Linear Bandits in High Dimension and Recommendation Systems

Yash Deshpande and Andrea Montanari

***Abstract*—A large number of online services provide automated recommendations to help users to navigate through a large collection of items. New items (products, videos, songs, advertisements) are suggested on the basis of the user's past history and –when available– her demographic profile. Recommendations have to satisfy the dual goal of helping the user to explore the space of available items, while allowing the system to probe the user's preferences.**

We model this trade-off using linearly parametrized multi-armed bandits and prove upper and lower bounds that coincide up to constants in the data poor (high-dimensional) regime. We test (a variation of) the scheme used for estabilishing achievability on the Netflix dataset, and obtain results in agreement with the theory.

I. Introduction

Recommendation systems are a key technology for navigating through the ever-growing amount of data that is available on the Internet (products, videos, songs, scientific papers, and so on). Recommended items are chosen on the basis of the user's past history and have to strike the right balance between two competing objectives:

Serendipity i.e. allowing accidental pleasant discoveries. This has a positive –albeit hard to quantify– impact on user experience, in that it naturally limits the recommendations monotony. It also has a quantifiable positive impact on the systems, by providing fresh independent information about the user preferences.

Relevance i.e. determining recommendations which are most valued by the user, given her past choices.

While this trade-off is well understood by practitioners, as well as in the data mining literature [1], [2], [3], rigorous and mathematical work has largely focused on the second objective [4], [5], [6], [7], [8], [9]. In this paper we address the first objective, building on recent work on linearly parametrized bandits [10], [11], [12].

Y. Deshpande is with the Department of Electrical Engineering, Stanford University, Email: yashd@stanford.edu

A. Montanari is with the Departments of Electrical Engineering and Statistics, Stanford University, Email: montanari@stanford.edu

In a simple model, the system recommends items $i(1), i(2), i(3), \dots$ sequentially at times $t \in \{1, 2, 3, \dots\}$. The item index at time t is selected from a large set $i(t) \in [M] \equiv \{1, \dots, M\}$. Upon viewing (or reading, buying, etc.) item $i(t)$, the user provides feedback y_t to the system. The feedback can be explicit, e.g. a one-to-five-stars rating, or implicit, e.g. the fraction of a video's duration effectively watched by the user. We will assume that $y_t \in \mathbb{R}$, although more general types of feedback also play an important role in practice, and mapping them to real values is sometimes non-trivial.

A large body of literature has developed statistical methods to estimate the feedback a user will provide on a specific item, given past data concerning the same and other users (see the references above). A particularly successful approach consists in 'low rank' or 'latent space' models. These models postulate that the rating $y_{i,u}$ provided by user u on item i is approximately given by the scalar product of two feature vectors θ_u and $x_i \in \mathbb{R}^p$ characterizing, respectively, the user and the item. In formulae

$$y_{i,u} = \langle x_i, \theta_u \rangle + z_{i,u} \, ,$$

where $\langle a, b \rangle \equiv \sum_{i=1}^{p} a_i b_i$ denotes the standard scalar product, and $z_{i,u}$ captures unexplained factors.

The items feature vectors x_i can be either constructed explicitly, or derived from users' feedback using matrix factorization methods. Throughout this paper we will assume that they have been computed in advance using either one of these methds and are hence given. We will use the shorthand $x_t = x_{i(t)}$ for the feature vector of the item recommended at time t.

Since the items' feature vectors are known in advance, distinct users can be treated independently, and we will hereafter focus on a single users, with feature vector θ. The vector θ can encode demographic information known in advance or be computed from the user's feedback. While the model can easily incorporate the former, we will focus on the most interesting case in which no information is known in advance.

