Empirical Likelihood for Contextual Bandits

Nikos Karampatziakis
Microsoft Dynamics 365 AI
nikosk@microsoft.com

John Langford
Microsoft Research
jcl@microsoft.com

Paul Mineiro
Microsoft Research
pmineiro@microsoft.com

Abstract

We propose optimization-based techniques for off-policy evaluation, confidence intervals (CIs), and learning in contextual bandits, using the framework of empirical likelihood (EL). Each of these is empirically evaluated to show good performance against strong baselines in finite sample regimes.

1 Introduction

Contextual Bandits [ACFS02, LZ07] are now in widespread practical use ([LCLS10, CABJ17, PGC+14]). Key to their value is the ability to do *counterfactual estimation* [HT52] of the value of any policy enabling sound train/test regimes similar to supervised learning. A limiting factor on the lower (data) scale of application is the variance of counterfactual estimation. How can we find the tightest-possible CI on counterfactual estimates? And since tight CIs are deeply dependent on the form of their estimate, how can we find a tight estimate? And given what we discover, how can we leverage this for improved learning algorithms?

We discover good answers to these questions through the application of EL [Owe01].

Applying this first to estimation, we construct a simply-specified estimator in section 2.1 by solving a low dimensional convex problem. The estimator has substantially lower mean squared error than widely used alternatives and the gap is most pronounced in regimes where the number of samples n is of the same order as the smallest inverse probability $1/p$ of an action. Next we elaborate a computationally tractable asymptotically exact CI in section 2.2. Typically confidence intervals are either small but undercover, or guarantee prescribed coverage but are too wide to be useful. Our interval is both small and (despite having only an asymptotic guarantee) empirically honors prescribed coverage. Turning to learning in section 2.3, we use our CI to construct a robust counterfactual learning objective with which we experiment in section 3.

Our contributions: The estimator, CI, and learning objective presented here are all new. Of these, the estimator and learning objective are useful improvements, while the CI is a large improvement over previous approaches as shown in figure 1.

1.1 Related Work

The EL framework [Owe01] forms the basis for our approach. It is a non-parametric maximum likelihood approach that treats the sample as a realization from a multinomial distribution with an infinite number of categories. Surprisingly, EL results in both efficient algorithms and efficient estimators with guarantees similar to those of parametric maximum likelihood with a well specified model. Precise asymptotic optimality statements are in [Owe01] section 13.5 and [Kit01].

There are many previous estimators for contextual bandits. The simplest one is the "Inverse Propensity Score" (IPS) approach [HT52] which is unbiased, but suffers from high variance. The Self-Normalized IPS (SNIPS) [SJ15b] estimate is a simple modification which is biased but has superior mean squared error. An orthogonal way to reduce variance is to incorporate a reward estimator. This can be done via doubly robust (DR) estimation [RR95, DLL11] which is unbiased and has lower

Preprint. Under review.

(a) 95% CI width

(b) Empirical coverage @ 95% nominal

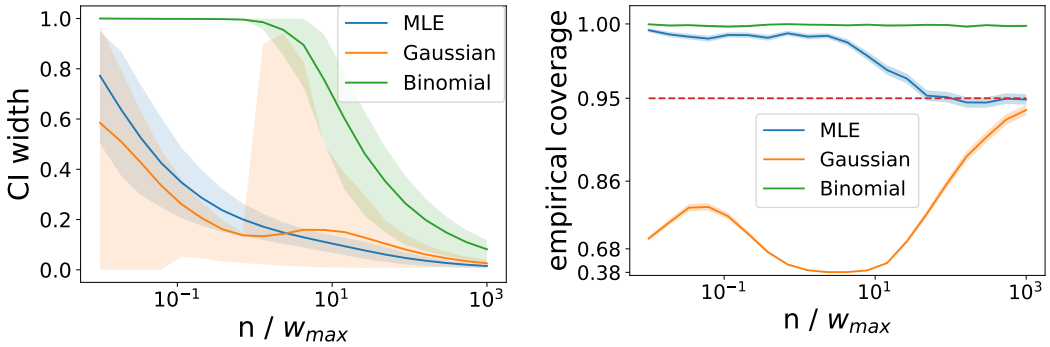


Figure 1: A comparison of CIs on contextual bandit data. The MLE CI is dramatically tighter than an approach based on a binomial CI while avoiding chronic undercoverage as per the asymptotic Gaussian CI. Note that in some regimes, the asymptotic Gaussian CI both undercovers *and* has greater average width. This is possible as the MLE CI has a different functional form than a multiplier on the Gaussian CI. On the left, shaded area represents 90% of the empirical distribution indicating the MLE CI width varies less over realizations. On the right, shaded area represents 4 times the standard error of the mean indicating coverage differences are everywhere statistically significant.

variance when the reward estimator is good. The SWITCH estimator [WAD17] provides a method for switching between a DR estimator and direct application of a reward estimator to optimize mean square error. The estimator presented here is a natural alternative to IPS and SNIPS, can be seamlessly combined with DR or SWITCH (replacing their IPS part), and provides lower mean squared error. We briefly discuss how to incorporate a reward predictor.

There is less work on CIs for contextual bandits. A simple asymptotically motivated approach fits the observations to a normal distribution and uses the CI for the normal [LCKG15]. This results in a too-tight interval, e.g., when all observed rewards are zero. Our CIs are also asymptotically motivated but we incorporate *missing mass* to empirically mitigate undercoverage.

There are many contextual bandit learning algorithms including theoretical [ACFS02, LZ07], reduction oriented [DLL11], optimization-based [SJ15a], and Bayesian style [MKLL12] algorithms. A recent paper about empirical contextual bandit learning [BAL18] informs our experiments.

Ideas from EL have been applied to robust supervised learning [DGN16]. Our combination of CIs with learning is a contextual bandit analogue to robust supervised learning. Regularizing counterfactual learning via lower-bound optimization has also been previously considered [SJ15a, SLA+15, KL02].

