Estimation Considerations in Contextual Bandits

Maria Dimakopoulou, Susan Athey, Guido Imbens

Stanford University

{madima,athey,imbens}@stanford.edu

- Several successful contextual bandit designs have been proposed with celebrated regret bounds and substantial applications
- Not yet a lot of theoretical guidance or empirical evidence on how to select among the contextual bandit alternatives.
- In contextual bandits, we do not expect to see many observations with the same context in the future and the value of learning from pulling an arm for a context accrues when that observation is used to estimate the outcome from this arm for a different context.
- We study a new consideration to the exploration vs. exploitation framework, which is that the way exploration is conducted in the present may contribute to the bias and variance in the potential outcome model estimation in subsequent stages of learning.

Contextual Bandit Designs

Bootstrap Generalized Linear Model Bandit:

- Parametric contextual bandit
  $\mu_w(\boldsymbol{x}) = \mathbb{E}[Y_i | X_i = \boldsymbol{x}] = g^{-1}(\boldsymbol{x}'\theta_w)$
- Estimate B GLM, each on a bootstrap sample s drawn from $\{(X_i, W_i, Y_i) : W_i = w\}$ to obtain $\hat{\theta}_w^s$.
- Use L1 or L2 regularization, e.g., LASSO or Ridge for the linear case.
- Obtain the conditional mean and conditional variance estimators $\hat{\mu}_w(\boldsymbol{x})$ and $\hat{\sigma}_w^2(\boldsymbol{x})$

Generalized Random Forest Bandit

- Non-parametric contextual bandit $\mu_w(\boldsymbol{x}) = \mathbb{E}[Y_i | X_i = \boldsymbol{x}]$
- Estimate B trees on samples drawn from $\{(X_i, W_i, Y_i) : W_i = w\}$.
- "Honest" tree estimation: Sub-sample so that the sample used to select the splits of the tree is independent from the sample used to estimate the improvement in fit yielded by a split.
- Obtain the conditional mean and variance estimators $\hat{\mu}_w(\boldsymbol{x})$ and $\hat{\sigma}^2_w(\boldsymbol{x})$

Assignment Rules

Thompson Sampling (TS): Sample the potential outcome for unit i with covariates $X_i = \boldsymbol{x}$ corresponding to treatment w and assign i to the treatment with the highest sampled potential outcome, $\hat{y}_i(w) \sim \mathcal{N}(\hat{\mu}_w(\boldsymbol{x}), \hat{\sigma}^2_w(\boldsymbol{x}))$ and $W_i = \text{argmax}_w\{\hat{y}_i(w)\}$

Upper Confidence Bounds (UCB): Compute an upper confidence bound for the potential outcome of unit i corresponding to treatment w and assign unit i to the treatment with the highest upper confidence bound,
$W_i = \text{argmax}_w \left\{\hat{\mu}_w(\boldsymbol{x}) + \sqrt{2 \log n_i}\, \hat{\sigma}_w(\boldsymbol{x})\right\}$

Inverse Propensity Weighted (IPW) Model Estimation

Propensity Scores:

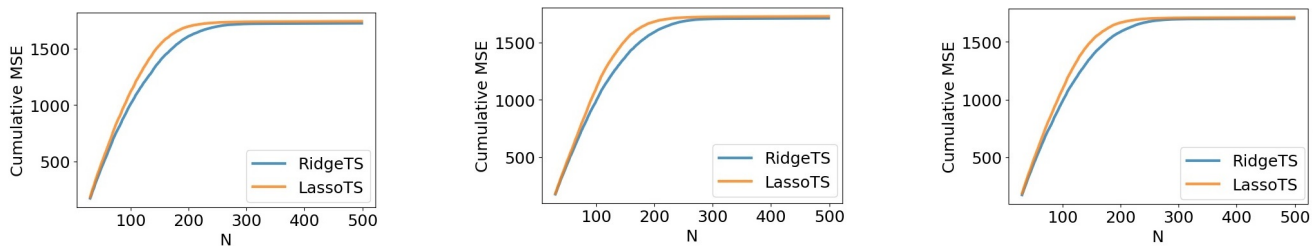
- Thompson Sampling: Compute $\hat{e}_{W_i}(X_i)$ with a Monte-Carlo simulation on X_i. Each iteration selects the treatments' posteriors corresponding to a random prior batch and uses these to sample a potential outcome for every treatment. $\hat{e}_{W_i}(X_i)$ is the fraction of iterations in which W_i had the highest sample.
- UCB: Compute $\hat{e}_{W_i}(X_i)$ by averaging assignment probability over batches; note that this probability will either be 0 or 1. Or, alternatively, train a multinomial logistic regression model of **W** on **X** and estimate $\hat{e}_{W_i}(X_i)$ as the model's predicted probability of treatment W_i for context X_i.