We are therefore led to consider the linear bandit model

$$y_t = \langle x_t, \theta \rangle + z_t \,, \tag{1}$$

where, for simplicity, we will assume $z_t \sim \mathsf{N}(0, \sigma^2)$ independent of θ, $\{x_i\}_{i=1}^t$ and $\{z_i\}_{i=1}^{t-1}$. At each time t, the recommender is given to choose a item feature vector $x_t \in \mathcal{X}_p \subseteq \mathbb{R}^p$, with $\mathcal{X}_p$ the set of feature vectors of the available items. A recommendation policy is a sequence of random variables $\{x_t\}_{t\geq 1}$, $x_t \in \mathcal{X}_p$ whereby x_{t+1} is a function of the past history $\{y_\ell, x_\ell\}_{1\leq \ell \leq t}$ (technically, x_{t+1} has to be measurable on $\mathcal{F}_t \equiv \sigma(\{y_\ell, x_\ell\}_{1\leq \ell\leq t})$). The system is rewarded at time t by an amount equal to the user appreciation y_t, and we let r_t denote the expected reward, i.e. $r_t \equiv \mathbb{E}\{\langle x_t, \theta\rangle\}$.

As mentioned above, the same linear bandit problem was already studied in several papers, most notably by Rusmevichientong and Tsitsiklis [11]. However, the theory developed in that work has two limitations that are important in the present context. First, the main objective of [11] is to construct policies with nearly optimal 'regret', and the focus is on the asymptotic behavior for t large with p constant. In this limit the regret per unit time goes to 0. In a recommendation system, typical dimensions of the latent feature vector are $p = 20 \sim 50$ (significantly larger dimensions are obtained if x_i includes explicitly constructed features). As a consequence, existing theory requires $t \gtrsim 100$ ratings, which is unrealistic for many recommendation systems and a large number of users.

Second, the policies that have been analyzed in [11] are based on an alternation of pure exploration and pure exploitation. In exploration phases, recommendations are completely independent of the user profile. This is of course unrealistic (and potentially harmful) in practice because it might correspond to a very negative user experience.

We aim at developing a policy with the following properties:

1) *Constant-optimal cumulative reward:* For all time t, $\sum_{\ell=1}^t r_\ell$ is within a constant factor of the maximum achievable reward.
2) *Constant-optimal regret:* Let the maximum achievable reward be $r^{\rm opt} \equiv \sup_{x\in\mathcal{X}_p}\langle x, \theta\rangle$, then the 'regret' $\sum_{\ell=1}^t (r^{\rm opt} - r_\ell)$ is within a constant of the optimal.
3) *Approximate monotonicity:* For any $0 \leq t \leq s$, we have $\mathbb{P}\{\langle x_s, \theta\rangle \geq c_1 r_t\} \geq c_2$ for c_1, c_2 as close as possible to 1.

We will describe a simple policy that achieves the first two objectives under the assumption that $\mathcal{X}_p = \{x \in \mathbb{R}^p : \|x\|_2 \leq 1\}$, and that the 'signal to noise ratio' of each observation y_t is of order one. While we will not state any formal result on point 3, the policy has interesting monotonicity properties and is a good candidate in that respect as well.

In Section II we formally state our main results. In Section III we discuss further related work. In Section IV we provide numerical simulations of our policy on synthetic as well as realistic data and also provide a comparison with prior work. Proofs are omitted for the conference version of this paper.

II. Main results

We will consider a specific form for the set of possible arms, namely $\mathcal{X}_p = \mathsf{Ball}(1) \equiv \{x | x \in \mathbb{R}^p, \|x\|_2 \leq 1\}$ is the unit ℓ_2 ball in p dimensions. This is of course a very crude model for the set of feature vectors corresponding to movies in a given database, since the latter is a cloud of M points in $\mathbb{R}^p$. However, numerical simulations presented in section IV suggest that, already for M as small as $20,000$, this cloud can be 'dense enough' to make the unit ball model qualitatively correct.