2 Empirical Likelihood Applications to Contextual Bandits

We consider the standard contextual bandit problem, with contexts $x \in \mathcal{X}$, a finite set of actions A, and bounded rewards $r \in A \to [r_{\min}, r_{\max}]$. The environment generates $(x, r) \sim D$ i.i.d. and reveals x to the policy, the policy samples $a \in A$ from a context-conditional distribution $\pi : X \to \mathcal{P}(A)$ and observes reward $r(a)$. We denote the all ones vector as $\vec{1}$ and the indicator function as $\mathbb{1}$.

2.1 Off-Policy Evaluation

We assume a dataset $\{(x_n, a_n, r(a)_n)\}_{n \in N}$, generated from a fixed historical policy h, with which we want to estimate the value of another fixed policy π. The value of π is $V(\pi) = \mathbb{E}_{\substack{(x,r)\sim D \\ a\sim\pi(x)}}[r(a)] = \mathbb{E}_{\substack{(x,r)\sim D \\ a\sim h(x)}}\left[\frac{\pi(a|x)}{h(a|x)} r(a)\right]$, where $\pi(a|x) = \mathbb{E}_{a'\sim\pi(x)}[\mathbb{1}_{a=a'}]$ and analogously for $h(a|x)$. Define $w \doteq \frac{\pi(a|x)}{h(a|x)}$, and assume the joint distribution of w and r has (possibly infinite) discrete support.[1] Then we

[1] This is for ease of exposition; EL is applicable to data coming from continuous distributions.

can represent the joint distribution of w and r for data generated from h as a matrix Q via

$$Q_{w,r} \doteq \mathbb{E}_{\substack{(x,r)\sim D \\ a\sim h}}\left[\mathbb{1}_{w=\frac{\pi(a|x)}{h(a|x)}}\mathbb{1}_{r_a=r}\right], \quad V(\pi) = \sum_{w,r} wQ_{w,r}r = \vec{w}^\top Q\vec{r}.$$

We assume w is bounded in $[w_{\min}, w_{\max}]$; given h and π these limits are easily determined in practice.[2] We assume $0 \leq w_{\min} < 1$ and $w_{\max} > 1$. This precludes $w_{\min} = w_{\max} = 1$, which is the case of on-policy evaluation. To estimate V we first estimate Q and then use $\hat{V}(\pi) = w^\top \hat{Q} r$. To estimate Q we solve the following empirical maximum likelihood optimization:

$$\begin{aligned} \max_{Q\succeq 0} \quad & \sum_n \log(Q_{w_n,r_n}), && (1)\\ \text{subject to} \quad & \vec{w}^\top Q\vec{1} = 1, && (\beta)\\ & \vec{1}^\top Q\vec{1} = 1. && (\gamma) \end{aligned}$$

We defer the motivation for this procedure to section 2.2. The dual variable for each constraint is shown in parentheses. The constraints (β) and (γ) respectively normalize the counterfactual and factual distribution. Theorem 1 characterizes the solution.

Theorem 1. *The solution to equation* (1) *satisfies for every observed* (w, r) *pair*

$$\hat{Q}_{w,r} = \sum_n \frac{\mathbb{1}_{w=w_n, r=r_n}}{\beta^*(w_n - 1) + N}, \tag{2}$$

where β^* *is the solution to the dual problem*

$$\sup_\beta \sum_n \log\left(\beta(w_n - 1) + N\right) \quad \textit{subject to} \qquad \forall w : \beta(w-1) + N \geq 0. \tag{3}$$

Moreover, if $w_{\min}$ *or* $w_{\max}$ *are not observed the solution to* (1) *puts mass on these according to the solution of the non-negative linear feasibility program*

$$w_{\min}\hat{q}_{\min} + w_{\max}\hat{q}_{\max} = 1 - \sum_n \frac{w_n}{\beta^*(w_n-1)+N}, \quad \hat{q}_{\min} + \hat{q}_{\max} = 1 - \sum_n \frac{1}{\beta^*(w_n-1)+N},$$

where $\hat{q}_{\min} \geq 0$ *and* $\hat{q}_{\max} \geq 0$ *are associated with* $w_{\min}$ *and* $w_{\max}$ *respectively. This additional mass can be distributed arbitrarily over* $r \in [r_{\min}, r_{\max}]$*, implying the value estimate is an interval.*

Proof. See appendix A.1. □

When EL estimators are subject to additional constraints they can place mass on unobserved data[GŠ+17]. In our case the additional mass is due to the β constraint. Once both $w_{\min}$ and $w_{\max}$ have been observed, all mass is placed upon the sample. Until then, it might be possible to increase the likelihood of the observed data while satisfying the constraint by placing mass on an unobserved extreme value. The dual is a one dimensional convex problem which we solve to accuracy ϵ in $O(N\log(\frac{N}{\epsilon}))$ time via bisection[3] within $(\frac{N}{1-w_{\max}}, \frac{N}{1-w_{\min}})$. The resulting value estimate is

$$\hat{V}^{(\text{MLE})} = \rho + \sum_n \frac{w_n(r_n - \rho)}{\beta^*(w_n-1)+N}, \tag{4}$$

where $\rho \in [r_{\min}, r_{\max}]$ is arbitrary and only affects $\hat{V}^{(\text{MLE})}$ if mass is placed outside the sample.

Comparing the MLE with the standard IPS [HT52] and SNIPS [SJ15b] estimates in the same notation,

$$\hat{V}^{(\text{IPS})} = \sum_n \frac{w_n r_n}{N}, \quad \hat{V}^{(\text{SNIPS})} = \sum_n \frac{w_n r_n}{\sum_m w_m},$$

[2] For example, if π is deterministic, then $w_{\min} = 0$ and $w_{\max} = \frac{1}{\min_{x,a} h(a|x)}$.

[3] The factor N can be refined to be the size of a histogram of (w, r) pairs. On a laptop, a C++ implementation finds the MLE of a histogram with $N = 10^8$ in a second (to single precision).

and assuming $\hat{q}_{\max} = \hat{q}_{\min} = 0$, reveals that IPS corresponds to $\beta^* = 0$. This implies the β constraint is not active at the optimum, i.e., IPS is the MLE when the sum of the importance weights of the realization equals the number of examples. In that case SNIPS is also the MLE.

