Imitation-Regularized Offline Learning (Supplemental Materials)

A Related work

Clipped-IPWE was originally used in offline evaluation, which, however, leaves the remaining gaps to be estimated. [5] estimated the remainders via Bernstein's inequality, which may empirically be a loose estimation, and [33] used reward modeling, which assumes nonexistence of confounding variables and universal approximators, which are under our scrutiny.

Some theoretical justification for the large variance in IPWE is provided in [7, 33]. Doubly robust estimation (DR) was initially proposed to take advantage of a possibly inaccurate Q-learning model for estimating the expected rewards [25, 8, 13], yet DR does not fundamentally solve the large variance inverse probabilities [18]. Another variant is to clip IPWE into a head average and a tail bound [5, 4], but this approach has not been formulated as an stochastic optimization objective. As an extension, SWITCH method [33] was proposed to tighten the mean-squared error (MSE) bounds in clipped IPWE via Q-learning estimates to achieve minimax-optimality. However, SWITCH was not directly used for optimization, either. On policy imitation, a related topic is to learn the probabilities through propensity fitting [29, 1]. We use similar techniques, but only for regularizing safe exploration, where we also assume knowledge of the exact logged probabilities. Our solution connects to other variance reduction techniques [31, 14] and we show that ours are more stable with varying hyperparameter choices.

In the RL literature, Munos et al [21] uses importance weighting to improve Q-learning convergence. When choosing the importance weighting as a soft-Q policy [10], they proved a contraction property that leads to final convergence to true Q*. While the method works well for tabular settings, it is unclear whether Q* is fundamentally easy to estimate in continuous settings from data with confounding variables.

Similarly, trust-region policy gradient method [26] used Pinsker's inequality to motivate a worst-case KL-divergence as regularization. Differently, we directly motivated expected KL-divergence as regularization. This was indeed their empirical choice as well, but with less clear intuitions. The follow-up proximal policy gradient method [27] used direct IPWE with heavy clipping on both sides. Our method leads to a smoother policy that tend to generalize better (Section 3.3).

B Example on IPWE large variance and statistical analysis

Example 8 (Epsilon-greedy). *Suppose the logging policy (ignores the context and) tosses a biased coin to choose between two actions: a rare action $\mu(A) = \epsilon \ll 1$ and a default action $\mu(C) = 1 - \epsilon$. Suppose the rewards are noiseless and only given to the rare action, $r(a) = 1_{\{a=A\}}$. With n examples, the rare action will be included in the logging dataset at least once with high probability $(1 - (1 - \epsilon)^n) \geq 1 - e^{-n\epsilon}$.*

An (optimal) deterministic policy that always chooses A has expected reward 1. While IPWE is unbiased, its mean-squared error (MSE) can be at least $\frac{1}{n}\mathbb{E}_\mu[(\frac{\pi}{\mu}r)^2 - (\mathbb{E}_\pi r)^2] = \frac{1}{n}[\epsilon(\frac{1}{\epsilon})^2 - 1] \geq \frac{1}{2n\epsilon}$, because the rare event has significant weight $\frac{1}{\epsilon}$. On the other hand, Q-learning has zero variance but a unit bias if A is not observed, with probability at least $e^{-n\epsilon}$. Comparatively, the MSE in Q-learning is exponentially smaller than IPWE in fully observed but imbalanced datasets.

C Concentration inequalities for PIL-IML

The concentration for the weight clipping-based PIL straightforwardly follows from the boundedness and Hoeffding's inequality, as was observed in [5].

In this section, we derive the concentration inequality for the cross-entropy based PIL.

Lemma 9 (Exponential tails for the cross-entropy weights). *Let $\log(\pi(x,a)/\mu(x,a))$ be a random variable induced by $x \sim \mathcal{D}, a|x \sim \mu(a|x)$. For all $t > 0$*

$$\mathbb{P}\left[|\log\frac{\pi(x,a)}{\mu(x,a)}| \geq t\right] \leq e^{-t}(1 + \mathbb{E}_\pi[(\mu/\pi)^2])$$

where $\mathbb{E}_\pi[(\mu/\pi)^2] =: D_{\chi^2}(\mu\|\pi)$ *is the* χ^2*-divergence.*

Let $\bar{w}(x,a) = \begin{cases} \log\frac{\pi(x,a)}{\mu(x,a)} & \textit{if } \pi >= \mu; \\ \frac{\pi(x,a)}{\mu(x,a)} - 1 & \textit{otherwise.} \end{cases}$ *then*

$$\mathbb{P}\left[|\log\frac{\pi(x,a)}{\mu(x,a)}| \ge t\right] \le 2e^{-t}$$

Proof. For readability, we use π and μ as shorthands for random variables $\pi(x,a)$ and $\mu(x,a)$. By Chernoff's argument and Markov's inequality:

$$\begin{aligned}
\mathbb{P}\left[|\log\frac{\pi}{\mu}| \ge t\right] \le& \mathbb{P}\left[e^{|\log\frac{\pi}{\mu}|} \ge e^t\right] \\
\le& e^{-t}\mathbb{E}_\mu[e^{|\log\frac{\pi}{\mu}|}] \\
=& e^{-t}\left\{\mathbb{E}_\mu[e^{\log\frac{\pi}{\mu}}\mathbf{1}(\pi \ge \mu)] + \mathbb{E}_\mu[e^{\log\frac{\mu}{\pi}}\mathbf{1}(\pi < \mu)]\right\} \\
=& e^{-t}\left\{\mathbb{E}_\mu[\frac{\pi}{\mu}\mathbf{1}(\pi \ge \mu)] + \mathbb{E}_\mu[\frac{\mu}{\pi}\mathbf{1}(\pi < \mu)]\right\} \\
\le& e^{-t}\left(1 + \mathbb{E}_\pi[\frac{\mu^2}{\pi^2}]\right).
\end{aligned}$$