Following [11] we will also assume $\theta \in \mathbb{R}^p$ to be drawn from a Gaussian prior $\mathsf{N}(0, \mathrm{I}_{p\times p}/p)$. This roughly corresponds to the assumption that nothing is known a priori about the user except the length of its feature vector $\|\theta\| \approx 1$. Under this assumption, the scalar product $\langle x_1, \theta\rangle$, is also Gaussian with mean 0 and variance $1/p$ and hence $\Delta = p\sigma^2$ is noise-to-signal ratio for the problem. Our results are completely explicit and apply to any value of Δ. However they are constant-optimal for Δ of order one.

We will refer to our strategy as SmoothExplore . Let $\hat\theta_t$ be the the maximum likelihood estimate of θ at time t, namely

$$\hat\theta_t \equiv \arg\min_{\theta\in\mathbb{R}^p} \Big\{ \frac{1}{2\sigma^2}\sum_{\ell=1}^{t-1} \big(y_\ell - \langle x_\ell, \theta\rangle\big) + \frac{1}{2p}\|\theta\|^2 \Big\}\,. \tag{2}$$

We also denote by $\mathsf{P}_t^\perp \equiv \mathrm{I}_{p\times p} - \hat\theta_t\hat\theta_t^{\mathsf{T}}/\|\hat\theta_t\|^2$ the projector orthogonal to $\hat\theta_t$. Finally let $u_t \in \mathbb{R}^p$ a uniformly random unit vector independent of θ and $\mathcal{F}_t$. We then let

$$x_t = \sqrt{1-\beta_t^2}\,\frac{\hat\theta_t}{\|\hat\theta_t\|} + \beta_t \mathsf{P}_t^\perp u_t\,, \tag{3}$$

$$\beta_t = \sqrt{\frac{2}{3}}\min\left(\frac{p\Delta}{t},\ 1\right)^{1/4}\,. \tag{4}$$

For $t = 1$, $\hat\theta_t = 0$ and we interpret $\hat\theta_t/\|\hat\theta_t\|$ as

an arbitrary unit norm vector. In words, we combine greedy exploitation as indicated by the current maximum likelihood estimate, and random exploration in the orthogonal direction. As is natural to expect, the amount of exploration β_t is monotone decreasing, with $\lim_{t\to\infty}\beta_t = 0$, and depends smoothly on t. Hereafter we will adopt the notation $\bar{\beta}_t^2 = 1 - \beta_t^2$.

Out main result characterizes the cumulative reward

$$R_t \equiv \sum_{\ell=1}^{t} r_t = \sum_{\ell=1}^{t} \mathbb{E}\{\langle x_\ell, \theta\rangle\}$$

Under the current model the oracle reward is $r^{\rm opt} \equiv \mathbb{E}\{\|\theta\|\} \le 1$ (and indeed $r^{\rm opt} = 1 - \exp(-\Theta(p))$ for large p).

Theorem 1. *Consider the linear bandits problem with $\theta \sim \mathsf{N}(0, \mathrm{I}_{p\times p}/p)$, $x_t \in \mathcal{X}_p \equiv \mathsf{Ball}(1)$ and $p\sigma^2 = \Delta$. Then there exist constants $C_a = C_a(\Delta)$ bounded for Δ bounded away from 0 and ∞, such that* SMOOTHEXPLORE *achieves reward*

$$\text{for } 1 < t \le p\Delta, \quad R_t \ge C_1\, t^{3/2} p^{-1/2} - C_2\, t^{1/2} p^{-1/2},,$$
$$\text{for } t > p\Delta, \quad R_t \ge r^{\rm opt}\, t - C_3\, (pt)^{1/2+\omega(p)}\,.$$

where $\omega(p) = 1/(2(p+2))$.

Further, the cumulative reward of any *strategy is bounded as follows*

$$\text{for } 1 < t \le p\Delta, \quad R_t \le C_5\, t^{3/2} p^{-1/2}\,.$$

For $t > p\Delta$, we can obtain a matching upper bound by a simple modification of the arguments in [11].

