Focused Context Balancing for Robust Offline Policy Evaluation

Hao Zou[*]
Tsinghua University
ahio@163.com

Kun Kuang[†]
Tsinghua University
kkun2010@gmail.com

Boqi Chen[‡]
Boston University
bqchen@bu.edu

Peng Cui[†]
Tsinghua University
cuip@tsinghua.edu.cn

Peixuan Chen
Tencent
noahchen@tencent.com

ABSTRACT

Precisely evaluating the effect of new policies (e.g. ad-placement models, recommendation functions, ranking functions) is one of the most important problems for improving interactive systems. The conventional policy evaluation methods rely on online A/B tests, but they are usually extremely expensive and may have undesirable impacts. Recently, Inverse Propensity Score (IPS) estimators are proposed as alternatives to evaluate the effect of new policy with offline log data that was collected from a different policy in the past. They tend to remove the distribution shift induced by past policy, but ignore the distribution shift that would be induced by the new policy. Moreover, their performances rely on accurate estimation of propensity score, which can not be guaranteed or validated in practice. In this paper, we propose a non-parametric method, named Focused Context Balancing (FCB) algorithm, to learn sample weights for context balancing, so that the distribution shift induced by the past policy and new policy can be eliminated respectively. To validate the effectiveness of our FCB algorithm, we conduct extensive experiments on both synthetic and real world datasets. The experimental results clearly demonstrate that our FCB algorithm outperforms existing estimators by achieving more precise and robust results for offline policy evaluation.

KEYWORDS

Policy Evaluation; Context Balancing; Distribution Shift

ACM Reference Format:
Hao Zou, Kun Kuang, Boqi Chen, Peng Cui, and Peixuan Chen. 2019. Focused Context Balancing for Robust Offline Policy Evaluation. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330852

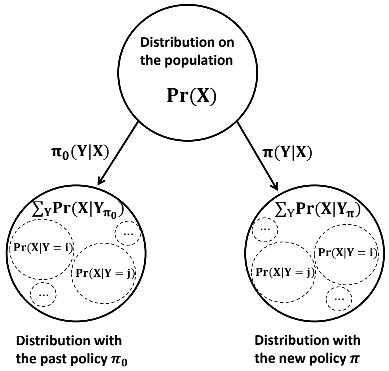


Figure 1: The distribution shift among the population, the logged data with the past policy π_0, and the future data with the new policy π.

1 INTRODUCTION

Policy evaluation is an extremely important problem in interactive systems, such as recommender systems, ad-placement systems, and search engines. For example, the policy (recommendation function) in a recommendation system decides the specific items (e.g. Ads) exposed to each user. The effect of the policy can be observed through feedbacks, such as click rate in the above example. The goal of policy evaluation is to select the well-performed ones among all proposed polices. Conventional approaches rely on online A/B tests [19], where researchers run each proposed policy on a fraction of randomly sampled users. Unfortunately, online A/B tests have two major drawbacks [1, 13]. One is long turnaround time, since each A/B test needs to be run on a certain fraction of the overall traffic and should ideally cover any cycles in user behavior. The other is that they can be detrimental to the user experience if the policy to be evaluated performs poorly.

To overcome these drawbacks, counterfactual estimators are proposed as alternatives to A/B tests by offline policy evaluation [6, 18, 21, 28] with only historical data produced by a past policy, which consists of contexts (e.g user characteristics), actions (e.g the Ads that are placed) and feedbacks (e.g click or not). The motivation of counterfactual estimators is to address the counterfactual reasoning problem of how a new policy would perform, if it has been applied instead of the past policy that has generated the observed data.

In offline policy evaluation problem, the policy assigns actions to units based on their contexts, leading to the context distribution in each action group become different with the one on the

[*] Beijing National Research Center for Information Science and Technology (BNRist).
[†] Corresponding authors
[‡] This work was finished during the author's visiting in Tsinghua.

KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330852

population as shown in Figure 1. We denote that distribution difference as distribution shift. The main challenge of offline policy evaluation is the distribution discrepancy in each action group (i.e., $Pr(\mathbf{X}|Y_{\pi_0} = k) \neq Pr(\mathbf{X}|Y_{\pi} = k)$) induced by the distribution shifts from different polices, including (1) **Distribution shift induced by the past policy** π_0. The past policy π_0 induces the distribution of context $\mathbf{X}$ in each action group (i.e., $Y = i$) becomes different with the one on the population, that is $Pr(\mathbf{X}|Y_{\pi_0} = i) \neq Pr(\mathbf{X})$. Moreover, the distribution of context $\mathbf{X}$ might become different between different action groups, say $Pr(\mathbf{X}|Y_{\pi_0} = i) \neq Pr(\mathbf{X}|Y_{\pi_0} = j)$ for $i \neq j$. (2) **Distribution shift induced by the new policy** π. When adopting a new policy π on the population, there will also be distribution shift among the action groups that are assigned by π, and those distributions also might be different with the one on the population. Hence, to achieve precise policy evaluation, one needs to remove that distribution discrepancy induced by the distribution shifts from both the past and the new polices.

Recent work on counterfactual estimators evaluate the new policy by either estimating the feedback function [8] or removing distribution shift in data by sample reweighting with the inverse of propensity score [20, 27]. The feedback function estimation [8] heavily relies on the correct model specification and may be affected by the distribution shift in the historical data. Propensity score reweighting methods [20, 27] can be applied to remove the distribution shift induced by the past policy, but they need accurate estimation on propensity score, which can not be guaranteed or validated in many real applications. More importantly, propensity score reweighting methods only focus on removing the distribution shift induced by the past policy, while ignoring the distribution shift induced by the new policy (as shown in Figure 1), leading to imprecise and high variance estimation on the new policy.

In this paper, we propose a novel algorithm, named Focused Context Balancing (FCB), to address the challenges above. More specifically, we introduce covariate balancing [3, 10, 15, 31], a well proven non-parametric method for bias removal in causal inference field, into offline policy evaluation. By learning sample weights to balance the moments of contexts in any two groups, the weighted sample distributions in these groups will become more identical. To address the distribution shift problem in policy evaluation, we first incorporate knowledge of the new policy to infer the context distribution in each action group under the new policy, i.e. the shifted distribution induced by the new policy. Then, given an action, we conduct covariate balancing on the two action groups formed under the past and new policy respectively, to remove their distribution discrepancy. Thereafter, with the sample weights learned from covariate balancing, we can easily evaluate the effect of the new policy based on the historical data. We validate our FCB algorithm with extensive experiments on both synthetic and real world datasets. The experimental results demonstrate that our algorithm outperforms existing methods.

The main contributions of this paper are as following:

- We introduce covariate balancing into offline policy evaluation to avoid model mis-specification and inaccurate estimation issues in traditional IPS based estimators.
- We propose a novel FCB estimator, which optimizes sample weights for context balancing to directly remove distribution shift from the past policy and new policy respectively, making the action group distributions under past and new policies become more identical for each action.
- The advantages of FCB estimator are demonstrated in both synthetic and real world datasets.

2 RELATED WORK

Existing counterfactual estimators for offline policy evaluation can be mainly categorized into two classes, direct method (DM) [8] and inverse propensity score (IPS) estimator [11, 12, 20, 23, 27].

