# Off-policy evaluation for slate recommendation

**Adith Swaminathan**
Cornell University
adith@cs.cornell.edu

**Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík,**
**John Langford, Damien Jose, Imed Zitouni**
Microsoft
{akshaykr,alekha,mdudik,jcl,dajose,izitouni}@microsoft.com

## Abstract

This paper studies the evaluation of policies that recommend an ordered set of items (e.g., a ranking) based on some context—a common scenario in web search, ads and recommender systems. We develop the first practical technique for evaluating page-level metrics of such policies offline using *logged past data*, alleviating the need for online A/B tests. Our method models the observed quality of the recommended set (e.g., time to success in web search) as an additive decomposition across items. Crucially, the per-item quality is not directly observed or easily modeled from the item's features. A thorough empirical evaluation reveals that this model fits many realistic measures of quality and theoretical analysis shows exponential savings in the amount of required data compared with prior off-policy evaluation approaches.

## 1 Introduction

In recommendation systems for e-commerce, online advertising, search, or news, we would like to use the data collected during operation to test new content-serving algorithms (called *policies*) along metrics such as revenue and number of clicks [4, 17]. This task is called *off-policy evaluation* and standard approaches, namely *inverse propensity scores* (IPS) [9, 11], require unrealistically large amounts of past data to evaluate whole-page metrics that depend on multiple recommended items, such as when showing ranked lists. Therefore, the industry standard for evaluating new policies is to simply deploy them in weeks-long A/B tests [13]. Replacing or supplementing A/B tests with accurate off-policy evaluation, running in seconds instead of weeks, would revolutionize the process of developing better recommendation systems. For instance, we could perform automatic *policy optimization* (i.e., learn a policy that scores well on whole-page metrics), a task which is currently plagued with bias and an expensive trial-and-error cycle.

The data we collect in these recommendation applications provides only partial information, which is formalized as *contextual bandits* [2, 9, 15]. We study a combinatorial generalization of contextual bandits, where for each *context* a policy selects a list, called a *slate*, consisting of component *actions*. In web search, the context is the search query augmented with a user profile, the slate is the search results page consisting of a list of retrieved documents, and actions are the individual documents. Example metrics are page-level measures such as time-to-success, NDCG (position-weighted relevance) or more general measures of user satisfaction.

The key challenge in off-policy evaluation and optimization is the fact that a new policy, called the *target policy*, recommends different slates than those with recorded metrics in our logs. Without structural assumptions on the relationship between slates and observed metrics, we can only hope to evaluate the target policy if its chosen slates occur in the logged past data with a decent probability. Unfortunately, the number of possible slates is combinatorially large, e.g., when recommending $\ell$ of

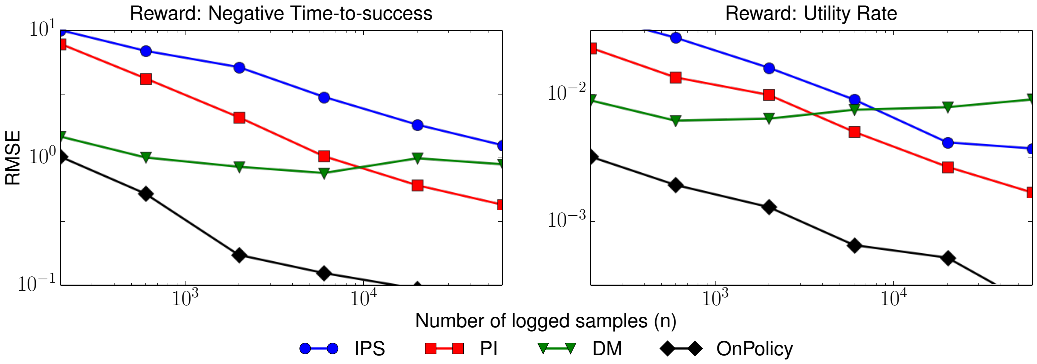

Figure 1: Off-policy evaluation of two whole-page user-satisfaction metrics on proprietary search engine data. Average RMSE over 50 runs on a log-log scale. Our method (pseudoinverse or PI) achieves the best performance for moderate data sizes. The unbiased IPS method suffers high variance, and direct modeling (DM) suffers high bias. OnPolicy is the expensive alternative of deploying the policy. *Improvements of PI are significant, with $p$-values in text.* Details in Sec. 4.3.

$m$ items, there are $m^{\Omega(\ell)}$ ordered sets, so the likelihood of even one match in past data with a target policy is extremely small, leading to a complete breakdown of fully general techniques such as IPS.

To overcome this limitation, some authors [4, 22] restrict their logging and target policies to a parameterized stochastic policy class. Others assume specific parametric (e.g., linear) models relating the observed metrics to the features describing a slate [1, 21, 10, 6, 19]. Yet another paradigm, called *semi-bandits*, assumes that the slate-level metric is a sum of *observed* action-level metrics [12, 14].

We seek to evaluate arbitrary policies, while avoiding strong assumptions about user behavior, as in parametric bandits, or the nature of feedback, as in semi-bandits. We relax restrictions of both parametric and semi-bandits. Like semi-bandits, we assume that the slate-level metric is a sum of action-level metrics that depend on the context, the action, and the position on the slate, but not on the other actions in the slate. Unlike semi-bandits, these per-action metrics are *unobserved by the decision maker*. This model also means that the slate-level metric is linearly related with the unknown vector listing all the per-action metrics in each position. However, this vector of per-action metric values can depend arbitrarily on each context, which precludes fitting a single linear model of rewards (with dimensionality independent of the number of contexts) as usually done in linear bandits.

This paper makes the following contributions:

1. The *additive decomposition assumption* (ADA): a realistic assumption about the feedback structure in combinatorial contextual bandits, which generalizes contextual, linear, and semi-bandits.
2. The *pseudoinverse estimator* (PI) for off-policy evaluation: a general-purpose estimator for any stochastic logging policy, unbiased under ADA. The number of logged samples needed for evaluation with error $\varepsilon$ when choosing $\ell$ out of $m$ items is typically $\mathcal{O}(\ell m/\varepsilon^2)$—an exponential gain over the $m^{\Omega(\ell)}$ complexity of other unbiased estimators. We provide careful distribution-dependent bounds based on the overlap between logging and target policies.
3. Experiments on a real-world search ranking dataset: The strong performance of the PI estimator provides, to our knowledge, the first demonstration of high quality off-policy evaluation of whole-page metrics, comprehensively outperforming prior baselines (see Fig. 1).
4. Off-policy optimization: We provide a simple procedure for learning to rank (L2R) using the PI estimator. Our procedure tunes L2R models directly to online metrics by leveraging pointwise supervised L2R approaches, without requiring pointwise feedback.

Without contexts, several authors have studied a similar linear dependence of the reward on action-level metrics [7, 21]. Their approaches compete with the *best fixed slate*, whereas we focus on evaluating arbitrary *context-dependent* policies. While they also use the pseudoinverse estimator in their analysis (see, e.g., Lemma 3.2 of Dani et al. [7]), its role is different. They construct specific distributions to optimize the explore-exploit trade-off, while we provide guarantees for off-policy evaluation with arbitrary logging distributions, requiring a very different analysis and conclusions.

## 2 Setting and notation

In combinatorial contextual bandits, a decision maker repeatedly interacts as follows:

1. the decision maker observes a *context* $x$ drawn from a distribution $D(x)$ over some space $X$;
2. based on the context, the decision maker chooses a *slate* $\mathbf{s} = (s_1, \ldots, s_\ell)$ consisting of *actions* $s_j$, where a position $j$ is called a *slot*, the number of slots is $\ell$, actions at position $j$ come from some space $A_j(x)$, and the slate $\mathbf{s}$ is chosen from a set of allowed slates $S(x) \subseteq A_1(x) \times \cdots \times A_\ell(x)$;
3. given the context and slate, the environment draws a reward $r \in [-1, 1]$ from a distribution $D(r \mid x, \mathbf{s})$. Rewards in different rounds are independent, conditioned on contexts and slates.

The context space $X$ can be infinite, but the set of actions is of finite size. For simplicity, we assume $|A_j(x)| = m_j$ for all contexts $x \in X$ and define $m \coloneqq \max_j m_j$ as the maximum number of actions per slot. The goal of the decision maker is to *maximize the reward*.

The decision maker is modeled as a *stochastic policy* $\pi$ that specifies a conditional distribution $\pi(\mathbf{s} \,|\, x)$ (a deterministic policy is a special case). The *value* of a policy $\pi$, denoted $V(\pi)$, is defined as the expected reward when following $\pi$:

$$V(\pi) \coloneqq \mathbb{E}_{x\sim D}\mathbb{E}_{\mathbf{s}\sim\pi(\cdot|x)}\mathbb{E}_{r\sim D(\cdot|x,\mathbf{s})}\big[r\big] \ .$$

To simplify derivations, we extend the conditional distribution $\pi$ into a distribution over triples $(x, \mathbf{s}, r)$ as $\pi(x, \mathbf{s}, r) \coloneqq D(r \mid x, \mathbf{s})\pi(\mathbf{s} \mid x)D(x)$. With this shorthand, we have $V(\pi) = \mathbb{E}_\pi[r]$.

To finish this section, we introduce notation for the expected reward for a given context and slate, which we call the *slate value*, and denote as $V(x, \mathbf{s}) \coloneqq \mathbb{E}_{r\sim D(\cdot|x,\mathbf{s})}[r]$.

**Example 1** (Cartesian product)**.** Consider whole-page optimization of a news portal where the reward is the whole-page advertising revenue. The context $x$ is the user profile, the slate is the news-portal page with slots corresponding to news sections or topics,[1] and actions are the news articles. It is natural to assume that each article can only appear in one of the sections, so that $A_j(x) \cap A_k(x) = \emptyset$ if $j \neq k$. The set of valid slates is the Cartesian product $S(x) = \prod_{j\le\ell} A_j(x)$. The number of valid slates is exponential in $\ell$, namely, $|S(x)| = \prod_{j\le\ell} m_j$.

**Example 2** (Ranking)**.** Consider information retrieval in web search. Here the context $x$ is the user query along with user profile, time of day etc. Actions correspond to search items (such as webpages). The policy chooses $\ell$ of $m$ items, where the set $A(x)$ of $m$ items for a context $x$ is chosen from a large corpus by a fixed filtering step (e.g., a database query). We have $A_j(x) = A(x)$ for all $j \le \ell$, but the allowed slates $S(x)$ have no repeated actions. The slots $j \le \ell$ correspond to positions on the search results page. The number of valid slates is exponential in $\ell$ since $|S(x)| = m!/(m - \ell)! = m^{\Omega(\ell)}$. A reward could be the *negative time-to-success*, i.e., negative of the time taken by the user to find a relevant item, typically capped at some threshold if nothing relevant is found.

### 2.1 Off-policy evaluation and optimization

In the *off-policy* setting, we have access to the *logged data* $(x_1, \mathbf{s}_1, r_1), \ldots, (x_n, \mathbf{s}_n, r_n)$ collected using a past policy $\mu$, called the *logging policy*. *Off-policy evaluation* is the task of estimating the value of a new policy $\pi$, called the *target policy*, using the logged data. *Off-policy optimization* is the harder task of finding a policy $\hat{\pi}$ that improves upon the performance of $\mu$ and achieves a large reward. We mostly focus on off-policy evaluation, and show how to use it as a subroutine for off-policy optimization in Sec. 4.2.

There are two standard approaches for off-policy evaluation. The *direct method* (DM) uses the logged data to train a (parametric) model $\hat{r}(x, \mathbf{s})$ to predict the expected reward for a given context and slate. $V(\pi)$ is then estimated as

$$\hat{V}_{\text{DM}}(\pi) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{\mathbf{s}\in S(x)} \hat{r}(x_i, \mathbf{s})\pi(\mathbf{s} \mid x_i) \ . \tag{1}$$

[1] For simplicity, we do not discuss the more general setting of showing multiple articles in each news section.

The direct method is frequently biased because the reward model $\hat{r}(x, \mathbf{s})$ is typically misspecified.

The second approach, which is provably unbiased (under modest assumptions), is the *inverse propensity score* (IPS) estimator [11]. The IPS estimator reweights the logged data according to ratios of slate probabilities under the target and logging policy. It has two common variants:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} r_i \cdot \frac{\pi(\mathbf{s}_i|x_i)}{\mu(\mathbf{s}_i|x_i)} \quad , \quad \hat{V}_{\text{wIPS}}(\pi) = \sum_{i=1}^{n} r_i \cdot \frac{\pi(\mathbf{s}_i|x_i)}{\mu(\mathbf{s}_i|x_i)} \Big/ \left(\sum_{i=1}^{n} \frac{\pi(\mathbf{s}_i|x_i)}{\mu(\mathbf{s}_i|x_i)}\right) \quad . \tag{2}$$

The two estimators differ only in their normalizer. The IPS estimator is unbiased, whereas the weighted IPS (wIPS) is only asymptotically unbiased, but usually achieves smaller error due to smaller variance. Unfortunately, the variance of both estimators grows linearly with the magnitude of $\pi(\mathbf{s} \mid x)/\mu(\mathbf{s} \mid x)$, which can be as bad as $\Omega(|S(x)|)$. This is prohibitive when $|S(x)| = m^{\Omega(\ell)}$.

## 3 Our approach

To reason about the slates, we consider vectors in $\mathbb{R}^{\ell m}$ whose components are indexed by pairs $(j, a)$ of slots and possible actions in them. A slate is then described by an indicator vector $\mathbf{1_s} \in \mathbb{R}^{\ell m}$ whose entry at position $(j, a)$ is equal to 1 if the slate $\mathbf{s}$ has action $a$ in the slot $j$, i.e., if $s_j = a$. At the foundation of our approach is an assumption relating the slate value to its component actions:

**Assumption 1** (ADA). A combinatorial contextual bandit problem satisfies the *additive decomposition assumption* (ADA) if for each context $x \in X$ there exists a (possibly unknown) *intrinsic reward* vector $\boldsymbol{\phi}_x \in \mathbb{R}^{\ell m}$ such that the slate value decomposes as $V(x, \mathbf{s}) = \mathbf{1}_\mathbf{s}^T \boldsymbol{\phi}_x = \sum_{j=1}^{\ell} \phi_x(j, s_j)$.

ADA only posits the existence of intrinsic rewards, not their observability. This distinguishes it from semi-bandits where $\{\phi_x(j, s_j)\}_{j=1}^{\ell}$ can be observed for the $s_j$'s chosen in context $x$. The slate value is described by a linear relationship between $\mathbf{1_s}$ and the unknown "parameters" $\boldsymbol{\phi}_x$, but we do not require that $\boldsymbol{\phi}_x$ be easy to fit from features describing contexts and actions, which is the key departure from the direct method and parametric bandits.