Theorem 2 (Rusmevichientong and Tsitsiklis)**.** *Under the described model, the cumulative reward of any policy is bounded as follows*

$$\text{for } t > p\Delta, \quad R_t \le r^{\rm opt}\, t - \sqrt{pt\Delta} + \frac{p\Delta}{2}\,.$$

The above results characterize a sharp dichotomy between a low-dimensional, data rich regime for $t > p\Delta$ and a high-dimensional, data poor regime for $t \le p\Delta$. In the first case classical theory applies: the reward approaches the oracle performance with a gap of order $\sqrt{pt}$. This behavior is in turn closely related to central limit theorem scaling in asymptotic statistics. Notice that the scaling with t of the risk of SMOOTHEXPLORE for large t is suboptimal, namely $(pt)^{1/2+\omega(p)}$. Since however $\omega(p) = \Theta(1/p)$ the difference can be seen only on exponential time scales $t \ge exp\{\Theta(p)\}$ and is likely to be irrelevant in the context considered here (see Section IV for a demonstration).

In the high-dimensional, data poor regime $t \le p\Delta$ the number of observations is smaller than the model parameters and the vector θ can only be partially estimated. Nevertheless, such partial estimate can be exploited to produce a cumulative reward scaling as $t^{3/2}p^{1/2}$. In this regime performances are not limited by central limit theorem fluctuations in the estimate of θ. The limiting factor is instead the dimension of the parameter space that can be effectively explored in time t.

In order to understand this behavior, it is convenient to consider the noiseless case $\sigma = 0$. This is a somewhat degenerate case that is not covered by the above theorem and, nevertheless, yields useful intuition. In the noiseless case, acquiring t observations $y_1, \dots y_t$ is equivalent to learning the projection of θ on a t-dimensional subspace spanned by $x_1, \dots, x_t$. Equivalently, we learn t coordinates of θ in a suitable basis. Since the mean square value of each component of θ is $1/p$, this yields an estimate of θ (the restriction to these coordinates) with $\mathbb{E}\|\hat{\theta}_t\|_2^2 = t/p$. By selecting x_t in the direction of $\hat{\theta}_t$ we achieve instantaneous reward $r_t \approx \sqrt{t/p}$ and hence cumulative reward $R_t = \Theta(t^{3/2}p^{-1/2})$ as stated in the theorem.

Due to space limitations, the proof of Theorem 1 is deferred to the journal version of this paper.

III. Related work

Auer in [13] first considered a model similar to ours, wherein the parameter θ and noise z_t are bounded almost surely. The work assumes $\mathcal{X}_p$ finite and introduces an algorithm based on upper confidence bounds. Dani et al. [10] extended the policy of [13] for arbitrary compact decision sets $\mathcal{X}_p$. For finite sets, [10] prove an upper bound on the regret that is logarithmic in its cardinality $|\mathcal{X}_p|$, while for continuous sets the authors proved an upper bound of $O(\sqrt{pt}\log^{3/2} t)$. This result was further improved by logarithmic factors in [12]. The common theme throughout this line of work is the use of upper confidence bounds and least-squares estimation. The algorithms typically construct ellipsoidal confidence sets around the least-squares estimate $\hat{\theta}$ which, with high probability, contain the parameter θ. The algorithm then chooses optimistically the arm that appears the best with respect to this ellipsoid. As the confidence ellipsoids are initialized to be large, the bounds are only useful for $t \gg p$. In particular, in the high-dimensional data-poor regime $t = O(p)$, the bounds typically become trivial.

In light of Theorem 2 this is not surprising. Even after normalizing the noise-to-signal ratio while scaling the dimension, the $O(\sqrt{pt})$ dependence of the risk is relevant only for large time scales of $t \ge p\Delta$. This is the regime in which the parameter θ has been estimated fairly well.