DM regresses historical data on an estimated feedback function given context and action, then uses the estimated feedback in place of the actual feedback to evaluate the effect of new policy. Even though many machine learning algorithms can be employed for feedback function learning, DM often suffers from large bias since it ignores the distribution shift problem in historical data and requires correct model specification on the feedback function.

For IPS estimator, it is proposed to correct the distribution shift induced by the past policy by sample reweighting with the inverse of propensity score [4, 5, 16, 17],. Propensity score (PS) was proposed by Rosenbaum and Rubin [24] in causal inference, where it means the conditional probability of receiving treatment (i.e. the action) given the confounders (i.e. the context). In many policy evaluation applications, propensity score (i.e. the probability of a specified action) is unknown. It can be estimated with many machine learning algorithms, such as logistic regression [24, 26], lasso [7, 9], bagged CART and neural network [30], producing different variants of IPS estimators. However, the performance of IPS estimators rely on accurate estimation on propensity score, which can not be guaranteed or validated in many real applications. Further, IPS estimators ignore the distribution shift problem induced by the new policy. Because of the reliance on accurate propensity score and the negligence of distribution shift from the new policy, the policy evaluation based on IPS can be imprecise with high variance [29].

To reduce the high variance of IPS based estimators, some improved IPS methods have been proposed to incorporate control variate, such as Doubly Robust(DR) estimator [8, 22], and Self-Normalized IPS [14, 25, 29]. However, these methods still exist the above issues, including reliance on accurate propensity score and negligence of distribution shift from the new policy.

The direct bias removal method in our algorithm is inspired by the literatures in causal inference [3, 10, 15, 31], where researchers proposed to directly correct the distribution shift via confounder balancing, bypassing propensity score estimation. Hainmueller [10] proposed entropy balancing to adjust the sample weights as little as possible to match the target sample moments. Athey et al. [3] proposed approximate residual balancing by combining confounder balancing and lasso regression. Zubizarreta [31] learnt the balancing weights via minimizing its variance and directly adjust for confounder balancing. Kuang [15] differentiated the confounders and balanced the confounders unequally when learning the balancing weights. In this paper, we adopt the covariates balancing technique to correct the distribution shifts from both the past and new policies for offline policy evaluation.

3 PROBLEM STATEMENT

In this section, we give the basic concepts and notations of offline policy evaluation, and revisit some existing approaches.

3.1 Concepts and Notations

In this paper, we focus on estimating the effect of new policy with the offline logged data from a past policy in interactive learning systems. Based on the context information, which typically encodes users' characteristics, denoted as $\mathbf{X} \in \mathcal{X}$, the interactive learning system assigns an action $Y \in \mathcal{Y}$ according to its policy π. It will receive the feedback of the action in the form of a cardinal utility value $\delta : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$. Taking ad-placement systems for example, the context $\mathbf{X}$ consists of user's characteristics and the web page content, the policy π refers to the ad-placement algorithm, and the action Y represents the ads that displayed to users. Finally, the system will receive a feedback $\delta(\mathbf{X}, Y)$, depending on whether the user clicks on the ad or not.

In this paper, the context $\mathbf{X}$ are assumed to be randomly drawn from a fixed but unknown distribution $\mathbf{X} \sim Pr(\mathcal{X})$. We consider the policy π which we aim to evaluate is stochastic policy that defines a probability distribution over the action space $\mathcal{Y}$. Then the actions are assigned based on $Y \sim \pi(\mathcal{Y}|\mathbf{X})$. Our goal is to evaluate the utility of a new policy π based on the logged data from a past policy π_0. The utility of a policy $U(\pi)$ is defined as the expected utility of its feedback over both the context distribution and the action distribution. Formally,

$$U(\pi) = \mathbb{E}_{\mathbf{X} \sim Pr(\mathcal{X}), Y \sim \pi(\mathcal{Y}|\mathbf{X})} \left[\delta(\mathbf{X}, Y)\right]. \tag{1}$$

The logged data from a past policy π_0 consists of the contexts, the actions assigned by the past policy, and the feedback. The process of collecting that logged data is:

- A context $\mathbf{X}$ is sampled according to the distribution $Pr(\mathcal{X})$.
- The policy chooses an action $Y \sim \pi_0(\mathcal{Y}|\mathbf{X})$.
- Given context $\mathbf{X}$ and action Y, a feedback δ is revealed.

In many scenarios, we can hardly know the exact mechanism of the past policy. In this paper, therefore, we focus on a more general setting of offline policy evaluation where neither the context distribution $Pr(\mathcal{X})$ nor the past policy π_0 is known.

3.2 Basic Approaches

There are two kinds of basic approaches for offline policy evaluation, including direct methods (DM) and Inverse Propensity Score (IPS) estimators.

DM methods directly estimate the feedback function $\widehat{\delta}(\mathbf{X}, Y)$ by utilizing the logged data from the past policy. Then, they estimate the utility function of the new policy by

$$\widehat{U}_{DM}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{Y_j \in \mathcal{Y}} \widehat{\delta}(X_i, Y_j)\pi(Y_j|X_i), \tag{2}$$

where the feedback function $\widehat{\delta}$ can be estimated through any regression methods. The DM methods are unbiased if and only if the estimated feedback function is an accurate approximation of the expected feedback. In practical, however, we can hardly know the underlying feedback function, leading to inaccurate estimation on feedback function. Moreover, due to the affection of the past policy, the logged data used to learn the feedback function might have different distribution on context $\mathbf{X}$ among action groups, and might be different with the one on the population as we shown in Figure 1. Actually, DM methods suffer from large bias [8].

The other kind of methods are IPS estimators. Instead of estimating the feedback function, IPS estimators attempt to discover the underlying mechanism of the past policy by estimating the probability of each action given context (i.e. $ps = \hat{\pi}_0(Y|\mathbf{X})$). The motivation of these methods is that the distribution shift in logged data induced by the past policy can be removed by sample reweighting with the inverse of propensity score.

The estimated utility of the new policy π by IPS estimators can be written as:

$$\widehat{U}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\pi(Y_i|X_i)}{\widehat{\pi}_0(Y_i|X_i)}. \tag{3}$$

The performance of IPS estimators depends on the accuracy of estimated propensity score $\widehat{\pi}_0(Y|\mathbf{X})$. Those estimators are unbiased [6] for estimating $U(\pi)$, if the estimated propensity score is exactly the true action assignment mechanism of the past policy, that is $ps = \widehat{\pi}_0(Y|\mathbf{X}) = \pi_0(Y|\mathbf{X})$. However, in many real applications, we have no prior knowledge of the past policy, hence cannot guarantee the accuracy of estimated propensity score. Moreover, the IPS estimators ignore the distribution shift that would be induced by the new policy as we described in Figure 1, resulting in imprecise evaluation of the new policy.

4 OUR ESTIMATORS

In this section, we introduce the details of our proposed estimators, including Context Balancing (CB) estimator and Focused Context Balancing (FCB) estimator, for offline policy evaluation.

4.1 Context Balancing Estimator

To overcome the drawbacks of propensity score based methods, we propose a context balancing method to directly balance the context distributions among action groups with the past policy. From the knowledge of moments, we know that the distribution of each variable can be uniquely determined by the collection of all its moments. Hence, we propose to correct the distribution shift by directly moment balancing via sample weights learning.