Incorporating a reward predictor The MLE estimator is analogous to the IPS estimator, which can be augmented with a reward predictor via the DR estimator [DLL11]. Analogues to the DR predictor exist in the EL literature [LYLL16]. The simplest approach is to apply EL on the residual, and then add back the expected shift. Given predictor $\hat{r} : \mathcal{X} \times A \to [r_{\min}, r_{\max}]$, we construct data

$$(w_n, \tilde{r}_n) \leftarrow \left(\frac{\pi(a_n|x_n)}{h(a_n|x_n)}, r_n - \hat{r}(x_n, a_n) \right),$$

apply the MLE on this data (with modified $\tilde{r}_{\min}$ and $\tilde{r}_{\max}$), and then adjust the result via

$$\hat{V}^{(\text{rpmle})} = \hat{V}^{(\text{mle})} + \sum_n \sum_a \pi(a_n|x_n)\hat{r}(x_n, a_n).$$

2.2 Confidence Intervals

The MLE comes with an asymptotically exact coverage interval defined by a likelihood level set. Let

$$\begin{aligned}
\log R_w &= \sup_Q \left\{ \sum_n \log(N Q_{w_n, r_n}) \middle| \vec{w}^\top Q \vec{1} = 1, \vec{1}^\top Q \vec{1} = 1 \right\} \\
\log R_{wr}(v) &= \sup_Q \left\{ \sum_n \log(N Q_{w_n, r_n}) \middle| \vec{w}^\top Q \vec{1} = 1, \vec{1}^\top Q \vec{1} = 1, \vec{w}^\top Q \vec{r} = v \right\}
\end{aligned}$$

with $Q^{(\text{prof})}$ the maximizer of $\log R_{wr}(V(\pi))$ and $Q^{(\text{mle})}$ the maximizer of $\log R_w$. Then we can apply Theorem 3.5 of [Owe01] which states that $-2(\log R_{wr}(V(\pi)) - \log R_w) \to \chi^2_{(1)}$ in distribution as $n \to \infty$. Letting $\chi^{2,(1-\alpha)}_{(1)}$ be the $1-\alpha$ quantile of the χ-squared distribution with one degree of freedom, we have that for all α

$$\lim_{n\to\infty} \Pr\left(\sum_n \log Q^{(\text{mle})}_{w_n, r_n} - \sum_n \log Q^{(\text{prof})}_{w_n, r_n} \le \frac{1}{2} \chi^{2,(1-\alpha)}_{(1)} \right) = 1 - \alpha$$

To get the lower bound of the CI for $V(\pi)$ we just search for the minimum possible $\hat{V}(\pi)$ that still respects the prescribed coverage $1-\alpha$. This leads to the optimization problem

$$\begin{aligned}
\min_{Q \succeq 0} \quad & \vec{w}^\top Q \vec{r}, && (5) \\
\text{subject to} \quad & \vec{w}^\top Q \vec{1} = 1, && (\beta) \\
& \vec{1}^\top Q \vec{1} = 1, && (\gamma) \\
& \Delta + \sum_n \log Q_{w_n, r_n} \ge \sum_n \log Q^{(\text{mle})}_{w_n, r_n}, && (\kappa)
\end{aligned}$$

where Δ should asymptotically be $\frac{1}{2}\chi^{2,1-\alpha}_{(1)}$. Considerations from the proof in [Owe01] suggest that setting Δ to be half the $1-\alpha$ quantile $F^{1-\alpha}_{(1,N-1)}$ of the corresponding F-distribution leads to better coverage for small samples. Theorem 2 characterizes the solution.

Theorem 2. *The solution to equation* (5) *satisfies for every observed* (w, r) *pair*

$$\hat{Q}_{w,r} = \kappa^* \sum_n \frac{\mathbb{1}_{w=w_n, r=r_n}}{\gamma^* + \beta^* w_n + w_n r_n}, \tag{6}$$

where $(\kappa^*, \beta^*, \gamma^*)$ *is the solution to the dual problem*

$$\begin{aligned}
\sup_{\beta,\gamma} \quad & N \exp\left(-\frac{\Delta}{N} + \frac{1}{N} \sum_n \log \frac{\gamma + \beta w_n + w_n r_n}{\beta^{(mle)}(w_n - 1) + N} \right) - \gamma - \beta, && (7) \\
\textit{subject to} \quad & \forall w, r : \gamma + \beta w + w r \ge 0, \\
& \kappa = \exp\left(-\frac{\Delta}{N} + \frac{1}{N} \sum_n \log \frac{\gamma + \beta w_n + w_n r_n}{\beta^{(mle)}(w_n - 1) + N} \right).
\end{aligned}$$

where $\beta^{(mle)}$ is the solution of (3)*. Mass may be allocated to unobserved extreme values $(w_{\min}, r_{\min})$ and $(w_{\max}, r_{\min})$ according to the solution of the linear feasibility program*

$$w_{\min}\hat{q}_{\min} + w_{\max}\hat{q}_{\max} = 1 - \sum_n \frac{\kappa^* w_n}{\gamma^* + \beta^* w_n + w_n r_n}, \quad \hat{q}_{\min} + \hat{q}_{\max} = 1 - \sum_n \frac{\kappa^*}{\gamma^* + \beta^* w_n + w_n r_n},$$

where $\hat{q}_{\min} \geq 0$ and $\hat{q}_{\max} \geq 0$ correspond to $(w_{\min}, r_{\min})$ and $(w_{\max}, r_{\min})$ respectively.

Proof. See appendix A.3. □

Given $(\kappa^*, \beta^*, \gamma^*)$ from Theorem 2, the lower bound is

$$V^{(\text{lb})}(\pi) = r_{\min} + \kappa^* \sum_n \frac{w_n(r_n - r_{\min})}{\gamma^* + \beta^* w_n + w_n r_n}.$$

For the upper bound, we apply Theorem 2 to $-r$ placing additional mass at $r_{\max}$ instead of $r_{\min}$.