Now for the modified cross-entropy objective, by the same argument, we have

$$\begin{aligned}
\mathbb{P}\left[|\bar{w}| \ge t\right] \le& e^{-t}\left\{\mathbb{E}_\mu[e^{\log\frac{\pi}{\mu}}\mathbf{1}(\pi \ge \mu)] + \mathbb{E}_\mu[e^{\frac{\pi}{\mu}-1}\mathbf{1}(\pi < \mu)]\right\} \\
\le& e^{-t}\left\{\mathbb{E}_\mu[\frac{\pi}{\mu}\mathbf{1}(\pi \ge \mu)] + \mathbb{E}_\mu[1\mathbf{1}(\pi < \mu)]\right\} \le 2e^{-t}.
\end{aligned}$$

□

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*

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

Proof sketch. For every $\pi \in \Pi$, we use Lemma 9 to get that the empirical estimate converges its mean for both the objective that we optimizes over and the empirical estimate of the gap. Then we apply a union bound to make it uniform over the entire policy class. Since π is the argmax on the empirical estimates, it must also be close to the argmax on the population quantity, which concludes the proof. □

The assumption on discrete Π is not essential. It can be replaced with a uniform convergence bound for infinite alphabet.

Also note that the above lower bound is also something that we can hope to optimize without knowing what μ is. This motivates us to use policy immitation as a regularization term. R is a conservative regularization weight.

Now, assume that $\mu \in \Pi$, also Π is parameterized by a parameter vector θ and in addition π is a smooth function in θ. Then the standard policy gradient theorem tells us that there exists a policy in the neighborhood of μ that improves over μ provided that μ is not a local maxima. The maximizer of CE-IML objective resembles the natural policy gradient update which starts at μ and in fact, it magically get away with not need to know where to start.

To conclude the section, we note that this property of adaptivity in CE and CE-IML optimization implies that policy optimization might be an easier problem than policy evaluation, challenging the common wisdom that we need to know the objective function before we can optimize.

D Connections between IML and IPWE variance

Many counterfactual analysis papers maximize the lower bounds of expected rewards, as they are more reliable to estimate [31, 5]. Our intuition is similar.

However, we avoided direct optimization of the bounds like [31]:

$$\min_{\pi} \mathrm{IPWE}(\pi) + \alpha\sqrt{\mathbb{V}(\mathrm{IPWE}(\pi))}, \tag{16}$$

which is further optimized by tunning α such that $\alpha = \lambda'\sqrt{\mathbb{V}(\mathrm{IPWE}(\pi))}$. Instead, we connect the IPWE variance to the IML objective. Restate Theorem 2

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* Δ*IPWE variance*

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

Proof. First, we point out a key observation from the second-order Taylor expansion of the IML objective near $w = \frac{\pi}{\mu} \approx 1$, as

$$\begin{aligned} \mathbb{E}_\mu(\mathrm{IML}) = -\mathbb{E}_\mu \log(w) &= -\mathbb{E}_\mu \log\big(1 + (w-1)\big) && (17)\\ &= -\mathbb{E}_\mu\Big[(w-1) - \frac{1}{2}(w-1)^2 + o\big((w-1)^2\big)\Big] && (18)\\ &= \frac{1}{2}\mathbb{E}_\mu(w-1)^2 + o\big((w-1)^2\big), && (19) \end{aligned}$$

where $|o\big((w-1)^2\big)| \le B$ is the residual and $\mathbb{E}_\mu w = 1$ cancels the first-order term.

Then, suppose $|r| \le R$, we relax the policy-related variance estimation

$$\mathbb{V}_\mu\big((w-1)r\big) = \mathbb{E}_\mu\big((w-1)^2 r^2\big) \le \mathbb{E}_\mu(w-1)^2 R^2 \le \Big(2\mathbb{E}_\mu(\mathrm{IML}) + B\Big)R^2 \tag{20}$$

Finally, the variance ΔIPWE has an additional $\frac{1}{n}$ term, because ΔIPWE itself is a mean-value estimator. □

Note, the derived form is closely related to the variance of IPWE itself, by considering additional reward observation noise:

$$\mathbb{E}_\mu(wr) = \mathbb{E}_\mu\big((w-1)r\big) + \mathbb{E}_\mu(r) \quad \Rightarrow \quad \mathbb{V}_\mu(wr) \le 2\mathbb{V}_\mu\big((w-1)r\big) + 2\mathbb{V}_\mu(r). \tag{21}$$

Comparatively, IML can be more robust, because it avoids influence from large probabilities. On the other hand, IPWE/ΔIPWE as a mean value can sometimes be unstable due to large inverse propensities $\frac{1}{\mu_i}$ when weight-clipping is not involved ($\tau = \infty$). Additionally, IML does not involve reward estimation and thus the regularization directions may be orthogonal to the policy improvements, whereas empirical IPWE/ΔIPWE may directly penalize improvements.