We can separate the sample weights W into several parts and each part corresponds to the samples in one action group, that is $W = \{W_{Y=k} : k \in \mathcal{Y}\}$. Our algorithm learns the sample weights $W_{Y=k}$ with following objective function:

$$W_{Y=k} = \arg\min_{W_{Y=k}} \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{M}_i - \sum_{j:Y_j=k} W_j \cdot \mathbf{M}_j,\right\|_2^2, \tag{4}$$

where n is the total sample size over all action groups, and $j : Y_j = k$ refers to the samples that are assigned to action group $Y = k$. Here, term $\mathbf{M}$ represent the collection of all the moments of context $\mathbf{X}$, say $\mathbf{M} = \{\mathbf{X}, \mathbf{X}^2, \mathbf{X}_{,i}\mathbf{X}_{,j}, \mathbf{X}^3, \mathbf{X}_{,i}\mathbf{X}_{,j}\mathbf{X}_{,k}, \cdots\}$. Hence, term $\frac{1}{n}\sum_{i=1}^{n}\mathbf{M}_i$ refers to the context distribution on the population $Pr(\mathcal{X})$, and

$\sum_{j:Y_j=k} W_j \cdot \mathbf{M}_j$ represents the corrected context distribution in action group $Y = k$ with sample reweighting.

With the learned sample weights $W_{Y=k}$ for each action group $Y = k$ from our algorithm in Eq. (4), we can obtain the whole sample weights $W = \{W_{Y=k} : k \in \mathcal{Y}\}$. Then, the utility of the new policy π can be estimated by our Context Balancing algorithm as:

$$\widehat{U}_{CB}(\pi) = \sum_{i=1}^{n} \pi(Y_i|X_i) W_i \delta_i. \tag{5}$$

The following theoretical results show that our context estimator $\widehat{U}_{CB}(\pi)$ can unbiasedly estimate the utility of the new policy.

First, we have following corollaries on the sample weights W that learned from Eq. (4).

Corollary 1. *If the dimension of contexts p is finite, the distribution of contexts can be determined by finite order moments, and the sample size $n \longrightarrow \infty$, then $\exists \mathbf{W} \geq 0$ such that*

$$\lim_{n \to \infty} \Big\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}_i - \sum_{j:Y_j=k} W_j \cdot \mathbf{M}_{j,} \Big\|_2^2 = 0, \tag{6}$$

for each action group $Y = k$ with probability 1.

Corollary 2. *An ideal sample weights W learned from our algorithm can exactly ensure the context distribution of each action group is balanced, and equal to the distribution on the population. Formally, $\mathbb{E}\left[\sum_{i:Y_i=k, X_i=\mathbf{X}} W_i\right] = Pr(\mathbf{X}), \forall k \in \mathcal{Y}$.*

Next, we will theoretically prove that our proposed Context Balancing (CB) estimator is unbiased for offline policy evaluation based on Corollary 2.

Proposition 1. *Our Context Balancing estimator $\widehat{U}_{CB}(\pi)$ is unbiased for evaluating the utility of the new policy π.*

Proof.

$$\begin{aligned}
\mathbb{E}[\widehat{U}_{CB}(\pi)] &= \mathbb{E}\left[\textstyle\sum_{i=1}^{n} W_i \pi(Y = Y_i|X_i)\delta(X_i, Y_i)\right] \\
&= \mathbb{E}\left[\textstyle\sum_{\mathbf{X}\in\mathcal{X}} \sum_{k} \pi(Y = k|\mathbf{X})\delta(\mathbf{X}, k) \displaystyle\sum_{\substack{i:X_i=\mathbf{X} \\ Y_i=k}} W_i\right] \\
&= \textstyle\sum_{\mathbf{X}\in\mathcal{X}} Pr(\mathbf{X}) \sum_{k} \pi(Y = k|X)\delta(X, k) \\
&= U(\pi).
\end{aligned}$$

□

However, the proposed CB estimator also ignores the distribution shift that will be induced by the new policy. Hence, it would also suffer from high variance on the evaluation of the new policy.

4.2 Focused Context Balancing Estimator

From the distribution shift diagram as shown in Figure 1, we know that the new policy would also induce distribution shift among the action groups which are determined by the new policy. Existing estimators for offline policy evaluation did not recognize that distribution shift, hence wrongly balanced the context distribution of each action group to the same target distribution (i.e., context distribution on the population $Pr(\mathbf{X})$), leading to the high variance of offline policy evaluation.

To address the distribution shift induced by both past policy and new policy, we propose a new estimator, named Focused Context Balancing (FCB) algorithm, based on our previous proposed Context Balancing estimator. Comparing with Context Balancing estimator, our Focused Context Balancing estimator proposes a weighted balancing method by incorporating the prior knowledge of the new policy (i.e., $\pi(Y|\mathbf{X})$) to differentiate each action group. Hence, our FCB estimator can help the balancing process focus more on the samples that the new policy put larger probability on.

Actually, our proposed FCB estimator is motivated by following proposition.

Proposition 2. *To correct the distribution shift induced by both past and new policy, one should balance the context distribution in each action group to their own target distribution. And the target distribution of each action group is decided by the new policy.*

With Taylor's expansion on the context $\mathbf{X}$, we can represent the feedback function as $\delta(\mathbf{X}, Y = k) = \alpha_{Y=k} \cdot \mathbf{M}$, where $\mathbf{M} = \{\mathbf{X}, \mathbf{X}^2, \mathbf{X}_{,i}\mathbf{X}_{,j}, \mathbf{X}^3, \mathbf{X}_{,i}\mathbf{X}_{,j}\mathbf{X}_{,k}, \cdots\}$ is the collection of all the moments of context $\mathbf{X}$, and $\alpha_{Y=k}$ is a constant vector for each action group $Y = k$. Then, we can rewrite our weighted estimator in Eq. (5) as:

$$\begin{aligned}
\widehat{U}(\pi) &= \textstyle\sum_{i=1}^{n} W_i \pi(Y = Y_i|\mathbf{X}_{i,})\delta(\mathbf{X}_{i,}, Y_i) \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \left[\sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} - \frac{1}{n}\sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,}\right] \\
&\quad \textstyle +\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \frac{1}{n} \sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} B_k + \frac{1}{n}\sum_{i=1}^{n} \sum_{k\in\mathcal{Y}} \delta(\mathbf{X}_{i,}, Y = k)\pi(Y = k|\mathbf{X}_{i,}) \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} B_k + U(\pi),
\end{aligned}$$

where $U(\pi)$ is the true utility of new policy, $\sum_{k\in\mathcal{Y}} \alpha_{Y=k} B_k$ represents the bias between our weighted estimator $\widehat{U}(\pi)$ and the true utility $U(\pi)$, and $B_k = \sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} - \frac{1}{n}\sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,}$ refers to the bias from the action group $Y = k$. B_k indicates that when correcting distribution shift by weighting methods, the target distribution of the action group $Y = k$ should be $\frac{1}{n}\sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,}$, where π is the new policy. And for different action groups (i.e $Y = k$ and $Y = k'$), their target distributions are different, since $\pi(Y = k|\mathbf{X}) \neq \pi(Y = k'|\mathbf{X})$. Furthermore, we can observe that different moment variables contribute unequally to the bias due to the coefficient vector $\alpha_{Y=k}$. Hence, we could get a better result if we treat each moment variable unequally when balancing moments. We leave this extension to future work.