While ADA rules out interactions among different actions on a slate,[2] its ability to vary intrinsic rewards arbitrarily across contexts can capture many common metrics in information retrieval, such as the *normalized discounted cumulative gain* (NDCG) [5], a common reward metric in web ranking:

**Example 3** (NDCG). For a given slate $\mathbf{s}$ we first define a *discounted cumulative gain* value:

$$\text{DCG}(x, \mathbf{s}) := \sum_{j=1}^{\ell} \frac{2^{\mathit{rel}(x, s_j)} - 1}{\log_2(j+1)} \quad ,$$

where $\mathit{rel}(x, a) \geq 0$ is the relevance of document $a$ on query $x$. We define $\text{NDCG}(x, \mathbf{s}) := \text{DCG}(x, \mathbf{s})/\text{DCG}^\star(x)$ where $\text{DCG}^\star(x) = \max_{\mathbf{s} \in S(x)} \text{DCG}(x, \mathbf{s})$, so NDCG takes values in $[0, 1]$. Thus, NDCG satisfies ADA with $\phi_x(j, a) = \left(2^{\mathit{rel}(x,a)} - 1\right) \big/ \log_2(j + 1)\text{DCG}^\star(x)$.

In addition to ADA, we also make the standard assumption that the logging policy puts non-zero probability on all slates that can be potentially chosen by the target policy. This assumption is also required for the unbiasedness of IPS, otherwise off-policy evaluation is impossible [16].

**Assumption 2** (ABS). The off-policy evaluation problem satisfies the *absolute continuity* assumption if $\mu(\mathbf{s} \mid x) > 0$ whenever $\pi(\mathbf{s} \mid x) > 0$ with probability one over $x \sim D$.

### 3.1 The pseudoinverse estimator

Our estimator uses certain moments of the logging policy $\mu$, called *marginal values* and denoted $\boldsymbol{\theta}_{\mu,x} \in \mathbb{R}^{\ell m}$, and their empirical estimates, called *marginal rewards* and denoted $\hat{\boldsymbol{\theta}}_i \in \mathbb{R}^{\ell m}$:

$$\boldsymbol{\theta}_{\mu,x} := \mathbb{E}_\mu[r\mathbf{1_s} \mid x] \quad \text{and} \quad \hat{\boldsymbol{\theta}}_i := r_i \mathbf{1}_{\mathbf{s}_i} \quad .$$

Recall that $\mu$ is viewed here as a distribution over triples $(x, \mathbf{s}, r)$. In words, the components $\theta_{\mu,x}(j, a)$ accumulate the rewards only when the policy $\mu$ chooses a slate $\mathbf{s}$ with $s_j = a$. The random variable $\hat{\boldsymbol{\theta}}_i$ estimates $\boldsymbol{\theta}_{\mu,x}$ at $x_i$ by the observed reward for the slate $\mathbf{s}_i$ displayed for $x_i$ in our logs. The key

[2] We discuss limitations of ADA and directions to overcome them in Sec. 5.

insight is that the marginal value $\theta_{\mu,x}(j,a)$ provides an indirect view of $\phi_x(j,a)$, occluded by the effect of actions in slots $k \neq j$. Specifically, from ADA and the definition of $\boldsymbol{\theta}_{\mu,x}$, we obtain

$$\theta_{\mu,x}(j,a) = \mu(s_j = a \mid x)\phi_x(j,a) + \sum_{k\neq j} \sum_{a' \in A_k(x)} \mu(s_j = a, s_k = a' \mid x)\phi_x(k,a') \ . \tag{3}$$

Eq. (3) represents a linear relationship between $\boldsymbol{\theta}_{\mu,x}$ and $\boldsymbol{\phi}_x$, which is concisely described by a matrix $\mathbf{\Gamma}_{\mu,x} \in \mathbb{R}^{\ell m \times \ell m}$, with

$$\mathit{\Gamma}_{\mu,x}(j,a;\, k,a') \coloneqq \begin{cases} \mu(s_j = a \mid x) & \text{if } j = k \text{ and } a = a', \\ \mu(s_j = a, s_k = a' \mid x) & \text{if } j \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\boldsymbol{\theta}_{\mu,x} = \mathbf{\Gamma}_{\mu,x}\boldsymbol{\phi}_x$. If $\mathbf{\Gamma}_{\mu,x}$ was invertible, we could write $\boldsymbol{\phi}_x = \mathbf{\Gamma}_{\mu,x}^{-1}\boldsymbol{\theta}_{\mu,x}$ and use ADA to obtain $V(x,\mathbf{s}) = \mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x}^{-1}\boldsymbol{\theta}_{\mu,x}$. We could then replace $\boldsymbol{\theta}_{\mu,x_i}$ by its unbiased estimate $\hat{\boldsymbol{\theta}}_i$ to get an unbiased estimate of $V(x_i,\mathbf{s})$. In reality, $\mathbf{\Gamma}_{\mu,x}$ is not invertible. However, it turns out that the above strategy still works, we just need to replace the inverse by the pseudoinverse: [3]

**Theorem 1.** *If ADA holds and $\mu(\mathbf{s} \mid x) > 0$, then $V(x,\mathbf{s}) = \mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x}^\dagger\boldsymbol{\theta}_{\mu,x}$.*

This gives rise to the value estimator, which we call the *pseudoinverse estimator* or PI for short:

$$\hat{V}_{\text{PI}}(\pi) \ = \ \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{s}\in S}\pi(\mathbf{s} \mid x_i)\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x_i}^\dagger\hat{\boldsymbol{\theta}}_i \ = \ \frac{1}{n}\sum_{i=1}^{n} r_i \cdot \mathbf{q}_{\pi,x_i}^T\mathbf{\Gamma}_{\mu,x_i}^\dagger\mathbf{1}_{\mathbf{s}_i} \ , \tag{4}$$

where in Eq. (4), we have expanded the definition of $\hat{\boldsymbol{\theta}}_i$ and introduced the notation $\mathbf{q}_{\pi,x}$ for the expected slate indicator under $\pi$ conditional on $x$, $\mathbf{q}_{\pi,x} \coloneqq \mathbb{E}_\pi[\mathbf{1}_\mathbf{s} \mid x]$. The sum over $\mathbf{s}$ required to obtain $\mathbf{q}_{\pi,x_i}$ in Eq. (4) can be estimated with a small sample.

Theorem 1 immediately yields the unbiasedness of $\hat{V}_{\text{PI}}$:

**Theorem 2.** *If ADA and ABS hold, then the estimator $\hat{V}_{\text{PI}}$ is unbiased, i.e., $\mathbb{E}_{\mu^n}[\hat{V}_{\text{PI}}] = V(\pi)$, where the expectation is over the $n$ logged examples sampled i.i.d. from $\mu$.*

**Example 4** (PI when $\ell = 1$)**.** When the slate consists of a single slot, the policies recommend a single action chosen from some set $A(x)$ for a context $x$. In this case PI coincides with IPS since

$$\mathbf{\Gamma}_{\mu,x} = \operatorname{diag}\bigl(\mu(a \mid x)\bigr)_{a\in A(x)}, \ \ \mathbf{\Gamma}_{\mu,x}^\dagger = \operatorname{diag}\bigl(1/\mu(a \mid x)\bigr)_{a\in A(x)}, \ \text{ and } \mathbf{q}_{\pi,x} = \bigl(\pi(a \mid x)\bigr)_{a\in A(x)} \, .$$

**Example 5** (PI when $\pi = \mu$)**.** When the target policy coincides with logging, the estimator simplifies to the average of rewards: $\hat{V}_{\text{PI}}(\pi) = \frac{1}{n}\sum_{i=1}^n r_i$ (see Appendix C). For $\ell = 1$, this follows from the previous example, but it is non-trivial to show for $\ell \geq 2$.

**Example 6** (PI for a Cartesian product with uniform logging)**.** The PI estimator for the Cartesian product slate space when $\mu$ is uniform over slates simplifies to

$$\hat{V}_{\text{PI}}(\pi) = \tfrac{1}{n}\textstyle\sum_{i=1}^n r_i \cdot \left(\sum_{j=1}^{\ell} \frac{\pi(s_{ij}|x_i)}{1/m_j} - \ell + 1\right) \ ,$$

by Prop. 3 in Appendix D.1. Note that unlike IPS, which divides by probabilities of whole slates, the PI estimator only divides by probabilities of actions appearing in individual slots. Thus, the magnitude of each term of the outer summation is only $\mathcal{O}(\ell m)$, whereas the IPS terms are $m^{\Omega(\ell)}$.

**Example 7** (PI for rankings with $\ell = m$ and uniform logging)**.** In this case, the PI estimator equals

$$\hat{V}_{\text{PI}}(\pi) = \tfrac{1}{n}\textstyle\sum_{i=1}^n r_i \cdot \left(\sum_{j=1}^{\ell} \frac{\pi(s_{ij}|x_i)}{1/(m-1)} - m + 2\right) \ ,$$

by Prop. 4 in Appendix D.1. The magnitude of individual terms is again $\mathcal{O}(\ell m) = \mathcal{O}(m^2)$.

[3] A variant of Theorem 1 is proved in a different context by Dani et al. [7]. Our proof, alongside proofs of all other statements in the paper, is in Appendix.

### 3.2 Deviation analysis

We have shown that the pseudoinverse estimator is unbiased given ADA and have also given examples when it improves exponentially over IPS, the existing state-of-the-art. We next derive a distribution-dependent bound on finite-sample error and use it to obtain an exponential improvement over IPS for a broader class of logging distributions.

Our deviation bound is obtained by an application of Bernstein's inequality, which requires bounding the variance and range of the terms appearing in Eq. (4), namely $r_i \cdot \mathbf{q}_{\pi,x_i}^T \mathbf{\Gamma}_{\mu,x_i}^{\dagger} \mathbf{1}_{\mathbf{s}_i}$. We bound their variance and range, respectively, by the following distribution-dependent quantities:

$$\sigma^2 := \mathbb{E}_{x\sim D}\left[\mathbf{q}_{\pi,x}^T \mathbf{\Gamma}_{\mu,x}^{\dagger} \mathbf{q}_{\pi,x}\right] \ , \quad \rho := \sup_{x} \sup_{\mathbf{s}:\mu(\mathbf{s}|x)>0} \left|\mathbf{q}_{\pi,x}^T \mathbf{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}}\right| \ . \tag{5}$$

They capture the "average" and "worst-case" mismatch between the logging and target policy. They equal one when $\pi = \mu$ (see Appendix C), and in general yield the following deviation bound:

**Theorem 3.** *Assume that ADA and ABS hold, and let $\sigma^2$ and $\rho$ be defined as in Eq.* (5)*. Then, for any $\delta \in (0,1)$, with probability at least* $1-\delta$,

$$\left|\hat{V}_{\text{PI}}(\pi) - V(\pi)\right| \le \sqrt{\frac{2\sigma^2 \ln(2/\delta)}{n}} + \frac{2(\rho+1)\ln(2/\delta)}{3n} \ .$$

In Appendix D, Prop. 2, we show that $\sigma^2 \le \rho$, so to bound $\hat{V}_{\text{PI}}$, it suffices to bound $\rho$. We next show such a bound for a broad class of logging policies defined as follows:

**Definition 1.** *Let $\nu$ denote the uniform policy, that is, $\nu(\mathbf{s} \mid x) = 1/|S(x)|$. We say that a policy $\mu$ is* pairwise $\kappa$-uniform *for some $\kappa \in (0,1]$ if for all contexts $x$, actions $a, a'$, and slots $j, k$, we have*

$$\mu(s_j = a,\ s_k = a' \mid x) \ge \kappa\nu(s_j = a,\ s_k = a' \mid x) \ .$$

For the Cartesian product slate space, this means that $\mu(s_j = a,\ s_k = a' \,|\, x) \ge \kappa/(m_j m_k)$ for $j \neq k$. For rankings, $\mu(s_j = a,\ s_k = a' \mid x) \ge \kappa/(m(m-1))$ for $j \neq k$. Given any policy, we can obtain a pairwise $\kappa$-uniform policy by mixing in the uniform distribution with the weight $\kappa$.

**Proposition 1.** *Assume that valid slates form a Cartesian product space as in Example 1 or are rankings as in Example 2. Then for any pairwise $\kappa$-uniform logging policy, we have $\rho \le \kappa^{-1}\ell m$.*

Thus, using the fact that $\sigma^2 \le \rho$, Prop. 1 and Theorem 3 yield $|\hat{V}_{\text{PI}}(\pi) - V(\pi)| \le \mathcal{O}\big(\sqrt{\kappa^{-1}\ell m/n}\big)$, or equivalently $\mathcal{O}(\kappa^{-1}\ell m/\varepsilon^2)$ logging samples are needed to achieve accuracy $\varepsilon$.

## 4 Experiments

We now empirically evaluate the performance of the pseudoinverse estimator in the ranking scenario of Example 2. We first show that our approach compares favorably to baselines in a semi-synthetic evaluation on a public data set under the NDCG metric, which satisfies ADA as discussed in Example 3. On the same data, we further use the pseudoinverse estimator for *off-policy optimization*, that is, to learn ranking policies, competitively with a supervised baseline that uses more information. Finally, we demonstrate substantial improvements on proprietary data from search engine logs for two user-satisfaction metrics used in practice: *time-to-success* and *utility rate*, which are *a priori* unlikely to (exactly) satisfy ADA. More detailed results are deferred to Appendices E, F and G.

### 4.1 Semi-synthetic evaluation

Our semi-synthetic evaluation uses labeled data from the LETOR4.0 MQ2008 dataset [20] to create a contextual bandit instance. Queries form the contexts $x$ and actions $a$ are the available documents. The dataset contains 784 queries, 5–121 documents per query and relevance labels *rel*$(x, a) \in \{0, 1, 2\}$ for each query-document pair. Each pair $(x, a)$ has a 47-dimensional feature vector $\mathbf{f}(x, a)$, which can be partitioned into title features $\mathbf{f}_{\text{title}}$, and body features $\mathbf{f}_{\text{body}}$.