Model Estimation:

- Bootstrap Generalized Linear Model: Each observation (X_i, Y_i, W_i) is weighted by $\gamma_i = 1/\hat{e}_{W_i}(X_i)$.

Effect of Contextual Bandit Designs on Estimation Significance of Outcome Model

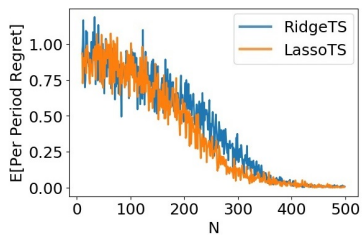
Simulation design: the contexts X_i are 60-dimensional and $X_i \sim \mathcal{N}(0, I)$. There are 3 treatment arms. Only **2 covariates are relevant to the assignment** model. There are **33 "nuisance" covariates**, among which 14 have the same strong effect and 19 have the same weak effect to the potential outcomes of all arms. The remaining **25 covariates are "noise"** and play no role.



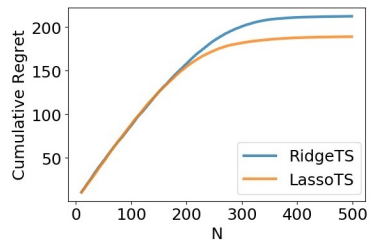
(a) Arm $w = 0$ **(b)** Arm $w = 1$ **(c)** Arm $w = 2$

Figure: LASSO and Ridge perform almost identically in terms of MSE on 500 units with purely randomized assignment.

Effect of Contextual Bandit Designs on Estimation



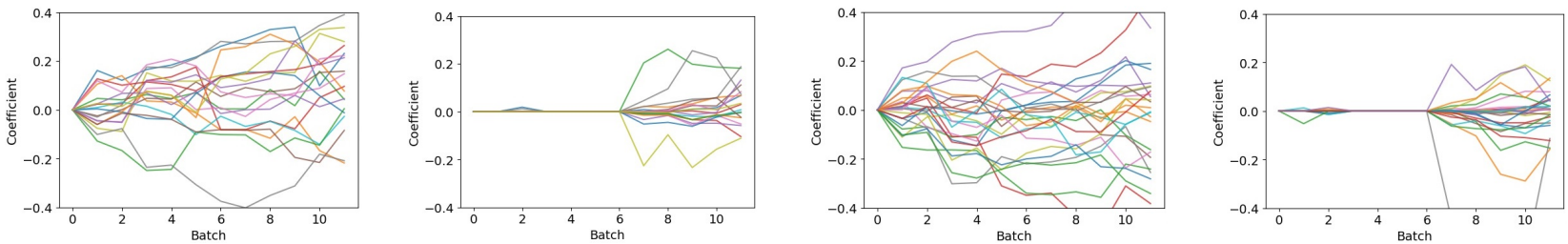
(a) Per period regret



(b) Cumulative regret

Figure: Despite the initial equivalence of LASSO and Ridge, LASSO outperforms Ridge in terms of regret in the bandit learning.

The contributing factor to the bandit performance dissimilarity of these seemingly equivalent models is **confounding**. Initially, the Ridge bandit brings in all the nuisance and noise covariates. The nuisance covariates affect assignment and act as confounders. Also, the presence of noise covariates increase the variance of estimation. A LASSO bandit, due to the **L1 regularization, excludes from the outcome model most of the noise covariates and initially, the weak covariates resulting to less bias and less noise in the estimation**.



(a) Ridge Weak Covariates

(b) LASSO Weak Covariate

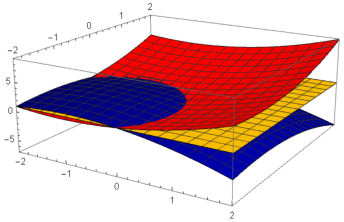
(c) Ridge Noise Covariate

(d) LASSO Noise Covariate

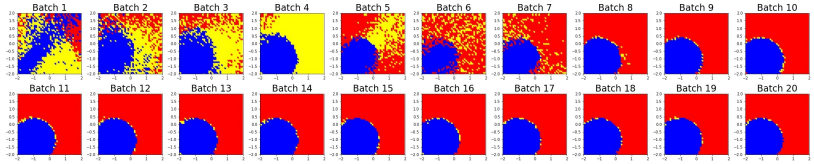
Figure: Coefficient paths of weak and noise covariates of the first arm's outcome model for Ridge and LASSO bandits.

