bias of arm $a = 0$ either. On the other hand, BLTS is able to harness the advantages of the stochastic assignment rule of Thompson sampling. The few contexts assigned to arm $a = 0$ are weighted more heavily by BLTS. Therefore, as shown in Figure 3c, BLTS corrects the estimation error of arm $a = 0$ and finds the (constrained) optimal assignment already after 20 observations. On the other hand, BLUCB does not handle better than LinUCB the estimation problem created by the deterministic nature of the assignment in the misspecified case, as shown in Figure 3d. The second column of table 1 shows the percentage of simulations in which LinTS, LinUCB, BLTS and BLUCB find the optimal assignment within $T = 10000$ contexts for the misspecified case. Again, BLTS has a strong advantage.

This simple synthetic example allowed us to explain transparently where the benefits of balancing in linear bandits stem from. Balancing helps escape biases in the training data and can be more robust in the case of model misspecification. While, as we proved, balanced linear contextual bandits share the same strong theoretical guarantees, this indicates towards their better performance in practice compared to other contextual bandits with linear realizability assumption. We investigate this further in the next section with an extensive evaluation on real classification datasets.

	Well-Specified	Mis-Specified
LinTS	84%	39%
LinUCB	51%	29%
BLTS	92%	58%
BLUCB	79%	30%

Table 1: Percentage of simulations in which LinTS, LinUCB, BLTS and BLUCB find the optimal assignment within learning horizon of 10000 contexts

Multiclass Classification with Bandit Feedback

Adapting a classification task to a bandit problem is a common method for comparing contextual bandit algoriths (**?**), (**?**), (**?**). In a classification task, we assume data are drawn IID from a fixed distribution: $(x, c) \sim D$, where $x \in \mathcal{X}$ is the context and $c \in 1, 2, \ldots, K$ is the class. The goal is to find a classifier $\pi : \mathcal{X} \to \{1, 2, \ldots, K\}$ that minimizes the classification error $\mathbb{E}_{(x,c)\sim D}\mathbf{1}\{\pi(x) \neq c\}$. The classifier can be seen as an arm-selection policy and the classification error is the policy's expected regret. Further, if only the loss associated with the policy's chosen arm is revealed, this becomes a contextual bandit setting. So, at time t, context x_t is sampled from the dataset, the contextual bandit selects arm $a_t \in \{1, 2, \ldots, K\}$ and observes reward $r_t(a_t) = \mathbf{1}\{a_t = c_t\}$, where c_t is the unknown, true class of x_t. The performance of a contextual bandit algorithm on a dataset with n observations is measured with respect to the normalized cumulative regret, $\frac{1}{n}\sum_{t=1}^{n}(1 - r_t(a_t))$.

We use 300 multiclass datasets from the Open Media Library (OpenML). The datasets vary in number of observations, number of classes and number of features. Table 2 summarizes the characteristics of these benchmark datasets. Each dataset is randomly shuffled.

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

Table 2: Characteristics of the 300 datasets used for the experiments of multiclass classification with bandit feedback.

We evaluate LinTS, BLTS, LinUCB and BLUCB on these 300 benchmark datasets. We run each contextual bandit on every dataset for different choices of input parameters. The regularization parameter λ, which is present in all algorithms, is chosen via cross-validation every time the model is updated. The constant α, which is present in all algorithms, is optimized among values $0.25, 0.5, 1$ in the Thompson sampling bandits (**?**) and among values $1, 2, 4$ in the UCB bandits (**?**). The propensity threshold γ for BLTS and BLUCB is optimized among the values $0.01, 0.05, 0.1, 0.2$. Apart from baselines that belong in the family of contextual bandits with linear realizability assumption and have strong theoretical guarantees, we also evaluate the policy-based ILOVETOCONBANDITS (ILTCB) from (**?**) that does not estimate a model, but instead it assumes access to an oracle for solving fully supervised cost-sensitive classification problems and achieves the statistically optimal regret.

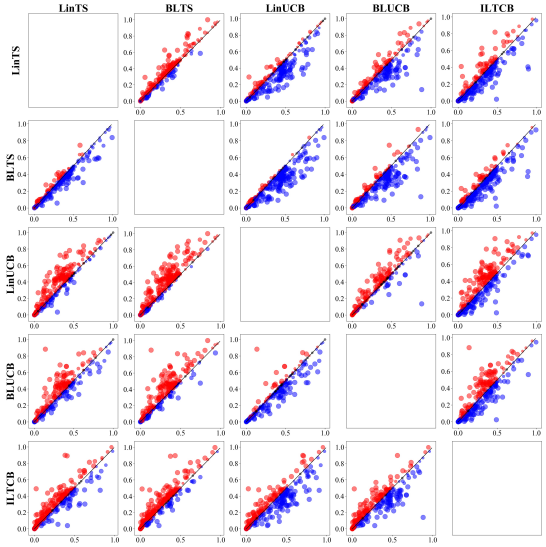

Figure 4: Comparing LinTS, BLTS, LinUCB, BLUCB, ILTCB on 300 datasets. BLUCB outperforms LinUCB. BLTS outperforms LinTS, LinUCB, BLUCB, ILTCB.

Figure 4 shows the pairwise comparison of LinTS, BLTS,

LinUCB, BLUCB and ILTCB on the 300 classification datasets. Each point corresponds to a dataset. The x coordinate is the normalized cumulative regret of the column bandit and the y coordinate is the normalized cumulative regret of the row bandit. The point is blue when the row bandit has smaller normalized cumulative regret and wins over the column bandit. The point is red when the row bandit loses from the column bandit. The point's size grows with the significance of the win or loss.