To derive a logging policy and a distinct target policy, we first train two lasso regression models, called *pred*$_{\text{title}}$ and *pred*$_{\text{body}}$, to predict relevances from $\mathbf{f}_{\text{title}}$ and $\mathbf{f}_{\text{body}}$, respectively. To create the

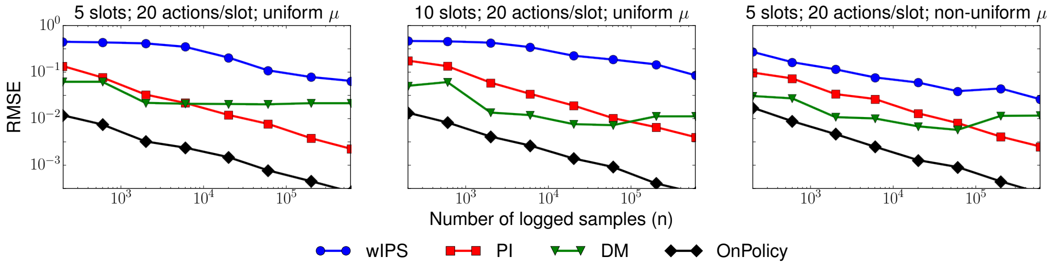

Figure 2: RMSE under uniform logging ($\alpha = 0$) and non-uniform logging ($\alpha = 10$).

logs, queries $x$ are sampled uniformly, and the set $A(x)$ consists of top $m$ documents according to $\textit{pred}_{\text{title}}$. The logging policy $\mu$ samples from a multinomial distribution over documents in $A(x)$, parameterized by $\alpha \geq 0$: $p_\alpha(a \mid x) \propto \exp(\alpha \cdot \textit{pred}_{\text{title}}(x, a))$. Slates are constructed slot-by-slot, sampling *without replacement* according to $p_\alpha$. Choosing $\alpha \in [0, \infty)$ interpolates between uniformly random and deterministic logging. Our target policy $\pi$ selects the slate of top $\ell$ documents according to $\textit{pred}_{\text{body}}$. The slate reward is the NDCG metric defined in Example 3.

We compare our estimator PI with the direct method (DM) and weighted IPS (wIPS, see Eq. 2), which out-performed IPS. Our implementation of DM concatenates per-slot features $\mathbf{f}(x, s_j)$ into $\mathbf{f}(x, \mathbf{s})$, training a reward predictor on the first $n/2$ examples and evaluating $\pi$ using Eq. (1) on the other $n/2$ examples. We experimented with regression trees, ridge and lasso regression for DM, and always report results for the choice with the smallest RMSE at $n = 10^6$ examples. We also include an aspirational baseline, OnPolicy. This corresponds to deploying the target policy as in an A/B test and returning the average of observed rewards. This is the expensive alternative we wish to avoid.

We plot the root mean square error (RMSE) of the estimators as a function of increasing data size over at least 20 independent runs.

In Fig. 2, the first two plots study the RMSE of estimators for two choices of $m$ and $\ell$, given the uniform logging policy $\mu$ (i.e., $\alpha = 0$). In both cases, the pseudoinverse estimator outperforms wIPS by a factor of 10 or more with high statistical significance, $p < 10^{-8}$ for both plots and for all $n$. The pseudoinverse estimator eventually also outperforms the biased DM with statistical significance, with $p \leq 7.3 \times 10^{-4}$ for both plots at $n = 600$K. The cross-over point occurs fairly early ($n \approx 10$K) for the smaller slate space, but is one order larger ($n \approx 100$K) for the largest slate space. Note that DM's performance can deteriorate with more data, likely because it optimizes the fit to the reward distribution of $\mu$, which is different from that of $\pi$.

As expected, OnPolicy performs the best, requiring between 10x and 100x less data. However, OnPolicy requires to fix the target policy $\pi$ for each data collection, while off-policy methods like PI take advantage of massive amounts of logged data to evaluate arbitrary policies. As an aside, since the user feedback in these experiments is simulated, we can also simulate semi-bandit feedback which reveals the intrinsic reward of each shown action, and use it directly for off-policy evaluation. This is a purely hypothetical baseline: with only page-level feedback, one cannot implement a semi-bandit solution. We compare against this hypothetical baseline in Appendix F.

In Fig. 2 (right panel), we study the effect of the overlap between the logging and target policies, by taking $\alpha = 10$, which results in a better alignment between the logging and target policies, While the RMSE of the pseudoinverse estimator is largely unchanged, both wIPS and DM exhibit some improvement. wIPS enjoys a smaller variance, while DM enjoys a smaller bias due to closer training and target distributions. PI continues to be statistically better than wIPS, with $p \leq 10^{-8}$ for all $n$, and eventually also better than DM, with $p \leq 4.4 \times 10^{-4}$ starting at $n = 200$K. See Appendices E and F for more results and the complete set of $p$-values.

### 4.2 Semi-synthetic policy optimization

We now show how to use the pseudoinverse estimator for off-policy optimization. We leverage pointwise learning to rank (L2R) algorithms, which learn a scoring function for query-document pairs by fitting to relevance labels. We call this the *supervised* approach, as it requires relevance labels.

Instead of requiring relevance labels, we use the pseudoinverse estimator to convert page-level reward into per-slot reward components—the estimates of $\phi_x(j,a)$—and these become targets for regression. *Thus, the pseudoinverse estimator enables pointwise L2R even without relevance labels.* Given a contextual bandit dataset $\{(x_i, \mathbf{s}_i, r_i)\}_{i \le n}$ collected by the logging policy $\mu$, we begin by creating the estimates of $\boldsymbol{\phi}_{x_i}$: $\hat{\boldsymbol{\phi}}_i = \boldsymbol{\Gamma}^{\dagger}_{\mu,x_i} \hat{\boldsymbol{\theta}}_i$, turning the $i$-th contextual bandit example into $\ell m$ regression examples. The trained regression model is used to create a slate, starting with the highest scoring slot-action pair, and continuing greedily (excluding the pairs with the already chosen slots or actions).

We used the MQ2008 dataset from the previous section and created a contextual bandit problem with 5 slots and 20 documents per slot, with a uniformly random logging policy. We chose a standard 5-fold split and always trained on bandit data from 4 folds and evaluated using the supervised data on the fifth. We compare our approach, titled PI-OPT, against the supervised approach, trained to predict the *gains*, equal to $2^{\mathit{rel}(x,a)} - 1$, computed using annotated relevance judgements in the training fold (predicting raw relevances was inferior). Both PI-OPT and SUP train regression trees. We find that PI-OPT is consistently competitive with SUP after seeing about 1K samples containing slate-level feedback, and gets a test NDCG of 0.450 at 1K samples, 0.451 at 10K samples, and 0.456 at 100K samples. SUP achieves a test NDCG of 0.453 by using approximately 12K annotated relevance judgements. We posit that PI-OPT is competitive with SUP because it optimizes the target metric directly, while SUP uses a surrogate (imperfect) regression loss. See Appendix G for detailed results.

### 4.3 Real-world experiments

We finally evaluate all methods using logs collected from a popular search engine. The dataset consists of search queries, for which the logging policy randomly (non-uniformly) chooses a slate of size $\ell = 5$ from a small pre-filtered set of documents of size $m \le 8$. After preprocessing, there are 77 unique queries and 22K total examples, meaning that for each query, we have logged impressions for many of the available slates. To control the query distribution in our experiment, we generate a larger dataset by bootstrap sampling, repeatedly choosing a query uniformly at random and a slate uniformly at random from those shown for this query. Hence, the conditional probability of any slate for a given query matches the frequencies in the original data.

We consider two page-level metrics: time-to-success (TTS) and UtilityRate. TTS measures the number of seconds between presenting the results and the first satisfied click from the user, defined as any click for which the user stays on the linked page for sufficiently long. TTS value is capped and scaled to $[0, 1]$. UtilityRate is a more complex page-level metric of user's satisfaction. It captures the interaction of a user with the page as a timeline of events (such as clicks) and their durations. The events are classified as revealing a positive or negative utility to the user and their contribution is proportional to their duration. UtilityRate takes values in $[-1, 1]$.

We evaluate a target policy based on a logistic regression classifier trained to predict clicks and using the predicted probabilities to score slates. We restrict the target policy to pick among the slates in our logs, so we know the ground truth slate-level reward. Since we know the query distribution, we can calculate the target policy's value exactly, and measure RMSE relative to this true value.

We compare our estimator (PI) with three baselines similar to those from Sec. 4.1: DM, IPS and OnPolicy. DM uses regression trees over roughly 20,000 slate-level features.

Fig. 1 from the introduction shows that PI provides a consistent multiplicative improvement in RMSE over IPS, which suffers due to high variance. Starting at moderate sample sizes, PI also outperforms DM, which suffers due to substantial bias. For TTS, the gains over IPS are significant with $p \le 3.7 \times 10^{-5}$ after 2K samples and for DM with $p \le 1.5 \times 10^{-3}$ after 20K samples. For UtilityRate, the improvements on IPS are significant with $p < 10^{-8}$ at 60K examples, and over DM with $p \le 4.3 \times 10^{-7}$ after 20K examples. The complete set of $p$-values is in Appendix E.

## 5 Discussion

In this paper we have introduced a new assumption (ADA), a new estimator (PI) that exploits this assumption, and demonstrated their significant theoretical and practical merits.

In our experiments, we saw examples of bias-variance trade-off with off-policy estimators. At small sample sizes, the pseudoinverse estimator still has a non-trivial variance. In these regimes, the biased direct method can often be practically useful due to its small variance (if its bias is sufficiently small). Such well-performing albeit biased estimators can be incorporated into the pseudoinverse estimator via the doubly-robust approach [8].

Experiments with real-world data in Sec. 4.3 demonstrate that even when ADA does not hold, the estimators based on ADA can still be applied and tend to be superior to alternatives. We view ADA similarly to the IID assumption: while it is probably often violated in practice, it leads to practical algorithms that remain robust under misspecification. Similarly to the IID assumption, we are not aware of ways for easily testing whether ADA holds.

One promising approach to relax ADA is to posit a decomposition over pairs (or tuples) of slots to capture higher-order interactions such as diversity. More generally, one could replace slate spaces by arbitrary compact convex sets, as done in linear bandits. In these settings, the pseudoinverse estimator could still be applied, but tight sample-complexity analysis is open for future research.

## References

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 2002.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

[3] Erik G Boman and Bruce Hendrickson. Support theory for preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 2003.

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 2013.

[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, 2005.

[6] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *Artificial Intelligence and Statistics*, 2011.

[7] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, 2008.

[8] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

[9] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

[10] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, 2010.

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

[12] Satyen Kale, Lev Reyzin, and Robert E Schapire. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems*, 2010.

[13] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Knowledge Discovery and Data mining*, 2009.

[14] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, 2015.

[15] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 2008.

[16] John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *International Conference on Machine Learning*, 2008.

[17] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, 2010.

[18] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 2008.

[19] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *International Conference on Data Mining*, 2014.

[20] Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *arXiv:1306.2597*, 2013.

[21] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 2010.

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

## A Proofs of Theorems 1 and 2

**Claim 1.** $\mathbf{\Gamma}_{\mu,x} = \mathbb{E}_\mu[\mathbf{1_s}\mathbf{1_s}^T \mid x]$.

*Proof.* Consider the matrix $\mathbf{1_s}\mathbf{1_s}^T$. Its element in the row indexed $(j,a)$ and column indexed $(k,a')$ equals

$$\mathbf{1}\{s_j = a,\, s_k = a'\} = \begin{cases} \mathbf{1}\{s_j = a\} & \text{if } j = k \text{ and } a = a', \\ \mathbf{1}\{s_j = a,\, s_k = a'\} & \text{if } j \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

The claim follows by taking a conditional expectation with respect to $\mu$. □

*Proof of Theorem 1.* Fix one $x$ for the entirety of the proof. Recall from Sec. 3.1 that

$$V(x,\mathbf{s}) = \mathbf{1_s}^T\boldsymbol{\phi}_x \ .$$

Let $N = |\operatorname{supp}\mu(\cdot \mid x)|$ be the size of the support of $\mu(\cdot \mid x)$ and let $\mathbf{M} \in \{0,1\}^{N\times m\ell}$ denote the binary matrix with rows $\mathbf{1_s}^T$ for each $\mathbf{s} \in \operatorname{supp}\mu(\cdot \mid x)$. Thus $\mathbf{M}\boldsymbol{\phi}_x$ is the vector enumerating $V(x,\mathbf{s})$ over $\mathbf{s}$ for which $\mu(\mathbf{s} \mid x) > 0$. Let $\operatorname{Null}(\mathbf{M})$ denote the null space of $\mathbf{M}$ and $\mathbf{\Pi}$ be the projection on $\operatorname{Null}(\mathbf{M})$. Let $\boldsymbol{\phi}_x^\star = (\mathbf{I} - \mathbf{\Pi})\boldsymbol{\phi}_x$. Then clearly, $\mathbf{M}\boldsymbol{\phi}_x = \mathbf{M}\boldsymbol{\phi}_x^\star$, and hence, for any $\mathbf{s} \in \operatorname{supp}\mu(\cdot \mid x)$,

$$V(x,\mathbf{s}) = \mathbf{1_s}^T\boldsymbol{\phi}_x^\star \ . \tag{6}$$

We will now show that $\boldsymbol{\phi}_x^\star = \mathbf{\Gamma}_{\mu,x}^\dagger\boldsymbol{\theta}_{\mu,x}$, which will complete the proof.

