Imitation-Regularized Offline Learning

Yifei Ma
yifeim@amazon.com

Yu-Xiang Wang
yuxiangw@cs.ucsb.edu[1]

Balakrishnan (Murali) Narayanaswamy
muralibn@amazon.com

Abstract

We study the problem of offline learning in automated decision systems under the contextual bandits model. We are given logged historical data consisting of contexts, (randomized) actions, and (nonnegative) rewards. A common goal is to evaluate what would happen if different actions were taken in the same contexts, so as to optimize the action policies accordingly. The typical approach to this problem, inverse probability weighted estimation (IPWE) [5], requires logged action probabilities, which may be missing in practice due to engineering complications. Even when available, small action probabilities cause large uncertainty in IPWE, rendering the corresponding results insignificant. To solve both problems, we show how one can use policy improvement (PIL) objectives, regularized by policy imitation (IML). We motivate and analyze PIL as an extension to Clipped-IPWE, by showing that both are lower-bound surrogates to the vanilla IPWE. We also formally connect IML to IPWE variance estimation [31] and natural policy gradients. Without probability logging, our PIL-IML interpretations justify and improve, by reward-weighting, the state-of-art cross-entropy (CE) loss that predicts the action items among all action candidates available in the same contexts. With probability logging, our main theoretical contribution connects IML-underfitting to the existence of either confounding variables or model mis-specification. We show the value and accuracy of our insights by simulations based on Simpson's paradox, standard UCI multiclass-to-bandit conversions and on the Criteo counterfactual analysis challenge dataset.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 Introduction

There are two types of offline learning approaches in automated decision systems (e.g. recommendation systems): Q-learning and policy learning. Q-learning uses reward-modeling (or supervised learning) to predict rewards from both the context features and the action features. Formally, we estimate $Q(a, x)$, the expected reward from taking an action a in context x; decisions are then implied by the *greedy policy* that selects actions with the highest expected reward in each decision context [15, 23, 11, 21]. Reward modeling suffers from biases due to unobserved confounding variables or model mis-specification. For example, items that are temporarily popular because of sales events may not be popular in general, but these sales can confuse the learning system by reinforcing any mistakes in the previous policies when they are mistaken as the causes of successful sales. Therefore, it is often desirable to build reward-model-free decision systems that directly estimate the *causal effects* of the candidate actions, robust to hidden biases in previous logging policies.

As a result, many decision systems use policy learning [5, 31, 19, 14, 1, 8, 12]. To directly optimize for the decision policy in the presence of confounders, one additional requirement is to have randomization in the logging process: every candidate action must have nonzero probability to be selected given any context. By logging these action probabilities, unbiased causal effects can be estimated via inverse probability weighted estimator (IPWE), which up- or down-weights the rewards according to the odds of choosing the same action in the same context, across the two policies.

Unfortunately, accurate probability logging is a significant practical challenge. More worryingly, even with probability logging, naive IPWE suffers from large variance in estimation of causal effects, due to the up-weighting of rare actions, some of which will, on balance, appear in the logged datasets at least once. For example, consider a logging policy with a 1% chance of sampling a rare action. The rare action will be included in a dataset of a hundred samples at least once with

[1]Most of the work done while at Amazon.

Table 1: Challenges tackled by different objectives

challenge	Q	IPWE	IML	PIL-IML
confounders	×	✓	✓	✓
small/no probs	✓	×	✓	✓
improvement	✓	✓	×	✓

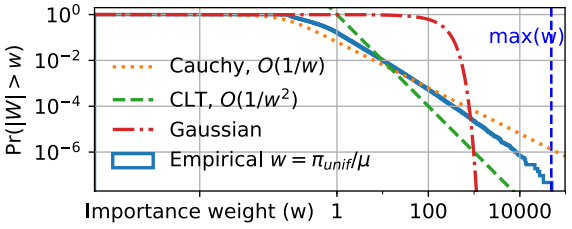


Figure 1: The importance weight distribution in Criteo counterfactual-analysis dataset [19] has unbounded variance due to its slower-than-central-limit-theorem (CLT) tail; $\max(w) = 49\,000$ in a total of 21MM examples.

probability $1-(1-1\%)^{100} = 63\%$. When this happens, IPWE weighs the single item as much as 100 examples– equivalent to half of the dataset. This increases the variance of any estimates that depend on that example significantly (See Example 8 in the appendix for more details and Figure 1 for a real-world example). Since this problem is caused by rare actions, one solution is to use biased estimators that conservatively estimate any potential lifts after up-weighting, [5, 31, 27, 26], which we show corresponds to estimating a lower bound on the eventual policy improvements.

In this paper, we show a connection between policy improvement lower bounds (PIL) and Clipped-IPWE [5] (Theorem 1). This connection opens up a number of extensions to Clipped-IPWE, and we focus on one in particular - the log-transformed-IPWE. We analyze this estimator, and further establish connections to policy gradients (PG) [30] using log-separability and Taylor approximations. In essence, we show that PG for contextual settings is equivalent to the cross-entropy (CE) objective in multi-class classification, where the label is whichever action that leads to the largest positive rewards in a particular context (Eq. 9). Since PG/CE does not require logged action probabilities, this provides a justification for their success even when logging is biased, particularly when compared with other offline learning objectives, e.g. Bayesian personalized ranking, sigmoid or triplet losses.

