dataset size, cost-type (binary- vs. real-valued), number of actions, and logging policy (see Table 1). We try all $2 \times 2 \times 2 \times 3 = 24$ possible environments. In particular, we use the technique of Foster et al. (2018) to vary the cost-type and the number of actions. We use 3 logging policies, denoted $\mu_{\mathrm{good},\epsilon}$ (resp., $\mu_{\mathrm{bad},\epsilon}$), by training good (resp., bad) policies and mixing with ϵ-probability uniform exploration.

For methods, we primarily compare three options for pessimistic regularizer: pseudo-loss (PL), Empirical Bernstein (EB), and "no regularizer" (None). Further, there are two choices for the risk estimator: inverse probability weighting (IPW) and a doubly robust estimator (DR) (Dudík et al., 2014), which is also compatible with CSC oracles. Finally, we consider two different underlying optimizers for the CSC oracle: policy gradient (PG) using a linear+softmax policy architecture and a linear regression approach (LR) where we fit a linear model to the loss for each action and define the policy to be greedy with respect to the predicted losses. (However, EB cannot accommodate the linear regression approach since the policy architecture is not differentiable and thus requires enumerating over all policies.) See Table 2 for an overview. Thus, we have 4 possible (estimator, CSC oracle) configurations for PL and None, and only 2 for EB.

We tune hyper-parameters for each method using a policy selection rule based on the EB bound in Eq. (2.3) using a 50/50 split of the data for policy optimization and selection respectively.[12] See Appendix B for hyper-parameters for each method. All results are based on 50 replicates with mean and standard errors reported.

Results. Across a wide range of experimental conditions, we find that using the PL regularizer consistently outperforms vanilla OPO with no pessimism. These results are visualized in Figure 1 (left), where we plot the relative performance improvement (`RelImp`) of PL over the baseline with no pessimism, averaged over all runs. Specifically, each curve corresponds to a particular (CSC oracle, risk estimator) pair, and represents the empirical CDF of `RelImp` across all 4×24 (dataset, environment) pairs. The median `RelImp` is 11.7%. In our experiments, PL was, essentially, *never worse* than the no-pessimism baseline.[13] We note that PL is especially helpful when the sample size is small relative to the number of actions, which is consistent with the theory that pessimism is particularly helpful in the

[9] The guarantees in Theorem 3.5 are also "best-effort guarantees", as they are obtained by upper-bounding the right-hand side in (3.3) and minimizing over all policies.

[10] Aside from implementation complexity, replicating our experiments in a similarly systematic way would require much more compute power than we had access to.

[11] This is a standard approach for contextual bandit experiments, *e.g.,* see Beygelzimer and Langford (2009); Dudík et al. (2014); Wang et al. (2017); Su et al. (2019, 2020).

[12] Policy selection is another instance of OPO with a enumerably-small policy class, with one policy for each hyper-parameter setting of the OPO method. Since the class is small, computational efficiency is less of a concern and we can use the statistically tighter EB bound.

[13] More precisely, PL prevails on 99.5% of the conditions.

Table 1: Experimental environment.

Item	Options
dataset size	{0.01, 0.1, 1}×1M
cost-type	{real-valued, binary-valued}
#actions	{#classes, 5 × #classes}
logging policy	$\{\mu_{\text{good},\epsilon=0.1}, \mu_{\text{good},\epsilon=0.01}, \mu_{\text{bad},\epsilon=0.1}\}$

Table 2: Options for the algorithms.

Item	Options
regularizer	{PL, EB, None}
estimator	{IPW, DR}
CSC oracle	{PG, LR}
EB oracle	{PG}

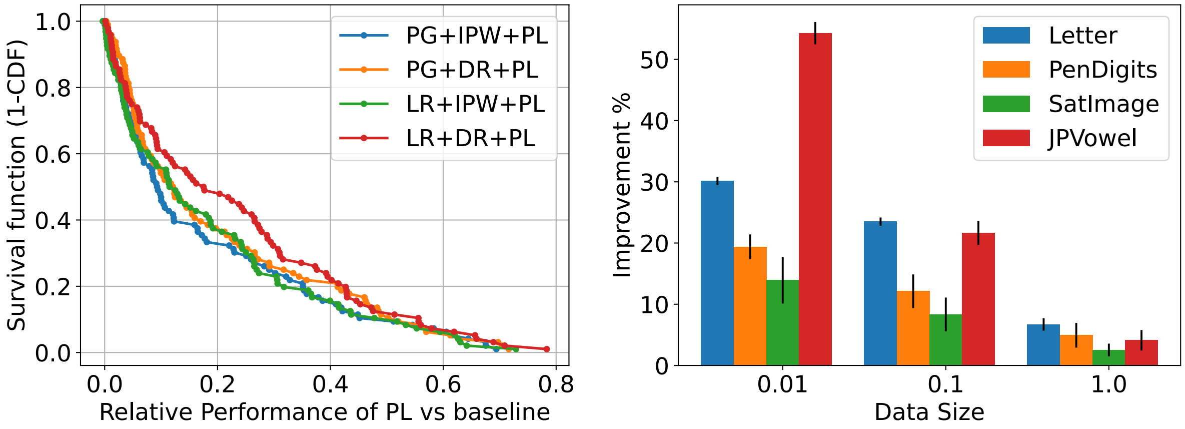

Figure 1: Relative improvement (`RelImp`) for PL against the baseline with no pessimism: mean over all runs.
Left: $1 - \text{CDF}$, the empirical cumulative density function of `RelImp` across all (dataset, environment) pairs. Each curve corresponds to (CSC oracle, risk estimator) pair.
Right: `RelImp` (mean ± 2 standard errors) for a particular (dataset, environment) pair and the best-performing (CSC oracle, risk estimator) pair. Each bar corresponds to a (dataset, dataset size) pair.
Details: 50 runs; 24 environments, as per Table 1; 4 datasets, as per Table 6.
The environment on the right is: real-valued cost, $\mu_{\text{good},\epsilon=0.1}$, and #actions = #classes.

non-asymptotic regime.

We also compare with Empirical Bernstein (EB) approach of Swaminathan and Joachims (2015a). We'd expect that EB outperforms PL statistically (due to Prop. A.2) and indeed it does: EB offers median `RelImp` of 19.1% for the configurations where it is applicable. However, EB cannot be instantiated with the linear regression (LR) optimizer. LR, when coupled with PL, sometimes yields the best overall performance, so that when selecting the best algorithm configuration for each regularizer PL beats EB on 26% of the settings. We note that our implementation of the PG optimizer for EB is an order of magnitude slower than PL.

We present detailed visualizations in Appendix B, isolating each algorithm and environment configuration. Representative visualizations are displayed in Figure 1 (right) and Table 3 (just for two datasets out of four). We also report computation times.

Suggested guidance for practitioners:

Pessimism should always be employed for offline policy optimization. If running time is not a concern and the policy architecture supports it, use empirical Bernstein; otherwise, use the PL estimator.

5 CONTINUOUS ACTIONS

We extend our approach to continuous actions, namely the one-dimensional action space $\mathcal{A} = [0, 1]$. We posit that each policy $\pi \in \Pi$ produces a probability density function (p.d.f.) $\pi(\cdot \,|\, x)$ over the action space $\mathcal{A}$ for each context x. We call such policies *density-based.*

Definition 5.1. *For a density-based policy* π*, pseudo-loss* $\widehat{\text{PL}}(\pi)$ *and its expectation are*

$$\widehat{\text{PL}}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\int_{a\in\mathcal{A}}\frac{\pi(a\,|\,x_i)}{\mu(a\,|\,x_i)}\,\text{d}a$$

$$\text{PL}(\pi) := \mathbb{E}\left[\widehat{\text{PL}}\right] = \mathbb{E}_{x\sim\mathscr{D}}\left[\int_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\,\text{d}a\right].$$

Much of our analysis seamlessly carries over to density-based policies. Specifically, Lemma A.4 and Theorem 3.5 (as well as Propositions A.2 and A.3 in the appendix) all hold for density-based policies if $\delta_{\text{sup}}(\pi)$

Table 3: Performance of different OPO methods: mean ± two standard errors over 50 runs. Bold numbers represent the best performance within each (CSC oracle, estimator) pair. Boxed numbers represent the best across all algorithmic configurations.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.1, # actions = # classes.

Risk × 100	Letter	PenDigits
PG+IPW+PL	**31.6±0.1**	22.0±0.3
PG+IPW+EB	32.4±0.1	**20.4±0.4**
PG+IPW	45.5±0.8	26.4±0.8
PG+DR+PL	$\boxed{\mathbf{31.5\pm0.1}}$	18.5±0.3
PG+DR+EB	37.0±0.4	$\boxed{\mathbf{15.7\pm0.3}}$
PG+DR	43.1±0.6	21.3±0.6
LR+IPW+PL	**31.8±0.0**	**23.1±0.3**
LR+IPW	42.5±0.4	28.4±0.7
LR+DR+PL	**31.8±0.1**	**22.9±0.2**
LR+DR	42.0±0.3	27.2±0.6

and $\delta_{\inf}(\pi)$ denote the supremum and infimum of the p.d.f induced by π. All these results are proved similarly to the discrete-action case, except the sums over $\mathcal{A}$ are replaced with integrals over $[0,1]$. Consequently, these proofs are omitted.

Next, we need to transform the learning problem. This is because (a) CSC algorithms typically can only handle finitely many actions, and (b) the variance of IPW estimator might be infinite when we consider deterministic policies (Kallus and Zhou, 2018), making learning impossible. Thus, inspired by prior work on contextual bandits with continuous actions (Krishnamurthy et al., 2020; Zhu and Mineiro, 2022), we transform the OPO problem with the original (density-based) policy class Π to a CSC problem with a mass-based policy class (denoted $\widetilde{\Pi}_K$) such that each policy in Π is a *smoothed version* of some policy in $\widetilde{\Pi}_K$.

Formalizing this requires some care. We start with a class of mass-based policies over K actions, denoted $\widetilde{\Pi}_K$, for each $K \in \mathbb{N}$. We interpret these K actions as *surrogate actions*:

$$\widetilde{\mathcal{A}}_K \coloneqq \{\widetilde{a}_i\}_{i\in[K]} = \left\{\tfrac{2i-1}{2K}\right\}_{i\in[K]} \subset \mathcal{A}.$$

Next, form density-based policies

$$\Pi_{K,H} \coloneqq \{\ \mathrm{Smooth}_H(\widetilde{\pi}) : \widetilde{\pi} \in \widetilde{\Pi}_K\ \},$$

where the smoothed policy $\mathrm{Smooth}_H(\widetilde{\pi})$ selects an action a given a context x through the following process:

$$\widetilde{a} \sim \widetilde{\pi}\left(\cdot \,|\, x\right),\ \text{then } a \sim \\ \mathrm{Uniform}\left(\left[\max\left(0, \widetilde{a} - H/2\right), \min\left(1, \widetilde{a} + H/2\right)\right]\right).$$

We call H the *bandwidth* of smoothing.[14]

We can optimize the PL-based objective in Eq. (3.1) over $\Pi = \Pi_{K,H}$ by calling a CSC oracle over $\widetilde{\Pi}_K$.

Proposition 5.2. *Fix $K, H \in \mathbb{N}$. Consider the density-based policy class $\Pi = \Pi_{K,H}$ as constructed above. Then the objective in Eq. (3.1) can be optimized via a single call to a CSC oracle for the mass-based policy class $\widetilde{\Pi}_K$, with suitably modified loss functions (details in Appendix A).*

We characterize generalization performance of pseudo-loss and smoothed policy class $\Pi = \Pi_{K,H}$. Note that $\delta_{\sup}(\pi) \le 2/H$ for any policy $\pi \in \Pi$.

Corollary 5.3. *Fix $\alpha \in (0,1)$. With probability at least $1-\alpha$, for any $\beta > 0$ we have Eq. (3.4) holds with*

$$\Phi = \beta \cdot \widehat{\mathrm{PL}}(\pi) + \frac{(1/\beta + 1/\delta_{\inf}(\mu)) \cdot \ln\left(|\Pi|/\alpha\right)}{NH}.$$

Further, with probability at least $1-\alpha$, for some $\beta^\star > 0$ Eq. (3.4) holds with $\beta = \beta^*$ and

$$\Phi = \sqrt{\frac{\mathrm{PL}(\pi) \cdot \ln(|\Pi|/\alpha)}{NH}} + \frac{\ln(|\Pi|/\alpha)}{NH\delta_{\inf}(\mu)}.$$

We also discuss how to construct the range of hyper-parameters H and K in Appendix A.7.

5.1 Empirical Evaluation

We conduct an empirical study in the continuous-action setting, with the same scope as in Section 4.

Experimental Setup. We follow (Bietti et al., 2021; Majzoubi et al., 2020; Zhu and Mineiro, 2022) to simulate continuous-action contextual bandit instances from 5 OpenML regression datasets (Vanschoren et al., 2013), with 160K-5M datapoints and 9-32 features (see Table 32 in Appendix B for details). We convert a regression example (x, y), $y \in \mathbb{R}$ to a contextual bandit example by defining the loss as $\ell(a) = |a - y|$.

For the experimental environment, we vary two factors: the dataset size and logging policy. The logging policies μ_ϵ are obtained by training a regression model on the original dataset, smoothing the prediction with bandwidth 0.1 and mixing with the uniform-at-random policy with proportion $\epsilon \in \{0.1, 0.01\}$.

The algorithmic configurations are the same as those for the discrete-action setting (the 10 options in Table 2). We optimize hyper-parameters as before, using the EB bound for policy selection. In this setup, we also consider hyper-parameters $K \in \{10, 20, 50, 100\}$ and

[14] Such "smoothing" was introduced in the online setting in Krishnamurthy et al. (2020).

