Balanced Linear Contextual Bandits

Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, Guido Imbens

NETFLIX
Stanford

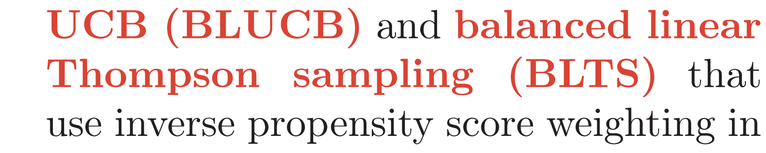
Introduction

Contextual bandit algorithms are sensitive to the estimation method of the outcome model as well as the exploration method used, particularly in the presence of rich heterogeneity or complex outcome models, which can lead to difficult estimation problems during learning. We develop algorithms for linear contextual bandits that integrate balancing methods from the causal inference literature in their estimation to make it less prone to problems of estimation bias. Our algorithms match the state of the art regret bound guarantees and have a strong advantage in practice.

Challenge

In contextual bandits, there is **inherent bias in estimation due to the adaptive assignment** of contexts to arms.

- Context assigned to arm with highest reward sample or confidence bound creates **systematically unbalanced data**.
- Complete randomization gives unbiased estimates, but this defeats the purpose.
- Aggravating sources of bias in practice: **model mis-specification**, **covariate shift** and **small samples at the initial stages** of learning.



Idea

- **We suggest the integration of balancing methods from the causal inference literature in online contextual bandits in order to make their estimation less prone to bias issues from the adaptive data collection.**
- We focus on linear online contextual bandits with provable guarantees, LinUCB (Li et al. 2010) and LinTS (Agrawal and Goyal 2013) and propose two new algorithms, **balanced linear UCB (BLUCB)** and **balanced linear Thompson sampling (BLTS)** that use inverse propensity score weighting in the training of the arms' reward models.
- **Why it works:** The propensity score in a contextual bandit is known and controlled by the policy, hence reweighting addresses model mis-specification thanks to **doubly-robustness.**

Regret Bounds

Theorem. *Assume that there exist parameters* $\{\theta_a\}_{a\in\mathcal{A}}$ *such that given any context* x, $\mathbb{E}[r_t(a)|x] = x^\top\theta_a, \forall a \in \mathcal{A}$, *that the noise* $r_t(a) - x_t^\top\theta_a$ *is conditionally sub-Gaussian and that the contexts* x_t *and parameters* θ_a *are bounded. If BLTS is run with* $\alpha = \sqrt{\frac{\log\frac{1}{\delta}}{\epsilon}}$, $Regret(T) = \tilde{O}\left(d\sqrt{\frac{KT^{1+\epsilon}}{\epsilon}}\right)$ *w.p.* $\geq 1-\delta$. *If BLUCB is run with* $\alpha = \sqrt{\log\frac{TK}{\delta}}$, $Regret(T) = \tilde{O}\left(\sqrt{TdK}\right)$ *w.p.* $\geq 1-\delta$.

