Sublinear Optimal Policy Value Estimation in Contextual Bandits

Weihao Kong
University of Washington

Gregory Valiant
Stanford University

Emma Brunskill
Stanford University

Abstract

We study the problem of estimating the expected reward of the optimal policy in the stochastic disjoint linear bandit setting. We prove that for certain settings it is possible to obtain an accurate estimate of the optimal policy value even with a number of samples that is sublinear in the number that would be required to *find* a policy that realizes a value close to this optima. We establish nearly matching information theoretic lower bounds, showing that our algorithm achieves near optimal estimation error. Finally, we demonstrate the effectiveness of our algorithm on joke recommendation and cancer inhibition dosage selection problems using real datasets.

1 Introduction

We consider how to efficiently estimate the best possible performance of the optimal representable decision policy in a disjoint linear contextual multi-armed bandit setting. Critically, we are interested in when it is possible to estimate this best possible performance using a sublinear number of samples, whereas a linear number of samples would typically be required to provide any such policy that can realize optimal performance.

Contextual multi-armed bandits (see e.g. Chu et al. (2011); Li et al. (2010); Agarwal et al. (2014)) are a well studied setting that is having increasing influence and potential impact in a wide range of applications, including customer recommendations (Li et al., 2010; Zhou and Brunskill, 2016), education (Lan and Baraniuk, 2016) and health (Greenewald et al., 2017). In contrast to simulated domains like games and robotics simulators, in many contextual bandit applications the best potential performance of the algorithm is unknown in advance. Such situations will often involve a human-in-the-loop approach to optimizing system performance, where a human expert specifies a set of features describing the potential contexts and a set of possible interventions/arms, and then runs a contextual bandit algorithm to try to identify a high performing decision policy for what intervention to automatically provide in which context. A key issue facing the human expert is assessing if the current set of context features and set of interventions/arms can yield sufficient performance. This can be challenging, because without prior knowledge about what optimal performance might be possible, the human may need to run the contextual bandit algorithm until it returns an optimal policy given the current set of arms and features, which may involve wasted time and effort if the best policy has mediocre performance. While there has been some limited algorithmic work on such human-in-the-loop settings for reinforcement learning (Mandel et al., 2017; Keramati and Brunskill, 2019) to our knowledge no formal analysis exists of how to efficiently estimate the average reward of the optimal policy representable with the current set of context features and arms.

The majority of prior work on multi-armed bandits has focused on online algorithms that minimize cumulative or per-step regret (see e.g. Auer et al. (2002); Agarwal et al. (2014) or Lattimore and Szepesvári (2018)). In simple multi-armed bandit settings (with no context) there has also been work on maximizing the probability of best arm identification given a fixed budget (Bubeck et al., 2009; Audibert et al., 2010; Gabillon et al., 2012; Karnin et al., 2013) or minimizing the number of samples needed to identify the best arm with high confidence (Even-Dar et al., 2006; Maron and Moore, 1994; Mnih et al., 2008; Jamieson et al., 2014). Note that in the simple multi-arm bandit setting, sample complexity bounds for ϵ-best arm identification will be equivalent to the bounds achievable for estimating the expected reward of the optimal policy as there is no sharing of rewards or information across arms.

In the case of contextual multi-armed bandits, there has been some limited work on single best arm identification when the arms are described by a high di-

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

mensional feature vector (Hoffman et al., 2014; Soare et al., 2014; Xu et al., 2018). However such work does not immediately include input context features (such as from a customer or patient), and would need to be extended to handle best policy identification over (as we consider here) a linear class of policies. A separate literature seeks to identify a good policy for future use given access to batch historical data in both bandit and reinforcement learning settings (Thomas et al., 2015; Athey and Wager, 2017; Gelada and Bellemare, 2019; Liu et al., 2019). In contrast to such work, we consider the setting where the algorithm may actively gather data, and the objective is to accurately estimate the performance of the optimal policy in the set, *without returning a policy that achieves such performance.*

In particular, in this work we consider disjoint linear contextual bandits (Li et al., 2010) (one parameter for each of a finite set of arms, such as a set of treatments) with a high dimensional, d, input context (such as a set of features describing the patient). We are interested in providing an accurate estimate of the expected performance of the best realizable decision policy. Here the decision policy class is implicitly defined by the input context feature space and finite set of arms. Following prior work on disjoint linear contextual bandits (see e.g. Li et al. (2010)) we assume that the reward for each arm can be expressed as a linear combination of the input features and an arm-specific weight vector.

Quite surprisingly, we present an algorithm that can estimate the potential expected reward of the best policy with a number of samples (pulls of the arms) that is sublinear in the input context dimension d. This is unintuitive because this is less than what is needed to estimate *any* fit of the d-dimensional arm weight vector, which would require at least d samples. Our approach builds on recent work Kong and Valiant (2018) that shows a related result in the context of regression, showing that the best accuracy of a regression algorithm can, in many situations, be estimated with sublinear sample size. A critical insight in that paper, which we leverage and build upon in our work, is the construction of a sequence of unbiased estimators for geometric properties of the data that can be used to estimate the best accuracy, without attempting to find the model achieving that accuracy. However, multiple additional technical subtleties arise when we move from the prediction setting to the control setting because we need to take the interaction between different arms into account while there is effectively only one "arm" in the prediction setting. Even assuming that we have learned the interaction between the arms, it is not immediately clear how such knowledge helps determine the potential expected reward of the best policy. We leverage a quantitative version of the Sudakov-Fernique inequality to answer the question. While in the classical (non-disjoint) stochastic linear bandit problem, it is crucial to use the information we learn from one arm to infer information for other arms, this does not hold in the non-disjoint setting. Nevertheless, we utilize the contexts across all the arms to reduce the estimation error, which yields a near optimal sample complexity dependency on the number of arms. Our approach can also leverage unsupervised prior data about the distribution of contexts, as may often be available (past patients' features or prior customers' features), in order to further improve the algorithm performance.