Once we identify PG/CE as an approximation we propose and analyze policy imitation learning (IML) as a regularizer (12), and show that this improves the tightness of estimates (Theorem 2). We connect IML to IPWE variance estimation [31]. In our experiments we see that IML is superior because it does not rely on unstable IPWE mean estimates, which is required for direct variance estimation (Section 7). We also connect IML to natural policy gradients [17, 27, 26, 21] without requiring knowledge of the model families. Similar to PG/CE, IML also works without logged action probabilities. The combined PIL-IML objective predicts the best next action that is ever taken in the logged data, with a weight that is large for very positive rewards and small but still positive for less-good rewards.

Finally, we show that when we have logged action probabilities, we can still benefit from IML by using it to diagnose a common problem in offline learning - when the logging policy is not in the class of optimization policies under consideration when learning. We show that IML-underfitting implies that the learning policy class does not have enough complexity or sufficient decision variables to imitate the original policy, which may lead to model biases. On the other hand, IML-underfitting can be used to our advantage by pointing out where we should collect additional data, through better action explorations (Theorem 7).

Notice, IML is different from propensity fitting, which is used as a plug-in replacement for the logging probabilities in the denominators of IPWE [8, 29]. On the other hand, we extend our methods to doubly robust approaches [25, 24, 2, 13, 8] and switching approaches [33, 18, 32] for additional, free variance reduction.

2 Offline Learning Objectives

While many methods have been proposed for learning from logged data, it is often unclear what the objective being maximized or minimized by different approaches. We introduce some clarifying definitions here. Let $(x_1, a_1, \mu_1, r_1), ..., (x_n, a_n, \mu_n, r_n) \in \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathbb{R}$ be a dataset containing n sample points of context, action, (possibly missing) probability of action given context, and reward, collected while running an automated decision system under a *logging policy*. Both contexts and actions have feature representations. For tabular actions, we use their indicator vectors, $a = e_a$.

Specifically, the data are generated i.i.d. as: $x_i, h_i \sim P(x, h), a_i \sim \mu(a|x_i, h_i)$ supported on $\mathcal{A}(x_i, h_i) \subset \mathcal{A}, \mu_i = \mu(a|x_i, h_i), r_i \sim P(r|x_i, h_i, a_i)$ Here, h_i is an unobserved confounding variable that affects both the action a_i and the reward r_i. $\mathcal{A}(x_i, h_i)$ is discrete set of candidate actions available in context (x_i, h_i). We can assume that $\mathcal{A}(x_i, h_i)$ is logged for every $i = [n]$. Note that most existing work implicitly assume that $\mu(a|x_i) = \mu(a|x_i, h_i)$, but it is common that certain decision variables are not logged, especially in publicly available data sets, due to proprietary features or human operators overwriting decisions every once in a while. In general, we do not assume that we know the

analytic form of μ besides having logged μ_i for the specific action taken. We will also consider the setting when even μ_i is not known, which makes it fundamentally impossible to do consistent off-policy evaluation (due to confounders), but we will show that often we can still do off-policy learning and adapt to the unknown propensities. To the best of our knowledge, this is the first time that a result of such flavor is presented.

The task of *offline learning* is to come up with a new policy, which is a distribution over candidate actions given context, $\pi(a \mid x), \forall a \in \mathcal{A}(x)$, such that the expected reward under π:

$$\mathbb{E}_\pi r = \mathbb{E} \sum\nolimits_{a \in \mathcal{A}(x)} \pi(a \mid x) r(x, a). \tag{1}$$

is as large as possible. This is difficult because it aims to estimate rewards for actions that may not have been logged in a particular context, unless they happen to coincide with the randomized action choices.

Inverse-probability weighted estimation (IPWE) [12, 5] is an unbiased offline evaluation method, which uses importance weights to estimate expectations under any new policy with samples generated from the original logging policy,

$$\begin{aligned} \mathbb{E}_\pi r &= \mathbb{E} \sum\nolimits_{a \in \mathcal{A}(x)} \pi(a \mid x) r(x, a) \qquad (2) \\ &= \mathbb{E} \sum\nolimits_{a \in \mathcal{A}(x)} \mu(a \mid x) \left[\frac{\pi(a \mid x)}{\mu(a \mid x)} r(x, a) \right] = \mathbb{E}_\mu \left[\frac{\pi}{\mu} r \right], \end{aligned}$$

where the last expectation is over the logging policy and can be estimated (without bias) by its sample mean. Define $w = \frac{\pi}{\mu}$ and $w_i = \frac{\pi(a_i|x_i)}{\mu_i}$ to be the importance weight in function form and instance form, respectively. Empirically, unbiased policy improvement can be maximized by

$$\max_\pi \Delta\text{IPWE}(\pi) = \frac{1}{n} \sum\nolimits_{i=1}^n (w_i - 1) r_i. \tag{3}$$

The variance of policy improvements is:

$$\mathbb{V}(\Delta\text{IPWE}(\pi)) = \frac{1}{n} \mathbb{V}_\mu((w-1)r), \tag{4}$$

which can similarly be estimated from the logged data.