Algorithms

Algorithm 1 Balanced Linear Thompson Sampling

Input: Regularization parameter $\lambda > 0$, propensity score threshold $\gamma \in (0,1)$, constant α (deafult is 1)
Set $\hat{\theta}_a \leftarrow$ **null**, $B_a \leftarrow$ **null**, $\forall a \in \mathcal{A}$
Set $X_a \leftarrow$ empty matrix, $r_a \leftarrow$ empty vector $\forall a \in \mathcal{A}$
for $t = 1, 2, \ldots, T$ **do**
  if $\exists a \in \mathcal{A}$ s.t. $\hat{\theta}_a =$ **null** or $B_a =$ **null then**
    Select $a \sim$ Uniform$(\mathcal{A})$
  else
    Draw $\tilde{\theta}_a$ from $\mathcal{N}\left(\hat{\theta}_a, \alpha^2\mathbb{V}(\hat{\theta}_a)\right)$ for all $a \in \mathcal{A}$
    Select $a = \arg\max_{a\in\mathcal{A}} x_t^\top\tilde{\theta}_a$
  end if
  Observe reward $r_t(a)$.
  Set $W_a \leftarrow$ empty matrix
  for $\tau = 1, \ldots, t$ **do**
    Compute $p_a(x_\tau)$ and set $w = \frac{1}{\max(\gamma, p_a(x_\tau))}$
    $W_a \leftarrow \text{diag}(W_a, w)$
  end for
  $X_a \leftarrow [X_a : x_t^\top]$
  $B_a \leftarrow X_a^\top W_a X_a + \lambda\mathbf{I}$
  $r_a \leftarrow [r_a : r_t(a)]$
  $\hat{\theta}_a \leftarrow B_a^{-1}X_a^\top W_a r_a$
  $\mathbb{V}(\hat{\theta}_a) \leftarrow B_a^{-1}\left((r_a - X_a^\top\hat{\theta}_a)^\top W_a(r_a - X_a^\top\hat{\theta}_a)\right)$
end for

Algorithm 2 Balanced Linear UCB

Input: Regularization parameter $\lambda > 0$, propensity score threshold $\gamma \in (0,1)$, constant α.
Set $\hat{\theta}_a \leftarrow$ **null**, $B_a \leftarrow$ **null**, $\forall a \in \mathcal{A}$
Set $X_a \leftarrow$ empty matrix, $r_a \leftarrow$ empty vector $\forall a \in \mathcal{A}$
for $t = 1, 2, \ldots, T$ **do**
  if $\exists a \in \mathcal{A}$ s.t. $\hat{\theta}_a =$ **null** or $B_a =$ **null then**
    Select $a \sim$ Uniform$(\mathcal{A})$
  else
    Select $a = \arg\max_{a\in\mathcal{A}}\left(x_t^\top\hat{\theta}_a + \alpha\sqrt{x_t^\top\mathbb{V}(\hat{\theta}_a)x_t}\right)$
  end if
  Observe reward $r_t(a)$.
  Set $W_a \leftarrow$ empty matrix
  for $\tau = 1, \ldots, t$ **do**
    Estimate $\hat{p}_a(x_\tau)$ and set $w = \frac{1}{\max(\gamma, \hat{p}_a(x_\tau))}$
    $W_a \leftarrow \text{diag}(W_a, w)$
  end for
  $X_a \leftarrow [X_a : x_t^\top]$
  $B_a \leftarrow X_a^\top W_a X_a + \lambda\mathbf{I}$
  $r_a \leftarrow [r_a : r_t(a)]$
  $\hat{\theta}_a \leftarrow B_a^{-1}X_a^\top W_a r_a$
  $\mathbb{V}(\hat{\theta}_a) \leftarrow B_a^{-1}\left((r_a - X_a^\top\hat{\theta}_a)^\top W_a(r_a - X_a^\top\hat{\theta}_a)\right)$
end for

Simulating Bias in Training Data

We simulate bias in the training data by under- and over-representation of certain regions of the context space and investigate how BLTS and BLUCB compare to LinTS and LinUCB when the outcome model is well- and mis-specified. Balancing combined with stochastic assignment rule helps escape biases much faster and can be more robust in the case of model mis-specification.