From the above theoretical analyses, we know the bias of weighted estimator, like CB estimator in Eq. (5), is induced by both past and new policy, which bring the distribution discrepancy on context $\mathbf{X}$ or its moments $\mathbf{M}$. Therefore, to fully reduce the bias and precisely evaluate the effect of the new policy, we propose a new weighted estimator, named Focused Context Balancing (FCB) algorithm, to optimize the sample weights $W_{Y=k}$ for each action group $Y = k$ by incorporating knowledge of the new policy $\pi(Y = k|\mathbf{X})$ as follow:

$$\begin{aligned}
&\min_{W_{Y=k}} \Big\| \sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} - \sum_{i=1}^{n} \frac{1}{n}\pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} \Big\|_2^2, \\
&s.t. \sum_{i:Y_i=k} W_i^2 \leq \lambda \sum_{i:Y_i=k} W_i = 1 \ \text{and} \ W \geq 0,
\end{aligned} \tag{7}$$

Algorithm 1 Focused Context Balancing Algorithm (FCB)

```
Input: Tradeoff parameters $\lambda > 0$, historical data $S = \{(X_i, Y_i, \delta_i)\}_{1 \le i \le n}$, and the new policy $\pi$.
Output: The estimated utility of policy $\widehat{U}_{FCB}(\pi)$.
for all action $k$ do
    Initialize sample weights in the group action $k$
    Calculate the current value of $J(W_{Y=k})^{(0)}$ with Equation (9)
    Initialize the iteration variable $t \leftarrow 0$
    repeat
        $t \leftarrow t + 1$
        Update $W^{(t)}$ by optimizing $J(W_{Y=k})^{(t-1)}$
        Calculate $J(W_{Y=k})^{(t)}$ with Equation (9)
    until $J(W_{Y=k})^{(t)}$ converges or max iteration is reached
end for
Calculate the effect of new policy $\widehat{U}_{FCB}(\pi)$ with Equation (8).
return $\widehat{U}_{FCB}(\pi)$.
```

where the formula $\sum_{i:Y_i=k} W_i = 1$ normalizes the weights of samples that were chosen action k by the past policy. $W \ge 0$ constraints each sample weight to be non-negative. Norm $\sum_{i:Y_i=j} W_i^2 \le \lambda$ reduces the variance of sample weights to achieves stability.

In our FCB algorithm, the balancing process will focus more on the samples that the new policy put large probability on in each action group, helping for better correcting the distribution shift induced by both past and new policy.

With the sample weights $W = \{W_{Y=k} : k \in \mathcal{Y}\}$ optimized by our FCB algorithm, we can estimate the effect as following:

$$\widehat{U}_{FCB}(\pi) = \sum_{i=1}^{n} \pi(Y_i|X_i) W_i \delta_i. \tag{8}$$

4.3 Optimization

For the action group k, the problem of Eq. (7) can be solved by optimizing the following objective functions $J(W_{Y=k})$:

$$J(W_{Y=k}) = \left\| W_{Y=k}\pi^k_{Y=k}\mathbf{M}_{Y=k} - \sum_{i=1}^{n} \tfrac{1}{n}\pi(Y=k|\mathbf{X}_{i,})\mathbf{M}_{i,} \right\|_2^2 + \lambda \left\| W_{Y=k} \right\|_2^2, \tag{9}$$
$$s.t. \quad W_{Y=k}\mathbf{1} = 1 \;\; and \;\; W_{Y=k} \ge 0,$$

where $W_{Y=k} \in \mathbb{R}^{1 \times n_k}$ is the vector of sample weights, $\pi^k_{Y=k} \in \mathbb{R}^{n_k \times n_k}$ is the diagonal matrix of $\pi(Y = k|X_{Y=k})$, and $\mathbf{M}_{Y=k} \in \mathbb{R}^{n_k \times l}$ is the matrix containing the moments, n_k is the number of samples chosen action k in historical data S, l is the dimension of moments. Since balancing all the moments is unrealistic, we focus on balancing the first-order moments of the contexts to make a trade-off between efficiency and feasibility.

Here, we use an iterative method to minimize the objective function $J(W_{Y=k})$.

The initial sample weights are set to be $W = 1/n_k$ for samples of action group $Y = k$. During each iteration, we update the weights using gradient descent. To ensure the non-negativity of W, we let $W = \omega \odot \omega$, where $\omega \in \mathbb{R}^{1 \times n}$. Symbol $\odot$ refers to Hadamard product. The problem can be rewritten as follow:

$$J(\omega_{Y=k}) = \left\| (\omega_{Y=k} \odot \omega_{Y=k})\pi^k_{Y=k}\mathbf{M}_{Y=k} - \sum_{i=1}^{n} \tfrac{1}{n}\pi(Y=k|\mathbf{X}_{i,})\mathbf{M}_{i,} \right\|_2^2$$
$$+\lambda \left\| \omega_{Y=k} \odot \omega_{Y=k} \right\|_2^2,$$
$$s.t. \; (\omega_{Y=k} \odot \omega_{Y=k})\mathbf{1} = 1.$$

The partial gradient of $J(\omega_{Y=k})$ with respect to $\omega_{Y=k}$ is:

$$\frac{\partial J(\omega_{Y=k})}{\partial \omega_{Y=k}} = 4((\omega_{Y=k} \odot \omega_{Y=k})\pi^k_{Y=k}\mathbf{M}_{Y=k} - \sum_{i=1}^{n} \tfrac{1}{n}\pi(Y=k|\mathbf{X}_{i,})\mathbf{M}_{i,})$$
$$(\pi^k_{Y=k}\mathbf{M}_{Y=k})^T \odot \omega_{Y=k} + 4\lambda\omega_{Y=k} \odot \omega_{Y=k} \odot \omega_{Y=k}.$$

We update $\omega_{Y=k}$ at the t^{th} iteration with step size a as:

$$\omega^{(t)}_{Y=k} = \omega^{(t-1)}_{Y=k} - a\frac{\partial J(\omega^{(t-1)}_{Y=k})}{\partial \omega^{(t-1)}_{Y=k}}.$$

With the constraint $(\omega_{Y=k} \odot \omega_{Y=k})\mathbf{1} = 1$, we normalize $\omega^{(t)}_{Y=k}$ at each iteration:

$$\omega^{(t)}_{Y=k} = \frac{\omega^{(t)}_{Y=k}}{\sqrt{(\omega^{(t)}_{Y=k} \odot \omega^{(t)}_{Y=k})\mathbf{1}}}.$$

We update $\omega_{Y=k}$ until the objective function $J(\omega_{Y=k})$ converges or the max iteration is reached. After optimizing objective function $J(\omega_{Y=k})$ for each action group, we can estimate the utility of policy π with sample weights W. The whole algorithm is summarized in Algorithm 1. The hype-parameter in our algorithm is tuned by grid searching.

4.4 Complexity Analysis

The time cost during optimization is mainly spent on calculating the loss $J(W_{Y=k})$ and updating the sample weights $W_{Y=k}$. For action group k, the time complexity of calculating loss $J(W_{Y=k})$ is $O(n_k l)$. The time complexity of updating sample weights $W_{Y=k}$ is dominated by calculating the partial gradient of function $J(\omega_{Y=k})$ with respect to $\omega_{Y=k}$, which is also $O(n_k l)$.

Hence, the time complexity of each iteration for action group k is $O(n_k l)$ and the total complexity is $\sum_k O(n_k l) = O(nl)$.