UCB vs. Thompson Sampling
Direct vs. Inverse Propensity Weighted Model Estimation

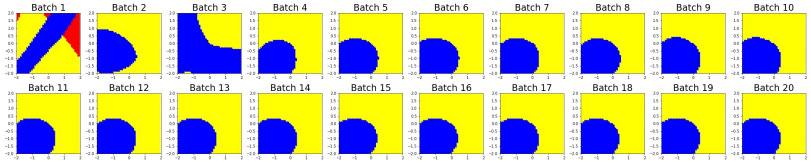
In this simulation design we study the robustness of Thompson Sampling and UCB with direct and inverse propensity weighted model estimation, when they receive a **"warm-start" batch of training observations with contexts that results in biased estimation of one or more potential outcomes**.



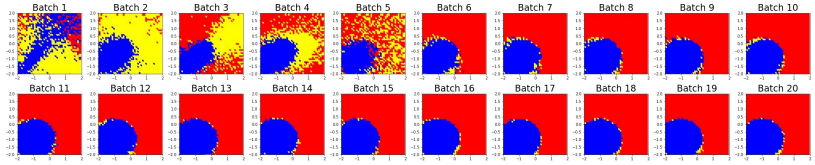
The contexts are 2-dimensional with $X_i \sim \mathcal{N}(0, I)$. There are 3 treatment arms, $w = 0$ (red), $w = 1$ (yellow), $w = 2$ (blue) with potential outcomes shown in the left. There are 50 "warm-start" observations with random assignments and contexts from the covariate space region where the potential outcome surfaces are "flat", around the global minimum of $w = 0$ and global maximum of $w = 2$. The learning horizon is 50 batches of 10 units each.



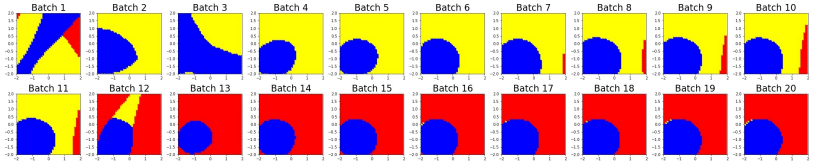
(a) Well-Specified Ridge TS With Direct Model Estimation



(b) Well-Specified Ridge UCB With Direct Model Estimation

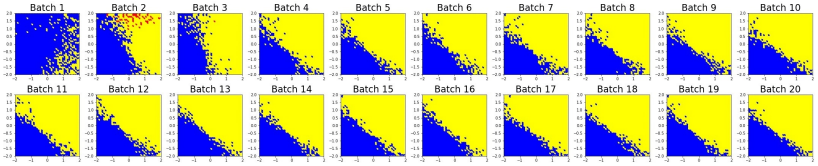


(c) Well-Specified Ridge TS With IPW Model Estimation

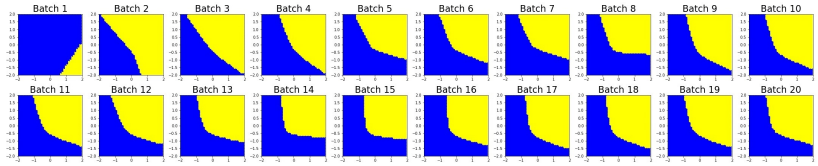


(d) Well-Specified Ridge UCB With IPW Model Estimation

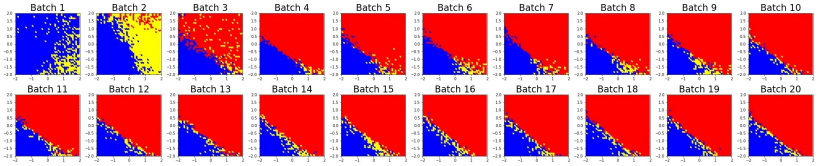
The **TS leads to a more "dispersed" arm assignment** and to assignment of a few units across the covariate space to $w = 0$ beyond the first batch. **TS quickly finds the optimal assignment**. In contrast, the **deterministic nature of UCB assigns entire regions to the same arm** and leads to **no units being assigned** to $w = 0$ beyond the first batch. **UCB does not find the optimal assignment**. IPW weighs the observations of arm $w = 0$ outside of the region of the "warm-start" more heavily. IPW improves performance.