While IPWE does not model the reward function and thus avoids modeling biases, it depends on randomized sampling of actions, which are usually the result of exploration/exploitation trade-offs. Lack of sufficient exploration, very common in practice, may lead to a large variance in the estimate, because data points with small action probabilities have large weights.

Any **objective** then must consider the trade-off between bias and variance. Our objective, which is often reasonable from a practical perspective, is to reliably maximize the policy improvement by a significant margin over the logging policy.

3 Proposed methods

To solve the large variance in IPWE, we propose to maximize a policy improvement lower-bound (PIL) regularized by policy imitation learning (IML). Assuming that the rewards are nonnegative, the general form is:

$$\max_\pi \text{PIL}(r, \pi) + \epsilon \text{IML}(\pi), \tag{5}$$

where ϵ is a tuning parameter to trade-off exploration/exploitation. One notable special case of the objective resembles a reward-weighted cross-entropy (CE) objective, written as:

$$\operatorname*{argmin}_\pi \sum\nolimits_{i=1}^n \left[r_i \log \frac{1}{\pi(a_i \mid x_i)} + \epsilon \log \frac{1}{\pi(a_i \mid x_i)} \right]. \tag{6}$$

We thus see the use of CE loss in offline learning as the result of a particular choice in the trade-offs of bias and variance (and ease of generalization and optimization).

3.1 Policy improvement lower-bounds (PILs)

One way to reduce the large IPWE variance is to clip its large weights by replacing w with $\bar{w}_\tau = \min(w, \tau)$ for a reasonable threshold $\tau > 0$. Here, τ is a bias-variance trade-off parameter. Fixing τ, Clipped-IPWE is

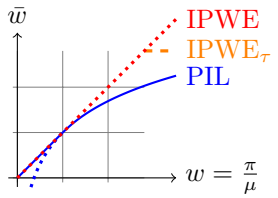


Figure 2: PIL and IPWE

$$\max_\pi \Delta\text{IPWE}_\tau(\pi) = \frac{1}{n} \sum\nolimits_{i=1}^n (\bar{w}_{\tau,i} - 1) r_i. \tag{7}$$

Instead, in offline learning, we take a different perspective on IPWE$_\tau$. Assuming that the rewards are nonnegative, $\bar{w}_\tau \le w$ lets IPWE$_\tau$ to be always a lower-bound on IPWE. Maximizing lower bounds as a surrogate objective is common practice. Generalizing this observation, we can arrive at many extensions of IPWE. One that we focus on is what we call the policy improvement lower bound estimator (PIL), which is based on the inequality $\log(w) \le w - 1, \forall w > 0$ (Figure 2). Depending on the logging scenario - with or without probabilities μ, we define the following objectives:

$$\begin{aligned} \text{PIL}_\mu(\pi) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n r_i \log w_i 1_{\{w_i \ge 1\}} + (w_i - 1) 1_{\{w_i < 1\}} \\ \text{PIL}_\emptyset(\pi) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n r_i \log w_i. \end{aligned} \tag{8}$$

Without the logged probabilities μ, we can also approximate IPWE up to a constant value. Consider the cross-entropy loss and its minimization,

$$\operatorname*{argmin}_\pi \text{CE}(\pi; r) = -\frac{1}{n} \sum\nolimits_{i=1}^n r_i \log \pi(a_i \mid x_i). \tag{9}$$

We see that, up to log-separable constant terms in the parameters of the logging policy, the PIL objective is equivalent to minimizing the reward-weighted cross-entropy loss for next-action predictions. That is $\mathrm{CE}(\pi; r) = \mathrm{CE}(\mu; r) - \mathrm{PIL}_\emptyset(\pi)$. As a result, CE may not require logged action probabilities yet enjoys the additional causality justification than other offline learning objectives, such as Bayesian personalized ranking and triplet loss. In particular, the other objectives are more likely to be biased by the logging policy and their negative sampling processes.

All these approximations of IPWE, come with an important indicator of the biases they induce - violations of the self-normalizing property. The self-normalizing property is that $\mathbb{E}_\mu(w) = \mathbb{E}(\pi - \mu) = 1 - 1 = 0$. When w is replaced with $\bar{w}$, the violation is empirically

$$\mathrm{Gap} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(1 - \bar{w}_i), \tag{10}$$

where $\bar{w}$ generalizes to any valid lower-bound surrogates. The theorem below shows the relationship between this observable quantity and the unobserved sub-optimality due to the use of a surrogate objective.

Theorem 1 (Probability gap)**.** *For any $\bar{w} \le w$, assuming $0 \le r \le R$, the approximation gap can be bounded by the probability gap, in expectation:*

$$0 \le \mathbb{E}_\mu[(w - \bar{w})r] \le \mathbb{E}_\mu[\mathrm{Gap}]R. \tag{11}$$

In particular, when $\bar{w} = 1 + \log w$, Gap has a simple form $-\frac{1}{n}\sum_{i=1}^{n}\log w_i$, which equals to one of the IML objectives that we introduce next.

3.2 Policy imitation for variance estimation

Another way to reduce IPWE variance is by adding regularization terms, that penalize the variance of the estimated rewards of the new policy. However, direct IPWE variance estimation [31] is problematic, because it requires the unreliable IPWE mean estimation in the first place. To this end, we propose to use policy imitation learning (IML) to bound the IPWE variance.