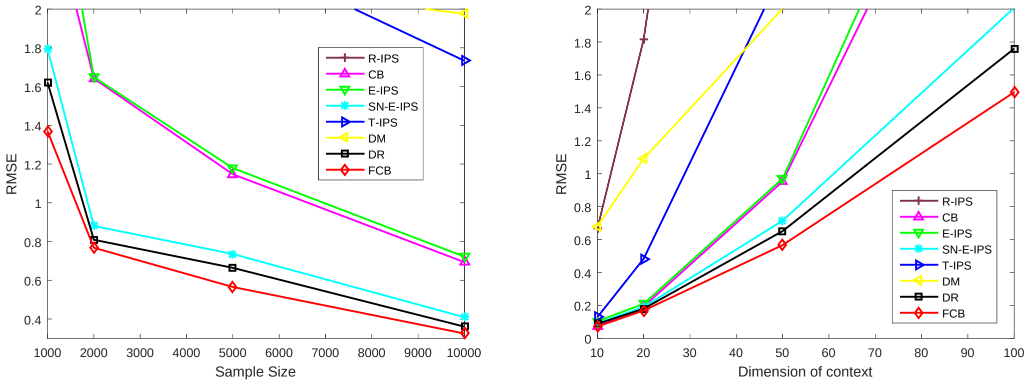
5 EXPERIMENT

In this section, we evaluate the effectiveness of our proposed estimator on both synthetic and real-world datasets.

5.1 Baseline Estimators

We implement the following baseline estimators for comparison. Parameter settings for baselines are as default.

- *Direct Method* $\widehat{U}_{DM}(\pi)$: It regresses historical data on an estimated feedback function given context and action to evaluate the effect of the new policy. The feedback function is estimated by elastic net.
- *Rough IPS* $\widehat{U}_{R\text{-}IPS}(\pi)$: It roughly assumes $\pi_0(Y_i|X_i)$ to be the proportion of samples with action Y_i in historical data, which ignore the association between contexts and actions.
- *IPS with estimated Propensity Score* $\widehat{U}_{E\text{-}IPS}(\pi)$[24]: It regresses on historical data to estimate $\pi_0(Y_i|X_i)$ in IPS estimator. In this paper, we choose logistic regression for this estimator.



(a) sample size n **(b) dimension of contexts p**

Figure 2: RMSE on policy evaluation when varying sample size and dimension of contexts under setting $\delta = \delta_{linear}$, $\pi_0 = \pi_{inv}$. In the figure (a), Rough IPS and DM make too huge error that the curves are out of the Y-axis range.

- *IPS with true Propensity Score* $\widehat{U}_{T\text{-}IPS}(\pi)$ [20, 27]: It assumes that the past policy is known to us and uses real value of $\pi_0(Y_i|X_i)$ to calculate IPS estimator.
- *Self-Normalized IPS* $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$ [23, 29]: It uses control variate to make the sample weights become more smooth by avoiding propensity overfitting, hence could reduce the variance of the estimator.
- *Doubly Robust Estimator* $\widehat{U}_{DR}(\pi)$ [8]: It evaluates the effect of the new policy with the combination of IPS and regression methods.
- *Context Balancing Estimator* $\widehat{U}_{CB}(\pi)$: It is a weaken version of our FCB estimator. It ignores the distribution shift induced by new policy when learning the sample weights.

5.2 Experiments on Synthetic Data

In this section, we introduce how to generate the synthetic datasets and demonstrate the effectiveness of our FCB estimator with extensive experiments.

5.2.1 Datasets. To test the robustness of our estimator, we generate the synthetic data with different settings. We first generate the context variables $\mathbf{X} = (x_1, x_2, ..., x_p)$ with considering of varying the population sample size $n = \{5,000, 10,000\}$ and the dimension of context variables $p = \{50, 100\}$. The context variables are generated with independent **Bernoulli** distribution as:

$$x_1, x_2, ..., x_p \overset{iid}{\sim} Bernoulli(0.5).$$

For each sample, we generate action Y_i by a controlled policy π_0 and observe the feedback δ_i. In this way, the historical data $S = \{(X_i, Y_i, \delta_i)\}_{i=1,\cdots,n}$ is accumulated.

As did in previous work [15], we also consider the action Y to be binary in our experiments, and we generate it from different functions as following:

$$\begin{aligned}
\pi_{inv}(Y = 1|\mathbf{X}) &= 1/(1 + 3\textstyle\sum_i x_i/p) + \mathcal{N}(0, 0.1),\\
\pi_{uni}(Y = 1|\mathbf{X}) &= 0.5 + \mathcal{N}(0, 0.1),\\
\pi_{lin}(Y = 1|\mathbf{X}) &= \textstyle\sum_i x_i/p + \mathcal{N}(0, 0.1).
\end{aligned}$$

These above policies are logging policies to generate the historical data. Here we generate the policy to be evaluated with sigmoid function (π_{sig}).

$$\pi_{sig}(Y = 1|\mathbf{X}) = 1/\left(1 + e^{-\sum_{i=1}^{p}(x_i - 0.5)}\right).$$

We generate the feedback δ from a linear function and a nonlinear function.

$$\begin{aligned}
\delta_{linear} &= Y + \textstyle\sum_{i=1}^{p}\left\{I(i \; mod \; 2 = 0)\cdot(\frac{i}{2} + Y)x_i\right\} + \mathcal{N}(0, 3),\\
\delta_{nonlin} &= Y + \textstyle\sum_{i=1}^{p}\left\{I(i \; mod \; 2 = 0)\cdot(\frac{i}{2} + Y)x_i\right\} + \mathcal{N}(0, 3)\\
&\quad + \textstyle\sum_{i=1}^{p-1}\left\{I(i \; mod \; 10 = 0)\cdot(\frac{i}{10} + Y)x_i x_{i+1}\right\},
\end{aligned}$$

where $I(\cdot)$ is the indicator function and function $mod(a, b)$ returns the modulus after division of a by b.

Under different settings on action Y and feedback δ, the ground truth (i.e. true utility of new policy) can be known as:

$$U(\pi) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{j=0}^{1}\delta(X_i, j)\pi(Y = j|X_i). \qquad (10)$$

We evaluate the utility of the new policy with our estimator, and comparing it with baseline estimators.

5.2.2 Results. To evaluate the performance of all estimators, we carry out the experiments 50 times independently under each experimental setting. Based on the estimated utilities and the ground truth, we calculate *Bias*, standard deviations (*SD*), mean absolute errors (*MAE*) and root mean square error (*RMSE*).

The results are reported in Table 1. From the results, we have the following observations and analyses:

- Rough IPS $\widehat{U}_{R\text{-}IPS}(\pi)$ fails when the past policy π_0 is π_{inv} or π_{lin}, where the contexts are associated with actions, but it ignores those associations.
- $\widehat{U}_{E\text{-}IPS}(\pi)$ achieves a more robust result than $\widehat{U}_{T\text{-}IPS}(\pi)$, and $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$ improves the robustness on $\widehat{U}_{E\text{-}IPS}(\pi)$. This is because extremely large or small value of propensity score would increase the variance of the estimator, while the estimated propensity score in $\widehat{U}_{E\text{-}IPS}(\pi)$ would be more smooth than the true propensity score, and $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$ uses control variates to further reduce the estimated variance.