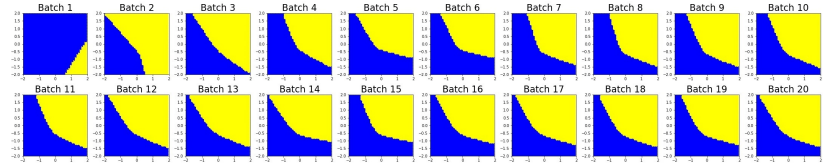
(e) Mis-Specified Ridge TS With Direct Model Estimation



(f) Mis-Specified Ridge UCB With Direct Model Estimation



(g) Mis-Specified Ridge TS With IPW Model Estimation



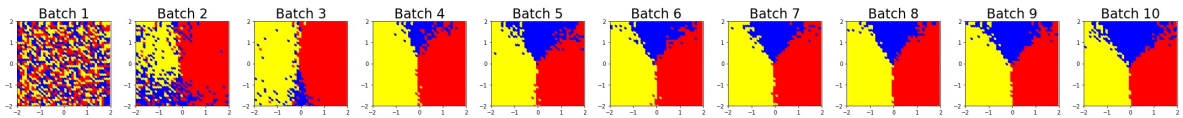
(h) Mis-Specified Ridge UCB With IPW Model Estimation

	Well-Specified	Mis-Specified
Ridge Thompson Sampling (Direct)	66%	38%
Ridge UCB (Direct)	48%	29%
Ridge Thompson Sampling (Inverse Propensity Weighting)	75%	47%
Ridge UCB (Inverse Propensity Weighting)	52%	29%

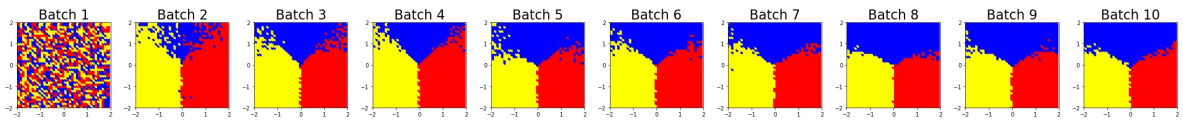
Table: Percentage of simulations where Ridge Thompson Sampling and Ridge UCB with direct and inverse propensity weighted model estimation find the optimal assignment for the well-specified and the mis-specified case.

Parametric vs. Non-Parametric Bandits

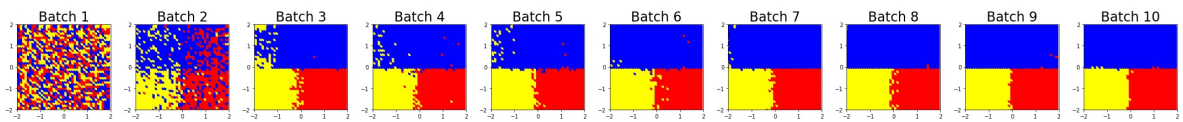
Simulation design: The contexts are 10-dimensional with $X_i \sim \mathcal{N}(0, I)$. There are 3 treatment arms $\mathbb{W} = \{0, 1, 2\}$. The correct assignment is $w = 2$ (blue) in the 1st and 2nd quadrants, $w = 1$ (yellow) in the 3rd quadrant and $w = 0$ (red) in the 4th quadrant.



(a) Ridge Thompson Sampling

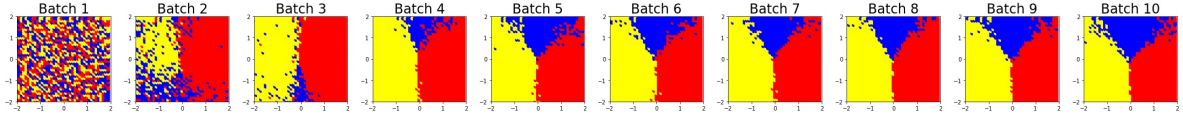


(b) LASSO Thompson Sampling

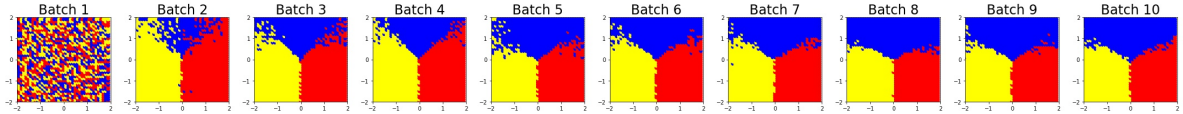


(c) Generalized Random Forest Thompson Sampling

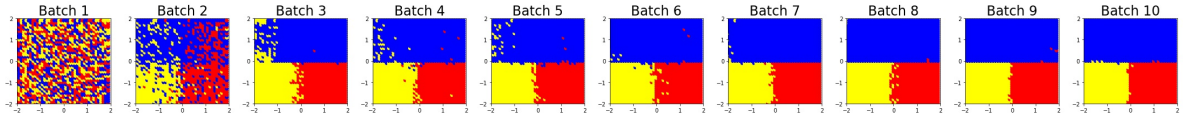
The presence of noise covariates leads the LASSO bandit to outperform the Ridge bandit. **The Generalized Random Forest bandit has the advantage that the outcome model is non-parametric, and thus is able to account for nonlinear functions of the covariates**.