Recall from Sec. 3.1 that

$$\boldsymbol{\theta}_{\mu,x} = \mathbf{\Gamma}_{\mu,x}\boldsymbol{\phi}_x \ . \tag{7}$$

Next note that $\mathbf{\Gamma}_{\mu,x}$ in symmetric positive semidefinite by Claim 1, so

$$\operatorname{Null}(\mathbf{\Gamma}_{\mu,x}) = \{\mathbf{v} :\ \mathbf{v}^T\mathbf{\Gamma}_{\mu,x}\mathbf{v} = 0\} = \{\mathbf{v} :\ \mathbf{1_s}^T\mathbf{v} = 0 \text{ for all } \mathbf{s} \in \operatorname{supp}\mu(\cdot \mid x)\} = \operatorname{Null}(\mathbf{M})$$

where the first step follows by positive semi definiteness of $\mathbf{\Gamma}_{\mu,x}$, the second step is from the expansion of $\mathbf{\Gamma}_{\mu,x}$ as in Claim 1, and the final step from the definition of $\mathbf{M}$. Since $\operatorname{Null}(\mathbf{\Gamma}_{\mu,x}) = \operatorname{Null}(\mathbf{M})$, we have from Eq. (7) that $\boldsymbol{\theta}_x = \mathbf{\Gamma}_{\mu,x}\boldsymbol{\phi}_x^\star$, but, importantly, this also implies $\boldsymbol{\phi}_x^\star \perp \operatorname{Null}(\mathbf{\Gamma}_{\mu,x})$, so by the definition of the pseudoinverse,

$$\mathbf{\Gamma}_{\mu,x}^\dagger\boldsymbol{\theta}_x = \boldsymbol{\phi}_x^\star.$$

This proves Theorem 1, since for any $\mathbf{s}$ with $\mu(\mathbf{s} \mid x) > 0$, we argued that $V(x,\mathbf{s}) = \mathbf{1_s}^T\boldsymbol{\phi}_x^\star = \mathbf{1_s}^T\mathbf{\Gamma}_{\mu,x}^\dagger\boldsymbol{\theta}_x$. □

*Proof of Theorem 2.* Note that it suffices to analyze the expectation of a single term in the estimator, that is

$$\sum_{\mathbf{s}\in S} \pi(\mathbf{s} \mid x_i)\mathbf{1_s}^T\mathbf{\Gamma}_{\mu,x_i}^\dagger\hat{\boldsymbol{\theta}}_i \ .$$

First note that $\mathbb{E}_{(\mathbf{s}_i,r_i)\sim\mu(\cdot,\cdot|x_i)}\big[\hat{\boldsymbol{\theta}}_i\big] = \boldsymbol{\theta}_{x_i}$, because

$$\mathbb{E}_{(\mathbf{s}_i,r_i)\sim\mu(\cdot,\cdot|x_i)}\big[\hat{\theta}_i(j,a)\big] = \mathbb{E}_{(\mathbf{s}_i,r_i)\sim\mu(\cdot,\cdot|x_i)}\big[r_i\mathbf{1}\{s_j = a\}\big] = \theta_{x_i}(j,a) \ .$$

The remainder follows by Theorem 1:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\mathbf{s}\in S}\pi(\mathbf{s}\mid x_i)\mathbf{1}_{\mathbf{s}}^T\boldsymbol{\Gamma}_{\mu,x_i}^{\dagger}\hat{\boldsymbol{\theta}}_i\right] &= \mathbb{E}_{x_i\sim D}\left[\sum_{\mathbf{s}\in S}\pi(\mathbf{s}\mid x_i)\mathbf{1}_{\mathbf{s}}^T\boldsymbol{\Gamma}_{\mu,x_i}^{\dagger}\ \mathbb{E}_{(\mathbf{s}_i,r_i)\sim\mu(\cdot,\cdot\mid x_i)}\left[\hat{\boldsymbol{\theta}}_i\right]\right]\\
&= \mathbb{E}_{x_i\sim D}\left[\sum_{\mathbf{s}\in S}\pi(\mathbf{s}\mid x_i)\mathbf{1}_{\mathbf{s}}^T\boldsymbol{\Gamma}_{\mu,x_i}^{\dagger}\boldsymbol{\theta}_{x_i}\right]\\
&= \mathbb{E}_{x_i\sim D}\left[\sum_{\mathbf{s}\in S}\pi(\mathbf{s}\mid x_i)V(x_i,\mathbf{s})\right] = V(\pi)\ . \qquad\square
\end{aligned}
$$

## B Proof of Theorem 3

*Proof.* The proof is based on an application of Bernstein's inequality to the centered sum

$$
\sum_{i=1}^{n}\left[\mathbf{q}_{\pi,x_i}^T\boldsymbol{\Gamma}_{\mu,x_i}^{\dagger}\hat{\boldsymbol{\theta}}_i - V(\pi)\right]\ .
$$

The fact that this quantity is centered is directly from Theorem 2. We must compute both the second moment and the range to apply Bernstein's inequality. By independence, we can focus on just one term, so we will drop the subscript $i$. First, bound the variance:

$$
\begin{aligned}
\operatorname{Var}\left[\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\hat{\boldsymbol{\theta}}\right] &\le \mathbb{E}_{\mu}\left[\left(\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\hat{\boldsymbol{\theta}}\right)^2\right]\\
&= \mathbb{E}_{\mu}\left[\left(\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\,r\mathbf{1}_{\mathbf{s}}\right)^2\right]\\
&\le \mathbb{E}_{\mu}\left[\left(\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\mathbf{1}_{\mathbf{s}}\right)^2\right]\\
&= \mathbb{E}_{x\sim D}\left[\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\ \mathbb{E}_{\mathbf{s}\sim\mu(\cdot\mid x)}\left[\mathbf{1}_{\mathbf{s}}\mathbf{1}_{\mathbf{s}}^T\right]\ \boldsymbol{\Gamma}_{\mu,x}^{\dagger}\mathbf{q}_{\pi,x}\right]\\
&= \mathbb{E}_{x\sim D}\left[\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\boldsymbol{\Gamma}_{\mu,x}\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\mathbf{q}_{\pi,x}\right]\\
&= \mathbb{E}_{x\sim D}\left[\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\mathbf{q}_{\pi,x}\right]\\
&= \sigma^2\ .
\end{aligned}
$$

Thus the per-term variance is at most $\sigma^2$. We now bound the range, again focusing on one term,

$$
\begin{aligned}
\left|\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\hat{\boldsymbol{\theta}} - V(\pi)\right| &\le \left|\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\hat{\boldsymbol{\theta}}\right| + 1\\
&= \left|\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}r\mathbf{1}_{\mathbf{s}}\right| + 1\\
&\le \left|\mathbf{q}_{\pi,x}^T\boldsymbol{\Gamma}_{\mu,x}^{\dagger}\mathbf{1}_{\mathbf{s}}\right| + 1\\
&\le \rho + 1
\end{aligned}
$$

The first line here is the triangle inequality, coupled with the fact that since rewards are bounded in $[-1,1]$, so is $V(\pi)$. The second line is from the definition of $\hat{\boldsymbol{\theta}}$, while the third follows because $r\in[-1,1]$. The final line follows from the definition of $\rho$.

Now, we may apply Bernstein's inequality, which says that for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$
\left|\sum_{i=1}^{n}\left[\mathbf{q}_{\pi,x_i}^T\boldsymbol{\Gamma}_{\mu,x_i}^{\dagger}\hat{\boldsymbol{\theta}}_i - V(\pi)\right]\right| \le \sqrt{2n\sigma^2\ln(2/\delta)} + \frac{2(\rho+1)\ln(2/\delta)}{3}\ .
$$

The theorem follows by dividing by $n$. □

## C Pseudo-inverse estimator when $\pi=\mu$

In this section we show that when the target policy coincides with logging (i.e., $\pi=\mu$), we have $\sigma^2=\rho=1$, i.e., the bound of Theorem 3 is independent of the number of actions and slots. Indeed, in Claim 3 we will see that the estimator actually simplifies to taking an empirical average of rewards which are bounded in $[-1, 1]$. Before proving Claim 3 we prove one supporting claim:

**Claim 2.** *For any policy $\mu$ and context $x$, we have $\mathbf{q}_{\mu,x}^T \mathbf{\Gamma}_{\mu,x}^\dagger \mathbf{1}_\mathbf{s} = 1$ for all $\mathbf{s} \in \operatorname{supp} \mu(\cdot \mid x)$.*

*Proof.* To simplify the exposition, write $\mathbf{q}$ and $\mathbf{\Gamma}$ instead of a more verbose $\mathbf{q}_{\mu,x}$ and $\mathbf{\Gamma}_{\mu,x}$.

The bulk of the proof is in deriving an explicit expression for $\mathbf{\Gamma}^\dagger$. We begin by expressing $\mathbf{\Gamma}$ in a suitable basis. Since $\mathbf{\Gamma}$ is the matrix of second moments and $\mathbf{q}$ is the vector of first moments of $\mathbf{1}_\mathbf{s}$, the matrix $\mathbf{\Gamma}$ can be written as

$$\mathbf{\Gamma} = \mathbf{V} + \mathbf{q}\mathbf{q}^T$$

where $\mathbf{V}$ is the covariance matrix of $\mathbf{1}_\mathbf{s}$, i.e., $\mathbf{V} := \mathbb{E}_{\mathbf{s}\sim\mu(\cdot|x)}\big[(\mathbf{1}_\mathbf{s} - \mathbf{q})(\mathbf{1}_\mathbf{s} - \mathbf{q})^T\big]$. Assume that the rank of $\mathbf{V}$ is $r$ and consider the eigenvalue decomposition of $\mathbf{V}$

$$\mathbf{V} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{u}_i^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \ ,$$

where $\lambda_i > 0$ and vectors $\mathbf{u}_i$ are orthonormal; we have grouped the eigenvalues into the diagonal matrix $\mathbf{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ and eigenvectors into the matrix $\mathbf{U} := (\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_r)$.

We next argue that $\mathbf{q} \notin \operatorname{Range}(\mathbf{V})$. To see this, note that the all-ones-vector $\mathbf{1}$ is in the null space of $\mathbf{V}$ because, for any valid slate $\mathbf{s}$, we have $\mathbf{1}_\mathbf{s}^T\mathbf{1} = \ell$ and thus also for the convex combination $\mathbf{q}$ we have $\mathbf{q}^T\mathbf{1} = \ell$, which means that

$$\mathbf{1}^T\mathbf{V}\mathbf{1} = \mathbb{E}_{\mathbf{s}\sim\mu(\cdot|x)}\big[\mathbf{1}^T(\mathbf{1}_\mathbf{s} - \mathbf{q})(\mathbf{1}_\mathbf{s} - \mathbf{q})^T\mathbf{1}\big] = 0 \ .$$

Now, since $\mathbf{1} \perp \operatorname{Range}(\mathbf{V})$ and $\mathbf{q}^T\mathbf{1} = \ell$, we have that $\mathbf{q} \notin \operatorname{Range}(\mathbf{V})$. In particular, we can write $\mathbf{q}$ in the form

$$\mathbf{q} = \sum_{i=1}^r \beta_i \mathbf{u}_i + \alpha\mathbf{n} = (\mathbf{U} \quad \mathbf{n}) \begin{pmatrix} \boldsymbol{\beta} \\ \alpha \end{pmatrix} \tag{8}$$

where $\alpha \neq 0$ and $\mathbf{n} \in \operatorname{Null}(\mathbf{V})$ is a unit vector. Note that $\mathbf{n} \perp \mathbf{u}_i$ since $\mathbf{u}_i \perp \operatorname{Null}(\mathbf{V})$. Thus, the second moment matrix $\mathbf{\Gamma}$ can be written as

$$\mathbf{\Gamma} = \mathbf{V} + \mathbf{q}\mathbf{q}^T = (\mathbf{U} \quad \mathbf{n}) \begin{pmatrix} \mathbf{\Lambda} + \boldsymbol{\beta}\boldsymbol{\beta}^T & \alpha\boldsymbol{\beta} \\ \alpha\boldsymbol{\beta}^T & \alpha^2 \end{pmatrix} (\mathbf{U} \quad \mathbf{n})^T \ . \tag{9}$$

Let $\mathbf{Q} \in \mathbb{R}^{(r+1)\times(r+1)}$ denote the middle matrix in the factorization of Eq. (9):

$$\mathbf{Q} := \begin{pmatrix} \mathbf{\Lambda} + \boldsymbol{\beta}\boldsymbol{\beta}^T & \alpha\boldsymbol{\beta} \\ \alpha\boldsymbol{\beta}^T & \alpha^2 \end{pmatrix} \ . \tag{10}$$

This matrix is a representation of $\mathbf{\Gamma}$ with respect to the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{n}\}$. Since $\mathbf{q} \notin \operatorname{Range}(\mathbf{V})$, the rank of $\mathbf{\Gamma}$ and that of $\mathbf{Q}$ is $r + 1$. Thus, $\mathbf{Q}$ is invertible and

$$\mathbf{\Gamma}^\dagger = (\mathbf{U} \quad \mathbf{n})\, \mathbf{Q}^{-1} (\mathbf{U} \quad \mathbf{n})^T \ . \tag{11}$$

To obtain $\mathbf{Q}^{-1}$, we use the following identity (see [18]):

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M}^{-1} & -\mathbf{M}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{M}^{-1} & \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{M}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1} \end{pmatrix} \ , \tag{12}$$

where $\mathbf{M} := \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ is the Schur complement of $\mathbf{A}_{22}$. The identity of Eq. (12) holds whenever $\mathbf{A}_{22}$ and its Schur complement $\mathbf{M}$ are both invertible. In the block representation of Eq. (10), we have $\mathbf{A}_{22} = \alpha^2 \neq 0$ and

$$\mathbf{M} = (\mathbf{\Lambda} + \boldsymbol{\beta}\boldsymbol{\beta}^T) - (\alpha\boldsymbol{\beta})\alpha^{-2}(\alpha\boldsymbol{\beta}^T) = \mathbf{\Lambda} \ ,$$

so Eq. (12) can be applied to obtain $\mathbf{Q}^{-1}$:

$$\begin{aligned}\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{\Lambda} + \boldsymbol{\beta}\boldsymbol{\beta}^T & \alpha\boldsymbol{\beta} \\ \alpha\boldsymbol{\beta}^T & \alpha^2 \end{pmatrix}^{-1} &= \begin{pmatrix} \mathbf{\Lambda}^{-1} & -\mathbf{\Lambda}^{-1}(\alpha\boldsymbol{\beta})\alpha^{-2} \\ -\alpha^{-2}(\alpha\boldsymbol{\beta}^T)\mathbf{\Lambda}^{-1} & \alpha^{-2}(\alpha\boldsymbol{\beta}^T)\mathbf{\Lambda}^{-1}(\alpha\boldsymbol{\beta})\alpha^{-2} + \alpha^{-2} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Lambda}^{-1} & -\alpha^{-1}\mathbf{\Lambda}^{-1}\boldsymbol{\beta} \\ -\alpha^{-1}\boldsymbol{\beta}^T\mathbf{\Lambda}^{-1} & \alpha^{-2}(1 + \boldsymbol{\beta}^T\mathbf{\Lambda}^{-1}\boldsymbol{\beta}) \end{pmatrix} \ . \end{aligned} \tag{13}$$

Next, we will evaluate $\boldsymbol{\Gamma}^\dagger\mathbf{q}$, using the factorizations in Eqs. (11) and (8), and substituting Eq. (13) for $\mathbf{Q}^{-1}$:

$$\begin{aligned}
\boldsymbol{\Gamma}^\dagger\mathbf{q} &= (\mathbf{U}\quad\mathbf{n})\,\mathbf{Q}^{-1}\,(\mathbf{U}\quad\mathbf{n})^T\,(\mathbf{U}\quad\mathbf{n})\begin{pmatrix}\boldsymbol{\beta}\\ \alpha\end{pmatrix}\\
&= (\mathbf{U}\quad\mathbf{n})\,\mathbf{Q}^{-1}\begin{pmatrix}\boldsymbol{\beta}\\ \alpha\end{pmatrix}\\
&= (\mathbf{U}\quad\mathbf{n})\begin{pmatrix}\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}-\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}\\ -\alpha^{-1}\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}+\alpha^{-1}(1+\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta})\end{pmatrix}\\
&= (\mathbf{U}\quad\mathbf{n})\begin{pmatrix}\mathbf{0}\\ \alpha^{-1}\end{pmatrix}\\
&= \alpha^{-1}\mathbf{n}\ .
\end{aligned}$$