Our key contribution is an algorithm for accurately estimating the expected performance of the optimal policy in a disjoint contextual linear bandit setting with an amount of samples that is sublinear in the input context dimension. We provide theoretical bounds when the input context distributions are drawn from Gaussians with zero mean and known or unknown covariances. We then examine the performance empirically, first in a synthetic setting. We then evaluate our method both in identifying the optimal reward for a joke recommendation decision policy, based on the Jester dataset (Goldberg et al., 2001), and on a new task we introduce of predicting the performance of the best linear threshold policy for selecting the dosage level to optimize cancer cell growth inhibition in the NCI-60 Cancer Growth Inhibition dataset. Encouragingly, our results suggest that our algorithm quickly obtains an accurate estimate of the optimal linear policy.

2 Problem Setting

A contextual multi-armed bandit (CMAB) can be described by a set of contexts $\mathcal{X} \in R^d$, a set of K arms $\mathcal{K}$ and a reward function. We consider the linear disjoint CMAB setting (Li et al., 2010), where there are a finite set of arms, and the reward y from pulling an arm a in a context $\mathbf{x}_j$ is

$$y_{a,j} = \beta_a^T \mathbf{x}_j + b_a + \eta_{a,j}. \tag{1}$$

For each arm a, β_a is an unknown d-dimensional real vector with bounded ℓ_2 norm and b_a is a real number, $\mathbf{E}[\eta_{a,j}] = 0$ and $\mathbf{E}[\eta_{a,j}^2]$ is bounded by a constant.

For simplicity, we focus primarily on the passive setting where for each arm a, we observe N iid samples $\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \ldots, \mathbf{x}_{a,N}$ drawn from $N(0, \Sigma)$, and each sample $\mathbf{x}_{a,j}$ is associated with a reward. Under this setting, we denote σ_a^2 as the variance of $y_{a,j}$, which is smaller than $\beta_a^\top \Sigma \beta_a + \mathbf{E}[\eta_{a,j}^2]$, and it is assumed that σ_a are all bounded by a constant. We define the total number of samples $T = K \cdot N$ to draw a connection to the adaptive setting where the algorithm can adaptively choose the action to play on each context. Interestingly,

in the worst case our approach of uniformly gathering samples across all actions is optimal up to a $\log^{3/2}(dK)$ factor (see Theorem 2).

Given a total of $T = K \cdot N$ samples $(\mathbf{x}_{a,j}, y_j)$, our goal is to predict the expected reward of the optimal policy realizable with the input definition of context features and finite set of actions, which is $OPT := \mathbf{E}_{\mathbf{x}}[\max_a(\beta_a^T \mathbf{x} + b_a)]$.

3 Summary of Results

Our first result applies to the setting where each context is drawn from a d-dimensional Gaussian distribution $N(0, \Sigma)$, with a *known* covariance matrix Σ, and the reward for the ath arm on context $\mathbf{x}_j$ equals $\beta_a^T \mathbf{x} + b_a + \eta_{a,j}$ where $\mathbf{E}[\eta_{a,j}] = 0, \mathbf{E}[\eta_{a,j}^2]$ is bounded by a constant.[1] Given $N = \Theta(\epsilon^{-2}\sqrt{d}\log K \log(K/\delta))$ samples for each arm, there is an efficient algorithm that with probability $1-\delta$ estimates the optimal expected reward with additive error ϵ.

Corollary 1 (Main result, known covariance setting)**.** *In the known covariance setting, for any $\epsilon \geq \frac{\sqrt{\log K}}{d^{1/4}}$, with probability $1-\delta$, Algorithm 1 estimates the optimal reward OPT with additive error ϵ using a total number of samples*

$$T = \Theta(\frac{\sqrt{d}K \log K}{\epsilon^2} \log(K/\delta)).$$

We prove a near matching lower bound, showing that in this passive setting, the estimation error can not be improved by more than a $\log K$ factor. The proof of Theorem 1 can be found in the supplementary material.

Theorem 1 (Lower bound for passive algorithms, known covariance setting)**.** *There exists a constant C such that for any $\epsilon > 0$ given*

$$T = C\frac{\sqrt{d}K \log K}{\epsilon^2}$$

samples (equivalently $N = C\frac{\sqrt{d}\log K}{\epsilon^2}$ samples for each arm), no algorithm can estimate the optimal reward with expected additive error less than ϵ with probability greater than $2/3$.

Comparing against the adaptive setting where the algorithm can adaptively choose the action to play on each context, we prove a surprising lower bound, showing that the estimation error can not be improved much. Specifically, our passive algorithm is minimax optimal even in the *adaptive setting* up to a polylog(dK) factor. The proof is deferred to the supplementary material.

Theorem 2 (Lower bound for fully adaptive algorithms, known covariance setting)**.** *There exists a constant C such that no algorithm can estimate the optimal reward with additive error ϵ and probability of success at least $2/3$ using a number of rounds that is less than*

$$T = C\frac{\sqrt{d}K}{\epsilon^2 \log^{3/2}(dK)}.$$

Our lower bound is novel, and we are not aware of similar results in this setting. It is curious that the standard approach by simply bounding the KL-divergence only yields a sub-optimal $\tilde{O}(\sqrt{dK})$ lower bound, since the divergence contribution of each arm scales with $\mathbf{E}[T_i^2]$ instead of $\mathbf{E}[T_i]$ in the classical (non-contextual) stochastic bandit setting. We apply a special conditioning to get around this issue.