We define IML by empirically estimating the Kullback-Leibler (KL) divergence between the logging and the proposed policies, $\mathrm{KL}(\mu\|\pi) = \mathbb{E}_\mu \log \frac{\mu}{\pi} = -\mathbb{E}_\mu \log w$. We consider three logging scenarios - full logging where we have access to the logged probabilities of all actions, partial logging where we only know the logged probability of the taken action and missing where no logging probabilities are available. Depending on the amount of logging, our definition of IML has the following forms:

$$\begin{aligned}
&\mathrm{IML}_{\mathrm{full}}(\pi) = -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{a\in\mathcal{A}(x_i)}\mu(a|x_i)\log w(a|x_i);\\
&\mathrm{IML}_{\mathrm{part}}(\pi) = -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log w_i;\\
&\mathrm{IML}_{\mathrm{miss}}(\pi) = -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log \pi(a_i|x_i) - \mathrm{CE}(\mu; 1),
\end{aligned} \tag{12}$$

where $\mathrm{CE}(\mu; 1) = -\frac{1}{n}\sum_{i=1}^{n}\log \mu_i$ is a log-separable constant term similar to (9), but without the reward weighting. The following theorem shows that IML is a reasonable surrogate for the IPWE variance.

Theorem 2 (IML and IPWE variance)**.** *Suppose $0 \le r \le R$ and a bounded second-order Taylor residual $|-\mathbb{E}_\mu \log w - \mathbb{V}_\mu(w-1)| \le B$, the IML objective is closely connected to the ΔIPWE variance*

$$\mathbb{V}(\Delta\mathrm{IPWE}) \le \frac{1}{n}\Big(2\mathbb{E}_\mu\big(\mathrm{IML}\big) + B\Big)R^2. \tag{13}$$

Proof by Taylor expansion around $w = 1$, $\mathbb{E}_\mu(\mathrm{IML}) = -\mathbb{E}_\mu \log(w) \approx \frac{1}{2}\mathbb{E}_\mu(w-1)^2$, where the first-order approximation term is exactly $\mathbb{E}_\mu(w-1) = \mathbb{E}(\pi - \mu) = 0$.

Corollary 3. *These two imitation methods are second-order similar:* $\min_\pi -\mathbb{E}_\mu \log(\pi) \approx \min_\pi \frac{1}{2}\mathbb{E}_\mu\big(\frac{\pi}{\mu} - 1\big)^2$.

3.3 Other Properties

Generalizability: In stochastic bandits, optimal policies are often greedy. While classical arguments suggest that IPWE can learn greedy policies unbiasedly, e.g., with a saturated softmax, such policies may not generalize well. For example, a greedy policy for binary selections may look like $\pi(y \mid x) = \frac{e^{y\theta^\top x}}{1+e^{\theta^\top x}}, \forall y \in \{0, 1\}$, which saturates with $\|\theta\|_2 \to \infty$. Unfortunately, learning models with large weights often suggests large model complexity, which easily leads to overfitting [3].

Instead, Bayesian decision theory suggests a two-step approach: (1) fit a (reward-posterior) probability distribution of the optimal actions, which leads to smoother functions and (2) apply greedy argmax (or probability sharpening to handle ties). Along this line, CE uses a log-softmax link function, which is a convex, margin-based loss function that further improves generalization. It yielded better empirical results in Section 7 and further motivates (6) as an offline learning objective.

Adaptivity to unknown μ. The fact that the CE objective (6) is independent to the logging policy μ indicates that we can optimize a causal objective without knowing or needing to estimate the underlying propensities, which avoids the potential pitfalls in model misspecification and confounding variables. The following theorem establishes that optimizing $CE(\pi; r)$ based on the observed samples is implicitly maximizing the lower bound $\mathbb{E}_\mu r \log(\pi^*/\mu)$ for an unknown μ.

Theorem 4 (Statistical learning bound)**.** *Let μ be the unknown randomized logging policy and Π be a policy class. Let $\pi^* = \mathrm{argmax}_{\pi\in\Pi}\mathrm{CE}(\pi; r)$. Then with*

probability $1-\delta$, π^* *obeys that*

$$\mathbb{E}_{\pi^*} r - \mathbb{E}_\mu r \geq \mathbb{E}_\mu r \log(\pi^*/\mu) \geq \max_{\pi\in\Pi}\{\mathbb{E}_\mu r \log(\pi/\mu)\} - O\left(\frac{\log(\max_{\pi\in\Pi} D_{\chi^2}(\mu\|\pi)) + \log(|\Pi|/\delta)}{\sqrt{n}}\right)$$

Please refer to Appendix C for proofs and discussions.

A caveat is that although we can optimize the lower bound, we cannot explicitly evaluate the resulting lower bound of the policy improvement (e.g., to tell whether it is positive), without knowing μ. A heuristic solution is to use $\hat{\mu} = \operatorname{argmin}_\pi \mathrm{IML}(\pi)$ as a surrogate of μ.