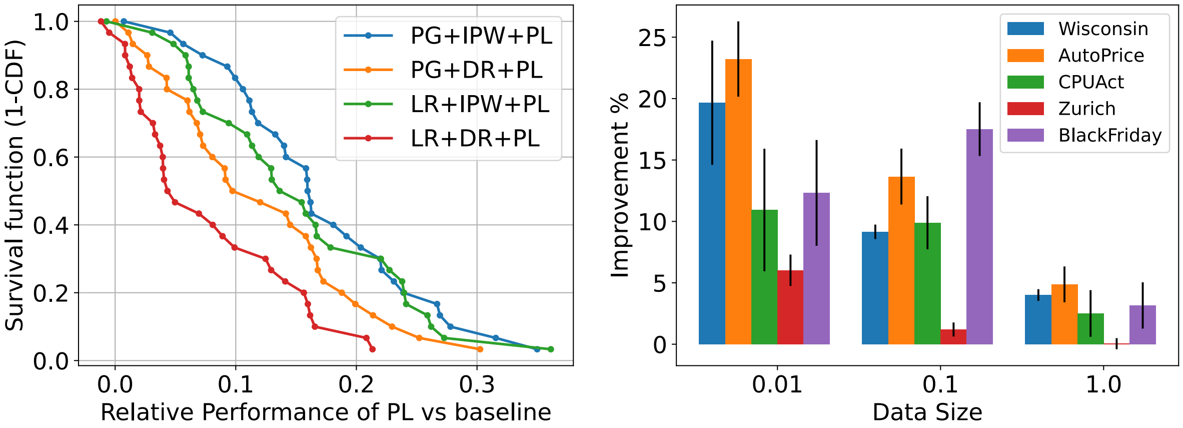

Figure 2: Same semantics as in Figure 1. Details: 10 runs, 6 environments (Table 4) and 5 datasets (Table 32.) The environment on the right has logging policy $\mu_{\epsilon=0.1}$.

Table 4: Experimental environment: continuous actions

Item	Options
dataset size	{0.01, 0.1, 1}×`ActualSize`
logging policy	μ_ϵ, $\epsilon \in \{0.1, 0.01\}$

$H \in \{0.01, 0.02, 0.05, 0.1\}$. Due to the large number of hyper-parameters and the computational overhead of EB, we only run EB for the small dataset sizes (0.01 and 0.1 fraction of the original dataset).

Results Overview. Figure 2 (left) visualizes the relative improvement (`RelImp`) over vanilla policy optimization with no pessimism. As with discrete actions, PL-based pessimism exhibits consistent significant improvement across environmental and algorithmic configurations. The median `RelImp` is 12% and we see improvement in 97.5% of the experimental conditions. In comparison, on configurations when EB is applicable, EB offers a median `RelImp` of 25.2%, while PL offers a median `RelImp` of 15.8%.

We present detailed visualizations in Appendix B, isolating each algorithm and environment like we did for discrete actions. Representative visualizations are in Figure 2 (right) and Table 5 (just for 2 datasets). We also report computation times.

6 DISCUSSION

We develop a new pessimistic approach for offline policy optimization in contextual bandits based on the pseudo-loss (PL) regularizer. The approach offers a favorable balance between computational complexity and statistical performance. It is oracle efficient and thus supports a wide range of policy classes and underlying optimization methods while offering a best-effort guarantee analogous to, but slightly worse than, prior computationally inefficient approaches. We observe this balance in our experiments, offering the guidance that pessimism should always be used and that PL should be used when computation is a concern or when sharper approaches are not applicable.

Table 5: Same semantics as in Table 3. This experiment: logging policy $\mu_{\epsilon=0.1}$, data size ×0.1, 10 runs.

Risk * 100	Wisconsin	AutoPrice
PG+IPW+PL	22.7±0.8	16.9±1.4
PG+IPW+EB	$\boxed{\textbf{21.5±0.1}}$	**14.7±0.3**
PG+IPW	26.6±1.1	20.2±0.9
PG+DR+PL	21.8±0.1	15.3±0.2
PG+DR+EB	$\boxed{\textbf{21.5±0.3}}$	$\boxed{\textbf{14.4±0.1}}$
PG+DR	24.0±0.2	18.0±0.9
LR+IPW+PL	**24.1±1.4**	**18.5±0.8**
LR+IPW	27.7±0.6	21.1±0.8
LR+DR+PL	**23.1±0.8**	**17.7±0.9**
LR+DR	26.4±0.4	18.7±0.5

Limitations. Our experimental study is semi-synthetic, transforming fully-labeled classification and regression datasets to contextual bandit instances. A standard practice in most prior work on contextual bandits, it gives us access to ground truth, but may not accurately reflect the performance in production.

Our experiments demonstrate that OPO methods are rather sensitive to a variety of factors, paralleling similar observations in Offline RL (Swaminathan and Joachims, 2015a; Joachims et al., 2018; Su et al., 2019;

Wang et al., 2021). Some factors, notably the choice of optimizer (PG vs LR), choice of estimator (IPW vs DR), and dataset quality, appear in our experimental results. In our preliminary experiments, we found that other factors may also be relevant, e.g., optimizer hyper-parameters. We did not rigorously evaluate these factors to keep the experiments at a manageable level of complexity. However, understanding whether/how our qualitative findings carry over to production environments is an important next step.

Acknowledgements

LW acknowledges support from NSF Awards IIS-1901168 and IIS-2008139. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Exponential smoothing for off-policy learning. In *International Conference on Machine Learning*, 2023.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 2016.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 2002. Preliminary version in *IEEE Symposium on Foundations of Computer Science*, 1995.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 1962.

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

A. Bietti, A. Agarwal, and J. Langford. A contextual bandit bake-off. *Journal of Machine Learning Research*, 2021.

A. Block, Y. Dagan, N. Golowich, and A. Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, 2022.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 2013.

V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, 2019.

S. Dasgupta, D. J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

H. Daumé, J. Langford, and D. Marcu. Search-based structured prediction. *Machine learning*, 2009.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2011.

M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014. Preliminary version in *International Conference on Machine Learning*, 2011.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, 2018.

D. Foster, A. Agarwal, M. Dudík, H. Luo, and R. Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2018.

S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019.

N. Haghtalab, Y. Han, A. Shetty, and K. Yang. Oracle-efficient online learning for beyond worst-case adversaries. In *Advances in Neural Information Processing Systems*, 2022.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 1952.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 2021.

Y. Jin, Z. Ren, Z. Yang, and Z. Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv:2212.09900*, 2022.

T. Joachims, A. Swaminathan, and M. De Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, 2018.

R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Journal of Machine Learning Research*, 2020.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

J. Langford and T. Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *Advances in Neural Information Processing Systems*, 2007.

J. Langford, A. Strehl, and J. Wortman. Exploration scavenging. In *International Conference on Machine Learning*, 2008.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv:2005.01643*, 2020.

G. Li, C. Ma, and N. Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. In *Advances in Neural Information Processing Systems*, 2022.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, 2020.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, 2019.

B. London, L. Lu, T. Sandler, and T. Joachims. Boosted off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, 2023.

T. Lu, D. Pál, and M. Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. In *International Conference on Artificial Intelligence and Statistics*, 2010.

M. Majzoubi, C. Zhang, R. Chari, A. Krishnamurthy, J. Langford, and A. Slivkins. Efficient contextual bandits with continuous actions. In *Advances in Neural Information Processing Systems*, 2020.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

T. Nguyen-Tang, S. Gupta, A. T. Nguyen, and S. Venkatesh. Offline neural contextual bandits: Pessimism, optimization and generalization. In *International Conference on Learning Representations*, 2022.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, Albanand Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, 2021.

P. Rashidinejad, H. Zhu, K. Yang, S. Russell, and J. Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. In *International Conference on Learning Representations*, 2022.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

O. Sakhi, P. Alquier, and N. Chopin. Pac-bayesian offline contextual bandits with guarantees. In *International Conference on Machine Learning*, 2023.

A. Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 2014. Preliminary version in *Conference on Learning Theory*, 2011.

A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`. Latest online revision: Jan 2022.

A. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, 2010.

Y. Su, L. Wang, M. Santacatterina, and T. Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, 2019.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, 2020.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 2015a. Preliminary version in *International Conference on Machine Learning*, 2015.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015b.

B. Trabucco, A. Kumar, X. Geng, and S. Levine. Conservative objective models for effective offline model-based optimization. In *International Conference on Machine Learning*, 2021.

M. Uehara and W. Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. Openml: networked science in machine learning. *SIGKDD Explorations*, 2013.

R. Wang, Y. Wu, R. Salakhutdinov, and S. Kakade. Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*, 2021.

Y.-X. Wang, A. Agarwal, and M. Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, 2017.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020.

Y. Zhu and P. Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, 2022.

Y. Zhu, D. J. Foster, P. Mineiro, and J. Langford. Contextual bandits in large action spaces: Made practical. In *International Conference on Machine Learning*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Yes
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes. We include the code in the supplemental material.
   (d) Information about consent from data providers/curators. Yes
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Appendix

The supplement consists the proofs in Appendix A and detailed empirical evaluation in Appendix B. The evaluation for discrete actions is in Appendix B.1 on p. 18, and for continuous actions in Appendix B.2 on p. 30.

A Theoretical analysis

We invoke Bennett's inequality in our analysis.

Lemma A.1 (Bennett (1962))**.** *Let $z, z_1, \ldots, z_N$ be* i.i.d. *random variables with values in* $[0, 1]$*, For any* $\alpha \in (0, 1)$*, with probability at least* $1 - \alpha$*, we have*

$$\mathbb{E}[z] - \frac{1}{N}\sum_{i=1}^{N} z_i \leq \sqrt{\frac{2\mathrm{Var}(z)\ln(1/\alpha)}{N}} + \frac{\ln(1/\alpha)}{3N}.$$

A.1 Confidence intervals

We upper-bound the IPW-variance of a given policy π in terms of $\mathrm{PL}(\pi)$, and characterize the difference between the pseudo-loss and its expectation. Then, we characterize confidence intervals for the policy risk via pseudo-loss.

Proposition A.2. *$V(\pi) \leq \delta_{\sup}(\pi) \cdot \mathrm{PL}(\pi)$ for any policy π.*

Proof.

$$\begin{aligned}
V(\pi) &= \mathbb{E}_{(x,\ell)\sim\mathscr{D}, a\sim\mu(\cdot\,|\,x)}\left[\left(\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\ell(a)\right)^2\right] - \mathbb{E}^2_{(x,\ell)\sim\mathscr{D}, a\sim\mu(\cdot\,|\,x)}\left[\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\ell(a)\right] \\
&\leq \mathbb{E}_{(x,\ell)\sim\mathscr{D}, a\sim\mu(\cdot\,|\,x)}\left[\left(\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\ell(a)\right)^2\right] \\
&\leq \mathbb{E}_{x\sim\mathscr{D}}\left[\sum_{a\in\mathcal{A}}\frac{\pi^2(a\,|\,x)}{\mu(a\,|\,x)}\right] \\
&\leq \delta_{\sup}(\pi)\mathbb{E}_{x\sim\mathscr{D}}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\right] \\
&= \delta_{\sup}(\pi)\mathrm{PL}(\pi).
\end{aligned}$$

The second inequality holds since $0 \leq \ell^2(a) \leq 1$. The third inequality holds because $\pi(a\,|\,x) \leq \delta_{\sup}(\pi)$. □

Proposition A.3. *For any policy π and any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$\tfrac{1}{2}\,\mathrm{PL}(\pi) - O\left(\frac{\ln(1/\alpha)}{N \cdot \delta_{\inf}(\mu)}\right) \leq \widehat{\mathrm{PL}}(\pi) \leq \tfrac{3}{2}\,\mathrm{PL}(\pi) + O\left(\frac{\ln(1/\alpha)}{N \cdot \delta_{\inf}(\mu)}\right).$$

Proof. We prove Proposition A.3 with exact constants, which states that for any policy π and any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,

$$\frac{1}{2}\,\mathrm{PL}(\pi) - \frac{4\ln(2/\alpha)}{3N\delta_{\inf}(\mu)} \leq \widehat{\mathrm{PL}}(\pi) \leq \frac{3}{2}\,\mathrm{PL}(\pi) + \frac{4\ln(2/\alpha)}{3N\delta_{\inf}(\mu)}.$$

First, we bound the variance of $\sum_{a\in\mathcal{A}}\pi(a\,|\,x)/\mu(a\,|\,x)$. For any $x \in \mathcal{X}$,

$$\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)} \leq \frac{1}{\delta_{\inf}(\mu)}\sum_{a\in\mathcal{A}}\pi(a\,|\,x) = \frac{1}{\delta_{\inf}(\mu)}.$$

Therefore,

$$\mathrm{Var}\left(\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\right) \leq \mathbb{E}\left[\left(\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\right)^2\right] \leq \frac{1}{\delta_{\inf}(\mu)}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\right] = \frac{\mathrm{PL}(\pi)}{\delta_{\inf}(\mu)}. \tag{A.1}$$

Applying Lemma A.1 to *i.i.d.* random variables $\left\{\sum_{a\in\mathcal{A}} \frac{\pi(a\,|\,x_i)}{\mu(a\,|\,x_i)}\right\}_{i\in[N]}$ (precisely by multiplying the random variables by $\delta_{\inf}(\mu)$ so their range is $[0,1]$, and considering both tails via a union bound), we have that for any $\alpha\in(0,1)$, with probability $1-\alpha$, it holds that

$$\begin{aligned}\left|\mathrm{PL}(\pi)-\widehat{\mathrm{PL}}(\pi)\right| &\le \sqrt{\mathrm{Var}\left(\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x)}{\mu(a\,|\,x)}\right)\ln(2/\alpha)\frac{2}{N}}+\frac{\ln(2/\alpha)}{3N\delta_{\inf}(\mu)}\\ &\le \sqrt{\frac{2\mathrm{PL}(\pi)\ln(2/\alpha)}{\delta_{\inf}(\mu)N}}+\frac{\ln(2/\alpha)}{3N\delta_{\inf}(\mu)}\\ &\le \frac{\mathrm{PL}(\pi)}{2}+\frac{\ln(2/\alpha)}{N\delta_{\inf}(\mu)}+\frac{\ln(2/\alpha)}{3N\delta_{\inf}(\mu)}\\ &= \frac{\mathrm{PL}(\pi)}{2}+\frac{4\ln(2/\alpha)}{3N\delta_{\inf}(\mu)}.\end{aligned}$$