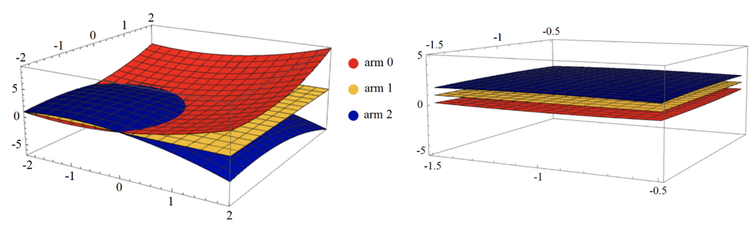
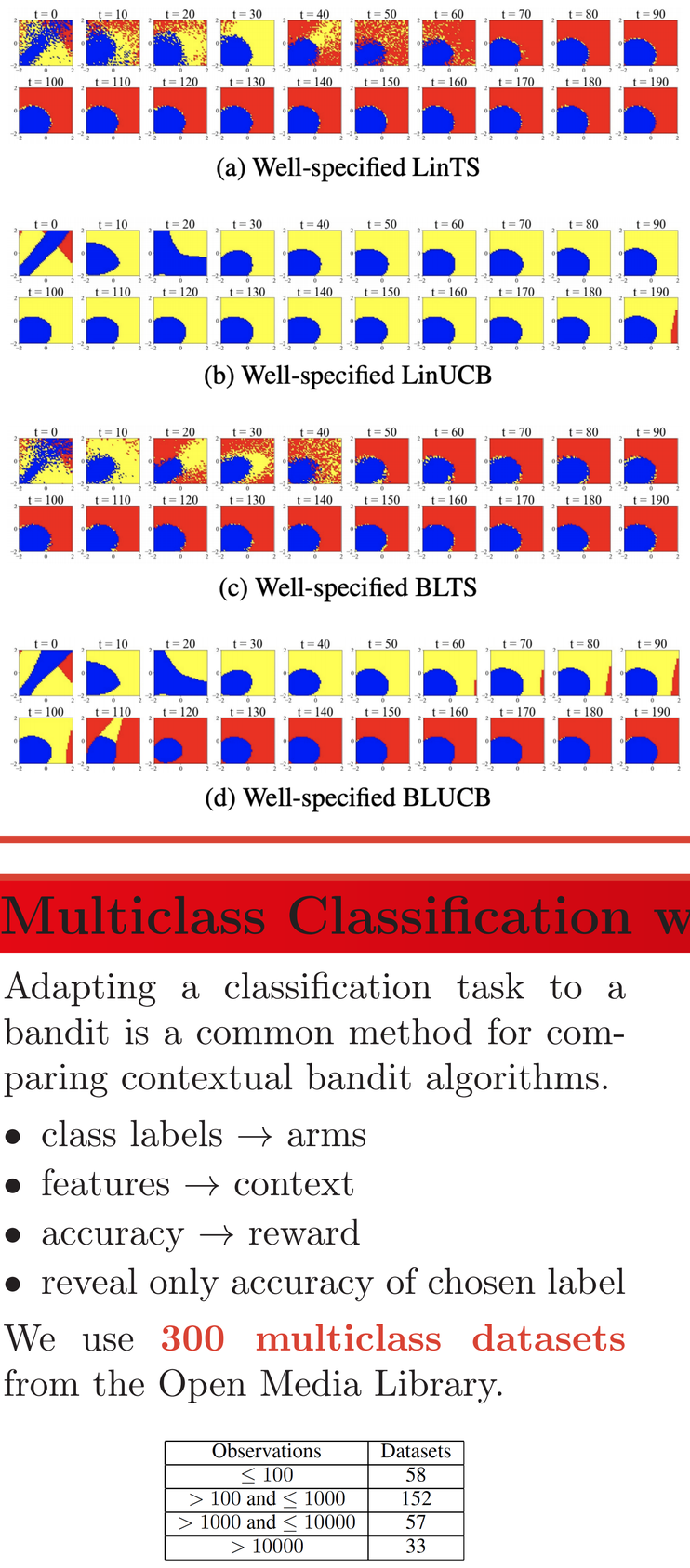


Figure 1: Expectation of each arm's reward. There is covariate shift in the initial data.

