Online Decision-Making with High-Dimensional Covariates

Hamsa Bastani, Mohsen Bayati

Online Decision-Making with High-Dimensional Covariates

Keywords: contextual multi-armed bandits, high-dimensional statistics, sequential decision-making

Consider a decision-maker choosing between an existing option with known reward, and a new option with unknown reward. For example, a marketer may be considering a new promotion, or a doctor may be considering a new treatment. For each customer or patient, the decision-maker has a choice: should she try the new, risky option, or stick with the existing, safe option? As studied in the classical multi-armed bandit model [1], the decision-maker faces a tradeoff between exploration (learning the reward for the new option) and exploitation (leveraging current knowledge to maximize reward).

However, the choice is complicated by the fact that individuals may respond differently depending on their specific characteristics: for instance, promotions for trendy apparel may be more effective when targeted at younger customers, and some treatments may be less effective on patients with certain pre-existing conditions. In order to tailor the choice to a specific individual, the decision-maker can leverage user-level data such as purchase history or medical records. Typically, the decision-maker learns a model predicting user-specific rewards for each choice conditioned on the observed covariates. This model can then be leveraged in future decisions. The decision-maker again faces an exploration-exploitation tradeoff: the quality of the model improves as more data on the new option becomes available, but the model must also be leveraged to maximize reward. This setting is considered in the *contextual* multi-armed bandit literature, which extends the traditional multi-armed bandit model to include covariates [2].

On the other hand, the recent explosion of big data has provided decision-makers access to user-level data at unprecedented levels of detail. For instance, web cookies give companies

access to a user's entire browsing history, including clicks, purchases, and websites visited. In healthcare, electronic health records provide information on a patient's entire medical history, including all past diagnoses, procedures, and medications. Thus, the observed user-specific covariates are *high-dimensional*. Consequently, it can be prohibitively expensive to learn traditional predictive models based on the observed covariates, especially when dealing with a new product for which there is no past data available. In particular, statistical models such as linear regression require the decision-maker to experiment on a large number of individuals (at least as many as the dimension of the observed covariates) before providing meaningful user-specific recommendations. Even worse, it has been shown that under a linear model, the regret of the bandit strategy scales with the cube of the dimension of the covariates [3], making it particularly ill-suited to high-dimensional problems.

Yet, in many cases, only a small subset of the observed covariates is relevant to evaluating the benefits of a given decision. For example, only a small number of a patient's diagnoses (which correspond to certain pre-existing conditions) may be relevant to the success of a specific treatment. The identities of these covariates are a priori unknown. However, it has been shown that techniques such as L1-regularized regression can successfully identify the relevant covariates with high probability using a very small number of samples [4].

Thus, we propose an approach that efficiently leverages high-dimensional data in an online setting by learning a L1-regularized sparse linear model. Our problem formulation builds on the contextual multi-armed bandit framework. At each time step, the decision-maker observes a new user with a high-dimensional covariate vector drawn independently and identically distributed from a fixed distribution. The decision-maker has access to two choices (or "arms"). Each arm is associated with an unknown parameter vector, which we assume is sparse (i.e., most

entries are zero). The reward for a given arm is the inner product of the user's covariate vector with the arm's parameter vector, plus zero-mean Gaussian noise.

Our primary contribution is an algorithm for the setting described above, and a corresponding regret analysis. Unlike previous work in [3] where the regret scales with the cube of the dimension of the covariates, our algorithm's regret depends only on the *logarithm* of the dimension of the covariates as well as the cube of the number of relevant covariates (i.e., covariates corresponding to non-zero arm parameters). Thus, our algorithm can efficiently learn the optimal arm choice conditional on observed covariates in a high-dimensional setting. We also note that our algorithm's regret grows logarithmically in the number of time steps, which is consistent with bounds achieved in [3]. We show that this is optimal (up to constants) by proving a matching worst-case lower bound on the regret. Our proofs rely on recent tail bounds on L1-regularized estimators studied in the compressed sensing literature [4]. Finally, we discuss several applications of our algorithm to retail and healthcare settings.

References

[1] Lai, Tze Leung and Robbins, Herbert (1985), "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics* **6**(1): 4-22.

[2] Slivkins, Alex (2014), "Contextual bandits with similarity information," *The Journal of Machine Learning Research*, **15**(1), 2533-2568.

[3] Goldenshluger, Alexander and Zeevi, Assaf (2013), "A linear response bandit problem," *Stochastic Systems* **3**(1): 230-261.

[4] Bühlmann, Peter and van de Geer, Sara (2011), "Statistics for high-dimensional data," Springer-Verlag Berlin Heidelberg.