Table 1: Results on synthetic datasets. The *Bias* is the absolute error between the ground truth and the mean estimated utility of policy π. The *SD*, *MAE* and *RMSE* refer to standard deviations, mean absolute errors and root mean square errors of the estimated policy utility in 50 times independent experiments. The smaller *Bias*, *SD*, *MAE* and *RMSE*, the better.

Setting 1:$\delta = \delta_{linear}$													
	n/p	$n = 5000, p = 50$			$n = 5000, p = 100$			$n = 10000, p = 50$			$n = 10000, p = 100$		
π_0	Estimator	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
π_{inv}	$\widehat{U}_{R\text{-}IPS}(\pi)$	7.306(1.632)	7.305	7.486	21.03(6.842)	21.03	22.11	7.083(1.399)	7.083	7.220	20.31(6.726)	20.31	21.40
	$\widehat{U}_{DM}(\pi)$	2.168(0.505)	2.168	2.226	3.612(1.274)	3.612	3.832	1.953(0.302)	1.953	1.975	3.439(1.104)	3.439	3.620
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.120(0.923)	0.787	0.927	0.577(3.865)	2.983	3.905	0.102(0.742)	0.641	0.746	**0.012**(3.015)	2.346	3.012
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.111(1.837)	1.496	1.839	0.058(7.736)	5.911	7.741	0.197(1.769)	1.486	1.780	0.360(7.382)	5.885	7.395
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.074(0.654)	0.540	0.659	**0.013**(1.696)	1.252	1.691	0.032(0.438)	0.350	0.438	0.430(1.299)	1.176	1.415
	$\widehat{U}_{DR}(\pi)$	0.056(0.576)	0.476	0.581	0.031(1.531)	1.079	1.512	0.021(0.398)	0.312	0.393	0.364(1.118)	0.974	1.197
	$\widehat{U}_{CB}(\pi)$	0.058(0.938)	0.755	0.942	0.093(3.363)	2.739	3.348	0.164(0.596)	0.499	0.620	0.256(2.681)	2.153	2.709
	$\widehat{U}_{FCB}(\pi)$	**0.008**(0.492)	**0.404**	**0.494**	0.128(1.250)	**0.904**	**1.295**	**0.014**(0.345)	**0.285**	**0.357**	0.213(0.935)	**0.775**	**0.972**
π_{uni}	$\widehat{U}_{R\text{-}IPS}(\pi)$	0.566(1.705)	1.439	1.796	2.033(7.426)	5.671	7.697	0.113(1.374)	1.120	1.378	0.977(5.715)	4.545	5.791
	$\widehat{U}_{DM}(\pi)$	0.571(0.522)	0.669	0.769	0.956(1.334)	1.484	1.680	0.558(0.372)	0.581	0.669	0.990(0.952)	1.207	1.407
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.267(0.879)	0.708	0.919	0.897(4.112)	3.461	4.217	0.113(0.604)	0.481	0.614	0.804(2.114)	1.805	2.263
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.487(1.652)	1.405	1.724	2.046(7.581)	5.777	7.854	0.120(1.422)	1.176	1.427	0.906(5.753)	4.577	5.827
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	**0.137**(0.730)	0.582	0.745	**0.053**(1.945)	1.448	1.920	0.013(0.433)	0.359	0.438	0.144(1.090)	0.763	1.054
	$\widehat{U}_{DR}(\pi)$	0.136(0.688)	0.550	0.700	0.006(1.759)	1.311	1.721	0.011(0.398)	0.325	0.402	0.114(0.919)	0.696	**0.956**
	$\widehat{U}_{CB}(\pi)$	0.255(0.878)	0.703	0.918	0.729(4.206)	3.517	4.264	0.121(0.603)	0.472	0.607	0.794(2.062)	1.760	2.206
	$\widehat{U}_{FCB}(\pi)$	0.148(0.633)	**0.528**	**0.652**	0.112(1.425)	**1.115**	**1.460**	**0.005**(0.385)	**0.316**	**0.384**	**0.058**(0.984)	**0.677**	0.962
π_{lin}	$\widehat{U}_{R\text{-}IPS}(\pi)$	15.44(1.778)	15.44	15.54	45.55(9.645)	45.55	46.56	15.78(1.026)	15.78	15.81	46.60(4.683)	46.60	46.83
	$\widehat{U}_{DM}(\pi)$	0.564(0.541)	0.662	0.784	0.625(1.713)	1.436	1.784	0.622(0.372)	0.631	0.719	1.089(1.030)	1.262	1.507
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.580(0.916)	0.860	1.086	2.508(4.127)	3.713	4.836	0.240(0.593)	0.466	0.646	1.824(2.586)	2.413	3.169
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.253(1.427)	1.192	1.449	1.116(5.573)	4.428	5.673	0.099(0.823)	0.682	0.833	**0.017**(4.387)	3.473	4.385
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.010(0.634)	0.518	0.633	0.450(1.879)	1.568	1.931	0.010(0.444)	0.338	0.437	0.030(1.346)	1.120	1.362
	$\widehat{U}_{DR}(\pi)$	0.001(0.583)	0.485	0.591	0.398(1.732)	1.478	1.779	0.004(0.417)	0.327	0.416	0.055(1.199)	1.008	1.243
	$\widehat{U}_{CB}(\pi)$	0.190(1.000)	0.818	1.016	0.685(3.657)	3.088	3.725	0.098(0.585)	0.475	0.594	0.458(2.698)	2.175	2.721
	$\widehat{U}_{FCB}(\pi)$	**0.005**(0.575)	**0.450**	**0.565**	**0.269**(1.571)	**1.330**	**1.634**	**0.009**(0.390)	**0.326**	**0.392**	0.102(1.132)	**0.890**	**1.107**