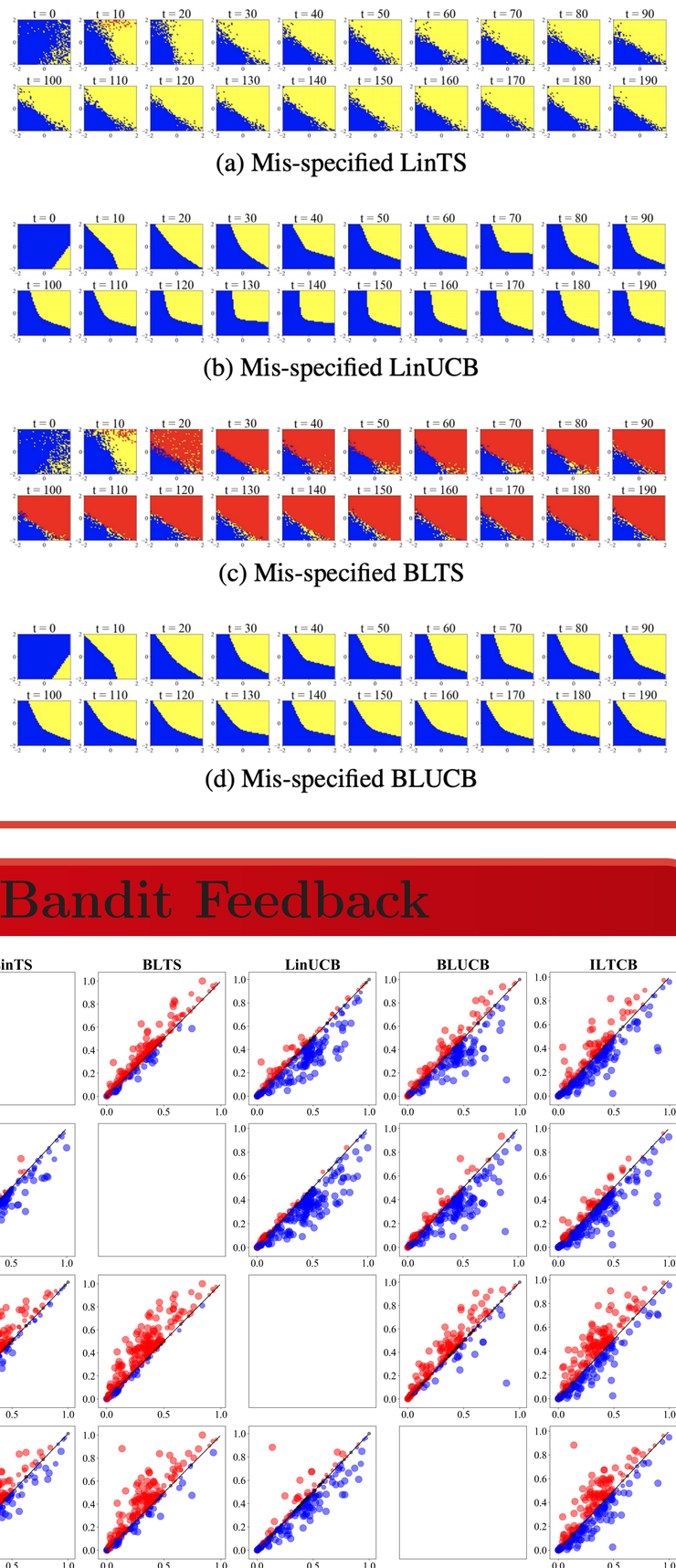


(a) Well-specified LinTS

(b) Well-specified LinUCB

(c) Well-specified BLTS

(d) Well-specified BLUCB



(a) Mis-specified LinTS

(b) Mis-specified LinUCB

(c) Mis-specified BLTS

(d) Mis-specified BLUCB

Multiclass Classification with Bandit Feedback

Adapting a classification task to a bandit is a common method for comparing contextual bandit algorithms.

- class labels → arms
- features → context
- accuracy → reward
- reveal only accuracy of chosen label

We use **300 multiclass datasets** from the Open Media Library.

Observations	Datasets
≤ 100	58
> 100 and ≤ 1000	152
> 1000 and ≤ 10000	57
> 10000	33

Classes	Count
2	243
> 2 and 10	48
> 10	9

Features	Count
≤ 10	154
> 10 and ≤ 100	106
> 100	40

BLUCB outperforms LinUCB. BLTS outperforms LinTS, LinUCB, BLUCB, ILTCB.

Figure 2: Comparing LinTS, BLTS, LinUCB, BLUCB, ILTCB on 300 datasets.