2.3 Learning From Logged Bandit Feedback

Here the goal is to learn a policy π based upon a dataset $\{(x_n, a_n, h(a_n|x_n), r(a)_n)\}_{n \in N}$ generated from a fixed historical policy h, i.e., without interacting with the data generating process. One strategy is to leverage a counterfactual estimator to reduce policy learning to optimization [LCKG15], suggesting the use of the MLE estimator in the objective. We can go one step further and use the lower bound of the MLE CI as the objective that policy π should maximize. This is similar to recent work which employs regularized learning, e.g., based upon empirical Bernstein bounds [SJ15a] or divergence-based trust regions [SLA+15, KL02, DGN16].

Suppose π is parameterized by θ. For each θ, π induces a set of importance weights $w_n(\theta)$ and solving (7) gives optimal values $(\kappa^*(\theta), \beta^*(\theta), \gamma^*(\theta))$. Reward lower bound maximization becomes:

$$\begin{aligned} \sup_\theta \quad & r_{\min} + \kappa^*(\theta) \sum_n \frac{w_n(\theta)(r_n - r_{\min})}{\gamma^*(\theta) + \beta^*(\theta) w_n(\theta) + w_n(\theta) r_n}, \\ \text{subject to} \quad & w_n(\theta) = \frac{\pi(a_n|x_n;\theta)}{h(a_n|x_n)}, \end{aligned} \tag{8}$$

We can view lower bound optimization as a game between two players: one controlling the distribution Q via the dual variables (κ, β, γ) and one controlling the policy π. Theorem 2 shows how to implement the first player given any policy π. The second player can be implemented by a learning algorithm that searches for the policy with the best reward under Q. Optimizing the MLE policy value estimate is analogous but leveraging the dual equation (3).

3 Experiments

Our code is available at `http://github.com/pmineiro/elfcb`. All details are in the appendix.

Off Policy Evaluation, Synthetic Data Here we first sample an environment, a set of examples is sampled from that environment, and the squared error of the value estimate is computed. Figure 2 shows the mean squared error (MSE) over 10,000 environments for various estimators: The best constant predictor of 1/2 ("Constant"), the DR estimator with a constant predictor of 1/2 clipped to $[0, 1]$ ("ClippedDR"), the self-normalized IPS estimator ("SNIPS"), and the MLE. When a small number of large importance weight events is expected ($n \approx w_{\max}$), both ClippedDR and SNIPS suffer due to their poor handling of the $\mathbb{E}[w] = 1$ constraint. Asymptotically all estimators are similar.

Off Policy Evaluation, Realistic Data We compare the mean square error of MLE, IPS, and SNIPS on 40 datasets from OpenML [VvRBT13]; using the true value of π on the evaluation set (available because the underlying dataset is fully observed and the action distribution of π is known). For each dataset we evaluate multiple times, with different actions chosen by the historical policy h. Table 1 shows the results of a paired t-test per dataset and 95% confidence level: "tie" indicates null result, and "win" or "loss" indicates significantly better or worse. IPS is clearly dominated, The MLE is overall superior. Additional details and results are presented in Table **??** of appendix B.2.

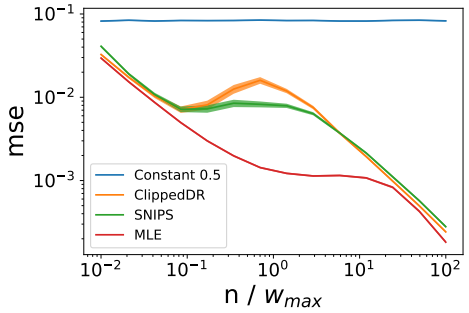


Figure 2: Mean squared error of MLE and other estimators on synthetic data. Asymptotics are similar while MLE dominates in the small sample regime. Line width is 4 times the standard error of the population mean.

MLE vs.	Exploration	Wins	Ties	Losses
IPS	$\epsilon = 0.25$	28	9	3
	bags=10	13	19	8
	cover=10	16	16	9
SNIPS	$\epsilon = 0.25$	2	37	1
	bags=10	7	30	3
	cover=10	7	33	0

Table 1: Off-policy evaluation results where $\epsilon = 0.25$ is ϵ-greedy exploration, bags=10 is bootstrap exploration with 10 replicas, and cover=10 is online cover [AHK+14] with 10 policies.

Confidence Intervals, Synthetic Data Figure 1 shows the mean width and empirical coverage over 10,000 environment samples for various CIs at 95% nominal coverage. Binomial CI is the Clopper Pearson interval on the random variable $\frac{w}{w_{\max}}R$. Asymptotic Gaussian is standard interval motivated by the central limit theorem and typically violates nominal coverage. The MLE interval is narrow and obeys nominal coverage despite only having asymptotic guarantees. Once again estimation is most challenging when $n \approx w_{\max}$ Appendix B.3 contains further details and two additional figures. The first shows that the MLE CI width does *not* depend upon the cardinality of the support. The second demonstrates that the MLE CI width adapts to the difficulty of the problem.

Confidence Intervals, Realistic Data We use the same datasets mentioned above. The Binomial Confidence Interval (Binom) overcovers and has wider intervals. MLE widths are comparable to asymptotic Gaussian (AG) on this data, but AG undercovers. A 95% binomial CI on the coverage of AG is $[90.0\%, 92.3\%]$, effectively concluding undercoverage.

Table 2: Off-Policy Confidence Intervals

Technique	Coverage (Average)	Width Ratio (Median)
MLE	0.975	n/a
Binom	0.996	2.89
AG	0.912	0.99

Learning From Logged Bandit Feedback We heuristically alternate between solving the dual problem with the policy fixed and then optimizing the policy with the dual variables fixed. When we optimize the policy we supply different importance weights on each example depending upon the learning objective. Specifically we use importance weights:

$$\nu_n(\theta) = w_n(\theta), \qquad \text{(baseline)}$$

$$\nu_n(\theta) = \frac{w_n(\theta)}{\beta^*(\theta)(w_n(\theta) - 1) + N}, \qquad \text{(MLE)}$$

$$\nu_n(\theta) = \frac{\kappa^*(\theta) w_n(\theta)}{\gamma^*(\theta) + \beta^*(\theta) w_n(\theta) + w_n(\theta) r_n}, \qquad \text{(CI LB)}$$

with $w_n(\theta) = \frac{\pi(a_n|x_n;\theta)}{h(a_n|x_n)}$. Details are in appendix B.5.