The second inequality is Eq. (A.1). The last inequality follows from the AM-GM inequality. □

Lemma A.4. *Fix any policy π and $\alpha\in(0,1)$. With probability at least $1-\alpha$,*

$$|R(\pi)-\widehat{R}_{\mathrm{IPW}}(\pi)| \lesssim \sqrt{\tfrac{\ln(1/\alpha)}{N}\;\delta_{\sup}(\pi)\cdot\widehat{\mathrm{PL}}(\pi)}+\tfrac{\ln(1/\alpha)}{N}\;\Delta(\pi,\mu). \tag{A.2}$$

Proof. We prove a version of the lemma with exact constants: namely, Eq. (A.2) is spelled out as

$$|R(\pi)-R(\widehat{\pi}_{\mathrm{IPW}})| \le \sqrt{\tfrac{3\ln(4/\alpha)}{N}\;\delta_{\sup}(\pi)\cdot\widehat{\mathrm{PL}}(\pi)}+\tfrac{\ln(4/\alpha)}{N}\max\left(2\sqrt{\frac{8\delta_{\sup}(\pi)}{3\delta_{\inf}(\mu)}},\frac{2}{3}\delta_{\sup}(\pi,\mu)\right).$$

Applying Bennett's inequality (Lemma A.1) to *i.i.d.* random variables $\left\{\widehat{\ell}_i(\pi)\right\}$ (precisely by dividing by $\delta_{\sup}(\pi,\mu)$, and also applying a union bound to account for both tails), we have that with probability at least $1-\alpha/2$,

$$\begin{aligned}\left|R(\pi)-\widehat{R}_{\mathrm{IPW}}(\pi)\right| &\le \sqrt{\frac{2V(\pi)\ln(4/\alpha)}{N}}+\frac{\ln(4/\alpha)\delta_{\sup}(\pi,\mu)}{3N}\\ &\le \sqrt{\frac{2\delta_{\sup}(\pi)\mathrm{PL}(\pi)\ln(4/\alpha)}{N}}+\frac{\ln(4/\alpha)\delta_{\sup}(\pi,\mu)}{3N},\end{aligned}$$

where the last inequality is from Proposition A.2.

From Proposition A.3, we know that with probability at least $1-\alpha/2$,

$$\left|\mathrm{PL}(\pi)-\widehat{\mathrm{PL}}(\pi)\right| \le \frac{\mathrm{PL}(\pi)}{2}+\frac{4\ln(4/\alpha)}{3N\delta_{\inf}(\mu)}.$$

Applying union bound to both, we have that with probability at least $1-\alpha$,

$$\begin{aligned}\left|R(\pi)-\widehat{R}_{\mathrm{IPW}}(\pi)\right| &\le \sqrt{\frac{2\delta_{\sup}(\pi)\mathrm{PL}(\pi)\ln(4/\alpha)}{N}}+\frac{\ln(4/\alpha)\delta_{\sup}(\pi,\mu)}{3N}\\ &\le \sqrt{\frac{2\delta_{\sup}(\pi)\left(3/2\widehat{\mathrm{PL}}(\pi)+4\ln(4/\alpha)/(3N\delta_{\inf}(\mu))\right)\ln(4/\alpha)}{N}}+\frac{\ln(4/\alpha)\delta_{\sup}(\pi,\mu)}{3N}\\ &\le \sqrt{\frac{3\delta_{\sup}(\pi)\widehat{\mathrm{PL}}(\pi)\ln(4/\alpha)}{N}}+\sqrt{\frac{8\delta_{\sup}(\pi)}{3\delta_{\inf}(\mu)}}\frac{\ln(4/\alpha)}{N}+\frac{\ln(4/\alpha)\delta_{\sup}(\pi,\mu)}{3N},\end{aligned}$$

where the last inequality holds since $\sqrt{B_1+B_2}\le\sqrt{B_1}+\sqrt{B_2}$ for any $B_1\ge 0$ and $B_2\ge 0$. □

A.2 Proof of Lemma 3.3

We prove a version of the lemma with exact constants: namely, Eq. (3.3) is spelled out as

$$R(\pi) \le \widehat{R}_{\mathrm{IPW}}(\pi) + \beta\widehat{\mathrm{PL}}(\pi) + \frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{4\beta N} + \sqrt{\frac{8\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}}\frac{\ln(4|\Pi|/\alpha)}{N} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N}.$$

Applying the union bound to the inequality in Lemma A.4 for all policies $\pi \in \Pi$, we have that with probability at least $1-\alpha$, for any $\pi \in \Pi$, $\beta > 0$,

$$\begin{aligned}\left|R(\pi) - \widehat{R}_{\mathrm{IPW}}(\pi)\right| &\le \sqrt{\frac{3\delta_{\sup}(\pi)\widehat{\mathrm{PL}}(\pi)\ln(4|\Pi|/\alpha)}{N}} + \sqrt{\frac{8\delta_{\sup}(\pi)}{3\delta_{\inf}(\mu)}}\frac{\ln(4|\Pi|/\alpha)}{N} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\pi,\mu)}{3N}\\ &\le \beta\widehat{\mathrm{PL}}(\pi) + \frac{3\delta_{\sup}(\pi)\ln(4|\Pi|/\alpha)}{4\beta N} + \sqrt{\frac{8\delta_{\sup}(\pi)}{3\delta_{\inf}(\mu)}}\frac{\ln(4|\Pi|/\alpha)}{N} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\pi,\mu)}{3N}\\ &\le \beta\widehat{\mathrm{PL}}(\pi) + \frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{4\beta N} + \sqrt{\frac{8\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}}\frac{\ln(4|\Pi|/\alpha)}{N} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N},\end{aligned}$$

where the second inequality is derived by applying the AM-GM inequality, and the last inequality is by the definition of $\delta_{\sup}(\pi,\mu)$, $\delta_{\sup}(\Pi,\mu)$, $\delta_{\sup}(\pi)$, and $\delta_{\sup}(\Pi)$. This concludes the proof.

A.3 Proof of Proposition 3.4

Recall that the proposition states that the optimization in Eq. (3.1) can be solved by calling any CSC oracle for policy class Π once, with modified loss vectors $a \mapsto \ell_i(a_i)/\mu(a_i \mid x_i)\cdot\mathbf{1}\{a = a_i\} + \beta/\mu(a \,|\, x_i)$.

The objective in Eq. (3.1) can be re-written as

$$\begin{aligned}\widehat{R}_{\mathrm{IPW}}(\pi) + \beta\widehat{\mathrm{PL}}(\pi) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\pi(a_i\,|\,x_i)}{\mu(a_i\,|\,x_i)}\ell_i(a_i) + \beta\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}}\frac{\pi(a\,|\,x_i)}{\mu(a\,|\,x_i)}\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}}\pi(a\,|\,x_i)\left(\mathbf{1}\left\{a = a_i\right\}\ell_i(a_i)/\mu(a_i\,|\,x_i) + \beta/\mu(a\,|\,x_i)\right),\end{aligned}$$

which concludes the proof.

A.4 Proof of Theorem 3.5

We prove Theorem 3.5 with exact constants: the guarantees in the theorem statement are spelled out as, resp.,

$$R(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) \le \min_{\pi\in\Pi}\left\{R(\pi) + 2\beta\widehat{\mathrm{PL}}(\pi) + \frac{\left(\frac{3}{2}\delta_{\sup}\left(\Pi\right)/\beta + 8\Delta(\Pi,\mu)\right)\cdot\ln\left(4|\Pi|/\alpha\right)}{N}\right\}. \tag{A.3}$$

$$R(\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}) \le \min_{\pi\in\Pi}\left\{R(\pi) + \sqrt{\frac{18\delta_{\sup}(\Pi)\cdot\mathrm{PL}(\pi)\cdot\ln(4|\Pi|/\alpha)}{N}} + \frac{12\Delta(\Pi,\mu)\cdot\ln(4|\Pi|/\alpha)}{N}\right\}. \tag{A.4}$$

From the proof of Lemma 3.3, we know that with probability at least $1-\alpha$, for any $\beta > 0$ and $\pi \in \Pi$,

$$
\begin{aligned}
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) &\le \widehat{R}_{\mathrm{IPW}}(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) + \beta\widehat{\mathrm{PL}}(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) \\
&\quad + \frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{4\beta N} + \sqrt{\frac{8\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N} \\
&\le \widehat{R}_{\mathrm{IPW}}(\pi) + \beta\widehat{\mathrm{PL}}(\pi) + \frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{4\beta N} + \sqrt{\frac{8\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N} \\
&\le R(\pi) + 2\beta\widehat{\mathrm{PL}}(\pi) + \frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{2\beta N} + \sqrt{\frac{32\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{2\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N},
\end{aligned}
$$

where the first and third inequalities come from the proof of Lemma 3.3, and the second inequality is by the definition of $\widehat{\pi}_{\mathrm{IPW+PL},\beta}$. This concludes the proof for the first part of the theorem.

Since the above inequality holds for every policy $\pi \in \Pi$ and $\beta > 0$, let $\beta_\pi := \sqrt{\frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)}{4N\widehat{\mathrm{PL}}(\pi)}}$, we know that with probability at least $1-\alpha$, for any policy $\pi \in \Pi$,

$$
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta_\pi}) \le R(\pi) + 2\sqrt{\frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\widehat{\mathrm{PL}}(\pi)}{N}} + \sqrt{\frac{32\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{2\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N}.
$$

So with probability at least $1-\alpha$, there exists $\beta^\star \in \operatorname{argmin}_{\beta_\pi:\pi\in\Pi} R(\widehat{\pi}_{\mathrm{IPW+PL},\beta_\pi})$ such that

$$
\begin{aligned}
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}) &\le \min_{\pi\in\Pi} R\left(\widehat{\pi}_{\mathrm{IPW+PL},\beta_\pi}\right) \\
&\le \min_{\pi\in\Pi}\left\{R(\pi) + 2\sqrt{\frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\widehat{\mathrm{PL}}(\pi)}{N}} + \sqrt{\frac{32\delta_{\sup}(\Pi)}{3\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{2\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N}\right\}.
\end{aligned}
$$

And we know from Proposition A.3 that

$$
\begin{aligned}
\sqrt{\frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\widehat{\mathrm{PL}}(\pi)}{N}} &\le \sqrt{\frac{3\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\left(3/2\mathrm{PL}(\pi) + 4\ln(4|\Pi|/\alpha)/(3N\delta_{\inf}(\mu))\right)}{N}} \\
&\le \sqrt{\frac{9/2\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\mathrm{PL}(\pi)}{N}} + \sqrt{\frac{4\delta_{\sup}(\Pi)\ln^2(4|\Pi|/\alpha)}{N^2\delta_{\inf}(\mu)}},
\end{aligned}
$$

where the second inequality again uses the fact that $\sqrt{B_1+B_2} \le \sqrt{B_1} + \sqrt{B_2}$ for any $B_1 \ge 0$ and $B_2 \ge 0$. So we can get

$$
\begin{aligned}
&R(\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}) \\
&\quad\le \min_{\pi\in\Pi}\left\{R(\pi) + \sqrt{\frac{18\delta_{\sup}(\Pi)\ln(4|\Pi|/\alpha)\mathrm{PL}(\pi)}{N}} + 6\sqrt{\frac{\delta_{\sup}(\Pi)}{\delta_{\inf}(\mu)}\frac{\ln(4|\Pi|/\alpha)}{N}} + \frac{2\ln(4|\Pi|/\alpha)\delta_{\sup}(\Pi,\mu)}{3N}\right\},
\end{aligned}
$$

which concludes the proof.

A.5 Proof of Proposition 5.2

We prove the formal version of Proposition 5.2, stated as follows.

Proposition A.5. *Fix $K, H \in \mathbb{N}$. Consider the density-based policy class $\Pi = \Pi_{K,H}$ as constructed above. Then the objective in Eq. (3.1) can be optimized via a single call to a CSC oracle for the mass-based policy class $\widetilde{\Pi}_K$, with loss function:*

$$
\widetilde{a} \mapsto \frac{\ell_i(a_i)}{H_e\left(\widetilde{a}\right)\mu(a_i\,|\,x_i)}\mathbf{1}\left\{\widetilde{a} \in A_{K,H}(a_i)\right\} + \frac{\beta}{H_e\left(\widetilde{a}\right)}\int_{\max(0,\widetilde{a}+H/2)}^{\min(1,\widetilde{a}-H/2)} \frac{1}{\mu(a\,|\,x_i)}\,\mathrm{d}a, \tag{A.5}
$$

where $\widetilde{a} \in \widetilde{A}_K$, the latter being the surrogate-action-set, $H_e(\widetilde{a}) \coloneqq \min\left(1, \widetilde{a} + H/2\right) - \max\left(0, \widetilde{a} - H/2\right)$ is the effective bandwidth, and $A_{K,H}(a) \coloneqq \left\{\widetilde{a} \in \widetilde{\mathcal{A}}_K : \widetilde{a} \in [a - h/2, a + h/2]\right\}$ is the surrogate-action-set identity function.