Setting 2:$\delta = \delta_{nonlin}$													
	n/p	$n = 5000, p = 50$			$n = 5000, p = 100$			$n = 10000, p = 50$			$n = 10000, p = 100$		
π_0	Estimator	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
π_{inv}	$\widehat{U}_{R\text{-}IPS}(\pi)$	7.631(2.060)	7.631	7.905	19.92(8.035)	19.92	21.49	7.710(1.709)	7.709	7.897	21.33(6.139)	21.33	22.20
	$\widehat{U}_{DM}(\pi)$	2.819(0.620)	2.819	2.886	4.506(1.392)	4.507	4.725	2.671(0.342)	2.671	2.694	5.038(1.132)	5.038	5.169
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.205(1.152)	0.873	1.173	0.230(4.027)	3.491	4.016	0.112(0.679)	0.561	0.695	**0.055**(3.016)	2.519	2.993
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.348(2.347)	1.921	2.374	2.743(8.166)	7.036	8.619	0.086(2.003)	1.635	2.004	0.532(6.791)	5.536	6.816
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.136(0.843)	0.704	0.844	0.237(2.000)	1.557	1.998	0.201(0.442)	0.394	0.492	0.233(1.510)	1.203	1.516
	$\widehat{U}_{DR}(\pi)$	0.121(0.753)	0.644	0.767	0.143(1.845)	1.437	1.821	0.164(0.410)	0.355	0.438	0.198(1.392)	1.102	1.377
	$\widehat{U}_{CB}(\pi)$	0.123(1.074)	0.872	1.085	0.075(4.065)	3.476	4.067	**0.009**(0.728)	0.520	0.729	0.067(2.704)	2.227	2.717
	$\widehat{U}_{FCB}(\pi)$	**0.087**(0.631)	**0.545**	**0.644**	**0.047**(1.714)	**1.404**	**1.704**	0.083(0.372)	**0.298**	**0.376**	0.154(1.212)	**0.990**	**1.213**
π_{uni}	$\widehat{U}_{R\text{-}IPS}(\pi)$	0.176(1.921)	1.599	1.930	0.584(8.051)	6.622	8.069	0.014(1.483)	1.121	1.486	0.145(5.599)	4.470	5.602
	$\widehat{U}_{DM}(\pi)$	0.807(0.534)	0.807	0.970	1.100(2.084)	1.806	2.344	0.916(0.286)	0.916	0.962	0.889(1.262)	1.229	1.532
	$\widehat{U}_{E\text{-}IPS}(\pi)$	**0.001**(0.970)	0.701	0.966	0.615(4.172)	3.546	4.238	0.029(0.736)	0.592	0.741	0.169(3.097)	2.379	3.108
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.261(1.936)	1.585	1.952	0.670(8.093)	6.677	8.112	0.080(1.473)	1.143	1.478	0.247(5.545)	4.372	5.554
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.009(0.740)	0.576	0.729	0.573(2.437)	2.061	2.500	**0.003**(0.598)	0.462	0.587	**0.033**(1.414)	1.127	1.401
	$\widehat{U}_{DR}(\pi)$	0.023(0.658)	0.540	0.659	0.529(2.165)	1.859	2.232	0.008(0.547)	0.436	0.550	0.002(1.250)	1.006	1.277
	$\widehat{U}_{CB}(\pi)$	0.032(0.950)	0.687	0.951	0.762(4.217)	3.542	4.290	0.030(0.747)	0.597	0.746	0.154(3.142)	2.382	3.126
	$\widehat{U}_{FCB}(\pi)$	0.028(0.552)	**0.465**	**0.550**	**0.475**(1.811)	**1.543**	**1.911**	0.055(0.508)	**0.412**	**0.513**	0.128(1.132)	**0.923**	**1.171**
π_{lin}	$\widehat{U}_{R\text{-}IPS}(\pi)$	16.33(2.043)	16.33	16.45	48.59(7.529)	48.60	49.17	16.77(1.536)	16.77	16.84	49.30(6.444)	49.30	49.73
	$\widehat{U}_{DM}(\pi)$	0.727(0.412)	0.747	0.841	1.188(1.811)	1.715	2.173	0.830(0.424)	0.840	0.932	0.975(1.075)	1.056	1.409
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.284(0.962)	0.790	1.005	3.721(3.992)	4.482	5.447	0.071(0.639)	0.511	0.643	1.421(2.926)	2.507	3.245
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.370(1.729)	1.400	1.769	0.501(6.801)	5.522	6.818	0.035(1.376)	1.109	1.374	0.260(5.769)	4.472	5.782
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.240(0.791)	0.612	0.826	0.423(1.668)	1.369	1.723	0.038(0.440)	0.327	0.435	**0.033**(1.489)	1.125	1.470
	$\widehat{U}_{DR}(\pi)$	0.229(0.736)	0.562	0.770	0.346(1.510)	1.255	1.584	0.045(0.410)	0.311	0.410	0.021(1.447)	1.017	1.367
	$\widehat{U}_{CB}(\pi)$	**0.120**(0.848)	0.678	0.852	0.272(4.412)	3.379	4.418	**0.013**(0.680)	0.551	0.687	0.496(2.331)	1.991	2.367
	$\widehat{U}_{FCB}(\pi)$	0.199(0.660)	**0.515**	**0.688**	**0.157**(1.581)	**1.232**	**1.569**	0.040(0.383)	**0.303**	**0.387**	0.039(1.311)	**0.973**	**1.295**

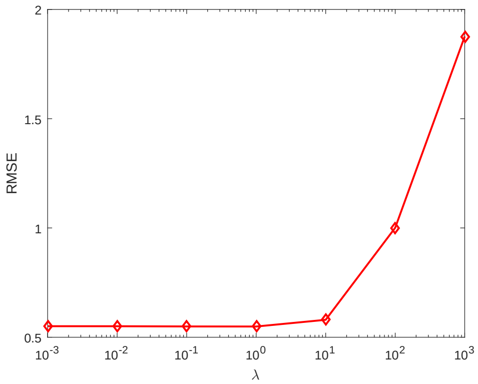


Figure 3: The effect of hyper-parameters λ on RMSE.

- DM $\widehat{U}_{DM}(\pi)$ makes huge bias because of the model misspecification. By combining DM and IPS, $\widehat{U}_{DR}(\pi)$ achieves better performance than both DM and IPS methods, even better than $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$.
- The CB estimator obtains a better performance than $\widehat{U}_{E\text{-}IPS}(\pi)$ in most of settings. This is because the incorrect estimated propensity score in $\widehat{U}_{E\text{-}IPS}(\pi)$ cannot completely correct the distribution shift induced by the past policy, while CB estimator directly correct that shift by context balancing.
- By incorporating the prior knowledge of the new policy and simultaneously correcting the distribution shift from both past and new policies, our proposed FCB estimator achieves a more precise and robust result than CB estimator and other baseline estimators under different settings.

We also demonstrate the robustness of our FCB estimator in Figure 2 by varying sample size n and dimension of contexts p under the setting $\delta = \delta_{linear}$, $\pi_0 = \pi_{inv}$. From Figure 2, we can observe that when increasing n or decreasing p, our FCB estimator always outperforms the baselines.

5.2.3 Parameter Analysis. In our FCB algorithm, we have hypeparameter λ. As mentioned before, we tuned the parameter with grid searching varying from {0.001,0.01,0.1,1,10,100,1000}. We displayed the RMSE of the estimated results with respect to λ under the setting $\delta = \delta_{linear}$, $\pi_0 = \pi_{inv}$, $n = 5000$, $p = 50$ in Figure 3. The RMSE does not change drastically and remains a low level. This means that our FCB estimator is a robust method. We can see that when λ is too large, the RMSE arises obviously since large λ would weaken the context balancing learning.

5.3 Experiments on Real World Data

It is challenging to evaluate the offline policy evaluation methods due to the lack of real benchmark datasets with ground truth of online policy performance. Fortunately, classifier evaluation problem can be regarded as a policy evaluation problem as demonstrated in [8]. Therefore, we also apply our proposed estimator on several public classification benchmark datasets to demonstrate the effectiveness of our estimator.

5.3.1 Dataset. In multiclass classification problem, features $\mathbf{X} \in \mathcal{X}$ and labels $Y^t \in \{1, 2, ..., K\}$ are assigned to each sample. A classifier can be defined as a function that receives features $\mathbf{X}$ and returns a probability distribution over the label space. The performance of the classifier can be simply evaluated by the classification accuracy.

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\pi(Y_i^t|X_i).$$

This is equivalent to policy evaluation problem where we define context $\mathbf{X}$ as features, action Y as the predicted label and the feedback function $\delta(\mathbf{X}, Y) = I(Y = Y^t)$. Hence, the goal is to predict the classification accuracy of a new classifier (i.e. new policy) π using historical data generated by a logging classifier (i.e. past policy) π_0 [2].