Table 3 in the appendix shows the results of a paired t-test with 95% confidence level: "tie" indicates null result, and "win" or "loss" indicates significantly better or worse evaluation value for CI lower bound or MLE. Using the CI lower bound overall yields superior results. Using the MLE estimate also provides some lift but is less effective than using the CI lower bound.

References

[ABC+16] Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[ACFS02] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[AHK+14] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

[BAL18] Alberto Bietti, Alekh Agarwal, and John Langford. Practical evaluation and optimization of contextual bandit algorithms. *CoRR*, abs/1802.04064, 2018.

[CABJ17] Ashok Chandrashekar, Fernando Amat, Justin Basilico, and Tony Jebara. Artwork personalization at netflix. *The Netflix Tech Blog*, 2017.

[DGN16] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[GŠ+17] Marian Grendár, Vladimír Špitalskỳ, et al. Multinomial and empirical likelihood under convex constraints: Directions of recession, fenchel duality, the pp algorithm. *Electronic Journal of Statistics*, 11(1):2547–2612, 2017.

[HT52] 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.

[Kit01] Yuichi Kitamura. Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 69(6):1661–1672, 2001.

[KL02] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

[LCKG15] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934. ACM, 2015.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *CoRR*, abs/1003.0146, 2010.

[LLS07] John Langford, Lihong Li, and Alexander Strehl. Vowpal wabbit open source project. *URL https://github. com*, 2007.

[LYLL16] Tianqing Liu, Xiaohui Yuan, Zhaohai Li, and Aiyi Liu. An efficient and doubly robust empirical likelihood approach for estimating equations with missing data. *arXiv preprint arXiv:1612.00922*, 2016.

[LZ07] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 817–824, 2007.

[MKLL12] Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012.

[Owe01] Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

[PGC+14] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. Poptherapy: coping with stress through pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2014, Oldenburg, Germany, May 20-23, 2014*, pages 109–117, 2014.

[RR95] 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.

[SJ15a] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[SJ15b] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems*, pages 3231–3239, 2015.

[SLA+15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[VvRBT13] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[WAD17] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3589–3597, 2017.

A Off-Policy Evaluation

A.1 Proof of Theorem 1

Theorem 1. *The solution to equation* (1) *satisfies for every observed* (w, r) *pair*

$$\hat{Q}_{w,r} = \sum_n \frac{\mathbb{1}_{w=w_n, r=r_n}}{\beta^*(w_n - 1) + N}, \tag{2}$$

where β^* *is the solution to the dual problem*

$$\sup_\beta \sum_n \log\left(\beta(w_n - 1) + N\right) \quad \textit{subject to} \qquad \forall w : \beta(w-1) + N \geq 0. \tag{3}$$

Moreover, if $w_{\min}$ *or* $w_{\max}$ *are not observed the solution to* (1) *puts mass on these according to the solution of the non-negative linear feasibility program*

$$w_{\min}\hat{q}_{\min} + w_{\max}\hat{q}_{\max} = 1 - \sum_n \frac{w_n}{\beta^*(w_n - 1) + N}, \quad \hat{q}_{\min} + \hat{q}_{\max} = 1 - \sum_n \frac{1}{\beta^*(w_n - 1) + N},$$

where $\hat{q}_{\min} \geq 0$ *and* $\hat{q}_{\max} \geq 0$ *are associated with* $w_{\min}$ *and* $w_{\max}$ *respectively. This additional mass can be distributed arbitrarily over* $r \in [r_{\min}, r_{\max}]$*, implying the value estimate is an interval.*

Mass on the Realization Starting from equation (1) we construct the Lagrangian:

$$L(Q, \beta, \gamma) = \beta(\vec{w}^\top Q\vec{1} - 1) + \gamma(\vec{1}Q\vec{1} - 1) + \sum_n -\log(Q_{w_n, r_n}).$$

The Lagrange dual function is

$$g(\beta, \gamma) = \inf_{Q \geq 0} L(Q, \beta, \gamma) = \inf_{Q \geq 0} \beta(\vec{w}Q\vec{1} - 1) + \gamma(\vec{1}^\top Q\vec{1} - 1) + \sum_n -\log(Q_{w_n, r_n})$$

$$= -\beta - \gamma + \inf_{Q \geq 0} \sum_{(w,r)} \left((\beta w + \gamma) Q_{w,r} - \sum_n \mathbb{1}_{w=w_n, r=r_n} \log(Q_{w,r}) \right).$$

This is a separable optimization and each term can be optimized separately. Observe that for $c \geq 0$ and $y \geq 0$

$$\inf_{q \geq 0} yq - c\log(q) = c - c\log(c) + c\log(y), \tag{9}$$

with the infimum attained at $q^* = c/y$ (and unbounded if $y < 0$). This, together with later simplifications establishes the form of Q. Using $c_{w,r} = \sum_n \mathbb{1}_{w=w_n, r=r_n}$ and (9) leads to

$$g(\beta, \gamma) = \begin{cases} -\beta - \gamma + N - \sum_{(w,r)} c_{w,r} \log(c_{w,r}) + \sum_n \log(w_n\beta + \gamma) & \text{if } \forall w : \beta w + \gamma \geq 0 \\ -\infty & \text{otherwise} \end{cases}.$$

The dual for equation (1) follows directly from this and strong duality. Ignoring constants yields

$$\sup_{\beta, \gamma} -\beta - \gamma + \sum_n \log(w_n\beta + \gamma) \quad \text{subject to} \qquad \forall w : \beta w + \gamma \geq 0.$$