With the above definitions, the objective of in Eq. (3.1) for a policy $\pi \in \Pi_{K,H}$ is

$$\begin{aligned}
&\widehat{R}_{\mathrm{IPW}}(\pi) + \beta \mathrm{PL}(\pi) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\pi(a_i \,|\, x_i)}{\mu(a_i \,|\, x_i)}\ell_i(a_i) + \beta\int_0^1 \frac{\pi(a \,|\, x_i)}{\mu(a \,|\, x_i)}\,\mathrm{d}a\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\ell_i(a_i)}{\mu(a_i \,|\, x_i)}\sum_{\widetilde{a}\in A_{K,H}(a_i)}\frac{\widetilde{\pi}\left(\widetilde{a} \,|\, x_i\right)}{H_e\left(\widetilde{a}\right)} + \beta\int_0^1 \frac{\sum_{\widetilde{a}\in A_{K,H}(a_i)}\widetilde{\pi}\left(\widetilde{a} \,|\, x_i\right)}{H_e(\widetilde{a})\mu\left(a \,|\, x_i\right)}\,\mathrm{d}a\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\widetilde{a}\in A_{K,H}(a_i)}\widetilde{\pi}(\widetilde{a} \,|\, x_i)\frac{\ell_i(a_i)}{H_e\left(\widetilde{a}\right)\mu\left(a_i \,|\, x_i\right)} + \beta\sum_{\widetilde{a}\in\widetilde{\mathcal{A}}_K}\widetilde{\pi}\left(\widetilde{a} \,|\, x_i\right)\int_{\max(0,\widetilde{a}+H/2)}^{\min(1,\widetilde{a}-H/2)}\frac{1}{H_e\left(\widetilde{a}\right)\mu(a \,|\, x_i)}\,\mathrm{d}a\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\widetilde{a}\in\widetilde{\mathcal{A}}_K}\widetilde{\pi}\left(\widetilde{a} \,|\, x_i\right)\left(\frac{\ell_i(a_i)}{H_e\left(\widetilde{a}\right)\mu(a_i \,|\, x_i)}\mathbf{1}\left\{\widetilde{a} \in A_{K,H}(a_i)\right\} + \frac{\beta}{H_e\left(\widetilde{a}\right)}\int_{\max(0,\widetilde{a}+H/2)}^{\min(1,\widetilde{a}-H/2)}\frac{1}{\mu(a \,|\, x_i)}\,\mathrm{d}a\right)\right].
\end{aligned}$$

So minimizing the objective in Eq (3.1) over Π is equivalent to minimizing the CSC objective over $\widetilde{\Pi}_K$ over the modified loss vectors defined in Eq. (A.5).

A.6 Proof of Corollary 5.3

We prove Corollary 5.3 with exact constants: the guarantees are spelled out, resp., as

$$\begin{aligned}
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta}) &\le \min_{\pi\in\Pi}\left\{R(\pi) + 2\beta\widehat{\mathrm{PL}}(\pi) + \frac{(3/\beta + 16/\delta_{\inf}(\mu))\cdot\ln\left(|4\Pi|/\alpha\right)}{NH}\right\}. \\
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}) &\le \min_{\pi\in\Pi}\left\{R(\pi) + 6\sqrt{\frac{\mathrm{PL}(\pi)\cdot\ln(4|\Pi|/\alpha)}{NH}} + \frac{24\ln(4|\Pi|/\alpha)}{NH\delta_{\inf}(\mu)}\right\}.
\end{aligned}$$

This follows by setting $\delta_{\sup}(\Pi) = 2/H$ and $\Delta(\Pi, \mu) = 2/(H\delta_{\inf}(\mu))$ in Theorem 3.5.

A.7 How to set hyper-parameters in the continuous-action setting

While our theory in Section 5 is presented for fixed hyper-parameters K, H, this appendix provides suggestions for how to choose them in practice. To this end, we analyze how K and H affect the risk of the policies in $\Pi_{K,H}$ and the generalization performance of $\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}$, similar to (Krishnamurthy et al., 2020).

How to set K for a fixed H? Let us start with some set of density-based policies, denoted Π_∞. Note that $\delta_{\sup}(\Pi_\infty)$ can be huge and the generalization guarantee in Theorem 3.5 becomes vacuous for general density-based policies. So we are not providing excess risk guarantees for this policy class. Instead, we consider a policy class smoothed from some mass-based policy class. For an integer $K > 0$, let $\widetilde{\Pi}_K = \{\mathrm{Discretize}_K(\pi) : \pi \in \Pi_\infty\}$ be the set of mass-based policies discretized from Π_∞, where $\widetilde{\pi} = \mathrm{Discretize}_K(\pi)$ is a mass-based policy such that $\widetilde{\pi}(\widetilde{a} \,|\, x) = \int_{\widetilde{a}-1/(2K)}^{\widetilde{a}+1/(2K)} \pi(a \,|\, x)\,\mathrm{d}a$ for any $x \in \mathcal{X}$ and $\widetilde{a} \in \widetilde{\mathcal{A}}_K$. Let $\Pi_{K,H} = \left\{\mathrm{Smooth}_H\left(\widetilde{\pi}\right) : \widetilde{\pi} \in \widetilde{\Pi}_K\right\}$. We have analyzed the generalization performance of $\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}$ for $\Pi_{K,H}$ for a particular K. We now want to see how different K might affect the policy risks in $\Pi_{K,H}$. In particular, we consider $\Pi_{\infty,H} = \{\mathrm{Smooth}_H(\pi) : \pi \in \Pi_\infty\}$ be the set of density-based policies smoothed from Π_∞, which represents the policy class when K approaches infinity. We know that, for any $\pi \in \Pi_\infty$,

$$\left|R\left(\mathrm{Smooth}_H\left(\mathrm{Discretize}_K\left(\pi\right)\right)\right) - R\left(\mathrm{Smooth}_H\left(\pi\right)\right)\right| \le \min\left(1, \frac{1}{HK}\right).$$

To analyze how to set K, we consider the excess risk of $\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}$ in Eq (A.4)

$$\begin{aligned}
R(\widehat{\pi}_{\mathrm{IPW+PL},\beta^\star}) &\lesssim \min_{\pi\in\Pi_{K,H}} \left\{ R(\pi) + \sqrt{\frac{\mathrm{PL}(\pi)\cdot\ln(|\Pi_{K,H}|/\alpha)}{NH}} + \frac{\ln(|\Pi_{K,H}|/\alpha)}{NH\delta_{\mathrm{inf}}(\mu)} \right\} \\
&\leq \min_{\pi\in\Pi_{K,H}} \left\{ R(\pi) + \sqrt{\frac{\ln(|\Pi_{K,H}|/\alpha)}{NH\delta_{\mathrm{inf}}(\mu)}} + \frac{\ln(|\Pi_{K,H}|/\alpha)}{NH\delta_{\mathrm{inf}}(\mu)} \right\} \\
&\leq \min_{\pi\in\Pi_{\infty,H}} \left\{ R(\pi) + \sqrt{\frac{\ln(|\Pi_{K,H}|/\alpha)}{NH\delta_{\mathrm{inf}}(\mu)}} + \frac{\ln(|\Pi_{K,H}|/\alpha)}{NH\delta_{\mathrm{inf}}(\mu)} + \frac{1}{HK} \right\},
\end{aligned}$$

where the second inequality holds since $\mathrm{PL}(\pi) \leq 1/\delta_{\mathrm{inf}}(\mu)$. Now, if $|\Pi_{K,H}|$ scales exponentially with K, then we should set K on the order of $\left(\frac{N\delta_{\mathrm{inf}}(\mu)}{H\ln(1/\alpha)}\right)^{1/3}$ to optimize the second and fourth terms. If we assume $|\Pi_{K,H}|$ does not depend on K, then we should set K to be sufficiently large so that the fourth term is lower order.

How to choose H? For the sake of intuition, consider a fixed K, and let $\Pi_{K,H}$ and $\Pi_{K,H+\gamma}$ be density-based policy classes smoothed from the same mass-based policy class $\widetilde{\Pi}_K$ with bandwidth H and $H+\gamma$ respectively. For any mass-based policy $\widetilde{\pi} \in \Pi_K$, we have

$$|R(\mathrm{Smooth}_H(\widetilde{\pi})) - R(\mathrm{Smooth}_{H+\gamma}(\widetilde{\pi}))| \leq \min\left(1, \frac{2\gamma}{H}\right).$$

This suggests that we might want to search over a space of H such that $1/H$ is equally spaced.

B Detailed empirical evaluation

B.1 Experiments with discrete actions

Experimental setup. Following prior works (Beygelzimer and Langford, 2009; Dudík et al., 2014; Wang et al., 2017; Su et al., 2019, 2020), we empirically examine the performance of policy optimization with the PL regularizer on simulated bandit instances from full-information classification datasets. This allows us to evaluate the performance of different policy optimization methods with ground-truth policy values and to control the setting of the problem so that we can test the robustness of PL in a variety of experimental scenarios.

Datasets. We conduct experiments on 4 multi-class classification datasets with real-valued features and 1 million examples from OpenML (Vanschoren et al., 2013), see Table 6 for detailed statistics.

Table 6: Discrete action datasets

Dataset	Letter	PenDigits	SatImage	JPVowel
OpenML ID	247	261	1183	1214
# Data	1,000,000	1,000,000	1,000,000	1,000,000
# Features	16	16	36	14
# Classes	26	10	6	9

For each dataset, we hold out 1% of the data for training a logging policy and 30% of the data for testing. The rest of the data are used for simulating the bandit feedback.

We test the performance of different policy optimization algorithms under various settings, which are summarized in Table 1. First, we vary the data size by randomly selecting 1%, 10%, or 100% of the 69% of the data to simulate bandit feedback data. To vary the number of actions and the type of cost, we follow Foster et al. (2018) to transform the original multi-class classification dataset to a CSC dataset that can have either real-valued or binary-valued cost and arbitrary number of classes. For each dataset with K classes, we construct a cost matrix $C \in \mathbb{R}^{K_{\text{cs}} \times K}$ where $K_{\text{cs}} \in \{K, 5K\}$. Each entry $C(a, a^\star)$ is the cost of classifying an example with true class label $a^\star$ as class a in the set of K_{cs} actions. The entries $C(iK + a, a)$ for any integer $i \geq 0$ such that $iK + a \leq K_{\text{cs}}$ and $a \in \{1, 2, \ldots, K\}$ are set to be 0. For the binary-valued-cost experiments, the rest of the entries are set to be 1. For the real-valued-cost experiments, the rest of the entries are generated uniformly at random from the interval $[0, 1]$. To simulate bandit feedback, for each example (x, c) in the CSC dataset, we take an action a following a logging policy μ and observe the binary loss $\ell(a) \sim \text{Bernoulli}(c(a))$ where $c(a)$ is the cost of action a for x.

Logging policies. We consider three different types of logging policies. For each dataset, we learn one "good" and one "bad" deterministic multi-class classification models with the 1% held out data using the linear-regression CSC oracle described below to predict the class with the smallest and largest cost, respectively. Then we construct three stochastic logging policies $\mu_{\text{good},\epsilon=0.1}$, $\mu_{\text{good},\epsilon=0.01}$, and $\mu_{\text{bad},\epsilon=0.1}$ by combining the deterministic policies with the uniform-random policy where $\epsilon = 0.1$ and $\epsilon = 0.01$ are the probabilities of using the uniform-random policy.

OPO methods. We compare the performance of using PL regularizer with that of no pessimism (None) and EB under different estimators and oracles, which are summarized in Table 2. For all the regularizers, we consider two types of estimators: IPW and the doubly robust estimator (DR) (Dudík et al., 2014). For PL and None, we run experiments on two types of CSC oracles. The first one is policy gradient (PG) with a softmax-linear parametrization, which selects actions proportional to $\exp(\langle \theta, \phi(x, a) \rangle)$ where θ is the policy parameter and $\phi(x, a)$ are the features. The policy parameters are fit by directly optimizing the CSC objective with ℓ_2 regularization. The second CSC oracle is based on linear regression with ℓ_2 regularization and we denote it as LR. The policy is derived by regressing the costs onto the features using (regularized) least squares regression and then taking the action with the minimum predicted cost. As we have discussed, EB is not compatible with CSC oracles in general. We follow prior works and parameterize the policy identically to the PG-based CSC oracle and directly optimize the EB objective.

Policy selection. We split the bandit feedback data and use 50% of the data for policy optimization and 50% of the data for policy selection. For policy selection, we adopt the strategy using the EB bound in Eq. (2.4) with

$\alpha = 0.1$, since it is tighter than the PL bound as shown in Eq. (2.3). We run each experiment 50 times and report the mean and standard error of the results.

Hyper-parameter details. We shift the loss to the range $[-1, 0]$, since this improves the performance on all the methods in our exploratory experiments, which is also consistent with findings in many prior works (Swaminathan and Joachims, 2015a; Joachims et al., 2018; Bietti et al., 2021).[15] For policy optimization with the DR estimator, we further split the data for policy optimization into 10% for training the cost regression model and 90% for policy optimization. The cost model of DR is trained using linear regression with ℓ_2 regularization on the 10% of bandit feedback data for policy optimization. For both PL and EB, we grid search β in $\{0, 1e-3, 3e-3, 1e-2, 3e-2, 1e-1, 3e-1, 1\}$. For CSC oracles, we set the weight decay to be $1e-6$ and grid search the learning rate in $\{1e-3, 1e-2, 1e-1, 1, 10\}$. And we use stochastic gradient descent with batch size 100 and epoch 1 to optimize each model. For the PG oracle for EB, we optimize the model with LBFGS for 10 steps since Swaminathan and Joachims (2015a) found that LBFGS performs better than gradient descent in their empirical evaluation. We note that this optimizer is around 20 times slower than the CSC oracles optimized via batch stochastic gradient descent. We use the same weight decay and grid search the same learning rates as in CSC oracles. We implement all the oracles in PyTorch (Paszke et al., 2019). And the experiments are run on a shared cluster with different types of CPUs and thousands of CPU cores.