Lemma 5 (Connections to natural policy gradients (NPGs) [17])**.** *Suppose the policy class is parametrized by* θ*, differentiable, and of the form* $\pi(a \mid x;\theta)$*. Suppose the logging policy also resides in the policy class, as* $\mu(a \mid x) = \pi(a \mid x;\theta_0)$*. The constrained optimization problem of natural policy gradient is a linear approximation to the PIL-IML in Lagrangian function form:*

$$\operatorname*{argmax}_{\Delta\theta} \mathbb{E}_\mu\left[r(x,a)\left(\frac{\pi(a \mid x;\theta_0+\Delta\theta)}{\mu(a \mid x)} - 1\right)\right] \quad (14)$$
$$\text{s.t. } \mathbb{E}\Big(\mathrm{KL}\big(\mu(a \mid x)\,\|\,\pi(a \mid x;\theta_0+\Delta\theta)\big)\Big) \leq \epsilon^2.$$

While PIL-IML can connect to NPG when the logging policy is included in the policy class, we should notice that the scope of PIL-IML is more general. In offline learning, we also consider problems where the logging policy may not be realizable in the policy class. In these problems, PIL-IML is still a valid objective, whereas NPG may not be properly evaluated.

Joachims et al., [14] showed some empirical successes using a Lagrangian form of a "self-normalized" SNIPS estimator, but did not provide much justification. We can show that the Lagrangian formulation of SNIPS transforms the original objective to a conservative one similar to our PIL-IML. To see this, notice that the optimal Lagrangian multipliers in [14] are always around 1, which equivalently means that the rewards are non-negative, agreeing with our intuitions.

Doubly-robust estimators can be natural extensions for PIL-IML. Please see Appendix H for details.

4 IML for causal exploration

In addition to variance control, IML has a number of useful properties and applications.

IML causality diagnosis: We define the IML training loss to be the objective values of (12) given partial or full logging probabilities. A positive IML training loss indicates logging biases such as confounding variables, likely due to the exclusion of engineered features that existed in the original logging policy, or policy misspecification when the true logging policy is not in the learned class. This property extends the classical propensity fitting methods [25, 24] that impute missing probabilities, which do not often check the feasiblity of the imputations when μ is assumed to be unknown.

Lemma 6 (IML diagnosis)**.** *Suppose the model family does not contain the logging policy* $\Pi \not\ni \mu$*, then* $\min_{\pi\in\Pi} \mathbb{E}\mathrm{KL}(\mu\|\pi) \geq 0$*. For example, if* μ *is a policy based on variables* $x = (x_1, x_2)$*, yet* Π *contains policies with only support on* x_1*, then* $\min_{\pi\in\Pi} \mathbb{E}\mathrm{KL}(\mu\|\pi) \geq \mathbb{E} I_\mu(a; x_2 \mid x_1) \geq 0$*, where* I_μ *is the mutual information between the logging policy and the confouding variable. Equality is found at* $\pi(a \mid x_1) = \mathbb{E}[\mu(a \mid x) \mid x_1], \forall x_1 \forall a$*.*

We often measure the IML feasibility gaps - i.e. indicators of how "far" the true logging policy is from the class of policies we are using - by perplexity (PPL), which is the exponent of the original IML loss. A perplexity of K implies that even after finding the policy in our class which best fits the logged data, the remaining uncertainty in the original action policy is equivalent to a uniform selection among K candidates. Perfect IML-fitting implies a CE of zero and a PPL of one.

IML for pure exploration: In batch offline cases, where we have multiple sequential opportunities to interact with the world - we suggest data recollection through online application of IML policies in (12). There are three benefits: the variance of each IML policy is small, due to the connection between IML objective and IPWE variance (Theorem 2); the performance of IML policy is predictable in offline evaluation and is typically comparable with the logging policy; lastly, with positive training loss and comparable performance, IML-resampling may greatly reduce model complexity by removing unexplained but unimportant decision factors from the logging policy. As a result, the new policies tend to be more exploratory. The improvements can be measured by increase in the entropy of the policy, which we quantify below (Proof in Appendix G is nontrivial due to action-induced distribution shifts).

Theorem 7 (Entropy increase)**.** *Let* $x = (x_1, x_2)^\top$ *be the vector of observed and confounding variables, respectively. If* π *is the marginalization of the logging policy,* $\pi(a \mid x_1) = \mathbb{E}[\mu(a \mid x) \mid x_1], \forall x_1 \forall a$*, we may guarantee an increase of expected entropy than that of the logging policy:*

$$\mathbb{E}H(\pi) - \mathbb{E}H(\mu) = \mathbb{E}\big(\mathrm{KL}(\mu\|\pi)\big) \geq 0. \quad (15)$$

5 Simpson's paradox and simulations

Simpson's paradox [28, 16, 5] is often used to explain the importance to remove confounders when modeling rewards. However, we use the example differently;

we simulate action randomization that also leads to correct action recommendations. Further, we validate the IML theoretical properties that detect confounders and improve exploration.