Remark 10. *The Taylor expansion in Theorem 2 relies on bounded central moments. This assumption may not be available in general offline learning scenarios, e.g., Example 8. However, the assumption is reasonable with variance-reduced approaches, e.g., Clipped-IPWE, IPWE-IML, and POEM[31].*

E Connections between PIL-IML and natural policy descent

Natural policy descent [17] is a steepest descent optimization approach to solve an approximate policy optimization problem:

$$\max_\theta \mathbb{E}_\mu\big(r(x,a)\log \pi(a \mid x;\theta)\big). \tag{22}$$

It computes natural policy gradients (NPGs), which use the first and second-order derivatives of the objective,

$$\Delta\theta = \Big[\mathbb{E}_\mu\big(\nabla_\theta \log \pi(a \mid x;\theta)\big)\big(\nabla_\theta \log \pi(a \mid x;\theta)\big)^\top\Big]^{-1}\mathbb{E}_\mu\big(r(x,a)\nabla_\theta \log \pi(a \mid x;\theta)\big). \tag{23}$$

The one-step NPG is also equivalent to the solution to the constrained optimization problem:

$$\operatorname*{argmax}_{\Delta\theta} \mathbb{E}_\mu\big(\Delta\theta^\top r(x,a)\nabla_\theta \log \pi(a \mid x;\theta)\big) \tag{24}$$

$$\text{s.t. } \mathbb{E}_\mu\Big[\Delta\theta^\top\big(\nabla_\theta \log \pi(a \mid x;\theta)\big)\big(\nabla_\theta \log \pi(a \mid x;\theta)\big)^\top \Delta\theta\Big] \le \epsilon^2, \tag{25}$$

where $\epsilon > 0$ is the step size.

We can establish the following connections between PIL-IML and NPG.

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:*

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

Proof. To show the equivalence of the objectives, we apply Taylor expansions and use the property

$$\nabla_\theta \log \pi(\theta) = \frac{\nabla_\theta \pi(\theta)}{\pi(\theta)} \quad \Rightarrow \quad \nabla_\theta \pi(\theta) = \pi(\theta)\nabla_\theta \log \pi(\theta). \tag{26}$$

The policy-related term in the objective becomes

$$\left(\frac{\pi(a \mid x;\theta_0+\Delta\theta)}{\mu(a \mid x)} - 1\right) \approx \left(\frac{\mu(a \mid x) + \Delta\theta^\top \nabla_\theta \pi(a \mid x;\theta_0)}{\mu(a \mid x)} - 1\right) \tag{27}$$

$$= \Delta\theta^\top \nabla_\theta \log \pi(a \mid x;\theta_0) \tag{28}$$

On the other hand, using the derivations in (19) and (28), the constraint is equivalent to

$$\mathbb{E}\Big(\mathrm{KL}\big(\mu(a \mid x) \,\|\, \pi(a \mid x;\theta_0+\Delta\theta)\big)\Big) \approx \mathbb{E}\left(\frac{\pi(a \mid x;\theta_0+\Delta\theta)}{\mu(a \mid x)} - 1\right)^2 \tag{29}$$

$$= \mathbb{E}\big(\Delta\theta^\top \nabla_\theta \log \pi(a \mid x;\theta_0)\big)^2. \tag{30}$$

Now, both the objective and the constraint are in the same form as NPG. □

F IML strictly positive due to confounding variables

Lemma 6 (IML diagnosis)**.** *Suppose the model family does not contain the logging policy $\Pi \not\ni \mu$, then $\min_{\pi\in\Pi} \mathbb{E}\mathrm{KL}(\mu\|\pi) \ge 0$. 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 $\min_{\pi\in\Pi} \mathbb{E}\mathrm{KL}(\mu\|\pi) \ge \mathbb{E}I_\mu(a; x_2 \mid x_1) \ge 0$, where I_μ is the mutual information between the logging policy and the confouding variable. Equality is found at $\pi(a \mid x_1) = \mathbb{E}[\mu(a \mid x) \mid x_1], \forall x_1 \forall a$.*

Proof. Without loss of generality, assume $x_1 = \emptyset$ and $x = x_2$. We can express the objective as

$$\mathbb{E}\mathrm{KL}(\mu\|\pi) = \int p(x) \sum_a \mu(a \mid x) \log \frac{\mu(a \mid x)}{\pi(a)}\, \mathrm{d}x = \mathrm{CE}(\pi, \mathbb{E}\mu(\cdot \mid x)) - \mathbb{E}H(\mu) \tag{31}$$

which is maximized by $\pi(a) = \mathbb{E}\mu(a \mid x) = \int p(x)\mu(a \mid x)\,\mathrm{d}x$.

Then we plug in the solution. Define $p_\mu = p(x)\mu(a \mid x)$, the solution becomes $\hat\pi(a) = \mathbb{E}\mu(a \mid x) = p_\mu(a)$ and the objective is

$$\mathbb{E}\mathrm{KL}(\mu\|\hat\pi) = \int p_\mu(x,a) \log \frac{p_\mu(a \mid x)}{p_\mu(a)}\, \mathrm{d}x\, \mathrm{d}a = \int p_\mu(x,a) \log \frac{p_\mu(a,x)}{p_\mu(a)p_\mu(x)}\, \mathrm{d}x\, \mathrm{d}a = I_\mu(a;x), \tag{32}$$

where $I_\mu(a;x)$ is the mutual information due to the logging policy between a and x, the variable that cannot be explained by the new policy. □