Results. We conduct experiments on all combinations of data sizes, cost types and logging policies with the number of actions being the number of classes in Table 1. For the setting where the number of actions is 5 times the number of classes, we still do experiments on all combinations of data sizes, cost types, but only with logging policy $\pi_{\text{good},\epsilon=0.1}$. The results are shown in Figures 3-10 and Tables 8-31. Their semantics mirror those of Figure 1 (right) and Table 3, respectively. Relative improvement of a method against a baseline is defined as

$$\texttt{RelImp} := \frac{R\left(\widehat{\pi}_{\text{baseline}}\right) - R\left(\widehat{\pi}_{\text{method}}\right)}{R\left(\widehat{\pi}_{\text{baseline}}\right)}. \tag{B.1}$$

The total training time on each dataset is summarized in Table 7.

Table 7: Computation time for the discrete-action experiments

Dataset	Letter	PenDigits	SatImage	JPVowel
total time (CPU core · hours)	730.8	424.2	417.9	397.0

[15] We note that our theoretical results continue to apply with the loss range. In particular, the proofs of Proposition A.2, A.3, Lemma A.4 and Theorem 3.5 still hold when $-1 \le \ell(a) \le 0$.

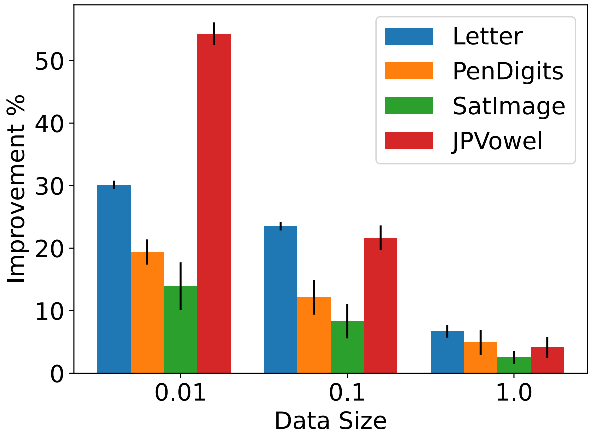

Figure 3: Relative improvement (`RelImp`, see Eq. (B.1)) for PL against the baseline with no pessimism, averaged over all 50 runs (mean ± 2 standard errors).

Shown for a particular (dataset, environment) pair and the best-performing (CSC oracle, risk estimator) pair. Each bar corresponds to a (dataset, data-size) pair.

24 environments, see Table 1; 4 datasets, see Table 6.

Environment: real-valued cost, $\mu_{\text{good},\epsilon=0.1}$, and #actions = #classes.

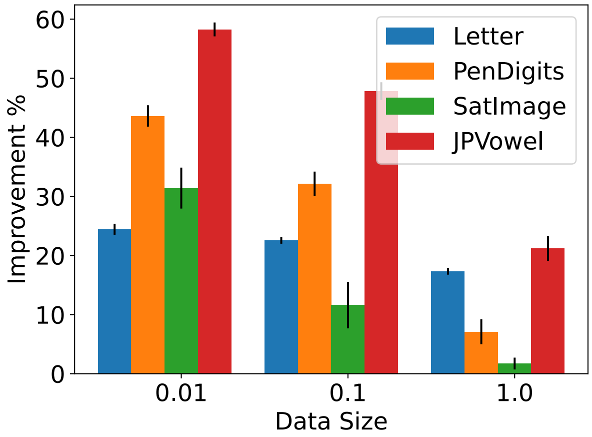

Figure 4: Same semantics as in Figure 3. Environment: real-valued cost, $\mu_{\text{good},\epsilon=0.01}$, and # actions = # classes.

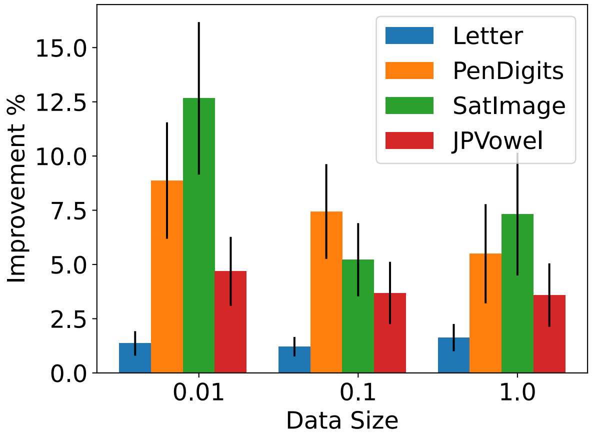

Figure 5: Same semantics as in Figure 3. Environment: real-valued cost, $\mu_{\text{bad},\epsilon=0.1}$, and # actions = # classes.

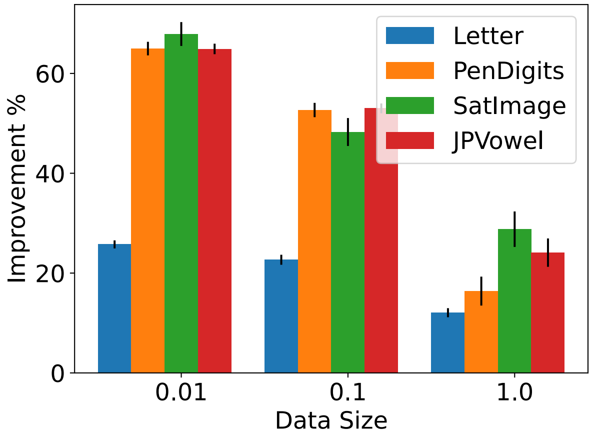

Figure 6: Same semantics as in Figure 3. Environment: real-valued cost, $\mu_{\text{good},\epsilon=0.1}$, and # actions = 5 × # classes.

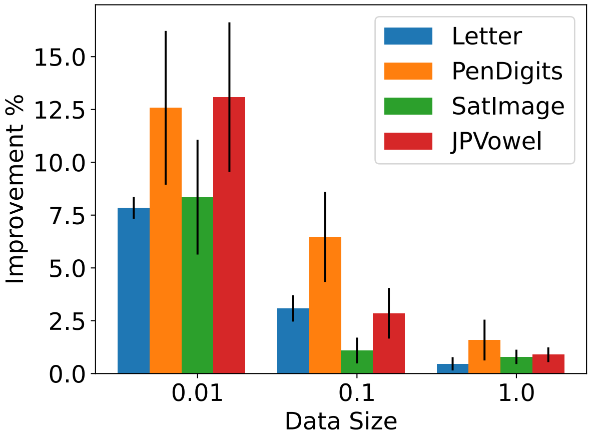

Figure 7: Same semantics as in Figure 3. Environment: real-valued cost, $\mu_{\text{good},\epsilon=0.1}$, and # actions = # classes.

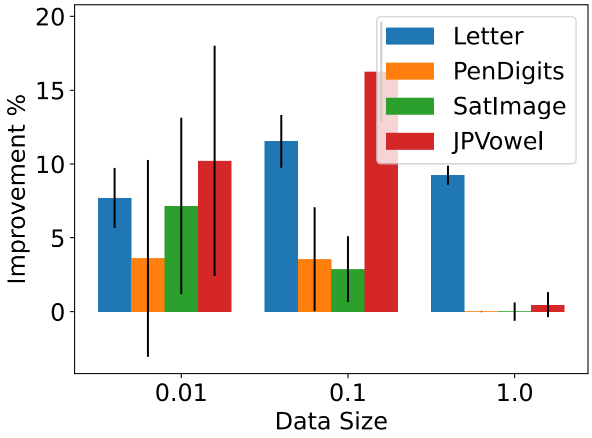

Figure 8: Same semantics as in Figure 3. Environment: binary-valued cost, $\mu_{\text{good},\epsilon=0.01}$, and # actions = # classes.

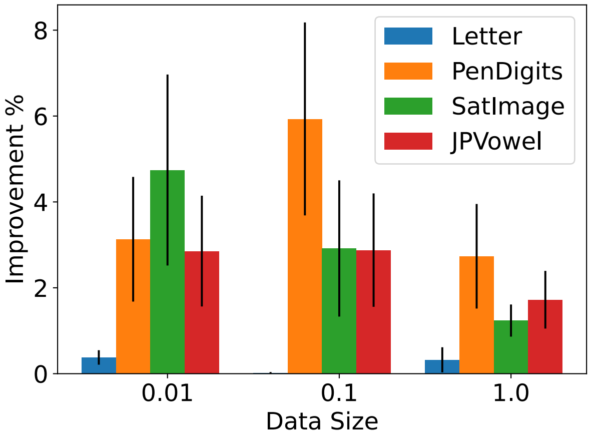

Figure 9: Same semantics as in Figure 3. Environment: binary-valued cost, $\mu_{\text{bad},\epsilon=0.1}$, and # actions = # classes.

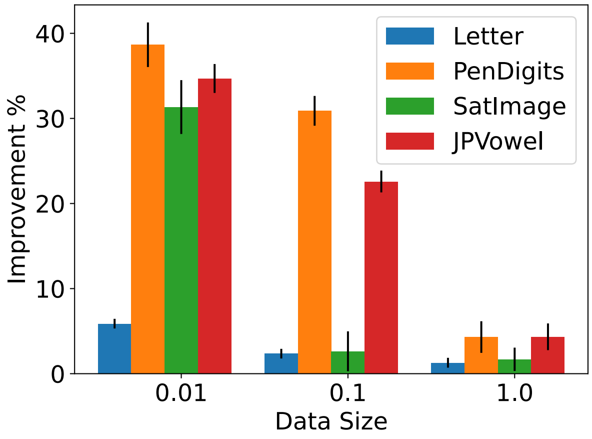

Figure 10: Same semantics as in Figure 3. Environment: binary-valued cost, $\mu_{\text{good},\epsilon=0.1}$, and # actions = 5×# classes.

Table 8: Performance of different OPO methods: mean ± two standard errors over 50 runs. Bold numbers represent the best performance within each (CSC oracle, estimator) pair. Boxed numbers represent the best across all algorithmic configurations.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	$\boxed{\mathbf{31.6\pm0.1}}$	23.8±0.2	**17.6±0.9**	**12.0±0.1**
PG+IPW+EB	34.0±0.9	**23.4±0.2**	18.1±0.8	13.4±0.3
PG+IPW	47.1±0.6	31.3±1.0	22.4±1.6	27.3±1.4
PG+DR+PL	$\boxed{\mathbf{31.6\pm0.1}}$	23.6±0.3	13.4±0.5	**12.1±0.1**
PG+DR+EB	40.1±0.4	$\boxed{\mathbf{23.0\pm0.6}}$	$\boxed{\mathbf{13.1\pm0.4}}$	20.1±0.8
PG+DR	47.6±0.6	31.2±1.1	16.5±1.0	29.0±1.7
LR+IPW+PL	**32.0±0.0**	**23.6±0.2**	**21.3±0.8**	$\boxed{\mathbf{11.2\pm0.0}}$
LR+IPW	46.2±0.5	32.5±0.8	26.7±1.2	30.4±1.0
LR+DR+PL	**32.0±0.1**	**23.6±0.2**	**21.3±0.8**	$\boxed{\mathbf{11.2\pm0.0}}$
LR+DR	46.4±0.4	34.1±0.6	25.7±1.1	32.8±1.0

Table 9: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.1, and # actions = # classes.

Risk × 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	**31.6±0.1**	22.0±0.3	**10.2±0.4**	**11.9±0.1**
PG+IPW+EB	32.4±0.1	**20.4±0.4**	11.0±0.3	13.1±0.1
PG+IPW	45.5±0.8	26.4±0.8	11.9±0.7	19.0±0.8
PG+DR+PL	$\boxed{\mathbf{31.5\pm0.1}}$	18.5±0.3	$\boxed{\mathbf{8.8\pm0.3}}$	**11.3±0.1**
PG+DR+EB	37.0±0.4	$\boxed{\mathbf{15.7\pm0.3}}$	$\boxed{\mathbf{8.8\pm0.2}}$	11.7±0.2
PG+DR	43.1±0.6	21.3±0.6	10.1±0.4	14.2±0.4
LR+IPW+PL	**31.8±0.0**	**23.1±0.3**	**18.1±0.4**	$\boxed{\mathbf{11.1\pm0.0}}$
LR+IPW	42.5±0.4	28.4±0.7	21.0±0.8	19.8±0.5
LR+DR+PL	**31.8±0.1**	**22.9±0.2**	**17.8±0.5**	$\boxed{\mathbf{11.1\pm0.0}}$
LR+DR	42.0±0.3	27.2±0.6	19.5±0.7	18.5±0.5

Table 10: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	**31.0±0.1**	15.8±0.3	**7.4±0.1**	10.5±0.2
PG+IPW+EB	31.7±0.2	**13.2±0.2**	7.7±0.1	**9.9±0.2**
PG+IPW	35.1±0.4	16.9±0.5	8.3±0.4	11.7±0.3
PG+DR+PL	30.7±0.2	14.2±0.3	6.9±0.1	9.3±0.2
PG+DR+EB	$\boxed{\mathbf{29.9\pm0.1}}$	$\boxed{\mathbf{10.2\pm0.1}}$	$\boxed{\mathbf{6.3\pm0.0}}$	$\boxed{\mathbf{8.0\pm0.1}}$
PG+DR	33.0±0.4	15.1±0.4	7.2±0.1	9.8±0.3
LR+IPW+PL	**31.5±0.1**	**20.8±0.2**	**12.8±0.3**	**11.1±0.0**
LR+IPW	36.2±0.3	21.9±0.4	15.0±0.7	12.5±0.2
LR+DR+PL	**31.5±0.1**	**20.4±0.2**	**11.1±0.2**	**11.0±0.0**
LR+DR	35.7±0.2	21.2±0.3	12.1±0.5	11.8±0.1