Table 2: Simpson's paradox for kidney stone treatments

Context	Open surgery	Small puncture
Small	**93% (81/87)**	87% (234/270)
Large	**73% (192/263)**	69% (55/80)
Hidden	78% (273/350)	**83% (289/350)**

In the Simpson's paradox example, a kidney stone treatment dataset with two actions is presented: an open surgery treatment or a small puncture treatment. The dataset was collected with an implicit bias where most people with large stones were treated with open surgery and most with small stones were treated with punctures, due to medical practices (such as risks and recovery times, which we do not model). An absolute majority of patients have small stones, which have higher cure rates with either treatment. As a result, despite the fact that an open surgery had a higher cure rate with either size of stones, regression on the treatment type without knowledge of the stone size would lead to the false conclusion that small punctures are correlated with higher cure rates (Table 2). It would seem that accurate reward modeling based action selection is impossible without the stone size contexts.

Table 3: Offline learning with logged probabilities.

Context	Action	Probability	Size	Cure
Hidden	Surgery	24%	87	93%
Hidden	Puncture	76%	270	87%
Hidden	Surgery	77%	263	73%
Hidden	Puncture	23%	80	69%

On the other hand, both treatment actions have nonzero frequencies given any stone size contexts. We could alternatively assume that the actions are randomized with their probabilities logged, given the hidden contexts (Table 3). Note that these probabilities are conditional on the internal states of the decision system and are different from the observed marginal probabilities, e.g., $\mu(\text{Surgery} \mid \text{Small Stone}) = \frac{87}{87+270} \approx 24\%, \mu(\text{Surgery} \mid \text{Large Stone}) = \frac{263}{263+80} \approx 77\%$.

In this way, unbiased offline learning is possible by weighting the action effects according to their inverse action probabilities, i.e., via IPWE. Thus logging would lead to the correct decision (surgery treatment). Doubly robust (DR) estimators may further reduce IPWE variance, but the amount of reduction depends on the quality of the reward model and can even be negative in some cases (Table 4 first column).

Table 4: Estimated cure rates of the surgery treatment

Method	Original data	IML-resampled
IPWE	83.3±5.0	83.3±3.7
Q-learning	78.0±1.6	**83.3±1.4**
DR	**83.3±2.6**	83.3±2.5
DR worst case	83.3±5.4	83.3±4.0

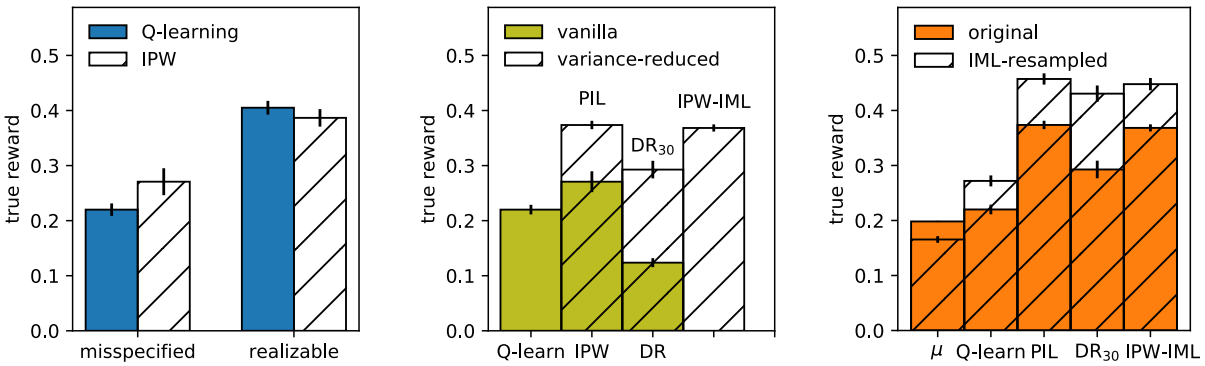
Based on our results, we use IML to first examine the logged action probabilities. In this case, IML is underfitted with $1.15(> 1)$ perplexity with a uniform policy, which suggests that there exist unobserved confounders. Since we do not have access to the additional confounders, we could simply resample data with the IML-fitted policy. We simulate resampling by weighing the examples according to the ratios between the IML policy and the logged probabilities. Due to self-normalization properties of the weights, the effective sample size (sum of all weights) remains the same for both small and large stone cases for fair comparisons. Table 4 shows that IML-resampling decreases the variance of all methods. This is because the new logging probabilities become more balanced (uniformly random trials), without depending on the hidden decision variable: the unknown stone sizes.

6 UCI bandit simulations

We use UCI multiclass-to-bandit conversion datasets that originally appeared in [8, 33] to simulate contextual bandit problems. For each data point, we sampled one class as the action and observed partial feedback whether the sampled class is the true class for that data point. Following previous literature, we constructed the logging policies by the softmax prediction of a linear logistic regression classifier trained on skewed datasets with induced covariate shifts. Since we have full knowledge of the original multiclass labels, we can exactly evaluate the learned policies during test time by multiclass fractional accuracy. We used 50% train-test splits where the training sets were converted to bandit datasets. Figure 3 also reports 95% confidence intervals from 100 repetitions. With this dataset, we show how reward modeling biases and IPWE variances affect offline learning, how to reduce variance using PIL-IML, and how to adapt IML into a batch-online method to collect better data for future offline learning.[2]

As discussed earlier, **Q-learning biases** can come from missing confounders and/or model underfitting, leading to variable under-utilization. We simulate this effect by using a second-order model, $\phi(x, a) = x^\top U V^\top a$, with

[2]Codes for these and experiments in the next section will be available at `https://github.com/yifeim/pil-iml`.