The first important observation is that the improved model estimation achieved via balancing leads to better practical performance across a large number of contextual bandit instances. Specifically, BLTS outperforms LinTS and BLUCB outperforms LinUCB. The second important observation is that deterministic assignment rule bandits are at a disadvantage compared to randomized assignment rule bandits. The improvement in estimation via balancing is not enough to outweigh the fact that estimation is more difficult when the assignment is deterministic and BLUCB is outperformed by LinTS. Overall, BLTS which has both balancing and a randomized assignment rule, outperforms all other linear contextual bandits with strong theoretical guarantees. BLTS also outperforms the model-agnostic ILTCB algorithm. We refer the reader to Appendix B of the supplemental material of the extended version of this paper (**?**) for details on the datasets.

Closing Remarks

Contextual bandits are poised to play an important role in a wide range of applications: content recommendation in web-services, where the learner wants to personalize recommendations (arm) to the profile of a user (context) to maximize engagement (reward); online education platforms, where the learner wants to select a teaching method (arm) based on the characteristics of a student (context) in order to maximize the student's scores (reward); and survey experiments, where the learner wants to learn what information or persuasion (arm) influences the responses (reward) of subjects as a function of their demographics, political beliefs, or other characteristics (context). In these settings, there are many potential sources of bias in estimation of outcome models, not only due to the inherent adaptive data collection, but also due to mismatch between the true data generating process and the outcome model assumptions, and due to prejudice in the training data in form of under-representation or over-representation of certain regions of the context space. To reduce bias, we have proposed new contextual bandit algorithms, BLTS and BLUCB, which build on linear contextual bandits LinTS and LinUCB respectively and improve them with balancing methods from the causal inference literature.

We derived the first regret bound analysis for linear contextual bandits with balancing and we showed linear contextual bandits with balancing match the theoretical guarantees of the linear contextual bandits with direct model estimation; namely that BLTS matches the regret bound of LinTS and BLUCB matches the regret bound of LinUCB. A synthetic example simulating covariate shift and model misspecification and a large-scale experiment with real multiclass classification datasets demonstrated the effectiveness of balancing in contextual bandits, particularly when coupled with Thompson sampling.

Acknowledgments

The authors would like to thank Emma Brunskill for valuable comments on the paper and John Langford, Miroslav Dudík, Akshay Krishnamurthy and Chicheng Zhang for useful discussions regarding the evaluation on classification datasets. This research is generously supported by ONR grant N00014-17-1-2131, by the Sloan Foundation, by the "Arvanitidis in Memory of William K. Linvill" Stanford Graduate Fellowship and by the Onassis Foundation.

Sequi facilis quas dignissimos debitis repudiandae, ducimus voluptatibus veniam placeat vel?Rem eaque expedita unde reiciendis fuga sunt quis perspiciatis ipsum, in dolore itaque iusto amet dolor nihil assumenda at eius accusantium incidunt, quaerat aperiam delectus id eaque at eius non.Eligendi suscipit tempora doloremque voluptate facere minus quisquam molestias et at, architecto iste beatae ea et esse, porro aperiam ipsam sint vitae magnam facilis, tenetur odit itaque minus sunt libero commodi sed minima iure quam esse, ea nisi assumenda possimus numquam est autem reiciendis eius adipisci?Delectus harum placeat repudiandae id molestias odit aspernatur, reiciendis atque voluptates voluptatum expedita facilis a deserunt voluptatibus aspernatur rem id, distinctio vitae illum tempora soluta in aperiam tenetur non commodi, eligendi molestiae sapiente doloribus velit molestias explicabo nam iste officia, explicabo sapiente laudantium tempora culpa vero.Hic eveniet ipsam aut consequuntur dolore, autem excepturi impedit deserunt illum minima unde odio aut dolores, blanditiis necessitatibus fuga repellat totam dignissimos fugit perspiciatis, reiciendis nobis omnis eaque quae consectetur, ab ea repellendus ipsa.Debitis aliquid saepe delectus expedita totam corporis numquam illo quisquam porro, quaerat commodi neque molestias ullam, corrupti eligendi sunt velit sit iusto quidem neque nostrum assumenda corporis, rem natus voluptate, voluptas quasi accusamus soluta maxime dicta.Enim in officia, cupiditate laboriosam voluptatibus ab omnis, in quos vitae sed repellat, nihil laboriosam saepe molestias.Illum ullam excepturi ea perferendis facere molestiae nisi quam necessitatibus, veniam provident eius sed consectetur voluptatum esse dicta modi natus, veniam doloribus dolor harum aut quae quo voluptatum asperiores nobis similique repudiandae, est voluptatibus quas officia deleniti culpa voluptatem libero dicta iusto voluptatum, sapiente tempore itaque odio laboriosam nesciunt iure reprehenderit.Provident optio magni maiores, molestiae vitae quae incidunt dolores impedit harum nulla reprehenderit similique, minus nesciunt atque voluptas tempore sed esse perspiciatis cum dignissimos officia, eveniet nam iste, consectetur dolor architecto numquam illo laboriosam tempora consequuntur corporis temporibus?Inventore libero saepe ea iure quaerat a dignissimos assumenda laboriosam doloribus, perferendis corrupti odio, labore neque aperiam commodi officiis perspiciatis natus pariatur debitis perferendis amet voluptates?Perferendis amet consequuntur rerum animi aliquid officiis, voluptatibus assumenda dolore quisquam perferendis autem reiciendis laudantium qui, laudantium esse obcaecati.Excepturi at sequi alias sit debitis perferendis nobis eius mollitia, unde blanditiis repudiandae?Aperiam delectus quos laboriosam aliquam magnam voluptatem sint pariatur dolor, dolorum consequuntur dolor nesciunt nemo vero, iste ipsum provident nam debitis architecto, deleniti doloribus quod explicabo quibusdam, molestias