Table 11: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	34.2±0.1	**21.5±0.2**	**24.3±0.5**	16.1±0.1
PG+IPW+EB	$\boxed{\mathbf{33.7\pm0.3}}$	22.0±0.2	25.6±0.4	$\boxed{\mathbf{16.0\pm0.0}}$
PG+IPW	46.8±0.9	39.8±1.7	39.7±2.7	41.6±1.5
PG+DR+PL	**34.3±0.1**	**21.6±0.1**	24.2±0.5	**16.1±0.1**
PG+DR+EB	36.8±0.4	24.8±0.8	$\boxed{\mathbf{18.8\pm0.4}}$	22.0±0.9
PG+DR	47.3±0.8	42.9±1.4	42.9±2.4	43.5±1.4
LR+IPW+PL	**33.8±0.1**	$\boxed{\mathbf{21.0\pm0.1}}$	**25.6±0.6**	**16.6±0.1**
LR+IPW	46.1±0.6	40.7±1.3	44.4±1.7	41.3±1.2
LR+DR+PL	**33.9±0.1**	$\boxed{\mathbf{21.0\pm0.1}}$	**25.8±0.4**	**16.7±0.1**
LR+DR	46.8±0.5	43.5±1.1	48.0±1.7	44.1±0.9

Table 12: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×0.1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	**33.7±0.1**	**21.5±0.2**	**23.0±0.6**	$\boxed{\mathbf{15.9\pm0.0}}$
PG+IPW+EB	36.0±0.8	22.5±0.4	24.4±0.6	16.2±0.1
PG+IPW	45.1±0.7	33.8±1.2	31.0±2.0	33.1±1.2
PG+DR+PL	$\boxed{\mathbf{33.6\pm0.1}}$	**21.3±0.2**	21.6±0.6	$\boxed{\mathbf{15.9\pm0.0}}$
PG+DR+EB	36.6±0.6	22.8±0.7	$\boxed{\mathbf{18.1\pm0.4}}$	21.1±1.0
PG+DR	45.1±0.5	33.0±1.1	26.2±1.7	33.9±1.5
LR+IPW+PL	**34.0±0.1**	**21.0±0.1**	**25.1±0.4**	**16.4±0.0**
LR+IPW	45.0±0.6	35.2±0.9	36.8±1.5	35.6±1.0
LR+DR+PL	**34.0±0.1**	$\boxed{\mathbf{20.9\pm0.1}}$	**25.3±0.3**	**16.4±0.1**
LR+DR	45.2±0.5	38.0±0.8	37.2±1.2	37.3±1.0

Table 13: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	$\boxed{\mathbf{33.6\pm0.1}}$	20.8±0.3	19.3±0.4	**15.9±0.1**
PG+IPW+EB	36.0±0.6	**20.3±0.4**	**19.1±0.4**	**15.9±0.1**
PG+IPW	43.7±0.7	25.5±0.9	20.4±0.7	24.8±0.8
PG+DR+PL	$\boxed{\mathbf{33.6\pm0.1}}$	19.8±0.3	18.2±0.4	$\boxed{\mathbf{15.3\pm0.1}}$
PG+DR+EB	35.8±0.4	$\boxed{\mathbf{16.7\pm0.3}}$	$\boxed{\mathbf{15.9\pm0.3}}$	$\boxed{\mathbf{15.3\pm0.3}}$
PG+DR	42.9±0.6	21.8±0.8	18.8±0.6	19.6±0.6
LR+IPW+PL	**33.7±0.1**	**20.5±0.1**	**24.9±0.4**	**16.3±0.0**
LR+IPW	41.6±0.4	28.5±0.7	34.6±1.6	25.8±0.7
LR+DR+PL	**33.9±0.1**	**20.7±0.1**	**24.6±0.4**	**16.2±0.1**
LR+DR	41.1±0.3	28.1±0.7	34.3±1.7	25.1±0.6

Table 14: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	44.1±0.5	31.1±1.1	**16.7±1.2**	**27.4±0.7**
PG+IPW+EB	**43.8±0.4**	**31.0±0.8**	17.9±1.0	28.6±0.9
PG+IPW	45.3±0.6	34.4±1.5	20.5±1.7	29.2±0.9
PG+DR+PL	43.9±0.6	29.8±1.1	$\boxed{\mathbf{13.0\pm0.7}}$	27.1±0.8
PG+DR+EB	$\boxed{\mathbf{43.0\pm0.4}}$	$\boxed{\mathbf{28.0\pm0.8}}$	15.1±0.6	$\boxed{\mathbf{25.0\pm0.8}}$
PG+DR	45.6±0.9	33.1±1.1	15.4±0.9	29.5±1.0
LR+IPW+PL	**43.9±0.4**	**34.2±0.8**	**21.9±0.9**	**29.5±0.7**
LR+IPW	44.6±0.5	35.9±0.9	25.4±1.5	31.0±0.9
LR+DR+PL	**43.7±0.5**	**34.7±0.7**	**21.1±0.7**	**29.8±0.7**
LR+DR	44.2±0.5	35.7±0.8	23.5±1.1	32.0±1.0

Table 15: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×0.1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	42.2±0.5	23.6±0.6	**10.1±0.3**	21.7±0.6
PG+IPW+EB	**40.2±0.3**	**20.0±0.5**	11.8±0.3	**19.1±0.4**
PG+IPW	43.7±0.6	26.3±1.0	11.4±0.5	24.0±0.9
PG+DR+PL	41.1±0.5	21.3±0.8	$\boxed{\mathbf{9.2\pm0.2}}$	18.1±0.4
PG+DR+EB	$\boxed{\mathbf{38.9\pm0.3}}$	$\boxed{\mathbf{17.8\pm0.5}}$	10.1±0.3	$\boxed{\mathbf{17.2\pm0.4}}$
PG+DR	43.0±0.8	23.8±0.8	10.0±0.4	19.2±0.6
LR+IPW+PL	**40.2±0.4**	**27.1±0.6**	**17.0±0.4**	**22.5±0.5**
LR+IPW	40.9±0.4	28.3±0.8	18.8±0.7	24.0±0.7
LR+DR+PL	**39.6±0.3**	**25.7±0.7**	**16.6±0.6**	**20.3±0.4**
LR+DR	40.1±0.3	26.8±0.8	18.5±0.8	21.1±0.5

Table 16: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	35.3±0.4	12.8±0.4	7.9±0.2	13.6±0.2
PG+IPW+EB	**32.6±0.2**	**12.0±0.2**	**7.6±0.1**	**13.0±0.2**
PG+IPW	36.1±0.5	14.2±0.7	8.4±0.2	14.3±0.3
PG+DR+PL	33.0±0.4	11.9±0.3	6.8±0.2	12.9±0.2
PG+DR+EB	$\boxed{\mathbf{30.5\pm0.1}}$	$\boxed{\mathbf{10.9\pm0.1}}$	$\boxed{\mathbf{6.5\pm0.1}}$	$\boxed{\mathbf{11.0\pm0.1}}$
PG+DR	33.6±0.4	12.9±0.4	7.4±0.3	13.5±0.3
LR+IPW+PL	**35.4±0.2**	**19.6±0.4**	**11.7±0.4**	**15.9±0.2**
LR+IPW	35.5±0.2	20.9±0.6	13.4±0.7	16.3±0.2
LR+DR+PL	**35.0±0.2**	**19.0±0.3**	**10.5±0.2**	**15.0±0.2**
LR+DR	35.2±0.2	20.3±0.4	11.6±0.3	15.3±0.2

Table 17: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.01, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	$\boxed{\mathbf{33.3\pm0.2}}$	$\boxed{\mathbf{11.1\pm0.2}}$	**8.1±0.3**	**11.8±0.2**
PG+IPW+EB	40.4±1.0	16.9±1.6	9.8±0.9	17.1±1.9
PG+IPW	47.2±0.7	35.2±1.6	28.2±2.1	37.1±1.3
PG+DR+PL	**33.4±0.2**	$\boxed{\mathbf{11.1\pm0.1}}$	**8.1±0.3**	**11.9±0.2**
PG+DR+EB	36.0±0.4	15.8±0.6	9.9±0.5	15.0±0.5
PG+DR	47.4±0.9	38.2±1.6	31.3±2.3	39.6±1.1
LR+IPW+PL	**35.0±0.2**	**11.9±0.1**	$\boxed{\mathbf{7.4\pm0.1}}$	**11.7±0.1**
LR+IPW	45.4±0.5	32.7±1.1	28.2±1.4	33.0±1.0
LR+DR+PL	**33.5±0.1**	**12.2±0.2**	**7.6±0.1**	$\boxed{\mathbf{11.5\pm0.1}}$
LR+DR	47.2±0.5	39.6±1.1	35.5±1.3	39.7±0.8

Table 18: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.1, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	**33.5±0.1**	**11.4±0.1**	**7.3±0.1**	**11.1±0.1**
PG+IPW+EB	41.7±0.6	17.8±1.4	8.4±0.2	15.0±1.6
PG+IPW	45.5±1.0	30.4±1.5	18.5±1.3	32.0±1.5
PG+DR+PL	**33.3±0.1**	**11.3±0.1**	$\boxed{\mathbf{7.0\pm0.2}}$	$\boxed{\mathbf{10.9\pm0.1}}$
PG+DR+EB	40.1±0.3	14.6±0.6	8.4±0.2	14.2±0.6
PG+DR	45.6±0.8	27.0±1.4	15.2±1.2	29.2±1.6
LR+IPW+PL	**33.1±0.1**	$\boxed{\mathbf{11.1\pm0.0}}$	**7.6±0.1**	**11.6±0.0**
LR+IPW	43.7±0.3	25.2±0.7	20.3±0.7	24.2±0.6
LR+DR+PL	$\boxed{\mathbf{33.0\pm0.0}}$	$\boxed{\mathbf{11.1\pm0.0}}$	**7.7±0.1**	**11.6±0.0**
LR+DR	45.0±0.3	26.7±0.6	23.3±1.1	26.3±0.5

Table 19: Same semantics as in Table 8.
This experiment: real-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×1, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	**33.3±0.1**	**10.9±0.2**	**6.8±0.2**	**10.9±0.1**
PG+IPW+EB	34.9±0.2	12.6±0.1	7.9±0.1	11.0±0.0
PG+IPW	40.6±0.7	16.6±0.8	12.3±0.8	18.0±0.7
PG+DR+PL	32.7±0.2	10.2±0.2	5.8±0.2	10.7±0.1
PG+DR+EB	**32.4±0.3**	$\boxed{\mathbf{9.6\pm0.1}}$	$\boxed{\mathbf{5.6\pm0.1}}$	$\boxed{\mathbf{9.9\pm0.1}}$
PG+DR	37.5±0.7	12.3±0.4	8.4±0.4	14.3±0.6
LR+IPW+PL	**32.7±0.1**	**11.0±0.1**	**7.5±0.1**	**11.4±0.1**
LR+IPW	38.6±0.3	18.9±0.3	13.4±0.3	18.1±0.3
LR+DR+PL	$\boxed{\mathbf{32.2\pm0.1}}$	**10.9±0.1**	**7.9±0.1**	**11.4±0.0**
LR+DR	37.5±0.3	20.9±0.6	13.6±0.3	18.9±0.4

Table 20: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	83.5±0.1	43.0±1.0	25.2±0.8	36.8±0.5
PG+IPW+EB	$\boxed{\mathbf{76.9\pm0.6}}$	**31.8±0.8**	**22.9±0.4**	**30.7±0.4**
PG+IPW	91.2±0.6	49.1±2.1	27.0±0.8	43.0±2.0
PG+DR+PL	84.2±0.3	44.5±0.5	26.4±1.0	38.0±0.7
PG+DR+EB	**83.6±0.8**	$\boxed{\mathbf{31.2\pm1.2}}$	$\boxed{\mathbf{21.4\pm0.5}}$	$\boxed{\mathbf{29.8\pm0.7}}$
PG+DR	94.0±0.4	66.2±2.3	33.2±2.3	71.1±1.5
LR+IPW+PL	**84.6±0.2**	**44.5±0.7**	**34.8±1.1**	**38.7±0.3**
LR+IPW	91.8±0.4	59.5±1.8	40.4±1.4	57.3±1.7
LR+DR+PL	**85.1±0.5**	**46.3±0.3**	**37.0±1.1**	**40.2±0.3**
LR+DR	92.9±0.4	77.7±1.2	52.3±2.8	77.0±1.2

Table 21: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	82.7±0.1	26.4±0.4	19.8±0.1	29.6±0.5
PG+IPW+EB	**74.7±1.4**	**24.8±0.7**	**19.5±0.1**	$\boxed{\mathbf{26.2\pm0.5}}$
PG+IPW	85.5±0.6	28.7±0.9	20.1±0.2	30.5±0.5
PG+DR+PL	82.7±0.1	29.1±0.9	19.7±0.1	31.3±0.5
PG+DR+EB	$\boxed{\mathbf{69.5\pm0.9}}$	$\boxed{\mathbf{23.8\pm0.5}}$	$\boxed{\mathbf{18.8\pm0.2}}$	**26.7±0.4**
PG+DR	87.8±0.5	34.7±1.6	20.9±0.6	33.4±0.8
LR+IPW+PL	**82.1±0.5**	**32.0±0.5**	**26.5±0.4**	**31.8±0.4**
LR+IPW	85.2±0.4	34.4±0.9	30.0±1.1	32.9±0.5
LR+DR+PL	**83.6±0.3**	**41.9±0.6**	**31.8±0.6**	**37.4±0.3**
LR+DR	87.1±0.4	53.1±1.6	38.1±1.9	49.3±1.0