γ can be eliminated by summing the KKT stationarity conditions. For this we introduce dual variables $\phi \succeq 0$ corresponding to $Q \succeq 0$, and leverage complementary slackness and primal feasibility:

$$\sum_n \frac{\mathbb{1}_{w_n=w, r_n=r}}{Q_{w,r}} = \phi_{w,r} + w\beta + \gamma \qquad \text{(KKT stationarity)},$$

$$\Rightarrow \sum_{w,r} Q_{w,r} \sum_n \frac{\mathbb{1}_{w_n=w, r_n=r}}{Q_{w,r}} = \sum_{w,r} Q_{w,r}\phi_{w,r} + \beta\vec{w}^\top Q\vec{1} + \gamma\vec{1}^\top Q\vec{1},$$

$$N = \sum_{w,r} Q_{w,r}\phi_{w,r} + \beta\vec{w}^\top Q\vec{1} + \gamma\vec{1}^\top Q\vec{1}$$

$$= 0 + \beta\vec{w}^\top Q\vec{1} + \gamma\vec{1}^\top Q\vec{1} \qquad \text{(complementary slackness)}$$

$$= \beta + \gamma. \qquad \text{(primal feasibility)}$$

Substitution results in:

$$\sup_{\beta} -N + \sum_n \log(w_n\beta + (N-\beta)) \quad \text{subject to} \qquad \forall w : \beta w_n + N - \beta \geq 0.$$

which ignoring constants gives

$$\sup_{\beta} \sum_n \log((w_n - 1)\beta + N) \quad \text{subject to} \qquad \forall w : \beta(w_n - 1) + N \geq 0,$$

as per equation (3). Equation (2) follows from the KKT stationarity conditions.

Additional Mass For an unobserved (w, r) pair with $Q_{w,r} > 0$ we have

$$\begin{aligned} 0 &= \phi_{w,r} + w\beta + \gamma && \text{(KKT stationarity)} \\ &= w\beta + \gamma && \text{(complementary slackness)} \\ &= (w-1)\beta + N, && \text{(dual variable relationship)} \end{aligned}$$

which due to the inequality constraints can only occur for a single value of w, either the smallest value $w_{\min}$ if $\beta > 0$ or the largest value $w_{\max}$ if $\beta < 0$; unless $\beta = 0$ in which case $1^\top Q 1 = 1$ and there is no missing mass.

If (w, r) is observed than

$$\begin{aligned} 0 &< \sum_n \frac{\mathbb{1}_{w_n=w, r_n=r}}{Q_{w,r}} && \text{(primal feasibility)} \\ &= \phi_{w,r} + w\beta + \gamma && \text{(KKT stationarity)} \\ &= \phi_{w,r} + (w-1)\beta + N && \text{(dual variable relationship)} \\ &= (w-1)\beta + N, && \text{(complementary slackness)} \end{aligned}$$

therefore additional mass can only be assigned to an unobserved importance weight. The distribution over r for this w is not determined, resulting in an interval corresponding to extreme values of r. □

A.2 Primal Recovery

Given the dual optimum β^* of equation (3) we can determine the mass assigned to unobserved w via primal feasibility. Introducing $q_{\min}$ and $q_{\max}$ to represent the mass at $w_{\min}$ and $w_{\max}$ respectively, we have

$$\begin{aligned} \max_{q_{\min}\geq 0, q_{\max}\geq 0} \quad & 1 \\ \text{subject to} \quad & w_{\min} q_{\min} + w_{\max} q_{\max} = 1 - \sum_n \frac{w_n}{\beta^*(w_n - 1) + N}, \\ & q_{\min} + q_{\max} = 1 - \sum_n \frac{1}{\beta^*(w_n - 1) + N}. \end{aligned} \tag{10}$$

Because the dual optimum is determined to finite precision, in practice (10) can be infeasible. Therefore we actually solve the non-negative least squares problem

$$\min_{q_{\min}\geq 0, q_{\max}\geq 0} \quad \left\| \begin{pmatrix} 1 & 1 \\ w_{\min} & w_{\max} \end{pmatrix} \begin{pmatrix} q_{\min} \\ q_{\max} \end{pmatrix} - \begin{pmatrix} 1 - \sum_n \frac{1}{\beta^*(w_n-1)+N} \\ 1 - \sum_n \frac{w_n}{\beta^*(w_n-1)+N} \end{pmatrix} \right\|^2,$$

which is equivalent when (10) is feasible but otherwise is more robust.

A.3 Proof of Theorem 2

Theorem 2. *The solution to equation* (5) *satisfies for every observed* (w, r) *pair*

$$\hat{Q}_{w,r} = \kappa^* \sum_n \frac{\mathbb{1}_{w=w_n, r=r_n}}{\gamma^* + \beta^* w_n + w_n r_n}, \tag{6}$$

where $(\kappa^*, \beta^*, \gamma^*)$ *is the solution to the dual problem*

$$\begin{aligned} \sup_{\beta,\gamma} \quad & N \exp\left(-\frac{\Delta}{N} + \frac{1}{N}\sum_n \log \frac{\gamma + \beta w_n + w_n r_n}{\beta^{(mle)}(w_n - 1) + N}\right) - \gamma - \beta, \qquad (7) \\ \textit{subject to} \quad & \forall w, r : \gamma + \beta w + wr \geq 0, \\ & \kappa = \exp\left(-\frac{\Delta}{N} + \frac{1}{N}\sum_n \log \frac{\gamma + \beta w_n + w_n r_n}{\beta^{(mle)}(w_n - 1) + N}\right). \end{aligned}$$

where $\beta^{(mle)}$ *is the solution of* (3)*. Mass may be allocated to unobserved extreme values* $(w_{\min}, r_{\min})$ *and* $(w_{\max}, r_{\min})$ *according to the solution of the linear feasibility program*

$$w_{\min}\hat{q}_{\min} + w_{\max}\hat{q}_{\max} = 1 - \sum_n \frac{\kappa^* w_n}{\gamma^* + \beta^* w_n + w_n r_n}, \quad \hat{q}_{\min} + \hat{q}_{\max} = 1 - \sum_n \frac{\kappa^*}{\gamma^* + \beta^* w_n + w_n r_n},$$

where $\hat{q}_{\min} \geq 0$ *and* $\hat{q}_{\max} \geq 0$ *correspond to* $(w_{\min}, r_{\min})$ *and* $(w_{\max}, r_{\min})$ *respectively.*