To finish the proof, we consider any $\mathbf{s}\in\operatorname{supp}\mu(\cdot\mid x)$ and consider the decomposition of $\mathbf{1_s}$ in the basis $\{\mathbf{u}_1,\dots,\mathbf{u}_r,\mathbf{n}\}$. First, note that $(\mathbf{1_s}-\mathbf{q})\perp\mathrm{Null}(\mathbf{V})$ since

$$\mathrm{Null}(\mathbf{V})=\bigl\{\mathbf{v}:\mathbb{E}_{\mathbf{s}\sim\mu(\cdot|x)}\bigl[\bigl((\mathbf{1_s}-\mathbf{q})^T\mathbf{v}\bigr)^2\bigr]=0\bigr\}=\bigl\{\mathbf{v}:(\mathbf{1_s}-\mathbf{q})^T\mathbf{v}=0\text{ for all }\mathbf{s}\in\operatorname{supp}\mu(\cdot|x)\bigr\}\ .$$

Thus, $(\mathbf{1_s}-\mathbf{q})\in\mathrm{Range}(\mathbf{V})$. Therefore, we obtain

$$\mathbf{q}^T\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{1_s}=\alpha^{-1}\mathbf{n}^T\mathbf{1_s}=\alpha^{-1}\mathbf{n}^T(\mathbf{1_s}-\mathbf{q})+\alpha^{-1}\mathbf{n}^T\mathbf{q}=0+\alpha^{-1}\alpha=1\ ,$$

where the third equality follows because $(\mathbf{1_s}-\mathbf{q})\perp\mathbf{n}$ and the decomposition in Eq. (8) shows that $\mathbf{n}^T\mathbf{q}=\alpha$. □

**Claim 3.** *If $\pi=\mu$ then $\sigma^2=\rho=1$ and $\hat{V}_{\mathrm{PI}}(\pi)=\hat{V}_{\mathrm{PI}}(\mu)=\frac{1}{n}\sum_{i=1}^n r_i$.*

*Proof.* From Claim 2

$$\mathbf{q}^T_{\mu,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{q}_{\mu,x}=\mathbb{E}_{\mathbf{s}\sim\mu(\cdot|x)}[\mathbf{q}^T_{\mu,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{1_s}]=1\ .$$

Taking expectation over $x$ then yields $\sigma^2=1$. Equality $\rho=1$ follows immediately from plugging Claim 2 into the definition of $\rho$. The final statement of Claim 3 follows by applying Claim 2 to a single term of $\hat{V}_{\mathrm{PI}}(\mu)$:

$$\mathbf{q}^T_{\mu,x_i}\boldsymbol{\Gamma}^\dagger_{\mu,x_i}r_i\mathbf{1}_{\mathbf{s}_i}=r_i\ .$$

□

## D Proof of Proposition 1

For a given logging policy $\mu$ and context $x$, let

$$\bar{\rho}_{\mu,x}\coloneqq\sup_{\mathbf{s}\in\operatorname{supp}\mu(\cdot|x)}\mathbf{1}_\mathbf{s}^T\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{1_s}\ .$$

This quantity can be viewed as a norm of $\boldsymbol{\Gamma}^\dagger_{\mu,x}$ with respect to the set of slates chosen by $\mu$ with non-zero probability. It can be used to bound $\sigma^2$ and $\rho$, and thus to bound an error of $\hat{V}_{\mathrm{PI}}$:

**Proposition 2.** *For any logging policy $\mu$ and target policy $\pi$ that is absolutely continuous with respect to $\mu$, we have*

$$\sigma^2\le\rho\le\sup_x\bar{\rho}_{\mu,x}\ .$$

*Proof.* Recall that

$$\sigma^2=\mathbb{E}_{x\sim D}\left[\mathbf{q}^T_{\pi,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{q}_{\pi,x}\right]\ ,\qquad\rho=\sup_x\sup_{\mathbf{s}\in\operatorname{supp}\mu(\cdot|x)}\left|\mathbf{q}^T_{\pi,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{1_s}\right|\ .$$

To see that $\sigma^2\le\rho$ note that

$$\mathbf{q}^T_{\pi,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{q}_{\pi,x}=\mathbb{E}_{\mathbf{s}\sim\pi(\cdot|x)}\left[\mathbf{q}^T_{\pi,x}\boldsymbol{\Gamma}^\dagger_{\mu,x}\mathbf{1_s}\right]\le\rho$$

where the last inequality follows by the absolute continuity of $\pi$ with respect to $\mu$. It remains to show that $\rho\le\sup_x\bar{\rho}_{\mu,x}$.

First, by positive semi-definiteness of $\boldsymbol{\Gamma}_{\mu,x}^{\dagger}$ and from the definition of $\bar{\rho}_{\mu,x}$, we have that for any slates $\mathbf{s}, \mathbf{s}' \in \operatorname{supp} \mu(\cdot \mid x)$ and any $z \in \{-1, 1\}$

$$z \mathbf{1}_{\mathbf{s}'}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}} \le \frac{\mathbf{1}_{\mathbf{s}}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}} + \mathbf{1}_{\mathbf{s}'}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}'}}{2} \le \max\{\mathbf{1}_{\mathbf{s}}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}},\ \mathbf{1}_{\mathbf{s}'}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}'}\} \le \bar{\rho}_{\mu,x} \ .$$

Therefore, for any $\pi$ absolutely continuous with respect to $\mu$ and any $\mathbf{s} \in \operatorname{supp} \mu(\cdot \mid x)$, we have

$$\left|\mathbf{q}_{\pi,x}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}}\right| = \max_{z \in \{-1,1\}} \mathbb{E}_{\mathbf{s}' \sim \pi(\cdot|x)} \left[z \mathbf{1}_{\mathbf{s}'}^T \boldsymbol{\Gamma}_{\mu,x}^{\dagger} \mathbf{1}_{\mathbf{s}}\right] \le \bar{\rho}_{\mu,x} \ .$$

Taking a supremum over $x$ and $\mathbf{s} \in \operatorname{supp} \mu(\cdot \mid x)$, we obtain $\rho \le \sup_x \bar{\rho}_{\mu,x}$. □

We next derive bounds on $\bar{\rho}_{\mu,x}$ for uniformly-random policies in the ranking and cartesian product examples. Then we prove a translation theorem, which allows translating of the bound for uniform distribution into a bound for pairwise $\kappa$-uniform distributions. Finally, we put these results together to prove Prop. 1.

### D.1 Bounds for uniform distributions

Let $\mathbf{1}_j \in \mathbb{R}^{\ell m}$ be the vector that is all-ones on the actions in the $j$-th position and zeros elsewhere. Similarly, let $\mathbf{1}_a \in \mathbb{R}^{\ell m}$ be the vector that is all-ones on the action $a$ in all positions and zeros elsewhere. Finally, let $\mathbf{1} \in \mathbb{R}^{\ell m}$ be the all-ones vector. We also use $\mathbf{I}_j = \operatorname{diag}(\mathbf{1}_j)$ to denote the diagonal matrix with all-ones on the actions in the $j$-th position and zeros elsewhere.

**Proposition 3.** *Consider the product slate space where $S(x) = A_1(x) \times \cdots \times A_\ell(x)$ with $|A_j(x)| = m_j$. Let $\nu$ be the uniform exploration policy, i.e., $\nu(\mathbf{s} \mid x) = 1/|S(x)|$. Then $\bar{\rho}_{\nu,x} = \sum_j m_j - \ell + 1$ and*

$$\boldsymbol{\Gamma}_{\nu,x}^{\dagger} = \sum_{j=1}^{\ell} \left(m_j \mathbf{I}_j - \mathbf{1}_j \mathbf{1}_j^T\right) + \left(\sum_{j=1}^{\ell} \frac{1}{m_j}\right)^{-2} \sum_{j,k} \frac{\mathbf{1}_j \mathbf{1}_k}{m_j m_k} \ .$$

*For any policy $\pi$, any $\mathbf{s} \in S(x)$, and any $r \in [-1, 1]$ we then have*

$$\mathbf{q}_{\pi,x}^T \boldsymbol{\Gamma}_{\nu,x}^{\dagger} r \mathbf{1}_{\mathbf{s}} = r \cdot \left[\sum_{j=1}^{\ell} \frac{\pi(s_j \mid x)}{1/m_j} - \ell + 1\right] \ . \tag{14}$$

*Proof.* Throughout the proof we will write $\boldsymbol{\Gamma}$ instead of the more verbose $\boldsymbol{\Gamma}_{\nu,x}$ and similarly $\bar{\rho}$ instead of $\bar{\rho}_{\nu,x}$. We will construct an explicit eigendecomposition of $\boldsymbol{\Gamma}$, which will immediately yield $\boldsymbol{\Gamma}^{\dagger}$. The remaining statements will follow by a direct calculation. From the definition of $\boldsymbol{\Gamma}$, we obtain

$$\boldsymbol{\Gamma} = \sum_{j=1}^{\ell} \frac{\mathbf{I}_j}{m_j} + \sum_{j,k} \frac{\mathbf{1}_j \mathbf{1}_k^T}{m_j m_k} - \sum_j \frac{\mathbf{1}_j \mathbf{1}_j^T}{m_j^2} \ . \tag{15}$$

Let $\mathbf{v} = \sum_j \mathbf{1}_j / m_j$ so that the second term on the right-hand side of Eq. (15) corresponds to $\mathbf{v}\mathbf{v}^T$. Thus, we can write

$$\boldsymbol{\Gamma} = \|\mathbf{v}\|_2^2 \cdot \frac{\mathbf{v}\mathbf{v}^T}{\|\mathbf{v}\|_2^2} + \sum_{j=1}^{\ell} \frac{1}{m_j} \left(\mathbf{I}_j - \frac{\mathbf{1}_j \mathbf{1}_j^T}{m_j}\right) \ . \tag{16}$$

We argue that this constitutes an eigendecomposition. Let $\mathbf{P}_j := \mathbf{I}_j - \mathbf{1}_j \mathbf{1}_j^T / m_j$ denote the terms appearing in the sum on the right-hand side of Eq. (16). Note that $\mathbf{P}_j$'s are projection matrices, i.e., their eigenvalues are in $\{0, 1\}$. Moreover, their ranges are orthogonal to each other, because $\operatorname{Range}(\mathbf{P}_j)$ is a subset of the span of the coordinates corresponding to the slot $j$. Finally, note that $\mathbf{v}$ is orthogonal to all of the ranges, because

$$\mathbf{v}^T \mathbf{P}_j \mathbf{v} = \mathbf{v}^T \mathbf{I}_j \mathbf{v} - (\mathbf{v}^T \mathbf{1}_j)^2 / m_j = 1/m_j - 1/m_j = 0 \ .$$

This shows that Eq. (16) is an eigendecomposition of $\mathbf{\Gamma}$, so

$$\begin{aligned}
\mathbf{\Gamma}^\dagger &= \|\mathbf{v}\|_2^{-2} \cdot \frac{\mathbf{v}\mathbf{v}^T}{\|\mathbf{v}\|_2^2} + \sum_{j=1}^{\ell} m_j \left( \mathbf{I}_j - \frac{\mathbf{1}_j \mathbf{1}_j^T}{m_j} \right) \qquad (17) \\
&= \|\mathbf{v}\|_2^{-4} \cdot \mathbf{v}\mathbf{v}^T + \sum_{j=1}^{\ell} \left( m_j \mathbf{I}_j - \mathbf{1}_j \mathbf{1}_j^T \right) \\
&= \left( \sum_{j=1}^{\ell} \frac{1}{m_j} \right)^{-2} \sum_{j,k} \frac{\mathbf{1}_j \mathbf{1}_k^T}{m_j m_k} + \sum_{j=1}^{\ell} \left( m_j \mathbf{I}_j - \mathbf{1}_j \mathbf{1}_j^T \right) \quad ,
\end{aligned}$$

where the last equality follows from the definition of $\mathbf{v}$. It remains to derive $\bar{\rho}$ and Eq. (14). Both will follow by analyzing the expression $\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}}$ for $\mathbf{s}, \mathbf{s}' \in S(x)$. To begin, note that $\mathbf{1}_j^T \mathbf{1}_{\mathbf{s}} = 1$ since any valid slate chooses exactly one action in each position. Thus,

$$\begin{aligned}
\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}} &= \left( \sum_{j=1}^{\ell} \frac{1}{m_j} \right)^{-2} \sum_{j,k} \frac{(\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_j)(\mathbf{1}_k^T \mathbf{1}_{\mathbf{s}})}{m_j m_k} + \sum_{j=1}^{\ell} \Big( m_j \mathbf{1}_{\mathbf{s}'}^T \mathbf{I}_j \mathbf{1}_{\mathbf{s}} - (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_j)(\mathbf{1}_j^T \mathbf{1}_{\mathbf{s}}) \Big) \\
&= \left( \sum_{j=1}^{\ell} \frac{1}{m_j} \right)^{-2} \sum_{j,k} \frac{1}{m_j m_k} + \sum_{j=1}^{\ell} \Big( m_j \mathbf{1}\{s_j' = s_j\} - 1 \Big) \\
&= \left( \sum_{j=1}^{\ell} \frac{1}{m_j} \right)^{-2} \left( \sum_{j=1}^{\ell} \frac{1}{m_j} \right)^{2} + \sum_{j=1}^{\ell} \frac{\mathbf{1}\{s_j' = s_j\}}{1/m_j} - \ell \\
&= 1 + \sum_{j=1}^{\ell} \frac{\mathbf{1}\{s_j' = s_j\}}{1/m_j} - \ell \quad .
\end{aligned}$$

Now the value of $\bar{\rho}$ follows by setting $\mathbf{s}' = \mathbf{s}$, and Eq. (14) follows by taking an expectation over $\mathbf{s}' \sim \pi(\cdot \mid x)$. □

**Proposition 4.** *Consider the ranking setting where for each $x$ there is a set $A(x)$ such that $A_j(x) = A(x)$ and where all slates $\mathbf{s} \in A(x)^\ell$ without repetitions are legal. Let $\nu$ denote the uniform exploration policy over these slates. If $\ell < m$, then $\bar{\rho}_{\nu,x} = m\ell - \ell + 1$ and*

$$\mathbf{\Gamma}_{\nu,x}^\dagger = \left( \frac{1}{\ell^2} - \frac{m-1}{m(m-\ell)} \right) \cdot \mathbf{1}\mathbf{1}^T + (m-1)\mathbf{I} - \frac{m-1}{m} \sum_j \mathbf{1}_j \mathbf{1}_j^T + \frac{m-1}{m-\ell} \sum_a \mathbf{1}_a \mathbf{1}_a^T \quad ,$$

*and for $\ell = m$, we have $\bar{\rho}_{\nu,x} = m^2 - 2m + 2$ and*

$$\mathbf{\Gamma}_{\nu,x}^\dagger = \frac{1}{m} \cdot \mathbf{1}\mathbf{1}^T + (m-1)\mathbf{I} - \frac{m-1}{m} \sum_j \mathbf{1}_j \mathbf{1}_j^T - \frac{m-1}{m} \sum_a \mathbf{1}_a \mathbf{1}_a^T \quad .$$