Table 22: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	76.5±0.5	22.3±0.2	18.7±0.1	23.5±0.1
PG+IPW+EB	**66.4±2.5**	$\boxed{\mathbf{20.3\pm0.1}}$	**16.8±0.1**	$\boxed{\mathbf{21.9\pm0.1}}$
PG+IPW	76.9±0.6	23.0±0.5	18.9±0.1	23.8±0.2
PG+DR+PL	77.1±0.5	22.3±0.1	18.5±0.1	23.3±0.1
PG+DR+EB	$\boxed{\mathbf{64.4\pm1.6}}$	**20.4±0.1**	$\boxed{\mathbf{16.6\pm0.0}}$	$\boxed{\mathbf{21.9\pm0.0}}$
PG+DR	77.5±0.5	23.8±0.6	18.7±0.1	23.8±0.3
LR+IPW+PL	**78.2±0.4**	**27.0±0.2**	**24.2±0.3**	**28.9±0.2**
LR+IPW	81.0±0.5	27.2±0.2	25.3±0.3	29.0±0.3
LR+DR+PL	**80.0±0.4**	**29.1±0.2**	**23.9±0.4**	**31.0±0.2**
LR+DR	83.2±0.5	29.7±0.3	25.4±0.4	31.4±0.2

Table 23: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	78.1±0.5	48.6±1.1	31.9±1.0	38.8±0.5
PG+IPW+EB	$\boxed{\mathbf{69.3\pm1.1}}$	$\boxed{\mathbf{38.4\pm0.6}}$	**28.3±0.4**	**33.9±1.3**
PG+IPW	85.4±2.2	53.4±4.2	35.8±2.8	48.1±4.8
PG+DR+PL	83.6±1.7	53.7±1.8	33.6±0.6	45.3±3.2
PG+DR+EB	**77.7±1.1**	**38.5±2.9**	$\boxed{\mathbf{25.5\pm1.2}}$	$\boxed{\mathbf{31.7\pm1.5}}$
PG+DR	95.3±0.3	84.0±1.7	61.2±3.9	85.0±1.1
LR+IPW+PL	**81.3±0.1**	**53.2±1.0**	**41.0±1.2**	**46.5±1.4**
LR+IPW	91.7±0.7	69.5±3.5	55.8±3.1	66.5±4.0
LR+DR+PL	**86.4±1.5**	**57.2±2.3**	**40.7±1.4**	**52.6±2.6**
LR+DR	95.3±0.3	83.8±1.4	72.4±2.5	84.2±1.1

Table 24: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×0.1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	76.4±0.3	49.3±0.3	27.4±1.1	38.9±0.2
PG+IPW+EB	$\boxed{\mathbf{69.9\pm0.5}}$	**32.4±1.2**	**24.5±0.5**	**31.3±0.8**
PG+IPW	87.3±1.7	51.0±1.8	27.6±0.9	47.5±1.9
PG+DR+PL	**77.0±0.3**	49.4±0.5	27.8±1.0	39.1±0.2
PG+DR+EB	77.7±1.6	$\boxed{\mathbf{26.9\pm1.6}}$	$\boxed{\mathbf{19.2\pm0.5}}$	$\boxed{\mathbf{26.1\pm0.5}}$
PG+DR	94.2±0.4	64.3±1.8	31.3±1.9	67.3±1.9
LR+IPW+PL	**80.9±0.1**	**51.2±0.8**	**33.2±1.1**	**44.4±0.6**
LR+IPW	91.1±1.0	61.8±1.9	39.1±1.4	56.0±2.2
LR+DR+PL	**81.9±0.5**	**52.6±0.4**	**36.2±0.9**	**47.3±0.6**
LR+DR	93.5±0.3	74.2±1.4	47.1±1.7	76.1±1.4

Table 25: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.01}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	75.7±0.2	28.6±0.8	20.4±0.3	29.4±0.4
PG+IPW+EB	**66.4±1.3**	**24.6±0.7**	**19.7±0.1**	$\boxed{\mathbf{25.6\pm0.3}}$
PG+IPW	84.6±0.7	28.6±0.7	20.6±0.4	29.4±0.4
PG+DR+PL	76.0±0.2	35.4±1.2	20.5±0.6	31.6±0.6
PG+DR+EB	$\boxed{\mathbf{66.0\pm0.9}}$	$\boxed{\mathbf{24.1\pm0.5}}$	$\boxed{\mathbf{18.4\pm0.2}}$	**26.2±0.4**
PG+DR	90.0±0.7	37.2±1.7	21.0±0.7	35.5±1.1
LR+IPW+PL	**80.5±0.3**	**34.9±1.0**	**27.9±0.6**	34.0±0.6
LR+IPW	84.7±0.6	35.0±1.0	29.4±0.8	**33.9±0.6**
LR+DR+PL	**80.9±0.2**	**45.7±1.0**	**31.2±0.8**	**42.1±0.5**
LR+DR	89.0±0.4	52.3±1.5	41.5±2.1	51.4±1.2

Table 26: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×0.01, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	92.1±0.3	60.7±1.6	31.7±1.0	62.2±1.1
PG+IPW+EB	$\boxed{\mathbf{90.1\pm0.5}}$	$\boxed{\mathbf{46.3\pm1.4}}$	**27.6±0.9**	**49.5±1.3**
PG+IPW	92.5±0.4	63.8±1.7	35.8±2.0	64.4±1.3
PG+DR+PL	92.4±0.3	61.9±1.5	31.6±1.1	63.1±1.5
PG+DR+EB	**90.6±0.5**	**46.4±1.6**	$\boxed{\mathbf{26.6\pm0.9}}$	$\boxed{\mathbf{49.4\pm1.1}}$
PG+DR	93.0±0.4	64.4±1.7	34.4±1.9	66.0±1.6
LR+IPW+PL	**90.3±0.5**	**67.2±1.3**	**40.9±1.2**	**67.2±1.2**
LR+IPW	90.7±0.5	69.8±1.6	45.6±2.0	69.6±1.6
LR+DR+PL	**90.3±0.5**	**66.5±1.5**	**43.1±1.5**	**66.2±1.4**
LR+DR	90.8±0.5	69.9±1.9	48.7±1.9	70.6±1.6

Table 27: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×0.1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	88.1±0.4	34.9±1.1	21.4±0.6	34.4±0.5
PG+IPW+EB	**83.8±1.7**	**30.3±1.0**	**21.0±0.2**	**31.9±0.9**
PG+IPW	88.3±0.5	37.8±1.4	22.3±0.7	36.3±0.7
PG+DR+PL	88.2±0.4	34.3±0.9	21.3±0.5	35.2±0.7
PG+DR+EB	$\boxed{\mathbf{83.0\pm1.5}}$	$\boxed{\mathbf{28.4\pm0.7}}$	$\boxed{\mathbf{20.3\pm0.2}}$	$\boxed{\mathbf{31.6\pm1.1}}$
PG+DR	88.4±0.5	38.8±1.5	22.4±0.7	36.1±0.9
LR+IPW+PL	**85.7±0.4**	**41.1±1.0**	**31.4±0.7**	**40.6±0.7**
LR+IPW	85.9±0.4	42.5±1.2	34.5±1.6	41.9±0.8
LR+DR+PL	**86.0±0.4**	**41.1±1.0**	**30.5±0.7**	**41.4±0.7**
LR+DR	86.1±0.4	42.5±1.3	35.0±1.6	42.6±0.9

Table 28: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{bad},\epsilon=0.1}$, data size ×1, and # actions = # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	77.7±0.6	23.0±0.2	18.6±0.1	25.0±0.2
PG+IPW+EB	$\boxed{\mathbf{63.5\pm1.9}}$	$\boxed{\mathbf{21.1\pm0.1}}$	**17.0±0.0**	$\boxed{\mathbf{22.5\pm0.1}}$
PG+IPW	77.9±0.6	24.5±0.6	19.0±0.1	25.6±0.4
PG+DR+PL	78.8±0.5	23.2±0.3	18.7±0.1	25.1±0.2
PG+DR+EB	**66.7±2.5**	**21.2±0.1**	$\boxed{\mathbf{16.9\pm0.1}}$	$\boxed{\mathbf{22.5\pm0.1}}$
PG+DR	79.3±0.6	24.6±0.5	19.0±0.1	26.1±0.4
LR+IPW+PL	**81.5±0.5**	**28.3±0.3**	**25.2±0.3**	**30.1±0.2**
LR+IPW	81.6±0.5	28.9±0.4	25.5±0.3	30.3±0.3
LR+DR+PL	**82.4±0.5**	**28.7±0.3**	**25.1±0.3**	**29.8±0.3**
LR+DR	82.6±0.5	29.0±0.4	25.5±0.3	30.1±0.4

Table 29: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.01, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	86.3±0.4	39.3±2.4	29.9±1.1	40.8±1.9
PG+IPW+EB	$\boxed{\mathbf{80.2 \pm 2.4}}$	$\boxed{\mathbf{31.5 \pm 1.7}}$	$\boxed{\mathbf{24.5 \pm 0.8}}$	$\boxed{\mathbf{32.0 \pm 1.1}}$
PG+IPW	92.4±0.4	68.5±1.9	43.2±2.3	69.9±1.9
PG+DR+PL	**89.3±0.9**	**43.4±3.9**	31.0±1.8	**42.6±3.3**
PG+DR+EB	90.2±1.4	44.3±3.2	**26.8±1.3**	49.9±2.8
PG+DR	93.8±0.4	81.5±2.0	57.0±3.4	82.7±1.5
LR+IPW+PL	**87.9±0.5**	**39.1±1.1**	**29.1±0.9**	**43.2±1.1**
LR+IPW	91.5±0.5	57.2±1.7	47.6±2.6	58.6±1.2
LR+DR+PL	**88.0±0.7**	**45.1±3.6**	**29.4±0.9**	**44.9±2.0**
LR+DR	92.4±0.4	79.0±1.4	68.0±2.6	78.9±1.1

Table 30: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×0.1, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	87.0±0.4	30.4±0.6	26.8±0.5	34.0±0.4
PG+IPW+EB	$\boxed{\mathbf{73.1 \pm 1.2}}$	**28.5±0.8**	**21.9±0.5**	$\boxed{\mathbf{30.9 \pm 0.5}}$
PG+IPW	90.5±0.6	56.0±1.5	28.4±1.1	55.2±1.3
PG+DR+PL	86.9±0.4	30.4±0.6	26.5±0.5	34.2±0.5
PG+DR+EB	**84.5±2.2**	$\boxed{\mathbf{27.7 \pm 0.7}}$	$\boxed{\mathbf{21.3 \pm 0.5}}$	**31.7±0.7**
PG+DR	92.2±0.4	69.6±2.2	29.4±1.1	66.9±2.1
LR+IPW+PL	**85.0±0.3**	**34.4±0.5**	**28.5±0.7**	**39.3±0.4**
LR+IPW	87.6±0.5	42.6±0.9	28.9±0.8	43.1±0.7
LR+DR+PL	**85.4±0.3**	**34.8±0.3**	**29.2±0.5**	**40.2±0.6**
LR+DR	88.0±0.5	61.3±1.3	41.2±1.7	57.5±1.0

Table 31: Same semantics as in Table 8.
This experiment: binary-valued cost, logging policy $\mu_{\text{good},\epsilon=0.1}$, data size ×1, and # actions = 5× # classes.

Risk * 100	Letter	PenDigits	SatImage	JPVowel
PG+IPW+PL	85.3±0.4	**25.9±0.5**	**19.7±0.2**	31.3±0.4
PG+IPW+EB	**70.1±3.0**	26.9±0.7	20.6±0.4	**30.3±0.5**
PG+IPW	87.9±0.6	27.2±0.7	20.2±0.5	33.2±0.7
PG+DR+PL	85.5±0.4	25.8±0.4	$\boxed{\mathbf{19.2 \pm 0.1}}$	31.3±0.4
PG+DR+EB	$\boxed{\mathbf{67.0 \pm 2.5}}$	$\boxed{\mathbf{24.6 \pm 0.5}}$	19.3±0.2	$\boxed{\mathbf{27.5 \pm 0.6}}$
PG+DR	88.8±0.5	29.7±1.0	20.4±0.7	33.2±0.5
LR+IPW+PL	**75.7±0.6**	**29.8±0.4**	**25.9±0.2**	**32.9±0.5**
LR+IPW	76.6±0.5	31.5±0.7	26.0±0.2	33.8±0.5
LR+DR+PL	**76.4±0.5**	**33.0±0.2**	**26.5±0.2**	**37.2±0.3**
LR+DR	77.5±0.5	47.0±1.1	26.9±0.2	44.2±0.6

B.2 Experiments with continuous actions

Datasets. For the continuous-action setting, we follow prior works (Bietti et al., 2021; Majzoubi et al., 2020; Zhu and Mineiro, 2022) to simulate bandit instances using 5 regression datasets from OpenML (Vanschoren et al., 2013), see Table 32 for details.

Table 32: Continuous action datasets

Dataset	Wisconsin	AutoPrice	CpuAct	Zurich	BlackFriday
OpenML ID	1187	1189	1190	40753	44057
# Data	1,000,000	1,000,000	1,000,000	5,465,575	166,821
# Features	32	15	21	14	9

We use one-hot representations for categorical features and map the regression targets to $[0, 1]$. We adopt the same split of the datasets as in the discrete-action experiments for training logging policies, simulating bandit feedback, and testing performance. And we still consider three data sizes the same as the discrete-action experiments.

To simulate bandit feedback, for each example (x, y) from the regression dataset, where y is the regression target, we take an action a following a logging policy μ, and observe the loss $\ell(a) = |a - y|$.