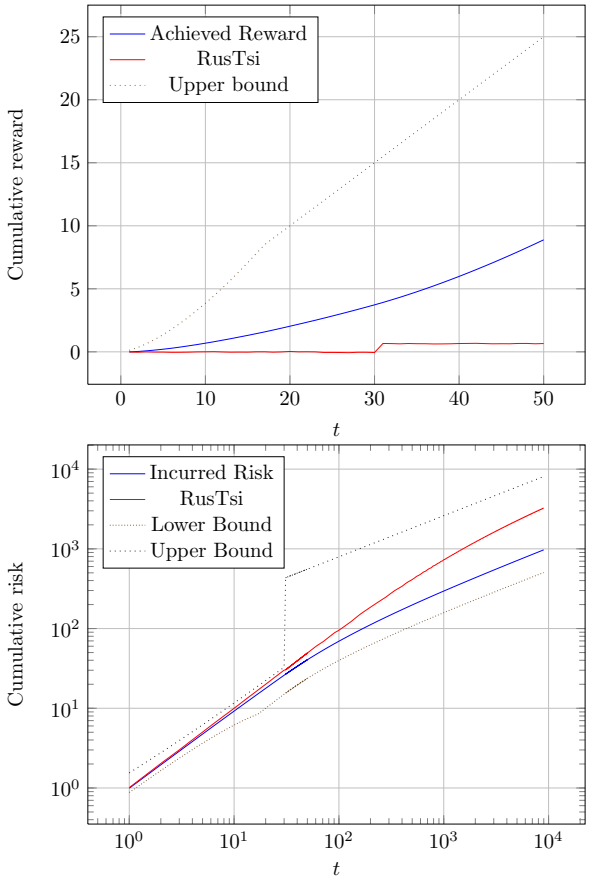

Fig. 1. Top frame: Cumulative reward R_t in the data poor regime $t \lesssim p\Delta$ as obtained through numerical simulations over synthetic data, together with analytical upper bound. Bottom frame: Cumulative risk in the data rich regime $t \gg p\Delta$.

Rusmevichientong and Tsitsiklis [11] propose a phased policy which operates in distinct phases of learning the parameter θ and earning based on the current estimate of θ. Although this approach yields order optimal bounds for the regret, it suffers from the same shortcomings as confidence-ellipsoid based algorithms. In fact, [11] also consider a more general policy based on confidence bounds and prove a $O(\sqrt{pt}\log^{3/2} t)$ bound on the regret.

Our approach to the problem is significantly different and does not rely on confidence bounds. It would be interesting to understand whether the techniques developed here can be use to improve the confidence bounds method.

IV. Numerical results

We will mainly compare our results with those of [11] since the results of that paper directly apply to the present problem. The authors proposed a phased exploration/exploitation policy, wherein they separate the phases of learning the parameter θ (exloration) and earning reward based on the current estimate of θ (exploitation).

In Figure 1 we plot the cumulative reward and the cumulative risk incurred by our policy and the phased policy, as well as analytical bounds thereof. We generated $\theta \sim \mathsf{N}(0, \mathrm{I}_{p\times p})$ randomly for $p = 30$, and produced observations y_t, $t \in \{1, 2, 3, \dots\}$ according to the general model (1) with $\Delta = p\sigma^2 = 1$. The curves presented here are averages over $n = 5000$ realizations and statistical fluctuations are negligible.

The top frame illustrates the performance of SmoothExplore in the data poor (high-dimensional) regime $t \lesssim p\Delta$. We compare the cumulative reward R_t as achieved in simulations, with that of the phased policy of [11] and with the theoretical upper bound of Theorem 1 (and Theorem 2 for $t > p\Delta$). In the bottom frame we consider instead the data rich (low-dimensional) regime $t \gg p\Delta$. In this case it is more convenient to plot the cumulative risk $tr^{\mathrm{opt}} - R_t$. We plot the curves corresponding to the ones in the top frame, as well as the upper bound (lower bound on the reward) from Theorem 1.