Our algorithmic techniques apply beyond the known covariance setting, and we prove an analog of Corollary 1 in the setting where the contexts $\mathbf{x}$ are drawn from a Gaussian distribution with arbitrary *unknown* covariance. Our general result, given in Corollary 7, is quite complicated. Here we highlight the special case where the desired accuracy ϵ and failure probability δ are small positive constants, and the covariance is well-conditioned:

Corollary 2 (Special case of main result, unknown covariance setting)**.** *Assuming that the covariance of the context $\mathbf{x}_i$ satisfies $\sigma_{\min} I_d \preceq \Sigma \preceq \sigma_{\max} I$ and $\sigma_{\max}/\sigma_{\min}$ is a constant, for constant ϵ, Algorithm 1 takes $\sigma_{\min}, \sigma_{\max}$, and a total number of*

$$T = O(d^{1-\frac{C}{\log\log K + \log(1/\epsilon)}} K^{\gamma} + \sqrt{d}K^{1+\gamma})$$

samples, where γ is any positive constant and C is a universal constant.

In the unknown covariance setting, the dependency on d of our algorithm is still sublinear, though is much worse than the $\sqrt{d}$ dependency in the known covariance setting. However this can not be improved by much as the lower bound result in Kong and Valiant (2018) implies that the dependency on d is at least $d^{1-\Theta(\frac{1}{\log 1/\epsilon})}$

It is worth noting that the techniques behind our result in the unknown covariance setting that achieves sublinear sample complexity essentially utilizes a set of unlabeled examples of size $O(T)$ to reduce the variance of the estimator, where unlabeled examples are the context vectors $\mathbf{x}_i$ drawn from $N(0, \Sigma)$. If one has an even

[1] The setting where the covariance, Σ, is known is equivalent to the setting where the covariance is assumed to be the identity, as the data can be re-projected so as to have identity covariance. While the assumption that the covariance is known may seem stringent, it applies to the many settings where there is a large amount of *unlabeled* data. For example, in many medical or consumer data settings, an accurate estimate of the covariance of $\mathbf{x}$ can be obtained from large existing databases.

larger set of unlabeled examples, the sample complexity for the labeled examples can be significantly reduced. For simplicity, we do not present a complete trade-off result between the labeled and unlabeled examples in this paper. Instead, we present one extreme case where there is a sufficiently large set of unlabeled examples (size $\Omega(d)$), and the problem essentially reduces to the known covariance problem.

Corollary 3 (Unknown covariance with a large set of unlabeled examples)**.** *In the unknown covariance setting, there is an algorithm that estimates the optimal reward OPT with additive error ϵ with probability $1-\delta$ using a total number of labeled examples*

$$T = \Theta(\frac{\sqrt{d}K\log K}{\epsilon^2}\log(K/\delta)),$$

and a set of unlabeled examples of size $\Theta((d + \log 1/\delta)\log^2 K/\epsilon^4)$

The algorithm that achieves the above result is straight forward. We first estimate the covariance of the context using the set of unlabeled examples up to ϵ spectral norm error, and let us denote $\hat{\Sigma}$ as the estimator. Given the covariance estimator, we will execute Algorithm 1 for the known covariance setting, and scale each context $\mathbf{x}_i$ as $\hat{\Sigma}^{-1/2}\mathbf{x}_i$. The covariance of the scaled context is not exactly the identity, hence our estimator is biased. However, it is straight forward to show that the bias is at most $O(\epsilon)$, which is on the same magnitude as the standard deviation of our estimator. The proof is deferred to the appendix.

Finally, we slightly generalize our results beyond the Gaussian context setting and show that if each context is drawn from a mixture of M Gaussian distributions which is completely known to the algorithm, then our algorithm can be applied to achieve ϵ estimation error while the sample complexity only increases by a factor of $\log M$. The proof is deferred to the appendix.

Theorem 3 (Extension to Gaussian Mixtures)**.** *Suppose each context $\mathbf{x}$ is drawn independently from a mixture of Gaussian distributions $\sum_{i=1}^{M}\alpha_i N(\mu_i,\Sigma_i)$, and the parameters μ_i,Σ_i,α_i are all known to the algorithm. In addition, let us assume that $\|\mu_i\|$, $\|\Sigma_i\|$ are all bounded by a constant. Then for any $\epsilon \geq \frac{\sqrt{\log K}}{d^{1/4}}$, with probability $1-\delta$, there is an algorithm that estimates the optimal reward OPT with additive error ϵ using a total number of samples*

$$T = \Theta(\frac{\sqrt{d}K\log K}{\epsilon^2}\log(KM/\delta)).$$

4 The Estimators

The basic idea of our estimator for the optimal reward of linear contextual bandits is as follows. For illustration, we assume that each context $\mathbf{x}$ is drawn from a standard Gaussian distribution $N(0, I_d)$. In the realizable setting where the reward for pulling arm a on context $\mathbf{x}_j$ is $\beta_a^T\mathbf{x} + b_a + \eta_{a,j}$ where β_a, b_a are the parameters associated with arm a and $\eta_{a,j}$ is random noise with mean 0, the expected reward of the optimal policy is simply $\mathbf{E}_{\mathbf{x}}[\max_a(\beta_a^T\mathbf{x} + b_a)]$. Let us define the K dimensional random variable $r = (\beta_1^T\mathbf{x} + b_1, \beta_2^T\mathbf{x} + b_2, \ldots, \beta_K^T\mathbf{x} + b_K)$. Notice that in the setting where $\mathbf{x} \sim N(0, I)$, r is a K dimensional Gaussian random variable with mean $\mathbf{b} = (b_1, \ldots, b_K)$ and covariance H where $H_{a,a'} = \beta_a^T\beta_{a'}$. Hence in this simplified setting, the optimal reward of the linear contextual bandit problem can be expressed as $\mathbf{E}_{\mathbf{r}\sim N(\mathbf{b},H)}[\max_i r_i]$ which is a function of $\mathbf{b}$ and H. Naturally, one can hope to estimate the optimal reward by first accurately estimating $\mathbf{b}$ and H. The bias $\mathbf{b}$ can be accurately estimated up to entry-wise error $O(\sqrt{\frac{1}{N}})$ by computing the average of the reward of each arm, simply because for any i, $y_{a,i}$ is an unbiased estimator of b_a.