(a) Unbiased IPWE is better than Q-learning with misspecified models.

(b) Variance reduction techniques further improve offline learning.

(c) Online application of IML improves future offline learning.

Figure 3: Multiclass-to-bandit conversion on UCI optdigits dataset. Proposed improvements are in hollow style. Results for the other UCI datasets are included in the Appendix L.

insufficient rank. For UCI optdigits, a rank-2 model could only realize 67% multiclass accuracy when trained with full information, compared with 95% accuracy for a full-rank model, i.e. $\phi(x, a) = x^\top W a$. We call rank-2 models misspecified and full-rank models realizable.

Figure 3a shows that Q-learning and IPWE policy learning behave differently for misspecified model families. This is because Q-learning studies the biased correlation effects between the rewards and context-action pairs, whereas IPWE studies the unbiased causal effects of the actions given the contexts. Therefore, IPWE lead to better actions. On the other hand, Q-learning and IPWE policy learning behaved similarly for realizable model families, as expected from our analysis.

Variance-reduced methods further improved offline learning. Figure 3b continues the experiments with misspecified models, where the solid boxes are carried over from Figure 3a, with the addition of the logging policy itself (μ), and doubly robust (DR).

We compare three different variance-reduction approaches: PIL, PIL with DR extensions, and the original IPWE with IML regularization. All three approaches improved the final policy. The results were not very sensitive to the parameter choices, which we picked $\epsilon = 10^{-4}$, after a coarse grid search.

IML causality diagnosis was able to detect model underfitting due to insufficient rank. With a theoretical limit of 0, the rank-2 model family could only achieve 0.60 training loss, which indicates that the best IML policy cannot explain $\exp(0.60) = 1.82$ perplexity in the logged actions. Full-rank action policies can achieve a near-zero (0.02) training loss.

IML-resampling to collect additional data improves the policies learned with all methods (Figure 3c, changes from solid to hollow boxes), despite a small cost during IML resampling (μ as a policy in the first box). This is because better exploration leads to smaller inverse probability weights. Besides, IML-fitting alleviates model underfitting biases in the new data. Finally, the cost of IML resampling can be estimated prior to applying IML online and is fundamentally unavoidable for all methods that use the same model class.

Additional results on the other UCI datasets are in the appendix. In those examples, we further observed that improvements from variance reduction are significant only when IML loss is above zero. IML loss at zero indicates that there were no confounding variables or model misspecification (e.g., Figure 3a full-rank model); and that both naive Q-learning and IPWE would perform similarly to the variance-reduced methods.

7 Large-scale experiments

We extend our study on Criteo counterfactual-analysis dataset [19]. This dataset is particularly interesting, because the logging policy is in fact unrealizable from the published features and models. We made novel discoveries on (1) the existence of large variance due to Cauchy-like importance weight distribution and (2) the existence of modeling biases and confounding variables. We communicated and confirmed our hypotheses with the original authors.

The dataset contains logs of display advertisements shown to users, the hidden action probabilities, the user context features as well as features for every candidate action. However, we observed some discrepancies when we reran the provided scripts (Table 6 in Appendix I).

First, we noticed that the importance weight distributions, e.g., between the uniform policy and the logging policy, resemble heavy-tail distributions with **unbounded variance**. I.e., $P(|W| > w)$ decays slower than $O(1/w^2)$ in Figure 1. This fact invalidates the confidence interval (CI) estimation in the original paper

Table 5: Criteo counterfactual analysis dataset [19].

Approach ⌈⌉=greedy	Offline Est. ($\times 10^4$)	Gap (10) (100%)	Paired $\hat{\Delta}$ ($\times 10^4$)
Logging	(53.3, 53.7)	(0.0, 0.0)	(0.1, 1.8)
IML	(51.5, 53.3)	(-0.3, 0.3)	(0.0, 0.0)
Uniform	(41.8, 52.6)	(7.0, 8.0)	(-10 , 0.1)
⌈Q-learn⌉	(49.3, 55.9)	(3.0, 4.0)	(-2.8, 3.1)
POEM [31]	(51.4, 53.7)	(0.1, 0.7)	(-1.0, 1.1)
IPWE_{100}	(51.9, 54.5)	(-0.2, 0.5)	(-0.6, 1.9)
PIL-IML	(52.3, 53.7)	(-0.2, 0.2)	(0.0, 0.8)
⌈IML⌉	(53.0, 55.1)	(-0.4, 0.2)	(0.2, 2.4)
⌈PIL-IML⌉	(53.1, 55.2)	(-0.3, 0.3)	(0.6, 2.9)

[19] based on Central Limit Theorem (CLT), which requires bounded variance. For example, while the IPWE of the uniform distribution is 44.7 ± 2.1 in the test split, it becomes 52.6 ± 18.7 in the training split, due to one observation with importance weight 4.9×10^4.