G Entropy increases with IML policies due to confounding variables

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:*

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

Proof. Starting from the left hand side and using the provided equation for π,

$$\begin{aligned}
\mathbb{E}H(\pi) - \mathbb{E}H(\mu) = \int \Big[&- \sum_a \pi(a \mid x_1) \log \pi(a \mid x_1) & (33)\\
&+ \int \sum_a \mu(a \mid x) \log \mu(a \mid x) p(x_2 \mid x_1)\,\mathrm{d}x_2 \Big] p(x_1)\,\mathrm{d}x_1 & (34)\\
= \int \Big[&- \sum_a \int \mu(a \mid x) p(x_2 \mid x_1)\,\mathrm{d}x_2 \log \pi(a \mid x_1) & (35)\\
&+ \int \sum_a \mu(a \mid x) \log \mu(a \mid x) p(x_2 \mid x_1)\,\mathrm{d}x_2 \Big] p(x_1)\,\mathrm{d}x_1. & (36)
\end{aligned}$$

With the key observation that $\pi(a \mid x_1) = \pi(a \mid x)$ is independent of $(x_2 \mid x_1)$, we may replace the summation and the integral in (35), which yields

$$\begin{aligned}
\mathbb{E}H(\pi) - \mathbb{E}H(\mu) = \int \Big[&- \int \sum_a \mu(a \mid x) \log \pi(a \mid x) p(x_2 \mid x_1)\,\mathrm{d}x_2 & (37)\\
&+ \int \sum_a \mu(a \mid x) \log \mu(a \mid x) p(x_2 \mid x_1)\,\mathrm{d}x_2 \Big] p(x_1)\,\mathrm{d}x_1 & (38)\\
= \int \sum_a &\mu(a \mid x) \log \Big(\frac{\mu(a \mid x)}{\pi(a \mid x)} \Big) p(x)\,\mathrm{d}x & (39)\\
= \mathbb{E}&\big(\mathrm{KL}(\mu \| \pi)\big).
\end{aligned}$$

□

H Doubly robust estimation

DR fixes the IPWE large variance without introducing biases. Instead, it uses two equivalent ways to find the expectation $\mathbb{E}_\pi[\hat{f} \mid x] = \mathbb{E}_\mu[\frac{\pi}{\mu} \hat{f} \mid x] = \sum_a \hat{f}(x, a)$ of a known function $\hat{f}$ that can be evaluated for any action, as

$$\max_\pi \mathrm{DR}(\pi) = \frac{1}{n} \sum_{i=1}^n \frac{\pi(a \mid x_i)}{\mu_i} \big(r_i - \hat{f}(x_i, a_i) \big) + \frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}(x_i)} \pi(a \mid x_i) \hat{f}(x_i, a). \tag{40}$$

DR uses a Q-learning estimator $\hat{f}$ to reduce the variance in the first term. Its optimization often yields at least as good performance as either Q-learning or IPWE, as long as Q-learning has positive correlation with the true rewards. However, vanilla DR can perform worse in extreme cases, as noticed by [18].

We extend PIL-IML to DR, given that the probabilities are logged for every action candidate, $\mu(a \mid x_i), \forall a \in \mathcal{A}(x)$, including the non-chosen ones. The PIL-DR is

$$\max_\pi \mathrm{DR}_\tau(\pi) = \frac{1}{n} \sum_{i=1}^n \bar{w}_i \big(r_i - \hat{f}(x_i, a_i) \big) + \frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{A}(x_i)} \bar{\pi}(a \mid x_i) \hat{f}(x_i, a) + \bar{R}(\pi), \tag{41}$$

where $\bar{\pi} = \bar{w}\mu$ is the lower-bounded policy induced by the lower-bounded weights and $\bar{R}(\pi) = \sum_{i=1}^n (w_i - \bar{w}_i) r_i$ reflects the bound error. With nonnegative rewards, we optimize the lower-bound of DR by removing the $\bar{R}(\pi) \geq 0$ term.

I Reproducing Criteo counterfactual analaysis

Table 6: Reproducing Criteo counterfactual analysis [19]

Approach	IPWE ($\times 10^4$)	Gap (%)
Logging policy	53.4	0.0
Uniform random	44.7 ± 2.1	1.7 ± 2.1
DM/Q-learning	49.6 ± 1.7	-0.5 ± 2.0
IPWE [12]	49.9 ± 1.8	0.1 ± 1.6
DR	53.7 ± 14.4	-1.5 ± 5.7
POEM [31]	52.7 ± 1.6	0.3 ± 0.6

A rerun of the provided scripts by [19] yielded results in Table 6. None of the approaches outperformed the logging policy using the released models and features, likely due to the existence of confounding variables and model misspecifications. Moreover, the confidence intervals are much smaller than their true values, as reflected in the inconsistency across data splits which is discussed in the main text and also show in Table 7&8. This is likely the combined results of fat-tailed importance weight distributions and a lack of weight-clipping in evaluations.

Table 7: Reproduce Criteo counterfactual dataset large variance (DM/Q-Learning)

Validation Rank	Test Rank	Validation IPWE ($\times 10^4$)	Test IPWE ($\times 10^4$)	P	L	Lambda
1	2	52.5 ± 3.1	49.6 ± 1.7	1	0.1	0
2	1	49.2 ± 3.4	53.5 ± 14.2	1	1	0
3	14	47.4 ± 7.0	41.0 ± 4.4	1	0.1	1E-8

Table 8: Reproduce Criteo counterfactual dataset large variance (IPS/IPWE)

Validation Rank	Test Rank	Validation IPWE ($\times 10^4$)	Test IPWE ($\times 10^4$)	P	L	Lambda
1	5.5	53.8 ± 3.6	49.9 ± 1.8	0	10	1E-8
2	5.5	53.8 ± 3.6	49.9 ± 1.8	0	10	0
3	4	53.8 ± 3.6	50.0 ± 1.8	0	10	1E-6
10	1	50.8 ± 2.1	53.7 ± 3.0	0.5	1	0

J Subsampling bootstrap

Subsampling bootstrap [22, 9] is a statistical approach to estimate confidence intervals with minimal assumptions. It is especially suitable for heavy-tail distributions or dependent samples. Our goal is to estimate asymptotic quantiles on independent sample draws from heavy-tail distributions.