Mass on the Realization The Lagrangian for equation (5) is

$$\begin{aligned} & L(\beta, \gamma, \kappa, Q) \\ & = \vec{w}^\top Q \vec{r} + \kappa\left(-\Delta - \sum_n \log Q_{w_n, r_n} + \sum_n \log Q^{(\text{mle})}_{w_n, r_n}\right) + \gamma\left(\vec{1}^\top Q \vec{1} - 1\right) + \beta\left(\vec{w}^\top Q \vec{1} - 1\right) \\ & = \kappa\left(-\Delta + \sum_n \log Q^{(\text{mle})}_{w_n, r_n}\right) - \gamma - \beta \\ & \quad + \sum_{w,r}\left(-\kappa\left(\sum_n \mathbb{1}_{w=w_n, r=r_n}\right)\log\left(Q_{w,r}\right) + \left(\gamma + \beta w + wr\right) Q_{w,r}\right), \end{aligned}$$

implying dual boundedness (primal feasibility) requires $\forall w, r : \gamma + \beta w + wr \geq 0$. Setting the derivative w.r.t. $Q_{w,r}$ to 0 gives us

$$Q_{w,r} = \frac{\kappa \sum_n \mathbb{1}_{w=w_n, r=r_n}}{\gamma + \beta w + wr}$$

Substituting back in, we get:

$$\begin{aligned} & \inf_{Q \succeq 0} L(\beta, \gamma, \kappa, Q) = L(\beta, \gamma, \kappa, Q)\Big|_{Q_{w,r} = \frac{\kappa \sum_n \mathbb{1}_{w=w_n, r=r_n}}{\gamma + \beta w + wr}} \\ & = \kappa\left(-\Delta + N - N\log\kappa\right) - \gamma - \beta + \kappa \sum_n \left(\log Q^{(\text{mle})}_{w_n, r_n} - \log \frac{\sum_m \mathbb{1}_{w_n = w_m, r_n = r_m}}{\gamma + \beta w_n + w_n r_n}\right) \\ & = N\left(-\kappa\log\kappa - \frac{\gamma + \beta}{N} + \kappa\left(-\frac{\Delta}{N} + 1 + \frac{1}{N}\sum_n \log \frac{\gamma + \beta w_n + w_n r_n}{\beta^{(\text{mle})}(w_n - 1) + N}\right)\right). \end{aligned}$$

Finally, we partially optimize over κ analytically resulting in equation (7).

Additional Mass If the realization is empty, a solution with the smallest possible lower bound can be constructed by placing mass solely on the 2 extreme values of $(w_{\min}, r_{\min})$ and $(w_{\max}, r_{\min})$. Therefore assume the realization is not empty.

Introducing dual variables $\phi \succeq 0$ corresponding to $Q \succeq 0$, for an unobserved (w, r) pair with $Q_{w,r} > 0$ we have

$$\begin{aligned} 0 &= \gamma + \beta w + wr + \phi_{w,r} && \text{(KKT stationarity)} \\ &= \gamma + \beta w + wr. && \text{(complementary slackness).} \end{aligned}$$

This condition can only exist at extreme points because $\gamma + \beta w + wr$ is linear in w and r and $\gamma + \beta w + wr \geq 0$ implies that the only points with equality can be on the boundary of the allowed set of w and r. When $w > 0$, only $r_{\min}$ is eligible, whereas for $w = 0$ all values of r are equivalent for the objective; there only considering $r_{\min}$ is sufficient.

□

Primal Recovery Given the dual optimum $(\beta^*, \gamma^*, \kappa^*)$ of equation (7) we can determine the mass assigned to unobserved (w, r) via primal feasibility by solving a linear program. Again, due to finite precision, non-negative least squares is recommended in practice.

B Experiments

Replication instructions and scripts are available at `http://github.com/pmineiro/elfcb`.

Our experimental design is inspired by the operational cycle of the Decision Service [ABC+16], in which an initial policy is deployed to a production endpoint which makes (randomized) decisions and collects rewards; the resulting data is used to produce a new policy initialized at the previous policy and trained via learning from logged bandit feedback; and then the new policy is optionally deployed if off-policy evaluation on additional collected data compares favorably with the initial policy. Consequently, each dataset is split into Initialize, Learn, and Evaluate sets. The Initialize set is used to produce a plausible initial policy; we use on-policy learning to achieve this. The Learn set corresponds to the off-policy step used to produce a new policy. The Evaluate set corresponds to the gated deployment step.

B.1 Off-Policy Evaluation, Synthetic Data

First, an environment is sampled. For all environments, the historical logging policy is ϵ-greedy with possible importance weights $(0, 2, 1000)$. We choose π to induce the maximum entropy distribution over importance weights consistent with $\mathbb{E}[w^2] = 100$. Rewards are binary with the conditional distribution of reward varying per environment draw such that the value of π is uniformly distributed on $[0, 1]$. Once an environment is drawn a set of examples is sampled from that environment, and the squared error of the value estimate is computed.

B.2 Off-Policy Evaluation, Realistic Data

.