Recently Kong and Valiant (2018) proposed an estimator for $\beta^T\beta$ in the context of learnability estimation, or noise level estimation for linear regression. In the setting where each covariate $\mathbf{x}_i$ is drawn from a distribution with zero mean and identity covariance, and response variable $y_i = \beta^T\mathbf{x}_i + \eta_i$ with independent noise η_i having zero mean, they observe that for any $i \neq j$, $y_iy_j\mathbf{x}_i^T\mathbf{x}_j$ is an unbiased estimator of $\beta^T\beta$. In addition, they showed that the error rate of estimating $\beta^T\beta$ using the proposed estimator $\frac{1}{\binom{N}{2}}\sum_{i\neq j} y_iy_j\mathbf{x}_i^T\mathbf{x}_j$ is $O(\frac{d+N}{N^2})$ which implies that one can accurately estimate $\beta^T\beta$ using $N = O(\sqrt{d})$ samples. Their estimator can be directly applied to estimate $\beta_a^T\beta_a$, and we extend their techniques to the contextual bandit setting for estimating $\beta_a\beta_{a'}$ for arbitrary a, a'. In order to estimate $\beta_a^T\beta_{a'}$ for $a \neq a'$, notice that for any i, j, $\mathbf{E}[\mathbf{y}_{a,i}\mathbf{y}_{a',j}\mathbf{x}_{a,i}^T\mathbf{x}_{a',j}] = \beta_a^T\mathbf{x}_{a,i}\mathbf{x}_{a,i}^T\mathbf{x}_{a',j}\mathbf{x}_{a',j}^T\beta_{a'} = \beta_a^T\beta_{a'}$, and we simply take the average of all these unbiased estimators of $\beta_a^T\beta_{a'}$. We show that $O(\sqrt{d})$ samples for each arm suffices for accurate estimation of $\beta_a^T\beta_{a'}$ for arbitrary pairs of arms a, a'.

Once we have estimates $\hat{\mathbf{b}}, \hat{H}$ for $\mathbf{b}$ and H, if $\hat{H}$ is a PSD matrix, our algorithm simply outputs $\mathbf{E}_{\mathbf{r}\sim N(\hat{\mathbf{b}},\hat{H})}(\max_i r_i)$, otherwise, let $\hat{H}^{(PSD)}$ be the projection of $\hat{H}$ on to the PSD cone and output $E_{\mathbf{r}\sim N(\hat{\mathbf{b}},\hat{H}^{(PSD)})}(\max_i r_i)$. Given an approximation of $\mathbf{b}, H$, it is not immediately clear how the errors in estimating $\mathbf{b}, H$ translate to the error in estimating $\mathbf{E}_{\mathbf{r}\sim N(\mathbf{b},H)}[\max_i r_i]$. Our Proposition 1 leverages a quantitative version of the Sudakov-Fernique inequality due to Chatterjee (2005) and shows that if each entry of H is perturbed by at most ϵ, the optimal re-

ward $\mathbf{E}_{\mathbf{r}\sim N(\mathbf{b},H)}[\max_i r_i]$ can only change by $2\sqrt{\log K}\epsilon$. Because $\mathbf{E}_{\mathbf{x}\sim N(\hat{\mathbf{b}},\hat{H})}(\max_i x_i + b_i)$ has no closed-form expression in general, we use Monte Carlo simulation to approximate $\mathbf{E}_{\mathbf{x}\sim N(\hat{\mathbf{b}},\hat{H})}(\max_i x_i + b_i)$ in the implementation.

Our estimator for the general unknown covariance setting is much more involved. Assuming each context $\mathbf{x}$ is drawn from a Gaussian distribution with zero mean and unknown covariance Σ, the optimal reward $\mathbf{E}_{\mathbf{x}\sim N(0,\Sigma)}[\max_a(\beta_a^T\mathbf{x} + b_a)]$ is equal to $\mathbf{E}_{\mathbf{r}\sim N(\mathbf{b},H)}[\max_i r_i]$ where $\mathbf{r} \sim N(\mathbf{b}, H)$ and $H_{a,a'} = \beta_a\Sigma\beta_{a'}$. Again, we extend the estimator proposed in Kong and Valiant (2018) for $\beta^T\Sigma\beta$ in the linear regression setting to the contextual linear bandit setting for estimating $\beta_a\Sigma\beta_{a'}$ for arbitrary a, a'. For each a, a', we design a series of unbiased estimators for $\beta_a^T\Sigma^2\beta_{a'}, \beta_a^T\Sigma^3\beta_{a'}, \beta_a^T\Sigma^4\beta_{a'}, \ldots$ and approximate $\beta_a^T\Sigma\beta_{a'}$ with a linear combination of these high order estimates. Our major contribution is a series of estimators which incorporate unlabeled examples. In the contextual bandit setting, especially when K is large, it is essential to incorporate unlabeled data, simply because when we estimate $\beta_a\Sigma^k\beta_{a'}$, the large number of examples which do not involve arm a or a' are effectively unlabeled examples and can be leveraged to significantly reduce the overall variance for estimating $\beta_a\Sigma^k\beta_{a'}$. We prove variance bounds in Corollary 6 for these novel estimators whose accuracy depends on both the number of labeled examples and unlabeled examples. As a side note, our estimator can also be applied to the setting of estimating learnability to better utilize the unlabeled examples. Proofs, where omitted, are in the appendix.

4.1 Main Algorithm

Our main algorithm is described in Algorithm 1. In line 1, we repeat the for loop body $\Theta(\log(K/\delta))$ times, and at each time, we collect n i.i.d. sample for each arm. Hence the total number of samples for each arm $N = \Theta(n(\log(K/\delta))$. For ease of notations, we will use n instead of N when we write down the error rate of the algorithm.