*For $\ell = m$, we have for any policy $\pi$, any $\mathbf{s} \in S(x)$, and any $r \in [-1, 1]$ that*

$$\mathbf{q}_{\pi,x}^T \mathbf{\Gamma}_{\nu,x}^\dagger r \mathbf{1}_{\mathbf{s}} = r \cdot \left[ \sum_{j=1}^{\ell} \frac{\pi(s_j \mid x)}{1/(m-1)} - m + 2 \right] \quad . \qquad (18)$$

*Proof.* Throughout the proof we will write $\mathbf{\Gamma}$ instead of the more verbose $\mathbf{\Gamma}_{\nu,x}$. Note that for ranking and the uniform distribution we have

$$\mathit{\Gamma}(j, a; k, a') = \begin{cases} \frac{1}{m} & \text{if } j = k \text{ and } a = a' \\ \frac{1}{m(m-1)} & \text{if } j \neq k \text{ and } a \neq a' \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for any $\mathbf{z}$

$$\begin{aligned}\mathbf{z}^T\boldsymbol{\Gamma}\mathbf{z} &= \sum_{j,a}\frac{z_{j,a}^2}{m} + \frac{1}{m(m-1)}\sum_{j\neq k, a\neq a'} z_{j,a}z_{k,a'} \\ &= \frac{1}{m}\|\mathbf{z}\|_2^2 + \frac{1}{m(m-1)}\left((\mathbf{z}^T\mathbf{1})^2 - \sum_j(\mathbf{z}^T\mathbf{1}_j)^2 - \sum_a(\mathbf{z}^T\mathbf{1}_a)^2 + \|\mathbf{z}\|_2^2\right) \\ &= \frac{1}{m(m-1)}\left((\mathbf{z}^T\mathbf{1})^2 - \sum_j(\mathbf{z}^T\mathbf{1}_j)^2 - \sum_a(\mathbf{z}^T\mathbf{1}_a)^2 + m\|\mathbf{z}\|_2^2\right) \quad . \qquad (19)\end{aligned}$$

Let $\mathbf{1}_\mathcal{J} \in \mathbb{R}^\ell$ and $\mathbf{1}_\mathcal{A} \in \mathbb{R}^m$ be all-ones vectors in the respective spaces and $\mathbf{I}_\mathcal{J} \in \mathbb{R}^{\ell\times\ell}$ and $\mathbf{I}_\mathcal{A} \in \mathbb{R}^{m\times m}$ be identity matrices in the respective spaces. We can rewrite the quadratic form described by $\boldsymbol{\Gamma}$ as

$$\begin{aligned} m(m-1)\boldsymbol{\Gamma} &= \mathbf{1}\mathbf{1}^T - \sum_j \mathbf{1}_j\mathbf{1}_j^T - \sum_a \mathbf{1}_a\mathbf{1}_a^T + m\mathbf{I} \\ &= (\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T)\otimes(\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T) - \mathbf{I}_\mathcal{J}\otimes(\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T) - (\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T)\otimes\mathbf{I}_\mathcal{A} + m\cdot\mathbf{I}_\mathcal{J}\otimes\mathbf{I}_\mathcal{A} \\ &= \ell m\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} - m\cdot\mathbf{I}_\mathcal{J}\otimes\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} - \ell\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\mathbf{I}_\mathcal{A} + m\cdot\mathbf{I}_\mathcal{J}\otimes\mathbf{I}_\mathcal{A} \\ &= \ell(m-1)\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} - m\cdot\mathbf{I}_\mathcal{J}\otimes\left(\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} - \mathbf{I}_\mathcal{A}\right) - \ell\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right) \\ &= \ell(m-1)\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} \\ &\qquad + m\cdot\left(\mathbf{I}_\mathcal{J} - \frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\right)\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right) + (m-\ell)\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right). \qquad (20)\end{aligned}$$

Next, we would like to argue that Eq. (20) is an eigendecomposition. For this, we just need to show that each of the three Kronecker products in Eq. (20) equals a projection matrix in $\mathbb{R}^{\ell m}$, and that the ranges of the projection matrices are orthogonal. The first property follows, because if $\mathbf{P}_1$ and $\mathbf{P}_2$ are projection matrices then so is $\mathbf{P}_1\otimes\mathbf{P}_2$. The second property follows, because for $\mathbf{P}_1, \mathbf{P}'_1$ (square of the same dimension) and $\mathbf{P}_2, \mathbf{P}'_2$ (square of the same dimension) such that either ranges of $\mathbf{P}_1$ and $\mathbf{P}'_1$ are orthogonal or ranges of $\mathbf{P}_2$ and $\mathbf{P}'_2$ are orthogonal, we obtain that the ranges of $\mathbf{P}_1\otimes\mathbf{P}_2$ and $\mathbf{P}'_1\otimes\mathbf{P}'_2$ are orthogonal.

Now we are ready to derive the pseudo-inverse. We distinguish two cases.

**Case $\ell < m$:** We directly invert the eigenvalues in Eq. (20) to obtain

$$\begin{aligned}\boldsymbol{\Gamma}^\dagger &= \frac{m}{\ell}\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m} + (m-1)\cdot\left(\mathbf{I}_\mathcal{J} - \frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\right)\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right) \\ &\qquad\qquad + \frac{m-1}{1-\ell/m}\cdot\frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{\ell}\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right) \\ &= \frac{1}{\ell^2}\cdot\mathbf{1}\mathbf{1}^T + (m-1)\cdot\left(\mathbf{I}_\mathcal{J} + \frac{\mathbf{1}_\mathcal{J}\mathbf{1}_\mathcal{J}^T}{m-\ell}\right)\otimes\left(\mathbf{I}_\mathcal{A} - \frac{\mathbf{1}_\mathcal{A}\mathbf{1}_\mathcal{A}^T}{m}\right) \\ &= \left(\frac{1}{\ell^2} - \frac{m-1}{m(m-\ell)}\right)\cdot\mathbf{1}\mathbf{1}^T + (m-1)\mathbf{I} - \frac{m-1}{m}\sum_j\mathbf{1}_j\mathbf{1}_j^T + \frac{m-1}{m-\ell}\sum_a\mathbf{1}_a\mathbf{1}_a^T \quad .\end{aligned}$$

Recall that Eq. (20) involves $m(m-1)\mathbf{\Gamma}$. To obtain $\bar{\rho}$, we again evaluate $\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}}$ for any $\mathbf{s} \in S(x)$. We write $A_{\mathbf{s}}$ for the set of actions appearing on the slate $\mathbf{s}$:

$$\begin{aligned}
\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}} &= \left(\frac{1}{\ell^2} - \frac{m-1}{m(m-\ell)}\right) \cdot (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1})(\mathbf{1}^T \mathbf{1}_{\mathbf{s}}) + (m-1)\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_{\mathbf{s}} - \frac{m-1}{m} \sum_j (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_j)(\mathbf{1}_j^T \mathbf{1}_{\mathbf{s}}) \\
&\qquad + \frac{m-1}{m-\ell} \sum_a (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_a)(\mathbf{1}_a^T \mathbf{1}_{\mathbf{s}}) \\
&= \left(\frac{1}{\ell^2} - \frac{m-1}{m(m-\ell)}\right) \cdot \ell^2 + \sum_j \frac{\mathbf{1}\{s'_j = s_j\}}{1/(m-1)} - \frac{m-1}{m} \cdot \ell \\
&\qquad + \frac{m-1}{m-\ell} \sum_a \mathbf{1}\{a \in A_{\mathbf{s}'}\} \mathbf{1}\{a \in A_{\mathbf{s}}\} \qquad (21) \\
&= 1 - \frac{(m-1)(\ell^2 + m\ell - \ell^2)}{m(m-\ell)} + \sum_j \frac{\mathbf{1}\{s'_j = s_j\}}{1/(m-1)} + \frac{m-1}{m-\ell} \cdot |A_{\mathbf{s}'} \cap A_{\mathbf{s}}| \\
&= 1 - \frac{m-1}{m-\ell} \cdot \ell + \sum_j \frac{\mathbf{1}\{s'_j = s_j\}}{1/(m-1)} + \frac{m-1}{m-\ell} \cdot |A_{\mathbf{s}} \cap A_{\mathbf{s}'}| \ ,
\end{aligned}$$

where Eq. (21) follows because $\mathbf{1}^T \mathbf{1}_{\mathbf{s}} = \ell$ and $\mathbf{1}_j^T \mathbf{1}_{\mathbf{s}} = 1$ for any valid slate $\mathbf{s}$. By setting $\mathbf{s}' = \mathbf{s}$, we obtain $\bar{\rho} = 1 + \ell(m-1) = m\ell - \ell + 1$.

**Case $\ell = m$:** Again, we directly invert the eigenvalues in Eq. (20) to obtain

$$\begin{aligned}
\mathbf{\Gamma}^\dagger &= \frac{1}{\ell^2} \cdot \mathbf{1}\mathbf{1}^T + (m-1) \cdot \left(\mathbf{I}_{\mathcal{J}} - \frac{\mathbf{1}_{\mathcal{J}}\mathbf{1}_{\mathcal{J}}^T}{\ell}\right) \otimes \left(\mathbf{I}_{\mathcal{A}} - \frac{\mathbf{1}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}^T}{m}\right) \\
&= \frac{1}{m} \cdot \mathbf{1}\mathbf{1}^T + (m-1)\mathbf{I} - \frac{m-1}{m} \sum_j \mathbf{1}_j \mathbf{1}_j^T - \frac{m-1}{m} \sum_a \mathbf{1}_a \mathbf{1}_a^T \ .
\end{aligned}$$

We finish the theorem by evaluating $\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}}$:

$$\begin{aligned}
\mathbf{1}_{\mathbf{s}'}^T \mathbf{\Gamma}^\dagger \mathbf{1}_{\mathbf{s}} &= \frac{1}{m} \cdot (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1})(\mathbf{1}^T \mathbf{1}_{\mathbf{s}}) + (m-1)\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_{\mathbf{s}} - \frac{m-1}{m} \sum_j (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_j)(\mathbf{1}_j^T \mathbf{1}_{\mathbf{s}}) \\
&\qquad - \frac{m-1}{m} \sum_a (\mathbf{1}_{\mathbf{s}'}^T \mathbf{1}_a)(\mathbf{1}_a^T \mathbf{1}_{\mathbf{s}}) \\
&= \frac{1}{m} \cdot m^2 + \sum_j \frac{\mathbf{1}\{s'_j = s_j\}}{1/(m-1)} - \frac{m-1}{m} \cdot m - \frac{m-1}{m} \cdot m \\
&= \sum_j \frac{\mathbf{1}\{s'_j = s_j\}}{1/(m-1)} - m + 2 \ .
\end{aligned}$$

We obtain $\bar{\rho} = m^2 - 2m + 2$ by setting $\mathbf{s}' = \mathbf{s}$ and Eq. (18) by taking an expectation over $\mathbf{s}' \sim \pi(\cdot \mid x)$. □

### D.2 Translation theorem and proofs for pairwise $\kappa$-uniform distributions

In this section we derive bounds on $\bar{\rho}_{\mu,x}$ when $\mu$ is not necessarily uniform, but only pairwise $\kappa$-uniform. The main component of the result is a translation theorem relating $\bar{\rho}_{\mu,x}$ to $\bar{\rho}_{\nu,x}$ for arbitrary $\mu$ and $\nu$. This lets us translate the bounds for uniform distributions in Appendix D.1 into bounds for pairwise $\kappa$-uniform distributions.

**Theorem 4.** *Let $\mu$ and $\nu$ be arbitrary stochastic policies such that $\mu$ is absolutely continuous with respect to $\nu$. Then*

$$\kappa \bar{\rho}_{\mu,x} \le \bar{\rho}_{\nu,x} \ , \quad \textit{where } \kappa = \min\left\{\frac{\mu(s_j = a, s_k = a' \mid x)}{\nu(s_j = a, s_k = a' \mid x)} : \nu(s_j = a, s_k = a' \mid x) > 0\right\} \ .$$

*Proof.* We consider a fixed $x$ throughout the proof and to abbreviate the set of tuples $(j, a, k, a')$ over which the minimum in the theorem is taken, we define the set

$$\mathcal{I} := \{(j, a) : \nu(s_j = a \mid x) > 0\}$$

and a symmetric relation $\sim$ on $\mathcal{I}$ by

$$(j, a) \sim (k, a') \text{ if and only if } j \neq k \text{ and } \nu(s_j = a, s_k = a' \mid x) > 0 \ .$$

By Claim 1, $\mathbf{\Gamma}_{\mu,x} = \mathbb{E}_\mu[\mathbf{1}_\mathbf{s}^T \mathbf{1}_\mathbf{s} \mid x]$, so the assumption that $\mu$ is absolutely continuous with respect to $\nu$ implies that the $\mathrm{Null}(\mathbf{\Gamma}_{\nu,x})$ is a subset of $\mathrm{Null}(\mathbf{\Gamma}_{\mu,x})$:

$$\begin{aligned}\mathrm{Null}(\mathbf{\Gamma}_{\nu,x}) = \{\mathbf{v} : \mathbf{v}^T\mathbf{\Gamma}_{\nu,x}\mathbf{v} = 0\} &= \{\mathbf{v} : \mathbf{1}_\mathbf{s}^T\mathbf{v} = 0 \text{ for all } \mathbf{s} \in \operatorname{supp}\nu(\cdot \mid x)\} \\ &\subseteq \{\mathbf{v} : \mathbf{1}_\mathbf{s}^T\mathbf{v} = 0 \text{ for all } \mathbf{s} \in \operatorname{supp}\mu(\cdot \mid x)\} = \mathrm{Null}(\mathbf{\Gamma}_{\mu,x}). \quad (22)\end{aligned}$$

The proof is based on analyzing the generalized maximum eigenvalue defined as

$$\lambda_{\max}(\mathbf{\Gamma}_{\nu,x}, \mathbf{\Gamma}_{\mu,x}) := \max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{\Gamma}_{\mu,x}),\, \mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{\Gamma}_{\nu,x}\mathbf{z}}{\mathbf{z}^T\mathbf{\Gamma}_{\mu,x}\mathbf{z}} \ .$$

Specifically,

$$\begin{aligned}\bar{\rho}_{\mu,x} &= \sup_{\mathbf{s}:\mu(\mathbf{s}|x)>0} \mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x}^\dagger\mathbf{1}_\mathbf{s} = \sup_{\mathbf{s}:\mu(\mathbf{s}|x)>0} \left(\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\nu,x}^\dagger\mathbf{1}_\mathbf{s} \cdot \frac{\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x}^\dagger\mathbf{1}_\mathbf{s}}{\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\nu,x}^\dagger\mathbf{1}_\mathbf{s}}\right) \\ &\leq \left(\sup_{\mathbf{s}:\nu(\mathbf{s}|x)>0} \mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\nu,x}^\dagger\mathbf{1}_\mathbf{s}\right) \cdot \left(\sup_{\mathbf{s}:\mu(\mathbf{s}|x)>0} \frac{\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\mu,x}^\dagger\mathbf{1}_\mathbf{s}}{\mathbf{1}_\mathbf{s}^T\mathbf{\Gamma}_{\nu,x}^\dagger\mathbf{1}_\mathbf{s}}\right) && (23) \\ &\leq \bar{\rho}_{\nu,x} \cdot \left(\sup_{\mathbf{z}\perp\mathrm{Null}(\mathbf{\Gamma}_{\mu,x}),\, \mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{\Gamma}_{\mu,x}^\dagger\mathbf{z}}{\mathbf{z}^T\mathbf{\Gamma}_{\nu,x}^\dagger\mathbf{z}}\right) && (24) \\ &\leq \bar{\rho}_{\nu,x} \cdot \left(\sup_{\mathbf{z}\perp\mathrm{Null}(\mathbf{\Gamma}_{\mu,x}),\, \mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{\Gamma}_{\nu,x}\mathbf{z}}{\mathbf{z}^T\mathbf{\Gamma}_{\mu,x}\mathbf{z}}\right) = \bar{\rho}_{\nu,x} \cdot \lambda_{\max}(\mathbf{\Gamma}_{\nu,x}, \mathbf{\Gamma}_{\mu,x}) \ . && (25)\end{aligned}$$

Eq. (23) follows because $\mathrm{Null}(\mathbf{\Gamma}_{\nu,x}) \subseteq \mathrm{Null}(\mathbf{\Gamma}_{\mu,x})$. Eq. (24) follows because, by Eq. (22), $\mathbf{1}_\mathbf{s} \perp \mathrm{Null}(\mathbf{\Gamma}_{\mu,x})$ for all $\mathbf{s} \in \operatorname{supp}\mu(\cdot \mid x)$. Finally, Eq. (25) follows by Claim 4 and the definition of generalized eigenvalue.