We use the following 40 datasets from OpenML [VvRBT13] identified by their OpenML dataset id: 1216, 1217, 1218, 1233, 1235, 1236, 1237, 1238, 1241, 1242, 1412, 1413, 1441, 1442, 1443, 1444, 1449, 1451, 1453, 1454, 1455, 1457, 1459, 1460, 1464, 1467, 1470, 1471, 1472, 1473, 1475, 1481, 1482, 1483, 1486, 1487, 1488, 1489, 1496, 1498. For each dataset we convert to Vowpal Wabbit format, shuffle the dataset, and utilize up to the first 10,000 examples as data. We utilize a 20%/60%/20% Initialize/Learn/Evaluate split sequentially by line number. Note the shuffle and split is done only once per dataset. We apply a supervised-to-bandit transform as in [DLL11]. We create a historical policy h using on-policy learning on the Initialize dataset, and then learn a new policy π on the Learn dataset using off-policy learning with data drawn from h. These Initialize and Learn steps are done once per dataset. Only the off-policy evaluation step is done multiple times per dataset, and the random variations are due to the different actions selected by h over the Evaluate set. For each evaluation, we compute the squared error of the different predictors, i.e., the squared difference between the off-policy value estimate and the true value of π. Note the true value of π can be computed (and is independent of the choices of h on the evaluation set) because the underlying datasets are fully observed. Using the squared error as the random variable, we apply a paired t-test between MLE and the other predictors to determine win, loss, or tie for each dataset. We use default settings for Vowpal Wabbit except for the choice of exploration strategy.

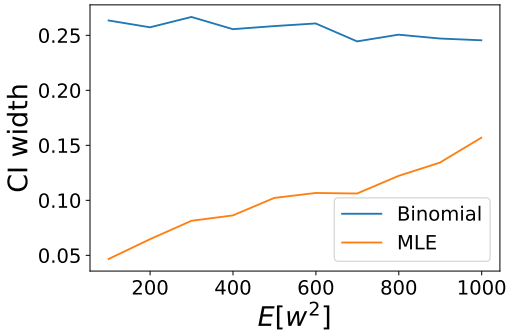
B.3 Confidence Intervals, Synthetic Data

We use the same synthetic ϵ-greedy data as described above.

Figure 3 demonstrates additional interesting properties of the MLE CI.

First, by holding the number of examples fixed but drawing examples from the maximum entropy distribution satisfying different $E[w^2]$, we can change the statistical difficulty of the problem. Larger $E[w^2]$ implies (slightly) more frequent use of the largest importance weight and (more pronounced) less frequent use of the smallest non-zero importance weight. Essentially the policy whose value is being estimated is "more off-policy" when $E[w^2]$ increases, and the MLE CI width is larger.

(a) Adapting to problem difficulty



(b) Independence from cardinality of support

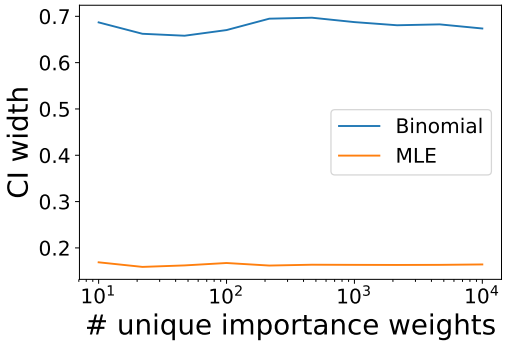


Figure 3: Additional properties of the MLE CI. Left: The MLE CI adapts to the statistical difficulty of the problem, while the Binomial CI does not. Right: The MLE CI does not depend upon the cardinality of the support despite having an equal number of latent parameters.

Second, by adding small magnitude noise to a dataset we can create a family of datasets that are nearly equivalent in all moments but have any desired cardinality. Under these conditions the MLE CI width does not degrade, indicating no fundamental dependence upon the cardinality of the support.

B.4 Confidence Intervals, Realistic Data

We use the same datasets mentioned above, but produce a 95% CI for off-policy evaluation rather than the maximum likelihood estimate. With 40 datasets and 60 evaluations per dataset we have 2400 CIs from which we compute the coverage and the ratio of the width of the interval to the MLE.

B.5 Learning from Logged Bandit Feedback

We use the same 40 datasets as above, but with a 20%/20%/60% Initialize/Learn/Evaluate split. For optimizing the policy parameters and the distribution dual variables, we alternate between solving the dual problem with the policy fixed and then optimizing the policy with the dual variables fixed. To optimize the policy we do a single pass over the data using Vowpal Wabbit as a black-box oracle for learning, supplying different importance weights on each example depending upon the learning

Exploration	CI LB			MLE		
	Wins	Ties	Losses	Wins	Ties	Losses
$\epsilon = 0.05$ greedy	16	18	6	11	26	3
$\epsilon = 0.1$ greedy	16	19	5	13	24	3
$\epsilon = 0.25$ greedy	15	22	3	3	34	3
bagging, 10 bags	21	18	1	11	28	1
bagging, 32 bags	4	26	10	7	31	2
cover, 10 policies	18	21	1	6	30	4
cover, 32 policies	9	29	2	6	34	0

Table 3: Learning From Logged Bandit Feedback. "CI LB" uses the lower bound dual problem of equation (7), while "MLE" uses the estimation dual problem of equation (3).

objective. Specifically we use importance weights:

$$\nu_n(\theta) = w_n(\theta), \tag{baseline}$$

$$\nu_n(\theta) = \frac{w_n(\theta)}{\beta^*(\theta)(w_n(\theta) - 1) + N}, \tag{MLE}$$

$$\nu_n(\theta) = \frac{\kappa^*(\theta) w_n(\theta)}{\gamma^*(\theta) + \beta^*(\theta) w_n(\theta) + w_n(\theta) r_n}, \tag{CI LB}$$

with $w_n(\theta) = \frac{\pi(a_n|x_n;\theta)}{h(a_n|x_n)}$. We use $\Delta = 0.5 F^{0.95}_{(1,N-1)}$ when solving (7). We do 4 passes over the learning set and update the dual variables before each pass.

The Initialize step is done once per dataset, then the Learn and Evaluate steps are done multiple times per dataset. Note the Evaluate step here is using the true value of π, i.e., is deterministic and independent of h given π. Using the evaluation score as the random variable, we apply a paired t-test between MLE and the other predictors to determine win, loss, or tie for each dataset. We use Vowpal Wabbit in IPS learning mode with default settings, and do 4 passes over the data. At the beginning of each pass, we optimize the dual variables holding the policy fixed, then use the resulting dual variables during the learning pass to compute importance weights.