In line 3, 4, 5, for each arm a we collect n i.i.d. samples and estimate the bias of that arm b_a. The estimation error of the bias vector $\mathbf{b}$ is bounded by the following corollary, and the claim holds by applying Chebyshev's inequality with the variance of $y_{a,i}$.

Corollary 4. *For each arm a, with probability $2/3$, $|\frac{1}{n}\sum_{i=1}^n y_{a,i} - b_a| \le 3\sqrt{\frac{1}{n}}\sigma_a$, where $\sigma_a^2 = \mathbf{Var}[y_{a,i}]$.*

After estimating b_a, we can subtract b_a from all the $y_{a,i}$. For sufficiently large n, our estimate of b_a is accurate enough such that we can assume that $y_{a,i} = \beta_a^T\mathbf{x}_{a,i} + \eta_{a,i}$. After collecting n i.i.d. samples from each arm, in the known covariance setting, we run Algorithm 2 to estimate the covariance H in line 8. In the unknown covariance setting, we need to split the n examples for each arm into one labeled example set and one unlabeled example set, and then run Algorithm 3 (see appendix) to estimate the covariance H. Bounds on Algorithm 2 and Algorithm 3 are formulated in the following two corollaries.

Corollary 5. *Given n independent samples for each arm, for a fixed pair a, a', with probability at least $2/3$, the output of Algorithm 2 satisfies*

$$|\hat{H}_{a,a'} - H_{a,a'}| \le 3\sqrt{\frac{9d+3n}{n^2}}\sigma_a\sigma_{a'},$$

where $\sigma_a^2 = \mathbf{Var}[y_{a,i}]$.

The above corollary follows from applying Chebyshev's inequality with the variance bound established in Proposition 3 and Proposition 2.

Corollary 6. *Given n independent samples for each arm, and s unlabeled examples, for a fixed pair a, a', with probability at least $2/3$, the output of Algorithm 3 satisfies*

$$|\hat{H}_{a,a'} - H_{a,a'}| \le \min(\frac{2}{k^2}, 2e^{-(k-1)\sqrt{\frac{\sigma_{min}}{\sigma_{max}}}}) + f(k)\max(\frac{d^{k/2}}{s^{k/2}}, 1)\sqrt{\frac{d+n}{n^2}},$$

where $f(k) = k^{O(k)}$.

The above corollary follows from applying Chebyshev's inequality with the variance bound established in Proposition 4. Notice that after sample splitting in Algorithm 1, the size of the set of the unlabeled examples $s = Kn/2$.

Since each entry of our estimate of $\mathbf{b}$, the output of Algorithm 2 and Algorithm 3, only satisfies the bound in Corollary 4, Corollary 5 or Corollary 6 respectively with probability $2/3$, we boost the entry-wise success probability to $1 - \delta/(K^2 + K)$ by repeating the estimation procedure $\Theta(\log(K/\delta))$ times and computing the median of our estimates (line 19 to line 20), such that the overall success probability is at least $1 - \delta$. We formalize the effect of this standard procedure in Fact 11.

Line 21 projects the matrix $\hat{H}$ onto the PSD cone and obtains the PSD matrix $\hat{H}_{PSD}$. This step is a convex optimization problem and can be solved efficiently. By the triangle inequality and the upper bound of $\max_{i,j}|\hat{H}_{i,j} - H_{i,j}|$, the discrepancy after this projection: $\max_{i,j}|\hat{H}_{i,j}^{(PSD)} - H_{i,j}|$ can be bounded with the upper bound in Corollary 5 and Corollary 6 up to a factor of 2.

Now that we have established upper bounds on the estimation error of $\mathbf{b}$ and H, we use these to bound the estimation error of the optimal reward.

Proposition 1. *Let $H \in R^{m\times m}$ and $H' \in R^{m\times m}$ be two PSD matrices, $\mathbf{b}, \mathbf{b}'$ be two d-dimensional real vectors. We have* $|\mathbf{E}_{\mathbf{x}\sim N(\mathbf{b},H)}[\max_i x_i] - \mathbf{E}_{\mathbf{x}\sim N(\mathbf{b}',H')}[\max_i x_i]| \leq 2\sqrt{\max_{i,j}|H_{i,j} - H'_{i,j}|\log K} + \max_i |b_i - b'_i|$.

Algorithm 1 Main Algorithm for Estimating OPT, the Optimal Reward [Corollary 1, Corollary 2]

for $i = 1$ **to** $\lceil 48(\log(K^2/\delta)+1)\rceil$ **do**
  for $a = 1$ **to** K **do**
    Pull the a'th arm n times, and let matrix $X_a = \left[\mathbf{x}_{a,1}^\top \cdots \mathbf{x}_{a,n}^\top\right]^\top$ consists of the n contexts, $\mathbf{y}_a = \left[y_{a,1}\cdots y_{a,n}\right]^\top$ consists of the n rewards.
    $\hat{b}_a^{(i)} \leftarrow \mathbf{1}^T\mathbf{y}_a/n$. {Estimate b_a.}
    $\mathbf{y}_a \leftarrow \mathbf{y}_a - \hat{b}_a^{(i)}\mathbf{1}$. {Subtract b_a off to make it zero mean.}
  end for
  if Known covariance **then**
    $\hat{H}^{(i)} \leftarrow$ **Algorithm 2**$(\{X_a\}_{a=1}^K, \{\mathbf{y}_a\}_{a=1}^K)$. {Corollary 5}
  else
    for $a = 1$ **to** K **do**
      $X_a^\top \leftarrow \left[\mathbf{x}_{a,1}^\top \cdots \mathbf{x}_{a,n/2}^\top\right]$.
      $\mathbf{y}_a^\top \leftarrow \left[y_{a,1}^\top \cdots y_{a,n/2}^\top\right]$.
      $S_a^\top \leftarrow \left[\mathbf{x}_{a,n/2+1}^\top \cdots \mathbf{x}_{a,n}^\top\right]$. {Split $\mathbf{x}$ into a labeled and an unlabeled example set}
    end for
    $S \leftarrow \left[S_1^\top \cdots S_K^\top\right]^\top$.
    $\hat{H}^{(i)} \leftarrow$ **Algorithm 3** $(\{X_a\}_{i=a}^K, \{\mathbf{y}_a\}_{a=1}^K, S, p(x))$. {Corollary 6}
  end if{Estimate H.}
end for
For all $1 \leq i,j \leq K$, $\hat{H}_{i,j} \leftarrow$ **median**$(\hat{H}_{i,j}^{(1)}, \ldots, \hat{H}_{i,j}^{(\lceil 48(\log(K^2/\delta)+1)\rceil)})$.
For all $1 \leq i \leq K$, $\hat{b}_i \leftarrow$ **median**$(\hat{b}_i^{(1)}, \ldots, \hat{b}_i^{(\lceil 48(\log(K^2/\delta)+1)\rceil)})$.
$\hat{H}^{(PSD)} \leftarrow \operatorname{argmin}_{M\succ 0}\max_{i,j}|\hat{H}_{i,j} - M_{i,j}|$ {Project onto the PSD cone under the max norm.}
Output:$\mathbf{E}_{\mathbf{r}\sim N(\hat{\mathbf{b}},\hat{H}^{(PSD)})}[\max_i r_i]$.