To finish the theorem, it remains to upper bound the generalized eigenvalue $\lambda_{\max}(\mathbf{\Gamma}_{\nu,x}, \mathbf{\Gamma}_{\mu,x})$. To that end, we will apply Claim 5, which means we must construct complex-valued matrices $\mathbf{U}, \mathbf{V} \in \mathbb{C}^{\ell m\times r}$ for some $r$ such that $\mathbf{\Gamma}_{\nu,x} = \mathbf{U}\mathbf{U}^T$ and $\mathbf{\Gamma}_{\mu,x} = \mathbf{V}\mathbf{V}^T$. Recall that

$$\mathbf{\Gamma}_{\mu,x}(j, a;\, k, a') = \begin{cases}\mu(s_j = a \mid x) & \text{if } j = k \text{ and } a = a', \\ \mu(s_j = a, s_k = a' \mid x) & \text{if } j \neq k, \\ 0 & \text{otherwise.}\end{cases}$$

Notice that for any fixed $j, a$ and $k \neq j$, we have $\sum_{a'} \mu(s_j = a, s_k = a' \mid x) = \mu(s_j = a \mid x)$. This means that for a fixed $(j, a)$:

$$\sum_{(k,a'):\, (k,a')\sim(j,a)} \mu(s_j = a, s_k = a' \mid x) = (\ell - 1)\mu(s_j = a \mid x) \ .$$

Using this fact and the definition of $\mathbf{\Gamma}_{\mu,x}$, we can write

$$\mathbf{\Gamma}_{\mu,x} = \frac{1}{2} \sum_{(j,a)\sim(k,a')} \mathbf{e}_{ja,ka'}\mathbf{e}_{ja,ka'}^T \mu(s_j = a, s_k = a' \mid x) \ - \ \frac{\ell - 3}{2} \sum_{(j,a)\in\mathcal{I}} \mathbf{e}_{ja}\mathbf{e}_{ja}^T \mu(s_j = a \mid x) \ .$$

Here $\mathbf{e}_{ja,ka'} \in \{0, 1\}^{\ell m}$ is a vector with two non-zeros, one in the $(j, a)$ coordinate and one in the $(k, a')$ coordinate, and $\mathbf{e}_{ja}$ is defined similarly, with just one non-zero. The notation $\sum_{(j,a)\sim(k,a')}$ denotes first summing over all pairs $(j, a)$ and then summing over all pairs $(k, a')$ with satisfying the $\sim$ relations. Thus each pair is counted twice.

Using this decomposition, we can write $\mathbf{\Gamma}_{\mu,x} = \mathbf{V}\mathbf{V}^T$ where $\mathbf{V}$ has one column for each $(j, a, k, a')$ tuple such that $(j, a) \sim (k, a')$, with vector $\sqrt{\mu(s_j = a, s_k = a' \mid x)/2} \cdot \mathbf{e}_{ja,ka'}$, and one column for each $(j, a) \in \mathcal{I}$, with vector $\sqrt{-(\ell - 3)\mu(s_j = a \mid x)/2} \cdot \mathbf{e}_{ja}$. Similarly, we can write $\mathbf{\Gamma}_{\nu,x} = \mathbf{U}\mathbf{U}^T$ with the same decomposition but using the coefficients $\sqrt{\nu(s_j = a, s_k = a' \mid x)/2}$ and $\sqrt{-(\ell - 3)\nu(s_j = a \mid x)/2}$ on the corresponding vectors. Notice that these matrices may have complex entries, since $\ell$ can be larger than 3.

Finally, we can write $\mathbf{W}$ as a diagonal matrix with entries $\sqrt{\frac{\nu(s_j=a|x)}{\mu(s_j=a|x)}}$ and $\sqrt{\frac{\nu(s_j=a,s_k=a'|x)}{\mu(s_j=a,s_k=a'|x)}}$; note that if any of the denominators is zero then $\kappa = 0$ and the theorem holds, so we are assuming that all denominators are non-zero. Aligning coordinates, it is easy to see that $\mathbf{U} = \mathbf{V}\mathbf{W}$, and by Claim 5, we have

$$\begin{aligned}\lambda_{\max}(\mathbf{\Gamma}_{\nu,x}, \mathbf{\Gamma}_{\mu,x}) \le \|\mathbf{W}\|_2^2 &\le \max\left\{\max_{(j,a)\in\mathcal{I}} \frac{\nu(s_j = a \mid x)}{\mu(s_j = a \mid x)},\ \max_{(j,a)\sim(k,a')} \frac{\nu(s_j = a, s_k = a' \mid x)}{\mu(s_j = a, s_k = a' \mid x)}\right\} \\ &= \max\left\{\frac{\nu(s_j = a, s_k = a' \mid x)}{\mu(s_j = a, s_k = a' \mid x)} : \nu(s_j = a, s_k = a' \mid x) > 0\right\} = \kappa^{-1}.\end{aligned}$$

Plugging this into Eq. (25) proves the theorem. □

*Proof of Prop. 1.* The proposition follows by Prop. 2 with the $\bar{\rho}_{\mu,x}$ bounded by Theorem 4, using the definition of pairwise $\kappa$-uniform distributions and the values of $\bar{\rho}_{\nu,x}$ obtained in Props. 3 and 4. □

### D.3 Supporting claims

**Claim 4.** *Let* $\mathbf{A}, \mathbf{B}$ *be two symmetric positive semi-definite matrices with* $\mathrm{Null}(\mathbf{A}) \subseteq \mathrm{Null}(\mathbf{B})$. *Then*

$$\max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{B}^\dagger\mathbf{z}}{\mathbf{z}^T\mathbf{A}^\dagger\mathbf{z}} \le \max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{A}\mathbf{z}}{\mathbf{z}^T\mathbf{B}\mathbf{z}} \ .$$

*Proof.* Let $\mathbf{U}$ be the square root of matrix $\mathbf{A}$, i.e., $\mathbf{U}$ is a symmetric positive semidefinite matrix with the same eigenvectors as $\mathbf{A}$, but with eigenvalues that are square root of the corresponding eigenvalues of $\mathbf{A}$. Similarly, let $\mathbf{V}$ be the square root of matrix $\mathbf{B}$. Thus, we have $\mathbf{A} = \mathbf{U}\mathbf{U}$ and $\mathbf{A}^\dagger = \mathbf{U}^\dagger\mathbf{U}^\dagger$ and similarly for $\mathbf{B}$ and $\mathbf{V}$. Let $\mathbf{\Pi}_\mathbf{A} = \mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger$ denote the projection onto the range of $\mathbf{A}$ and $\mathbf{\Pi}_\mathbf{B}$ denote the projection onto the range of $\mathbf{B}$. Since $\mathrm{Null}(\mathbf{A}) \subseteq \mathrm{Null}(\mathbf{B})$, we have $\mathrm{Range}(\mathbf{A}) \supseteq \mathrm{Range}(\mathbf{B})$. We prove the claim as follows:

$$\begin{aligned}\max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{B}^\dagger\mathbf{z}}{\mathbf{z}^T\mathbf{A}^\dagger\mathbf{z}} &= \max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\ \mathbf{U}^\dagger\mathbf{U}\ \mathbf{B}^\dagger\ \mathbf{U}\mathbf{U}^\dagger\ \mathbf{z}}{\mathbf{z}^T\ \mathbf{U}^\dagger\mathbf{U}^\dagger\ \mathbf{z}} && (26)\\
&\le \max_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T\mathbf{U}\ \mathbf{B}^\dagger\ \mathbf{U}\mathbf{y}}{\mathbf{y}^T\mathbf{y}} && (27)\\
&= \max_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T\mathbf{U}\ \mathbf{V}^\dagger\mathbf{V}^\dagger\ \mathbf{U}\mathbf{y}}{\mathbf{y}^T\mathbf{y}} = \max_{\mathbf{y}:\ \|\mathbf{y}\|_2=1} \|\mathbf{V}^\dagger\mathbf{U}\mathbf{y}\|_2^2 && (28)\\
&= \max_{\mathbf{y}:\ \|\mathbf{y}\|_2=1} \|\mathbf{U}\mathbf{V}^\dagger\mathbf{y}\|_2^2 && (29)\\
&= \max_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T\mathbf{V}^\dagger\mathbf{U}\ \mathbf{U}\mathbf{V}^\dagger\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \\
&= \max_{\mathbf{y}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T\ \mathbf{V}^\dagger\mathbf{A}\mathbf{V}^\dagger\ \mathbf{y}}{\mathbf{y}^T\mathbf{y}} && (30)\\
&= \max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{V}\ \mathbf{V}^\dagger\mathbf{A}\mathbf{V}^\dagger\ \mathbf{V}\mathbf{z}}{\mathbf{z}^T\ \mathbf{V}\mathbf{V}\ \mathbf{z}} && (31)\\
&= \max_{\mathbf{z}\perp\mathrm{Null}(\mathbf{B}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\mathbf{A}\mathbf{z}}{\mathbf{z}^T\mathbf{B}\mathbf{z}} \ . && (32)\end{aligned}$$

In Eq. (26) we substitute $\mathbf{U}^\dagger\mathbf{U}^\dagger = \mathbf{A}^\dagger$ and also use the fact that $\mathbf{U}\mathbf{U}^\dagger = \mathbf{\Pi}_\mathbf{A}$ and $\mathbf{\Pi}_\mathbf{A}\mathbf{z} = \mathbf{z}$ because $\mathbf{z} \in \mathrm{Range}(\mathbf{B}) \subseteq \mathrm{Range}(\mathbf{A})$. Eq. (27) is obtained by substituting $\mathbf{y} = \mathbf{U}^\dagger\mathbf{z}$ and relaxing the maximization to be over $\mathbf{y} \neq \mathbf{0}$. In Eq. (28) we substitute $\mathbf{V}^\dagger \mathbf{V}^\dagger = \mathbf{B}^\dagger$. In Eq. (29) we use the fact that the operator norm of a matrix and its transpose are equal. In Eq. (30) we substitute $\mathbf{A} = \mathbf{U}\mathbf{U}$ and note that it suffices to consider $\mathbf{y} \perp \mathrm{Null}(\mathbf{B})$ because $\mathrm{Null}(\mathbf{V}^\dagger \mathbf{A} \mathbf{V}^\dagger) = \mathrm{Null}(\mathbf{B})$. In Eq. (31) we use the fact that $\mathbf{z} \mapsto \mathbf{V}\mathbf{z}$ is a bijection on $\mathrm{Range}(\mathbf{B})$, which is an orthogonal complement to $\mathrm{Null}(\mathbf{B})$, so we can substitute $\mathbf{V}\mathbf{z} = \mathbf{y}$. Finally, in Eq. (32) we substitute $\mathbf{B} = \mathbf{V}\mathbf{V}$ and use the fact that $\mathbf{V}^\dagger \mathbf{V} = \mathbf{\Pi}_\mathbf{B}$ and $\mathbf{\Pi}_\mathbf{B} \mathbf{z} = \mathbf{z}$ because $\mathbf{z} \in \mathrm{Range}(\mathbf{B})$. □

The next claim is due to Boman and Hendrickson [3], although we provide a simple proof for completeness.