To reduce heavy-tail uncertainties, we not only used weight clipping [5, 33], but also novelly applied **subsampling bootstrap** [22], which only assumes that $n^{-\beta}\big(\text{IPWE}_n - \mathbb{E}(\text{IPWE})\big) \xrightarrow[\text{dist.}]{n\to\infty} F$ asymptotically converges to any fixed distribution F, where β can be fitted from data. By weight clipping, we showed that β improved from 0.3 to near 0.5, which would be the CLT ideal case. See Appendix J for more details.

Then, using IML (12), we estimated the KL-divergence between the logging policy and its realizable imitation to be $0.40 (\gg 0)$. I.e., the logging policy is **unrealizable** by an exponential-family model with raw features, which contradicts [19] and essentially implies confounders (Lemma 6). Even with a more complex second-order model with 256-dimensional embedding, IML improved to 0.35, but was still large. While having perfect imitation is a sufficient but not necessary condition, as long as key decision variables are included [29], we found the situation practically difficult.

Table 5 reports offline 95% CIs using IPWE with weights clipped at 500 and subsampling bootstrap. Similar to [5], the first column includes both IPWE uncertainties and any missing clicks from the self-normalization Gaps, which we report in the second column. Since the global click rate is very low (around 5%), using $0 < R < 1$ extreme values may overestimate the upper bounds. Reasonably, we used the additional assumption that the expected click rates is always between $0 \sim 1\%$ in any reasonable subsets, twich their global average. The last column reports any improvements $\hat{\Delta}$ compared with the realizable IML. Here, paired-tests were used to avoid reward estimation noise. The formula also used weight-clipping and Gap-filling combinations: $\hat{\Delta} = \frac{1}{n}\sum_{i=1}^n \Big[\text{clip}(\frac{\pi_i - \hat{\mu}_i}{\mu_i}, -\tau, \tau) r_i\Big] + \Big[\frac{1}{n}\sum_{i=1}^n \text{clip}(\frac{\pi_i - \hat{\mu}_i}{\mu_i}, -\tau, \tau)\Big] R$, s.t. $0 < R < 1\%$.

Beating the realizable IML are the logging policy with its secret features, $\text{PIL}_\mu + 0.8\,\text{IML}$ with a tight margin, and most interestingly, greedy policies by sharpening PIL-IML or even IML, which is reward-agnostic but near-optimal. Intuitively, optimal policies are often greedy in stochastic environments. An option is always desirable no matter how small its improvements are, as long as they are consistent. While IPWE_{100} would also yield greedy solutions, they tend to learn saturated softmax, which may not generalize well (Section 3.3).

Notice, this stochastic view may have taken advantage of the temporal overlaps between the train/test splits. In contrast, randomized policies are more popular in adversarial bandits to account for temporal uncertainties [6, 20]. Similarly, if we were allowed to collect new data, we would use PIL-IML (non-greedy) to improve exploration (Theorem 7). The perplexity would increase from 3.6 to 5.2 and the heavy-tail situations would alleviate.

8 Discussion and conclusion

Why should I imitate a policy when I already have the logged propensities? First, IPWE might suffer from high variance and DR is often not much better because we seldom observe all major variables that contribute to the variance of the reward. Second, whether the imitated policy produces probabilities that match the logged propensities or not reveals whether there are hidden decision variables used by the logging policy.

What can I use an imitated policy for? We can run it to collect new data, which guarantees that (if we cannot discover or log them for some reasons) the new data will have a realizable policy without unknown confounding decision variables. Besides, the IPWE estimate of the imitated policy should not suffer from high variance and we will have good evidence whether the imitated policy performs similarly to the unrealizable logging policy in most cases.

What are the take-home messages for data scientists and developers? The most important message is that one should be cautious about using and evaluating Q-learning approaches in problems where decisions are involved. We highlight the value of having randomized policies that allow one to marginalize over unobserved confounding variables and make statistically valid inferences despite possible confounders. Lastly, the probabilities corresponding to actions that are not taken are also useful and can be used to reduce variance in offline policy valuation and policy optimization.

Acknowledgements

We appreciate Haibin Lin for help with mxnet sparse matrix operators, Yuyang (Bernie) Wang, Tengyang Xie for detailed discussions, Adith Swaminathan for insights about the Criteo counterfactual challenge dataset, and the anonymous reviewers for their constructive comments.

References

[1] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

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

[3] Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel methods—support vector learning*, pages 43–54, 1999.

[4] Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *bepress*, 2008.

[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[6] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[7] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[8] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[9] Charles J Geyer. 5601 notes: The subsampling bootstrap. *Unpublished manuscript*, 2006.

[10] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361, 2017.

[11] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

[12] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[13] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

[14] Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. *International Conference on Learning Representations*, 2018.

[15] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 43–50. ACM, 2016.

[16] Steven A Julious and Mark A Mullee. Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481, 1994.

[17] Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

[18] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[19] Damien Lefortier, Adith Swaminathan, Xiaotao Gu, Thorsten Joachims, and Maarten de Rijke. Large-scale validation of counterfactual learning methods: A test-bed. *arXiv preprint arXiv:1612.00367*, 2016.

[20] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230. ACM, 2013.

[21] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2016.

[22] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

[23] Steffen Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57, 2012.

[24] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

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

[26] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1889–1897, 2015.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[28] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241, 1951.

[29] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.

[30] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[31] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[32] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[33] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597, 2017.