We are ready to state our main theorem for the known covariance setting.

Theorem 4 (Main theorem on Algorithm 1, known covariance setting)**.** *In the known covariance setting, with probability at least $1-\delta$, Algorithm 1 estimates the expected reward of the optimal policy with error bounded as follows:*

$$|OPT - \widehat{OPT}| = O(\sqrt{\log K}(\frac{d+n}{n^2})^{1/4})$$

For the following main theorem on the general unknown covariance setting, the proof is identical to the proof of Theorem 4.

Theorem 5 (Main theorem on Algorithm 1, unknown covariance setting)**.** *In the unknown covariance setting, for any positive integer k, with probability $1-\delta$, Algorithm 1 estimates the optimal reward OPT with additive error:*

$$|OPT - \widehat{OPT}| \leq O\Big(\sqrt{\log K}\big(\min(\frac{1}{k^2}, e^{-(k-1)\sqrt{\frac{\sigma_{min}}{\sigma_{max}}}}) + f(k)\max(\frac{d^{k/2}}{s^{k/2}}, 1)\sqrt{\frac{d+n}{n^2}}\big)^{1/2}\Big),$$

where $f(k) = k^{O(k)}$.

Choosing the optimal k in Theorem 5 yields the following Corollary 7 on the overall sample complexity in the unknown covariance setting.

Corollary 7. *For any $\epsilon > \frac{\sqrt{\log K}}{d^{1/4}}$, with probability $1-\delta$, Algorithm 1 estimates the optimal reward OPT with additive error ϵ using a total number of $T = \Theta\Big(\log(K/\delta)\max(k^{O(1)}d^{1-1/k}K^{2/k}, \frac{k^{O(k)}K\log K\sqrt{d}}{\epsilon^2})\Big)$ samples, where $k = \min(C_1\sqrt{\log K}/\epsilon + 2, \sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}}(\log(\log K/\epsilon^2) + C_2))$ for universal constants C_1, C_2.*

In the next two sections, we describe our estimators for H in the known covariance settings. The details of Algorithm 3 for the unknown covariance setting can be found in the appendix.

4.2 Estimating H in the Known Covariance Setting

In this section, we show that the output of Algorithm 2 satisfies Proposition 2 and Proposition 3. As stated earlier, we assume $\Sigma = I$ and $\mathbf{E}[\mathbf{x}] = 0$ in this section.

To bound the estimation error of H, first observe that $\hat{H}_{a,a} = \frac{\mathbf{y}_a^T A_{up}\mathbf{y}_a}{\binom{n}{2}}$ computed in Algorithm 2 is equal to $\frac{1}{\binom{n}{2}}\sum_{i<j} y_{a,i}y_{a,j}\mathbf{x}_{a,i}^T\mathbf{x}_{a,j}$. The following proposition on the estimation error of $\hat{H}_{a,a}$ is a restatement of Proposition 4 in Kong and Valiant (2018).

Proposition 2 (Restatement of Proposition 4 in Kong and Valiant (2018))**.** *For each arm a, define $\hat{H}_{a,a} = \frac{1}{\binom{n}{2}}\sum_{i<j} y_{a,i}y_{a,j}\mathbf{x}_{a,i}^T\mathbf{x}_{a,j}$ and $H_{a,a} = \beta_a^T\beta_a$. Then $\mathbf{E}[\hat{H}_{a,a}] = H_{a,a}$ and $\mathbf{E}[(\hat{H}_{a,a} - H_{a,a})^2] \leq \frac{9d+3n}{n^2}\sigma_a^4$.*

Algorithm 2 Estimating $\beta_a^T \beta_{a'}$, Identity covariance [Proposition 2, Proposition 3]

Input: $X_1 = [\mathbf{x}_{1,1}^\top, \cdots, \mathbf{x}_{1,n}^\top]^\top, \ldots, X_K = [\mathbf{x}_{K,1}^\top, \cdots, \mathbf{x}_{K,n}^\top]^\top, \mathbf{y}_1 = [y_{1,1}^\top, \cdots, y_{1,n}^\top]^\top, \ldots, \mathbf{y}_K = [y_{K,1}^\top, \cdots, y_{K,n}^\top]^\top$
for $a = 1$ **to** K **do**
  $A \leftarrow (X_a X_a^T)_{up}$ where $(X_a X_a^T)_{up}$ is the matrix $X_a X_a^T$ with the diagonal and lower triangular entries set to zero.
  $\hat{H}_{a,a} \leftarrow \mathbf{y}_a^T A_{up} \mathbf{y}_a / \binom{n}{2}$.
  for $a' = a + 1$ **to** K **do**
    $\hat{H}_{a,a'} \leftarrow \mathbf{y}_a^T X_a X_{a'}^T \mathbf{y}_{a'} / \binom{n}{2}$.
    $\hat{H}_{a',a} \leftarrow \hat{H}_{a,a'}$.
  end for
end for
Output: $\hat{H}$.