**Claim 5** (Boman and Hendrickson [3])**.** *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ *be positive semidefinite matrices with* $\mathrm{Null}(\mathbf{B}) \subseteq \mathrm{Null}(\mathbf{A})$*. Let* $\mathbf{U}, \mathbf{V} \in \mathbb{C}^{d \times r}$ *for some* $r$ *be any two matrices such that* $\mathbf{A} = \mathbf{U}\mathbf{U}^T$ *and* $\mathbf{B} = \mathbf{V}\mathbf{V}^T$*. Let* $\mathbf{W} \in \mathbb{C}^{r \times r}$ *satisfy* $\mathbf{U} = \mathbf{V}\mathbf{W}$*. Then,*

$$\lambda_{\max}(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{z} \perp \mathrm{Null}(\mathbf{B}),\, \mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{A} \mathbf{z}}{\mathbf{z}^T \mathbf{B} \mathbf{z}} \leq \|\mathbf{W}\|_2^2 \ .$$

*Proof.*

$$\lambda_{\max}(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{z} \perp \mathrm{Null}(\mathbf{B}),\, \mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{A} \mathbf{z}}{\mathbf{z}^T \mathbf{B} \mathbf{z}} = \max_{\mathbf{z} \perp \mathrm{Null}(\mathbf{B}),\, \mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{U} \mathbf{U}^T \mathbf{z}}{\mathbf{z}^T \mathbf{V} \mathbf{V}^T \mathbf{z}} = \max_{\mathbf{z} \perp \mathrm{Null}(\mathbf{B}),\, \mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T \mathbf{V} \mathbf{W} \mathbf{W}^T \mathbf{V}^T \mathbf{z}}{\mathbf{z}^T \mathbf{V} \mathbf{V}^T \mathbf{z}}$$

$$\leq \max_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T \mathbf{W} \mathbf{W}^T \mathbf{y}}{\mathbf{y}^T \mathbf{y}} = \|\mathbf{W}\|_2^2 \ .$$

□

## E The $p$-values for plots in Figures 1 and 2

Table 1: The $p$-values of a $t$-test between PI and IPS, and PI and DM on search engine data (Fig. 1). Results where DM performs better than PI are omitted.

| number of | PI vs IPS | | PI vs DM | |
|---|---|---|---|---|
| samples ($n$) | TTS | UTILITYRATE | TTS | UTILITYRATE |
| 200 | $2.5 \times 10^{-1}$ | $4.7 \times 10^{-3}$ | — | — |
| 600 | $3.8 \times 10^{-2}$ | $1.6 \times 10^{-3}$ | — | — |
| 2 000 | $1.3 \times 10^{-5}$ | $2.0 \times 10^{-2}$ | — | — |
| 6 000 | $3.7 \times 10^{-5}$ | $2.0 \times 10^{-2}$ | — | $1.8 \times 10^{-2}$ |
| 20 000 | $1.2 \times 10^{-5}$ | $1.9 \times 10^{-2}$ | $1.5 \times 10^{-3}$ | $4.3 \times 10^{-7}$ |
| 60 000 | $4.5 \times 10^{-6}$ | $< 10^{-8}$ | $8.1 \times 10^{-4}$ | $< 10^{-8}$ |

Table 2: The $p$-values of a $t$-test between PI and IPS, and PI and DM on semi-synthetic data (Fig. 2). Results where DM performs better than PI are omitted.

| number of | $\ell = 5$, $m = 20$, $\alpha = 0$ | | $\ell = 10$, $m = 20$, $\alpha = 0$ | | $\ell = 5$, $m = 20$, $\alpha = 10$ | |
|---|---|---|---|---|---|---|
| samples ($n$) | PI vs wIPS | PI vs DM | PI vs wIPS | PI vs DM | PI vs wIPS | PI vs DM |
| 200 | $< 10^{-8}$ | — | $< 10^{-8}$ | — | $< 10^{-8}$ | — |
| 600 | $< 10^{-8}$ | — | $< 10^{-8}$ | — | $1.0 \times 10^{-8}$ | — |
| 2 000 | $< 10^{-8}$ | — | $< 10^{-8}$ | — | $< 10^{-8}$ | — |
| 6 000 | $< 10^{-8}$ | — | $< 10^{-8}$ | — | $< 10^{-8}$ | — |
| 20 000 | $< 10^{-8}$ | $7.3 \times 10^{-2}$ | $< 10^{-8}$ | — | $< 10^{-8}$ | — |
| 60 000 | $< 10^{-8}$ | $5.6 \times 10^{-3}$ | $< 10^{-8}$ | — | $< 10^{-8}$ | — |
| 200 000 | $< 10^{-8}$ | $6.0 \times 10^{-5}$ | $< 10^{-8}$ | $6.1 \times 10^{-2}$ | $< 10^{-8}$ | $4.4 \times 10^{-4}$ |
| 600 000 | $< 10^{-8}$ | $< 10^{-8}$ | $< 10^{-8}$ | $7.3 \times 10^{-4}$ | $< 10^{-8}$ | $7.5 \times 10^{-5}$ |

## F Additional results for off-policy evaluation on semi-synthetic data

In Figure 3, we compare the performance of several variants of estimators in each family of baseline approaches (DM and IPS) plotted in Fig. 2. For the DM family of approaches, variants differ in

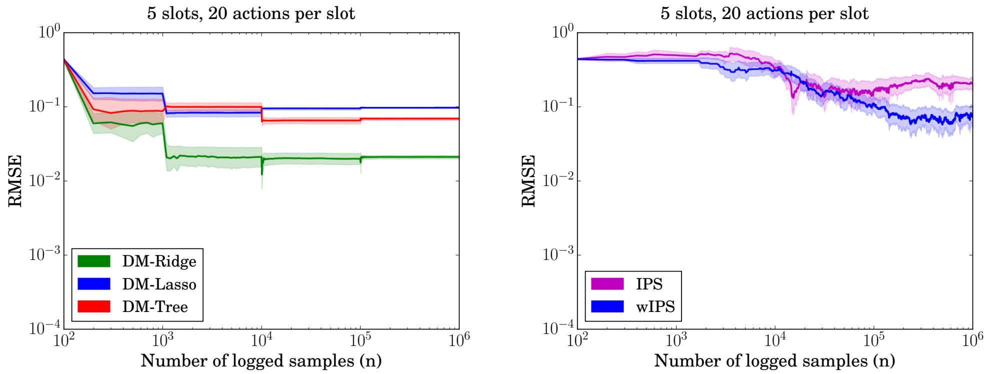

Figure 3: RMSE of value estimators for an increasing logged dataset under a uniform logging policy. $(m, l) = (20, 5), \alpha = 0$. Left: DM methods, Right: IPS estimators.

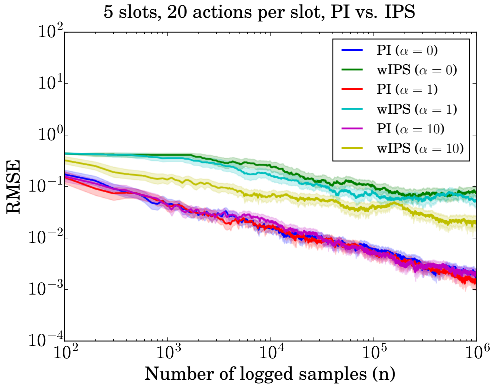

Figure 4: RMSE curves for pseudoinverse estimator and wIPS, $(m, \ell) = (20, 10), \alpha \in \{0, 1, 10\}$.

their choice of regression predictor $\hat{r}(x, \mathbf{s})$ that maps $\mathbf{f}(x, \mathbf{s})$ to $V(x, \mathbf{s})$. $\mathbf{f}(x, \mathbf{s})$ is defined as the concatenation of document features $\mathbf{f}(x, s_j)$ for all these variants. Regression hyper-parameters are selected via five-fold cross validation with each fold containing disjoint queries.

1. DM-tree: $\hat{r}(x, \mathbf{s})$ is implemented by a regression tree; maximum tree depth is the only hyper-parameter that is tuned.
2. DM-ridge: $\hat{r}(x, \mathbf{s})$ is obtained by ridge regression; $\ell_2$-regularization cross-validated
3. DM-lasso: $\hat{r}(x, \mathbf{s})$ is obtained by lasso regression; $\ell_1$-regularization cross-validated

For the IPS family, we compare standard inverse propensity scoring (IPS) against the weighted variant (wIPS). As theory predicts, RMSE of IPS is worse than that of wIPS since wIPS achieves a more favorable bias-variance trade-off.

Figure 4 shows a relative comparison of pseudoinverse estimator and IPS estimators as the discrepancy between the logging policy and the target policy is varied. The logging policy is as described in Section 4.1, parametrized by $\alpha \geq 0$. $\alpha = 0$ yields a uniform random logging policy and $\alpha \to \infty$ corresponds to a deterministic policy. As $\alpha$ is varied in $\{0, 1, 10\}$, pseudoinverse estimator remains stable while wIPS improves—this improvement is because the target policy and *pred*$_{\text{title}}$ (the deterministic extreme of $\mu$) overlap and the inverse propensity scores are better scaled and induce lower variance.

Finally, we also compare to the hypothetical semi-bandit approach, which uses more information than assumed by PI, IPS and DM. Semi-bandits assume that intrinsic values $\phi_{x_i}(j, s_{ij})$ are observed for $j \leq \ell$. Given these values, as defined in Example 3, i.e., $\phi_{x_i}(j, s_{ij}) = \left(2^{rel(x_i, s_{ij})} - 1\right) /$

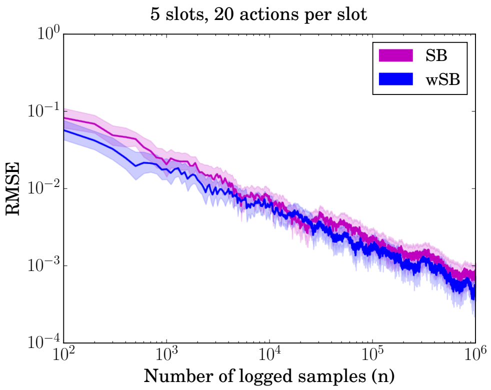

Figure 5: RMSE curves for SB using IPS estimator per slot (SB) and wIPS estimators (wSB), $(m,\ell) = (20,5), \alpha = 0$.

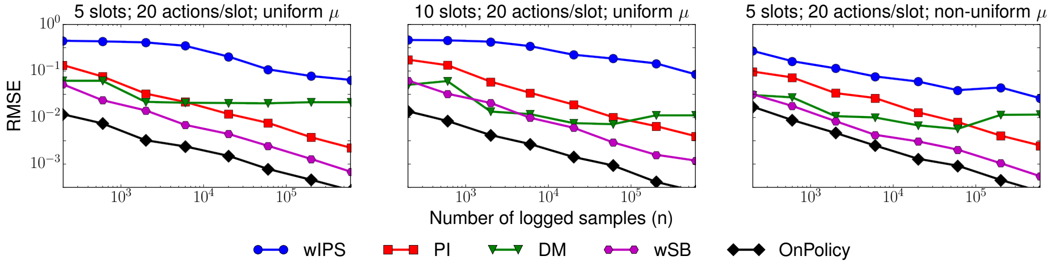

Figure 6: RMSE under uniform logging ($\alpha = 0$) and non-uniform logging ($\alpha = 10$).

$\log_2(j+1)\mathrm{DCG}^\star(x_i)$, the estimator $\hat{V}_{\mathrm{wSB}}$ sums wIPS estimates across slots:

$$\hat{V}_{\mathrm{wSB}}(\pi) := \sum_{j=1}^{\ell}\left[\sum_{i=1}^{n} \phi_{x_i}(j, s_{ij}) \cdot \frac{\pi(s_{ij} \mid x_i)}{\mu(s_{ij} \mid x_i)} \Big/ \left(\sum_{i=1}^{n} \frac{\pi(s_{ij} \mid x_i)}{\mu(s_{ij} \mid x_i)}\right)\right] .$$

It is only asymptotically unbiased, but it outperforms the unbiased variant based on standard IPS for each slot, as seen in Figure 5.

As Fig. 6 shows, the wSB approach requires somewhere between 4x and 10x less data than PI. So, in those cases when additional per-action feedback which relates to the page-level reward according to the semi-bandit model is available, this method is clearly preferred over PI. When the available feedback does not obviously satisfy the semibandit model, however, this approach will exhibit bias like the direct method. For instance, no obvious feedback of this nature was available in the search engine data from Section 4.3 and hence we could not evaluate the semi-bandits baseline in that setting.

## G Comparison of pointwise learning-to-rank approaches for off-policy optimization

Companion tables for Section 4.2 are provided in Tables 3, 4 and 5. Supervised pointwise learning-to-rank (L2R) algorithms typically regress to some monotone function of annotated relevance judgements $\mathit{rel}(x,a)$. Direct regression to $\mathit{rel}(x,a)$ gives the SUP-Rel approach, while regressing to $2^{\mathit{rel}(x,a)} - 1$ (which is well motivated by the fact that these methods are eventually trying to optimize NDCG) gives the SUP-Gain approach. Our PI-OPT approach is outlined in Section 4.2.

We studied the behavior of PI-OPT for three different model classes: decision tree regression, lasso and ridge regression. Since the MQ2008 dataset was already divided into 5 folds, for each fold we used the validation fold to tune hyper-parameters. After re-training on the train and validate

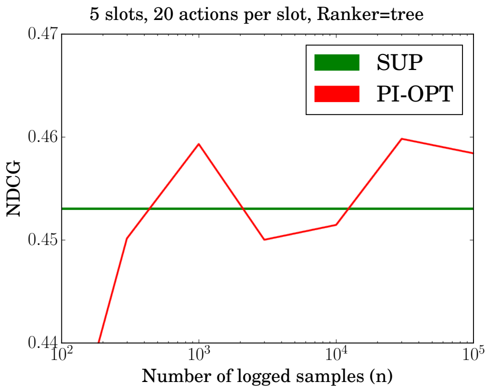

Figure 7: Test NDCG of off-policy optimization.

folds, we report the test fold NDCG. This procedure is repeated for 10 independent runs of $n = 10^5$ samples collected from the uniform random logging policy. Recall that SUP-Rel and SUP-Gain use approximately 12K annotated pairs. The average of the test set NDCG per fold, and the macro-average over folds is reported in these tables.

Table 3: Decision tree regression.

| Fold | Logger | SUP-Rel | SUP-Gain | PI-OPT |
|---|---|---|---|---|
| 1 | 0.273 | 0.455 | 0.461 | 0.473 |
| 2 | 0.285 | 0.426 | 0.427 | 0.421 |
| 3 | 0.289 | 0.415 | 0.420 | 0.426 |
| 4 | 0.273 | 0.470 | 0.469 | 0.469 |
| 5 | 0.259 | 0.480 | 0.489 | 0.492 |
| Avg | 0.276 | 0.449 | 0.453 | 0.456 |

We find that PI-OPT is able to compete and even outperform the best among SUP-Rel and SUP-Gain. We find that the number of samples needed to achieve parity is quite modest. Moreover, the variability across runs is negligible at $n = 10^5$ (standard error across 10 runs for each fold $< 0.002$).

Table 4: Lasso regression.

| Fold | Logger | SUP-Rel | SUP-Gain | PI-OPT |
|---|---|---|---|---|
| 1 | 0.273 | 0.466 | 0.459 | 0.467 |
| 2 | 0.285 | 0.427 | 0.427 | 0.413 |
| 3 | 0.289 | 0.425 | 0.423 | 0.413 |
| 4 | 0.273 | 0.468 | 0.462 | 0.484 |
| 5 | 0.259 | 0.492 | 0.486 | 0.517 |
| Avg | 0.276 | 0.456 | 0.451 | 0.459 |

Table 5: Ridge regression.

| Fold | Logger | SUP-Rel | SUP-Gain | PI-OPT |
|---|---|---|---|---|
| 1 | 0.273 | 0.456 | 0.455 | 0.451 |
| 2 | 0.285 | 0.418 | 0.416 | 0.418 |
| 3 | 0.289 | 0.418 | 0.417 | 0.413 |
| 4 | 0.273 | 0.460 | 0.457 | 0.454 |
| 5 | 0.259 | 0.487 | 0.486 | 0.476 |
| Avg | 0.276 | 0.448 | 0.446 | 0.442 |

Finally, in Fig. 7, we also depict the performance of PI-OPT as a function of increasing amount of logged samples for a particular run.