The key idea is to extend central limit theorem (CLT) convergence properties to a more general

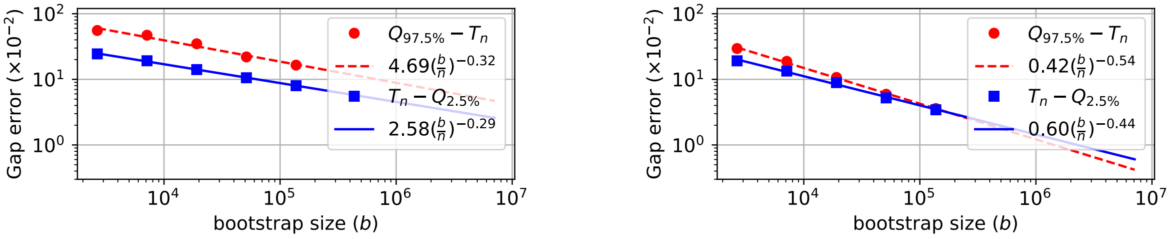
$$n^{-\beta}(T_n(\theta) - \theta) \overset{\text{dist.}}{\to} F, \tag{42}$$

where $T_n(\theta)$ is the finite-sample estimator of parameter θ, $\beta > 0$ is the generalized convergence rate, and F is any limiting distribution. A trivial special case is CLT with $\beta = 0.5$. A nontrivial example where $\beta \neq 0.5$ may be sample max estimator for the parameter of a uniform distribution, i.e., $T_n(\theta) = \max(X_1, \ldots, X_n)$ where $X \sim \text{Unif}(0, \theta)$. In this case, simple algebra shows $n(\theta - T_n(\theta)) \overset{dist.}{\to} \text{Exp}(\theta)$, i.e., $\beta = 1$. There are also examples where $\beta < 0.5$ for robust regression estimators. Additionally, the mean of a Cauchy distribution does not exist, so $\beta \geq 0$ for the sample mean estimator. (The median estimator does converges to its mode parameter at $\beta = 0.5$.) The following approach has been verified with both uniform and Cauchy distributions.

We are interested in finding the correct rate of β for the importance weight distribution, e.g., between the uniform policy and the logging policy, from Criteo counterfactual-analysis dataset. Shown in Figure 1 of the main text, the convergence rate of the raw weights (without clipping) must follow $\beta < 0.5$. We regressed β using subsampling bootstrap with varying size $0 \ll b \ll n$, where for each size, we bootstrapped $K = 10\,000 \sim 100\,000$ subsamples and recorded the distribution from the resulting estimators, $S_b = \{T_b^{(k)} : k = 1, \ldots, K\}$. Let $Q_q(S_b)$ be the distribution quantile at q, (42) suggests that

$$b^{-\beta}(Q_q(S_b) - T_n(\theta)) \rightarrow b^{-\beta}(Q_q(S_b) - \theta) \rightarrow F^{-1}(q), \tag{43}$$

where the difference $|T_n(\theta) - \theta|$ converges to zero at a faster rate $O(n^{-\beta}) < O(b^{-\beta})$ and can be ignored. For this purpose, we took $n^{0.5} \leq b \leq n^{0.75}$. Since the data is iid, we used subsampling with replacement to increase speed.

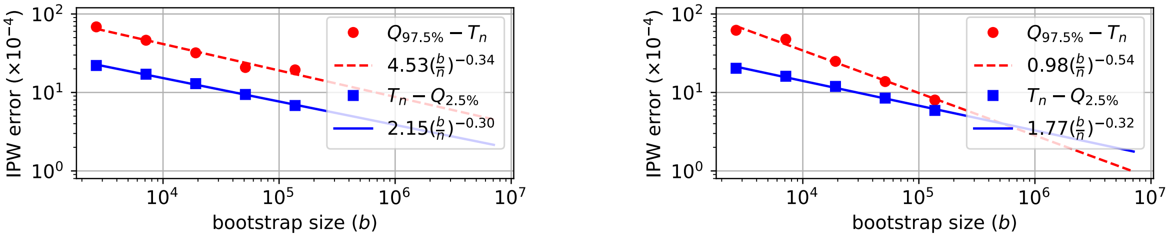


(a) Slower rates without weight clipping.

(b) Faster rates with weight clipped (at 500).

Figure 4: Error bounds of the self-normalization Gap estimator, as a function of subsample size. Using subsampling bootstrap, the final error can be extrapolated with the correct rates.

Figure 4 shows the log-log plots for the error bounds in the self-normalization estimator $T_b = \frac{1}{b}\sum_{i=1}^{b} \frac{\pi_{i_k}}{\mu_{i_k}}$, where $\{i_k : k = 1, \ldots, b\}$ is one subsample. Without weight-clipping, the estimator converged at a slower rate $\beta \approx 0.3$. With weight clipped at 500, the convergence rate was around $\beta = 0.5$. The final self-normalization quantity could be extrapolated as $(98.3 - 2.6, 98.3 + 4.7) = (95.7, 103.0)\%$ without clipping, i.e., 7.3% estimation error. On the other hand, weight clipping introduced 7∼8% bias, which is on a similar scale, but significantly improves stability. Intuitively, one click should neither be weighed more than 500 times.