The estimate $\hat{H}_{a,a'} = \mathbf{y}_a^T X_a X_{a'}^T \mathbf{y}_{a'} / \binom{n}{2}$ computed in Algorithm 2 is equivalent to $\frac{1}{n^2} \sum_{i,j} y_{a,i} y_{a',j} \mathbf{x}_{a,i}^T \mathbf{x}_{a',j}$, and the following proposition bounds the estimation error of $\hat{H}_{a,a'}$ for $a \neq a'$.

Proposition 3. *For a pair of arms a, a', define $\hat{H}_{a,a'} = \frac{1}{n^2} \sum_{i,j} y_{a,i} y_{a',j} \mathbf{x}_{a,i}^T \mathbf{x}_{a',j}$ and $H_{a,a'} = \beta_a^T \beta_{a'}$. Then $\mathbf{E}[\hat{H}_{a,a'}] = H_{a,a'}$ and $\mathbf{E}[(\hat{H}_{a,a'} - H_{a,a'})^2] \leq \frac{9d+3n}{n^2} \sigma_a^2 \sigma_{a'}^2$.*

5 Experiments

We now provide some empirical indication of the benefit of our approach. In these experiments, we consider the known covariance setting. As long as prior data about contexts is available, as is often the case for consumer, health and many other applications, it would be possible to estimate the covariance in advance.

We first present results in a synthetic contextual multi-armed bandits setting. There are $K = 5$ arms, and the input context vectors are drawn from a normal distribution with 0 mean and an identity covariance matrix. Our results are displayed in Figure 1 for context vectors of dimension 500, 2,000 and 50,000. Here our aim is to illustrate that we are able to estimate the optimal reward accurately after seeing significant fewer contexts than would be required by the standard alternative approach for contextual bandits which would try to estimate the optimal policy, and then estimate the performance of that policy. More precisely, in this setting we use the linear disjoint contextual bandits algorithm Li et al. (2010) to estimate the β and covariance for each arm (with an optimally chosen regularization parameter in the settings where $n < d$). We then define the optimal policy as the best policy given those empirical estimates. We show the true reward of this learned policy.

We also present results for a real-world setting that mimics a standard recommendation platform trying to choose which products to recommend to a user, given a high-dimensional featurization for that user. Our experiment is based on the Jester dataset (Goldberg et al., 2001). This is a well studied dataset which includes data for >70,000 individuals providing ratings for 100 jokes. We frame this as a multi-armed bandit setting by holding out the 10 most-rated jokes, and attempt to learn a policy to select which of these jokes to offer to a particular input user, based on a feature set that captures that user's preferences based on the ratings for the remaining 90 jokes. We keep a set of 48447 users who each rated all of the 10 most popular jokes. For each person, we create a $d = 2000$ dimensional feature vector by multiplying their 90-dimensional vector of joke ratings (with missing entries replaced by that user's average rating) by a random 90×2000 matrix (with i.i.d. $N(0, 1)$ entries), and then applying a sigmoid to each of the resulting values. The reward is the user's reported rating for the joke selected by the policy. We found that the optimal expected linear policy value using this featurization was 2.98 (out of a range of 0 to 5). For comparison, the same approach with $d = 100$ has optimal policy with value 2.81, reflecting the fact that linear functions of the lower dimensional featurization cannot capture the preferences of the user as accurately as the higher dimensional featurization. Even for $d = 2000$, the full dataset of $\approx 50,000$ people is sufficient to accurately estimate this "ground truth" optimal policy. Based on this $d = 2000$ representation of the user's context, we find that even with $n = 500$ contexts, we can accurately estimate the optimal reward of the best threshold policy, to within about 0.1 accuracy, which improves significant for $n \geq 1000$ (Figure 2). Note that this is significantly lower than we would need to compute any optimal policy.

We also evaluated our algorithm on NCI-60 Cancer Growth Inhibition dataset, where the cell growth inhibition effect is recorded for different types of chemical compounds tested on 60 different cancer cell lines with different concentration levels. We picked 26,555 types of chemicals that are tested on the NCI-H23 (non-small cell lung cancer) cell line with concentration level: $-4, -5, -6, -7, -8$ log10(M). We obtain the 1000-dimensional Morgan Fingerprints representation of each chemical from its SMILES representation using the Morgan algorithm implemented in RDKit. This is a standard method for featurizing chemical compounds. The task is to choose the most effective concentration level (among the five concentration levels) for the chemical compound, given the high-dimensional feature representation of the compound. We re-scaled the cancer

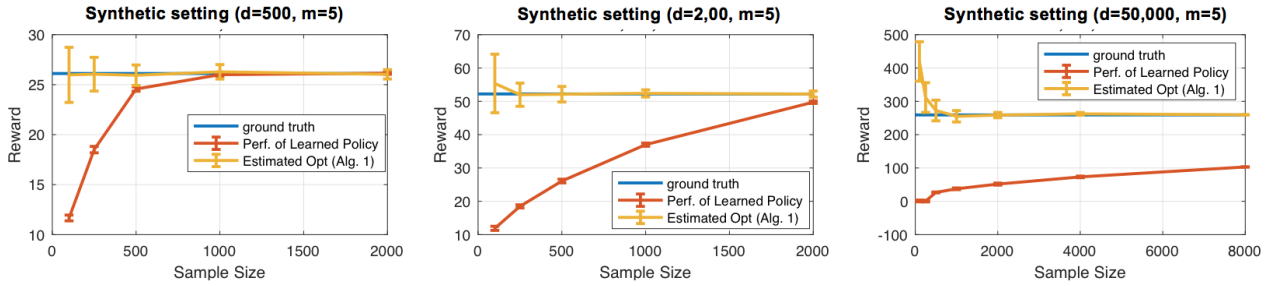

Figure 1: The three synthetic data plots depict our algorithm for estimating the optimal reward in a synthetic domain with dimension $d = 500$ (left), d=2,000 (center), and d =50,000 (right) in the setting with $m = 5$ arms corresponding to independently chosen vectors $\beta_1, \ldots, \beta_5 \in \mathbb{R}^d$ with entries chosen independently from $N(0, 1)$. Our estimated value of the optimal reward is accurate when the sample size is significantly less than d, a regime where the best learned policy does not accurately represent the optimal policy. In each plot the blue line corresponds to the true reward of the optimal policy, and the red lines depicts the performance of the learned policy at that sample size using LinUCB.