As did in previous work [8], we choose four classification benchmark datasets among which the sample sizes vary from 200 to 20000, and the class numbers vary from 2 to 6. Then for each dataset, we randomly split it into two parts, one for training new policy, the other for running the past policy to generate the historical data and running the new policy to provide the ground truth of the classification accuracy. In our experiments, we set the new policy π as a logistic regression, and the past policy π_0 as:

$$\pi_0(Y = j|\mathbf{X}) = \frac{S(x_r, \overline{x}_r^j)}{\sum_{q=1}^{K} S(x_r, \overline{x}_r^q)},$$

where $S(x_1, x_2) = min(x_1, x_2)/max(x_1, x_2)$, x_r is the r^{th} variable in $\mathbf{X}$ and $\overline{x}_r^j$ is the mean value of x_r belonging to class j. The r^{th} variable is chosen so that $\overline{x}_r^j$ differs significantly between classes.

5.3.2 Result. Similar to simulations, we repeat the experiments 50 times independently and calculate *Bias*, SD, MAE and RMSE for each estimator. The results are reported in Table 2. From the results, we have the following observations and analyses:

- Rough IPS $\widehat{U}_{R\text{-}IPS}(\pi)$ fails with huge error on all datasets, since it ignores the association between contexts and actions.
- $\widehat{U}_{E\text{-}IPS}(\pi)$ achieves a smaller variance on results than $\widehat{U}_{T\text{-}IPS}(\pi)$ in most datasets, and $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$ is even better than $\widehat{U}_{E\text{-}IPS}(\pi)$. This is because the estimated PS could be more smooth than the true PS, and $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$ uses control variate to make sample weights more smooth.
- DM $\widehat{U}_{DM}(\pi)$ performs differently in the different datasets, since the feedback function form may be different in these datasets. By combining DM and IPS methods, DR method $\widehat{U}_{DR}(\pi)$ reached a better performance than DM and IPS methods, and similar with $\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$
- Our CB estimator achieves similar performance with $\widehat{U}_{E\text{-}IPS}(\pi)$. Since the context balancing on first-order moment in our CB estimator achieves a comparable contribution with the estimated inverse of propensity score for correcting the distribution shift from the past policy.
- Our FCB algorithm performs the best on all datasets compared with baseline estimators. This is because that in our FCB algorithm, we incorporate knowledge of the new policy for correcting the distribution shift induced by both the past and new policies.

6 CONCLUSION

In this paper, we investigate the problem on how to better estimate the effect of a new policy based on historical data logged from a

Table 2: Results(i.e. *Bias*, SD, MAE and RMSE, the unit is percent) on classifier evaluation experiments in different datasets

Estimator	Dataset:glass			Dataset:wilt			Dataset:pageblock			Dataset:particle		
	Bias(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
$\widehat{U}_{R\text{-}IPS}(\pi)$	0.711(7.805)	5.961	7.837	0.750(1.090)	1.112	1.323	42.19(2.711)	42.19	42.28	4.093(0.432)	4.093	4.116
$\widehat{U}_{DM}(\pi)$	8.810(4.164)	8.810	9.744	0.096(0.380)	0.309	0.391	2.224(0.444)	2.224	2.267	0.741(0.228)	0.741	0.776
$\widehat{U}_{E\text{-}IPS}(\pi)$	1.648(5.707)	4.739	5.940	0.128(0.323)	0.267	0.347	4.723(3.991)	5.788	6.184	0.230(0.281)	0.287	0.362
$\widehat{U}_{T\text{-}IPS}(\pi)$	1.488(6.162)	4.866	6.339	0.175(1.205)	0.983	1.217	**0.324**(2.327)	1.794	2.348	**0.012**(0.553)	0.447	0.554
$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.315(5.455)	4.447	5.465	0.121(0.322)	0.265	0.343	1.539(2.326)	2.247	2.788	0.091(0.277)	0.222	0.293
$\widehat{U}_{CB}(\pi)$	**0.094**(6.364)	5.028	6.365	0.165(0.337)	0.318	0.372	4.660(2.810)	5.014	5.442	0.277(0.325)	0.347	0.429
$\widehat{U}_{DR}(\pi)$	1.035(5.334)	4.420	5.434	0.129(0.323)	0.269	0.347	1.734(1.978)	2.152	2.630	0.124(0.276)	0.228	0.303
$\widehat{U}_{FCB}(\pi)$	0.562(5.242)	**4.098**	**5.273**	**0.024**(0.329)	**0.250**	**0.328**	0.747(0.617)	**0.791**	**0.968**	0.080(0.261)	**0.215**	**0.272**

past policy. The main challenge of offline policy evaluation is the distribution shift problem from both the past and the new policies. However, the previous work only correct the distribution shift from the pact policy while ignoring the one from the new policy. By utilizing the prior knowledge of the new policy, we propose a Focused Context Balancing (FCB) algorithm, which learns the balancing weights to directly correct the distribution shift from both the past and the new policies. Extensive experimental results on both synthetic datasets and real world datasets demonstrate that our FCB algorithm achieves more precise and robust results on offline policy evaluation than other baselines.

7 ACKNOWLEDGEMENT

This work was supported in part by National Program on Key Basic Research Project (No. 2015CB352300), National Natural Science Foundation of China (No. 61772304, No. 61521002, No. 61531006, No. U1611461), Beijing Academy of Artificial Intelligence (BAAI), the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

REFERENCES

[1] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. 2017. Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 687–696.

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

[3] Susan Athey, Guido W Imbens, and Stefan Wager. 2016. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2016).

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

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

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

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey. 2016. *Double machine learning for treatment and causal parameters*. Technical Report. cemmap working paper, Centre for Microdata Methods and Practice.

[8] Miroslav Dudik, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *International Conference on International Conference on Machine Learning*. 1097–1104.

[9] Max H Farrell. 2015. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189, 1 (2015), 1–23.

[10] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 1 (2012), 25–46.

[11] John Hammersley. 2013. *Monte carlo methods.* Springer Science &amp; Business Media.

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

[13] Ron Kohavi and Roger Longbotham. 2011. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 31–35.

[14] Augustine Kong. 1992. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep* 348 (1992).

[15] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 265–274.

[16] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. 2017. Treatment effect estimation with data-driven variable decomposition. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[17] Kun Kuang, Meng Jiang, Peng Cui, and Shiqiang Yang. 2016. Steering social media promotions with effective strategies. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 985–990.

[18] John Langford and Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*. 817–824.

[19] Randall Lewis and David Reiley. 2009. Retail advertising works! measuring the effects of advertising on sales via a controlled experiment on yahoo! (2009).

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

[21] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 297–306.

[22] Art B Owen. 2013. Monte Carlo theory, methods and examples. *Monte Carlo Theory, Methods and Examples. Art Owen* (2013).

[23] Michael JD Powell and J Swann. 1966. Weighted uniform samplingï£¡ï£¡a Monte Carlo technique for reducing variance. *IMA Journal of Applied Mathematics* 2, 3 (1966), 228–236.

[24] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[25] Reuven Y Rubinstein and Dirk P Kroese. 2016. *Simulation and the Monte Carlo method.* Vol. 10. John Wiley &amp; Sons.

[26] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

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

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

[29] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*. 3231–3239.

[30] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology* 63, 8 (2010), 826–833.

[31] José R Zubizarreta. 2015. Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* 110, 511 (2015), 910–922.