(a) Slower rates without weight clipping.

(b) Faster rates with weight clipped (at 500).

Figure 5: Error bounds of IPWE as a function of subsample size. Using subsampling boostrap, the final error can be extrapolated with the correct rates.

Figure 5 shows a similar story that weight clipping could improve convergence quality and yield tighter IPWE error bounds. While most of the rate improvements were notable, one particular interesting point is that weight-clipping at 500 did not seem to help the lower-bound estimation of IPWE, which still had a convergence rate at $\beta = 0.32$. One possible explanation could be that $\tau = 500 \approx \sqrt{n\mathbb{E}R}$, i.e., square-root of the total number of clicks, which still left the worst-case IPWE variance potentially unbounded.

The final offline estimator should include both IPWE and self-normalization Gap errors [5]. Since the global click rate is around 0.5%, we may assume that the expected clicks in any self-normalization Gaps to be within $0 \sim 1\%$, i.e., at most twice their global average. Table 9 and Table 5 in the main text both report the offline estimates using the following estimator:

$$\text{Offline estimation} = \text{IPWE} + \text{Gap}R \text{ s.t. } 0 < R < 0.01. \tag{44}$$

Table 9: Offline estimates combining both IPWE and Gap errors. Weight clip contributed to smaller confidence intervals in the final offline estimates, without being overly pessimistic.

Weight Clip	Type	IPWE w/ Clip	Self-Normalization	Final Offline Estimate
τ		$\times 10^4$	100%	$\times 10^4$
∞	Mean	44.7	98.3	45.6
	CI	(42.5, 49.2)	(95.7, 103.0)	(38.2, 52.2)
500	Mean	43.6	92.6	47.3
	CI	(41.8, 44.6)	(92.0, 93.0)	(41.8, 52.6)

Judging by the amount of self-normalization Gaps, we found that clipped estimators to also be useful in (greedy application) of Q-learning (aka. direct method). On the other hand, IML-based methods, such as IPWE$_\tau$, PIL, and POEM, tend to have lighter tails. We applied weight clipping to all methods for fairness, while the light-tail distributions tend to suffer smaller clipping biases.

K Second-Order Model and Implementation

The second-order model which helped us reduce the IML Gap from 0.40 to 0.35 still follows an exponential family with potential function $\log \pi(a \mid x) \simeq \phi(x, a)$, but the potential scores become second-order:

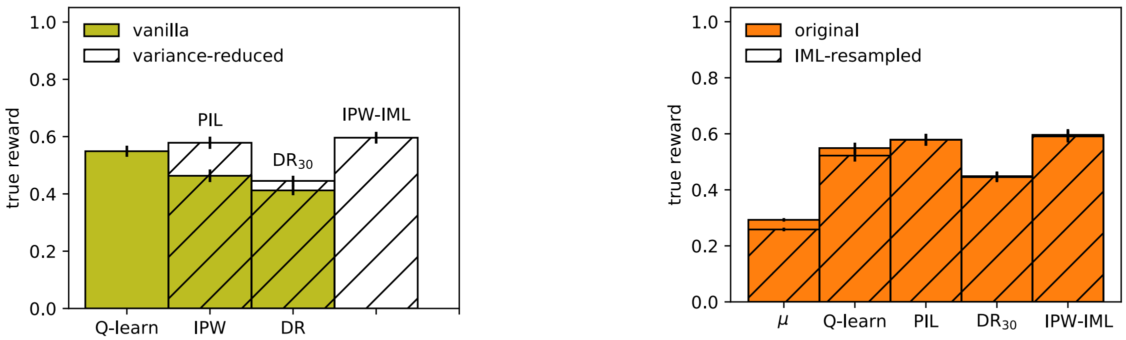
$$\phi(x,a) = x^\top U V^\top a + w^\top a, \tag{45}$$

where $w \in \mathbb{R}^p$ is the first-order coefficient vector and $U, V \in \mathbb{R}^{p\times r}$ are second-order coefficient matrices. We took $r = 256$. We also experimented with deeper models and nonlinear activation functions, but these approaches did not seem to improve IML much further.

While the IML loss improves 10%, the offline click rate improved much less, around 0.4×10^4. We used second-order models when available, except for the original first-order POEM for convenience. We believe its inferior performance may be partially due to lack of model depth as well as a possibility to yield policies that are more saturated, which may lead to inferior generalization. It is common for a policy to become much more complex to yield minimal improvements, but part of our conclusion also suggests keeping simpler models to improve exploration properties.

L Results on other UCI datasets

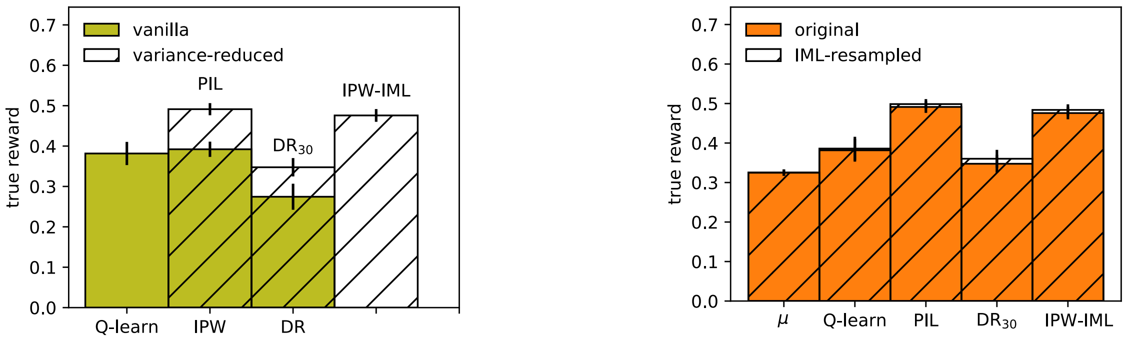
Results on UCI datasets largely follow similar patterns in Figure 3, with varying levels of benefits from variance reduction techniques. The effectiveness of clipped-IPWE, clipped-DR, and IPWE-IML depends on the amount of misspecification that we can artificially inject using rank-2 models (which underfit the problems). When the differences are small, we also observe that rank-2 models are sufficient to imitate the logging policy with near-zero IML losses. It often implies the true models are considerably simpler, the biases are considerably smaller, and the optimal strategy might still be Q-learning. The only negative evidence is from wdbc dataset, where the original problem was to identify benign against malign cancers but we modified the problem to classification of the actual cancer type, which are noisier labels.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