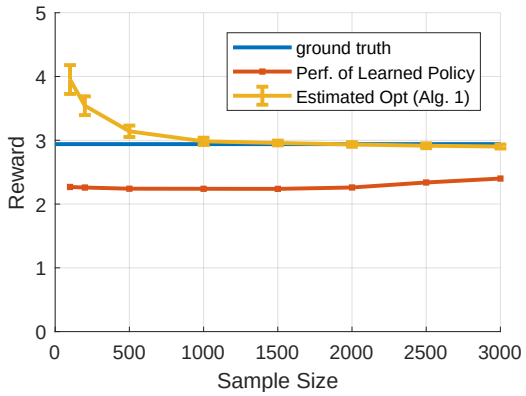

Figure 2: The plot depicts optimal reward estimation for a recommendation system that recommends 1 of 10 jokes (arms), where features are based on evaluations of 90 other jokes, represented in a $d = 2000$ space.

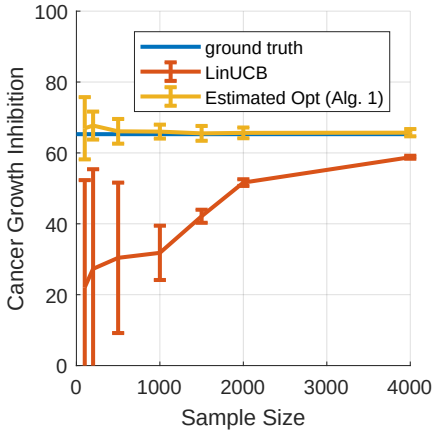

Figure 3: Evaluation on NCI-60 growth inhibition data. The blue line corresponds to the expected growth inhibition of the optimal policy. The red line is the reward and the confidence interval provided by LinUCB.

inhibition effect as between 0 and 200, where 0 means no growth inhibition, 100 means completion growth inhibition, and 200 means the cancer cells are all dead. Figure 3 depicts the result of running our algorithm and LinUCB algorithm (Li et al., 2010). The blue line depicts the true reward (65.29) of the optimal policy estimated from all 26,555 datapoints. The red line depicts the average reward and confidence interval over the last 100 rounds by executing the LinUCB algorithm with $\alpha = 1$ and different sample sizes. Notice that the LinUCB algorithm is fully adaptive and a given sample size n in Figure 3 actually corresponds to running the LinUCB algorithm for $5n$ rounds. Unlike our algorithm which achieves an accurate estimate with roughly 500 samples per arm, LinUCB is unable to learn a good policy even with $5 \times 4000 = 20000$ adaptive rounds. In this example, there is very little linear correlation between the features of the chemical compound and the inhibition effect, and simply always choosing the highest concentration achieves near-optimal reward. However, it takes thousands of rounds for the disjoint LinUCB algorithm to start playing near optimally.

6 Conclusion

To conclude, we present a promising approach for estimating the optimal reward in linear disjoint contextual bandits using a number of samples that is sublinear in the input contextual dimension. Without further assumptions, a linear number of samples is required to output a single potentially optimal policy. There exist many interesting directions for future work, including considering more generic contextual bandit settings with an infinite set of arms.

Acknowledgments

The contributions of Weihao Kong and Gregory Valiant were partially supported by a Google Faculty Fellowship, an Amazon Faculty Fellowship, and by NSF award 1704417 and an ONR Young Investigator award. Emma Brunskill was supported in part by a NSF CAREER award.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

S. Athey and S. Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

J. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory (COLT)*, 2010.

P. Auer, N. C. Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, 2009.

S. Chatterjee. An error bound in the sudakov-fernique inequality. *arXiv preprint math/0510424*, 2005.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3212–3220, 2012.

C. Gelada and M. G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. *AAAI*, 2019.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151, 2001.

K. Greenewald, A. Tewari, S. Murphy, and P. Klasnja. Action centered contextual bandits. In *Advances in neural information processing systems*, pages 5977–5985, 2017.

M. Hoffman, B. Shahriari, and N. Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374, 2014.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory (COLT)*, pages 423–439, 2014.

G. Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www.gautamkamath. com/writings/gaussian max. pdf*, 2015.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*, 2013.

R. Keramati and E. Brunskill. Value driven representation for human-in-the-loop reinforcement learning. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 176–180. ACM, 2019.

W. Kong and G. Valiant. Estimating learnability in the sublinear data regime. *arXiv preprint arXiv:1805.01626*, 2018.

A. S. Lan and R. G. Baraniuk. A contextual bandits framework for personalized learning action selection. In *EDM*, pages 424–429, 2016.

T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Off-policy policy gradient with state distribution correction. *UAI*, 2019.

T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović. Where to add actions in human-in-the-loop reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

O. Maron and A. W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 59–66, 1994.

V. Mnih, C. Szepesvári, and J.-Y. Audibert. Empirical bernstein stopping. In *International Conference on Machine Learning (ICML)*, pages 672–679. ACM, 2008.

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836, 2014.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388, 2015.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

L. Xu, J. Honda, and M. Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851, 2018.

L. Zhou and E. Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3646–3653. AAAI Press, 2016.