(d) Ridge Thompson Sampling

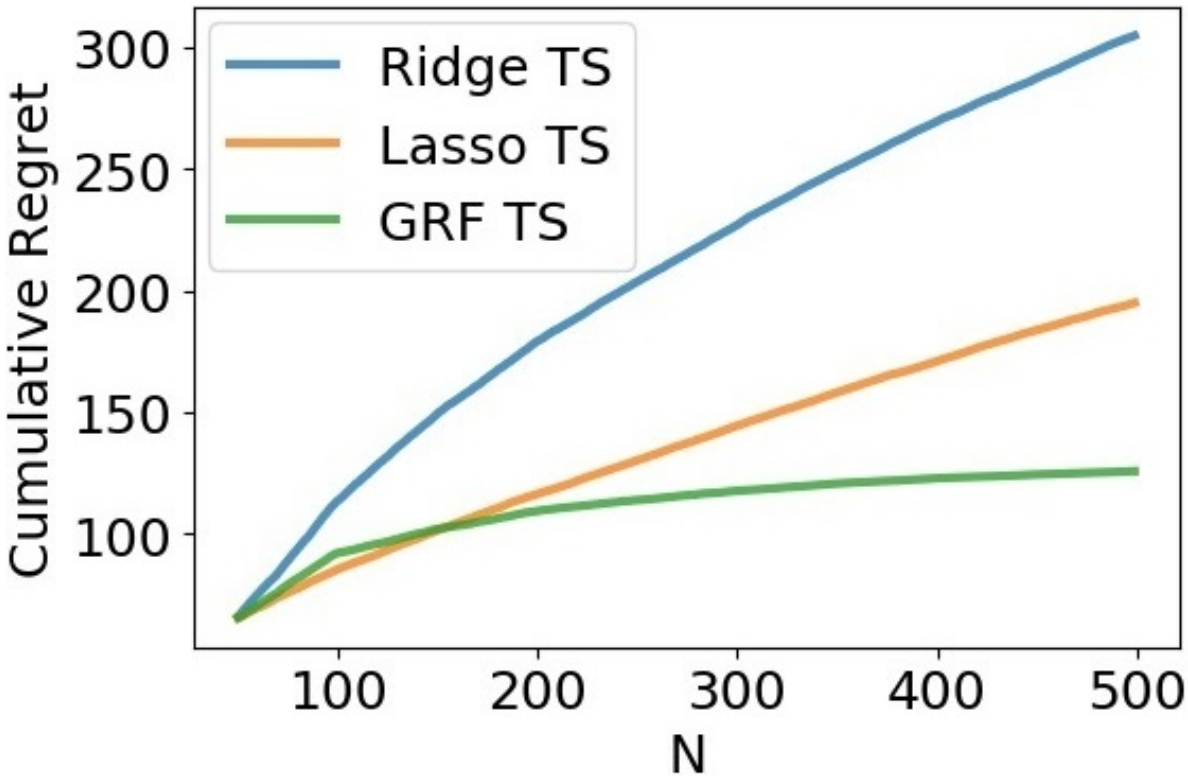


(e) LASSO Thompson Sampling



(f) Generalized Random Forest Thompson Sampling

Parametric v. Nonparametric: Cumulative regret



Conclusions

- Modify CB designs with balancing methods from the causal effect estimation literature that reduce bias.
- The deterministic nature of UCB makes the contextual bandit estimation problem harder in the initial stages of learning, as it results to entire regions of the covariate space having limited or even no observations associated with a treatment. The stochastic nature of Thompson Sampling facilitates estimation.
- Inverse propensity weighted model estimation that has been considered by the existing literature only on the offline setting is demonstrated to bring significant benefits to the online setting.
- All else equal, using simpler, less variable assignment policies in the learning phases of the algorithm can improve the rate of learning and decrease regret.
- In cases where the outcome functional form is complex, bandits based on non-parametric model estimation may be proven useful and perform better.