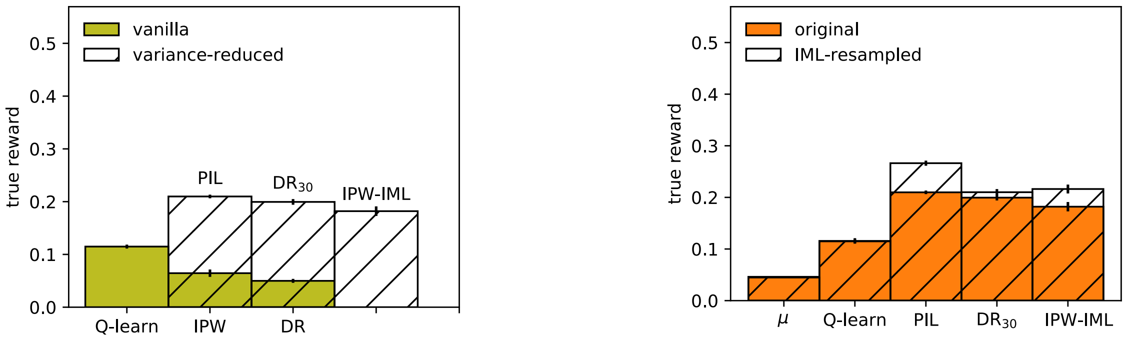
Figure 6: Multiclass-to-bandit conversion on UCI ecoli dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

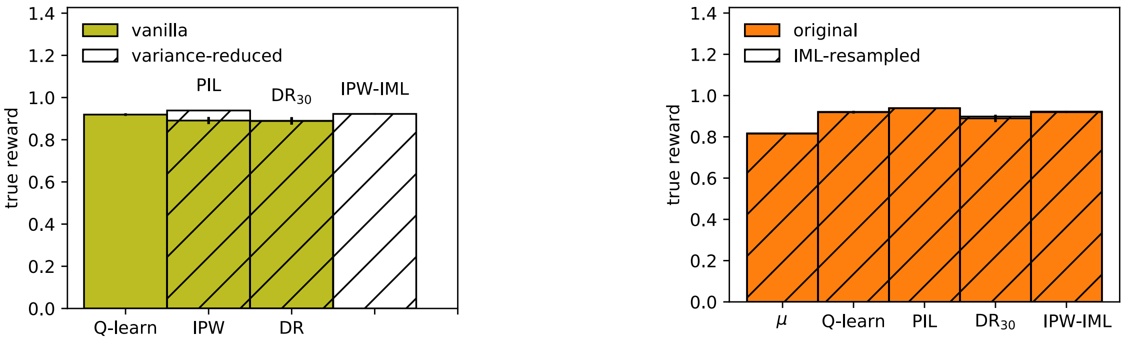
Figure 7: Multiclass-to-bandit conversion on UCI glass dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

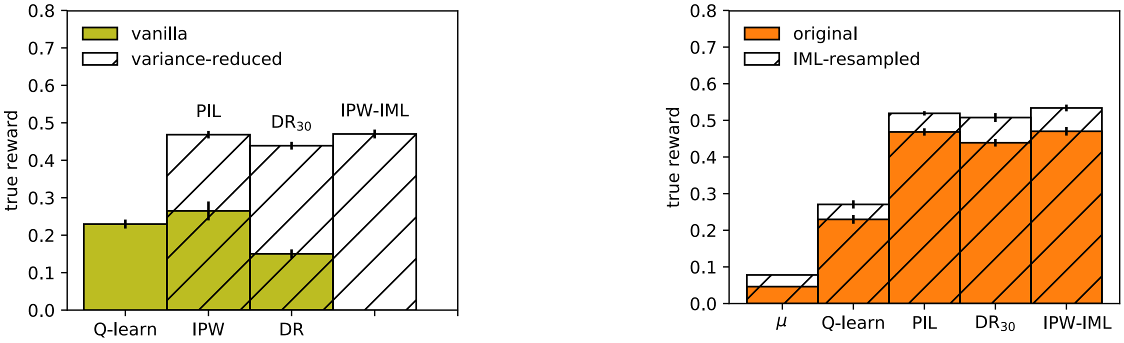
Figure 8: Multiclass-to-bandit conversion on UCI letter dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