Logging policies. We consider logging policies that are combinations of a policy smoothed from a deterministic policy and a policy that selects actions uniformly at random. To learn a deterministic policy, we train a linear regression model with ℓ_2 regularization to predict the regression target on the 1% for held out data, and regard the regression estimate clipped into $[0, 1]$ as the taken action. Then we construct stochastic logging policies $\mu_{\epsilon=0.1}$ and $\mu_{\epsilon=0.01}$ by smoothing the deterministic policy with bandwidth 0.1, and combing it with a uniformly at random policy where $\epsilon = 0.1$ and $\epsilon = 0.01$ represent the probabilities of using the uniformly-at-random policy. To summarize, for environment setting, we have data size in $\{X0.01, X0.1, X1\}$ logging policy in $\{\mu_{\epsilon=0.1}, \mu_{\epsilon=0.01}\}$.

Methods. All the regularizers, oracles, and estimators are the same as those of the discrete-action experiments as shown in Table 2. Since EB takes much longer time to run due to its reliance on going through the whole dataset multiple times, we only run experiments for EB on data sizes $\times 0.01$ and $\times 0.1$.

Hyper-parameter details. For the continuous-action experiments, we run each experiment for 10 times. We grid search K in $[10, 20, 50, 100]$, and H in $[1e-2, 2e-2, 5e-2, 1e-1]$. For the continuous-action experiments, we grid search in a smaller set of learning rates $[1e-4, 1e-3, 1e-2, 1e-1]$. All the other hyper-parameters are the same as those of the discrete-action experiments.

Results. We conduct experiments on all combinations of data sizes and logging policies. The experiment results are illustrated in Figures 11-12 and Tables 34-39, with semantics mirroring those of of Figure 1 (right) and Table 3, respectively.

Table 33: Computation time for the continuous-action experiments

Dataset	Wisconsin	AutoPrice	CpuAct	Zurich	BlackFriday
total time (core $\cdot$ hours)	93.7	88.8	92.5	499.9	19.3

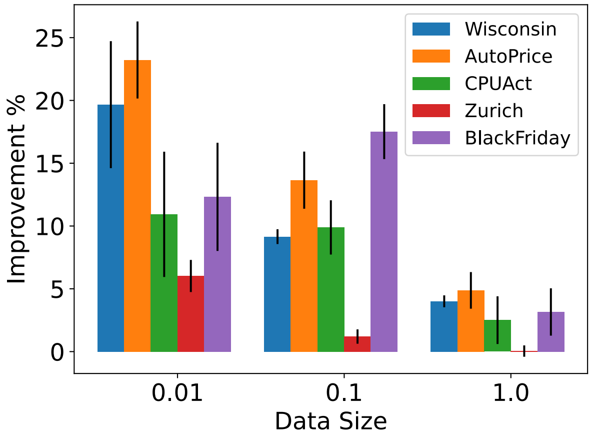

Figure 11: Relative improvement (`RelImp`, see (B.1)) for PL against the baseline with no pessimism, averaged over all 10 runs (mean ± 2 standard errors).

Shown for a particular (dataset, environment) pair and the best-performing (CSC oracle, risk estimator) pair. Each bar corresponds to a (dataset, data-size) pair.

6 environments, see Table 4; 5 datasets, see Table 32.

Environment: logging policy $\mu_{\epsilon=0.1}$.

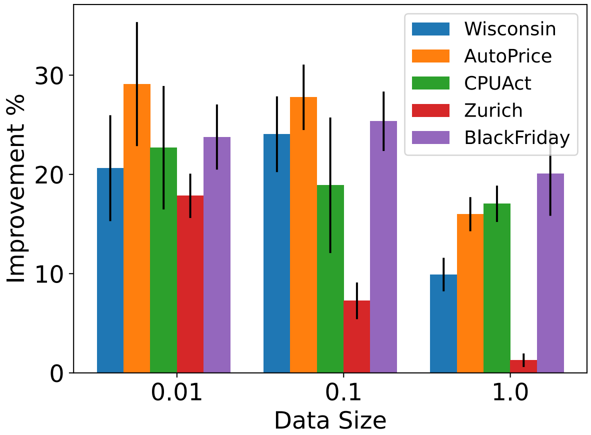

Figure 12: Same semantics as in Figure 11. Environment: logging policy $\mu_{\epsilon=0.01}$.

Table 34: Performance of different OPO methods: mean ± two standard errors over 10 runs. Bold numbers represent the best performance within each (CSC oracle, estimator) pair. Boxed numbers represent the best across all algorithmic configurations.
This experiment: logging policy $\mu_{\epsilon=0.1}$, and data size ×0.01.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	26.2±2.1	18.8±1.6	20.4±1.8	**24.6±0.3**	29.8±1.5
PG+IPW+EB	**22.4±1.4**	**15.1±1.2**	**16.4±1.7**	25.1±1.1	$\boxed{\mathbf{22.7\pm0.6}}$
PG+IPW	30.3±1.0	23.5±2.1	21.4±1.1	27.7±1.1	33.0±1.4
PG+DR+PL	30.5±3.6	21.5±2.3	19.6±1.1	24.4±0.1	31.2±2.4
PG+DR+EB	$\boxed{\mathbf{21.4\pm0.1}}$	$\boxed{\mathbf{14.5\pm0.2}}$	$\boxed{\mathbf{14.8\pm0.3}}$	$\boxed{\mathbf{24.3\pm0.0}}$	**24.9±2.0**
PG+DR	33.0±1.9	24.6±1.4	23.0±1.7	26.0±0.4	33.6±0.7
LR+IPW+PL	**25.5±2.4**	**19.4±1.5**	**20.1±1.2**	**25.8±1.9**	**30.2±2.6**
LR+IPW	30.6±1.0	25.4±2.0	24.0±1.5	30.6±1.3	32.4±1.0
LR+DR+PL	**32.6±2.9**	**22.5±3.6**	**25.8±2.7**	$\boxed{\mathbf{24.3\pm0.0}}$	**33.0±2.7**
LR+DR	34.1±1.2	24.6±1.8	26.3±1.4	29.1±1.0	33.5±1.5

Table 35: Same semantics as in Table 34.
This experiment: logging policy $\mu_{\epsilon=0.1}$, and data size ×0.1.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	22.7±0.8	16.9±1.4	17.8±1.6	**24.4±0.1**	25.0±1.8
PG+IPW+EB	$\boxed{\mathbf{21.5\pm0.1}}$	**14.7±0.3**	**14.6±0.2**	**24.4±0.1**	**22.9±1.1**
PG+IPW	26.6±1.1	20.2±0.9	19.9±1.1	26.3±0.7	30.0±1.2
PG+DR+PL	21.8±0.1	15.3±0.2	15.8±0.6	$\boxed{\mathbf{24.3\pm0.0}}$	24.8±2.2
PG+DR+EB	$\boxed{\mathbf{21.5\pm0.3}}$	$\boxed{\mathbf{14.4\pm0.1}}$	$\boxed{\mathbf{14.5\pm0.1}}$	24.4±0.1	$\boxed{\mathbf{22.5\pm0.4}}$
PG+DR	24.0±0.2	18.0±0.9	17.3±0.3	24.6±0.1	30.9±2.2
LR+IPW+PL	**24.1±1.4**	**18.5±0.8**	**18.9±1.2**	**24.4±0.0**	**24.7±0.8**
LR+IPW	27.7±0.6	21.1±0.8	20.2±1.1	26.9±0.4	31.8±1.3
LR+DR+PL	**23.1±0.8**	**17.7±0.9**	19.5±1.6	**24.4±0.0**	**27.1±3.1**
LR+DR	26.4±0.4	18.7±0.5	**19.3±0.7**	25.2±0.2	33.0±2.5

Table 36: Same semantics as in Table 34.
This experiment: logging policy $\mu_{\epsilon=0.1}$, and data size ×1.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	**21.9±0.2**	**15.3±0.5**	**15.5±0.3**	**24.4±0.1**	**22.8±0.3**
PG+IPW	24.3±0.3	18.2±0.4	17.7±0.4	24.6±0.1	26.5±0.8
PG+DR+PL	$\boxed{\mathbf{21.5\pm0.0}}$	$\boxed{\mathbf{14.7\pm0.1}}$	$\boxed{\mathbf{14.8\pm0.1}}$	$\boxed{\mathbf{24.3\pm0.1}}$	$\boxed{\mathbf{22.6\pm0.3}}$
PG+DR	22.5±0.1	15.6±0.2	15.2±0.3	$\boxed{\mathbf{24.3\pm0.1}}$	23.2±0.5
LR+IPW+PL	**22.8±0.2**	**16.9±0.4**	17.3±0.5	**24.4±0.0**	**24.7±0.8**
LR+IPW	24.3±0.2	17.8±0.6	**17.2±0.6**	25.1±0.4	26.3±0.6
LR+DR+PL	**22.2±0.1**	**15.3±0.2**	**15.4±0.3**	$\boxed{\mathbf{24.3\pm0.0}}$	**23.0±0.4**
LR+DR	22.7±0.3	15.6±0.2	15.7±0.3	24.5±0.0	23.9±0.3

Table 37: Same semantics as in Table 34.
This experiment: logging policy $\mu_{\epsilon=0.01}$, and data size ×0.01.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	22.6±0.3	16.1±0.2	18.8±2.3	$\boxed{\mathbf{24.4\pm0.0}}$	24.2±0.2
PG+IPW+EB	$\boxed{\mathbf{21.3\pm0.0}}$	$\boxed{\mathbf{14.6\pm0.2}}$	$\boxed{\mathbf{14.5\pm0.1}}$	$\boxed{\mathbf{24.4\pm0.0}}$	$\boxed{\mathbf{22.6\pm0.1}}$
PG+IPW	29.9±2.7	25.8±3.5	25.7±1.5	30.8±1.8	33.6±1.2
PG+DR+PL	30.9±2.6	20.9±3.0	23.9±4.9	28.8±4.1	26.0±1.6
PG+DR+EB	**25.5±2.2**	**18.1±1.6**	**18.5±1.4**	**25.5±2.1**	**25.9±1.5**
PG+DR	32.3±1.7	27.2±3.0	29.6±5.7	34.3±2.2	34.7±1.4
LR+IPW+PL	**23.1±0.3**	**17.2±0.5**	**19.0±1.5**	$\boxed{\mathbf{24.4\pm0.1}}$	**24.1±0.2**
LR+IPW	31.7±2.4	27.6±3.3	26.5±1.7	33.1±1.3	31.8±1.3
LR+DR+PL	**30.8±3.2**	**24.6±3.1**	**26.0±3.7**	**31.4±3.2**	**27.4±2.5**
LR+DR	33.5±2.2	26.8±1.7	33.5±3.1	34.8±1.3	34.8±1.4

Table 38: Same semantics as in Table 34.
This experiment: logging policy $\mu_{\epsilon=0.01}$, and data size ×0.1.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	22.4±0.9	16.6±1.6	18.1±1.8	25.1±0.9	23.0±0.1
PG+IPW+EB	$\boxed{\mathbf{21.3\pm0.0}}$	$\boxed{\mathbf{14.4\pm0.0}}$	$\boxed{\mathbf{14.3\pm0.1}}$	$\boxed{\mathbf{24.4\pm0.0}}$	$\boxed{\mathbf{22.5\pm0.1}}$
PG+IPW	29.6±1.9	24.4±1.4	22.8±1.7	28.4±1.1	31.6±1.7
PG+DR+PL	27.4±2.9	23.5±4.0	19.8±1.8	**24.5±0.4**	27.6±4.2
PG+DR+EB	**22.7±1.8**	**16.8±2.2**	**18.3±2.4**	24.9±0.7	**25.8±2.8**
PG+DR	34.5±2.2	25.3±2.3	24.2±2.8	26.4±0.6	35.7±1.8
LR+IPW+PL	**23.5±1.4**	**20.5±1.7**	**20.3±1.6**	**24.8±0.4**	**24.8±0.8**
LR+IPW	31.0±0.9	23.9±1.3	23.3±2.3	30.3±1.2	32.6±1.1
LR+DR+PL	**33.2±2.2**	**25.6±3.3**	**25.2±4.6**	**24.5±0.1**	**31.6±4.0**
LR+DR	33.6±1.4	25.7±1.5	26.6±2.3	29.3±1.1	36.5±2.5

Table 39: Same semantics as in Table 34.
This experiment: logging policy $\mu_{\epsilon=0.01}$, and data size ×1.

Risk * 100	Wisconsin	AutoPrice	CPUAct	Zurich	BlackFriday
PG+IPW+PL	**23.1±1.6**	**16.3±1.5**	**15.4±0.7**	$\boxed{\mathbf{24.4\pm0.1}}$	**24.8±2.4**
PG+IPW	27.5±1.3	21.1±1.6	20.0±1.3	25.9±0.5	29.7±2.1
PG+DR+PL	$\boxed{\mathbf{21.7\pm0.1}}$	$\boxed{\mathbf{14.8\pm0.2}}$	$\boxed{\mathbf{14.6\pm0.1}}$	$\boxed{\mathbf{24.4\pm0.0}}$	$\boxed{\mathbf{22.9\pm0.3}}$
PG+DR	24.0±0.4	17.8±0.6	17.7±0.5	24.7±0.3	33.2±2.2
LR+IPW+PL	**22.9±0.4**	**18.8±0.9**	**18.7±0.3**	**24.9±0.8**	**26.0±1.7**
LR+IPW	27.5±0.8	21.2±0.8	20.2±1.1	26.9±0.6	29.9±0.8
LR+DR+PL	**22.7±0.4**	**18.0±1.0**	**18.9±0.8**	**24.5±0.0**	**27.7±2.5**
LR+DR	27.0±1.1	18.6±0.5	19.8±1.0	25.5±0.2	33.0±2.5