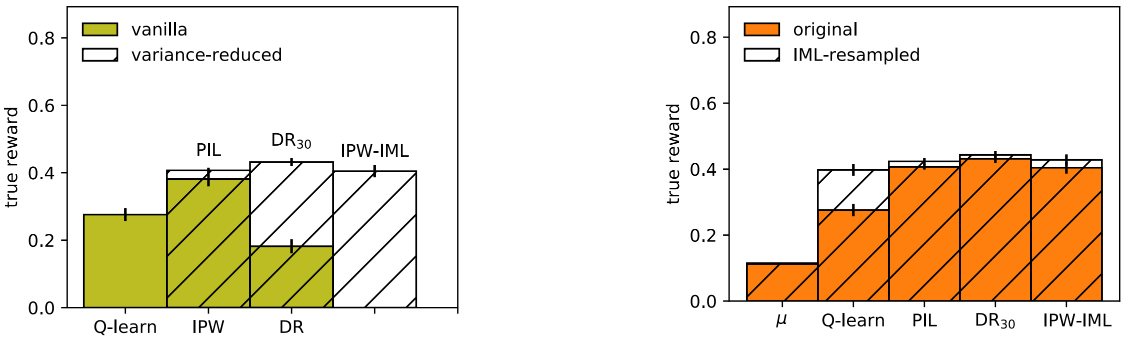
Figure 9: Multiclass-to-bandit conversion on UCI page-blocks dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

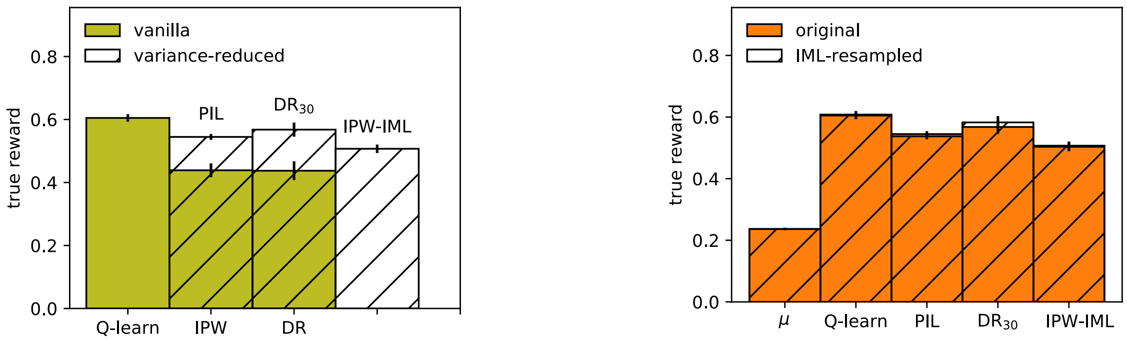
Figure 10: Multiclass-to-bandit conversion on UCI pendigits dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

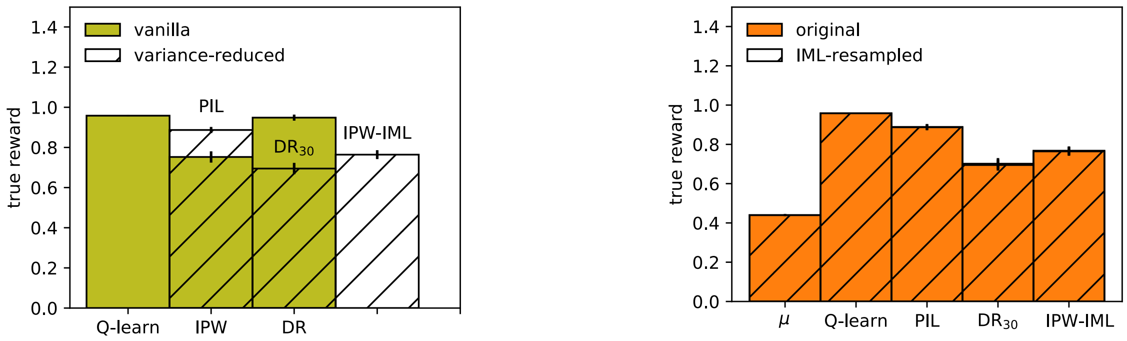
Figure 11: Multiclass-to-bandit conversion on UCI satimage dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

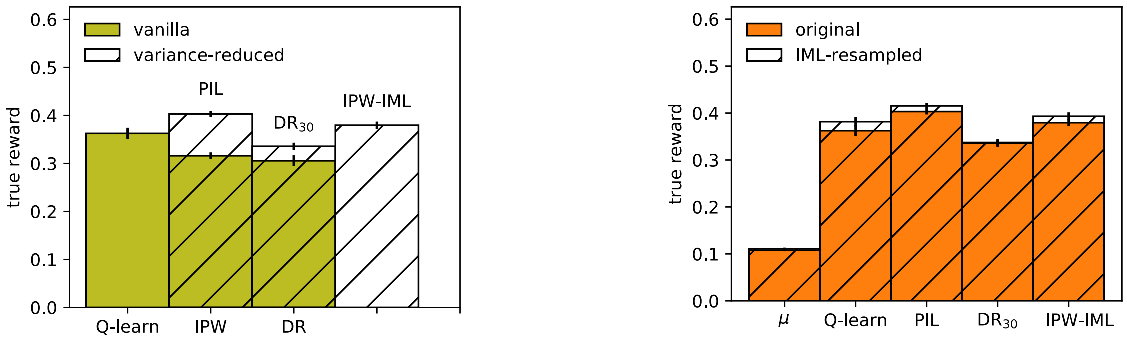
Figure 12: Multiclass-to-bandit conversion on UCI vehicle dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

Figure 13: Multiclass-to-bandit conversion on UCI wdbc dataset.



(a) Variance reduction compared with vanilla methods.

(b) Online application of IML-resampling

Figure 14: Multiclass-to-bandit conversion on UCI yeast dataset.