Note that the $O(\sqrt{pt})$ behavior of the risk of the phased policy can be observed only for $t \gtrsim 1000$. On the other hand, our policy displays the correct behavior for both time scales. The extra $\omega(p) = \Theta(1/p)$ factor in the exponent yields a factor larger than 2 only for $t \geq 2^{2(p+2)} \approx 2 \cdot 10^{19}$.

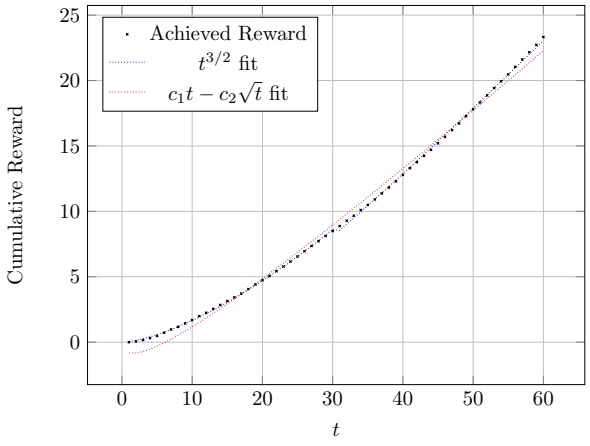

Fig. 2. Results using the Netflix datase. The policy SmoothExplore is effective in learning the user's preferences and is well described by the predicted behavior.

It is interesting to see that the policy adapts to real

datasets as well. We plot results obtained with the Netflix Prize dataset in Figure 2. Here the feature vectors x_i's for movies are obtained using the matrix completion algorithm of [8]. The user parameter vectors θ_u were obtained by regressing the rating against the movie feature vectors (the average user rating a_u was subtracted). As for synthetic data, we took $p = 30$. Regression also yields an estimate for the noise variance which is assumed known in the algorithm. We then simulated an interactive scenario by postulating that the rating of user u for movie i is given by

$$\widetilde{y}_{i,u} = \mathsf{Quant}(a_u + \langle x, \theta_u \rangle)\,,$$

where $\mathsf{Quant}(z)$ quantizes z to to $\{1, 2, \cdots, 5\}$ (corresponding to a one-to-five star rating). The feedback used for our simulation is the centered rating $y_{i,u} = \widetilde{y}_{i,u} - a_u$.

Notice that using the actual ratings in the dataset as $y_{i,u}$ is inconsistent. Indeed, first of all these ratings form (for each user) a very small subset (of the order of 100 movies) of the whole database. Second, this is a biased subset (since it is selected by the user itself).

In order to adapt to the fact that the decision set $\mathcal{X}_p$ is finite (comprising $M = 17,770$ movies) we modified the policy SMOOTHEXPLORE as follows. At each time we compute the maximum likelihood estimate of the user feature vector $\hat{\theta}_t$ and choose the "best" movie $\tilde{x}_t = \arg\max_{x \in \mathcal{X}_p} \langle x, \hat{\theta}_t \rangle$ assuming our estimate is error free. We then construct the ball in $\mathbb{R}^p$ with center $\bar{\beta}_t \tilde{x}_t$ and radius β_t. We list all the movies whose feature vectors fall in this ball, and recomend a randomly chosen one in this list. Notice that, if the decision region was indeed the unit ball, this strategy would be essentially the same (for large p) as the one we analyzed.

Classically bandit theory implies the reward behavior is described to be of type $c_1 t - c_2\sqrt{t}$ where c_1 and c_2 are (dimension-dependent) constants. Figure 2 presents the best fit of this type for $t \le 2p$. The description apears to be qualitatively incorrect in this regime. Indeed, in this regime, the reward behavior is better explained by a $c_3 t^{3/2}$ curve. These results suggest that our policy is fairly robust to the significant modeling uncertainty inherent in the problem. Remarkably, despite the fact that the decision set $\mathcal{X}_p$ is finite, the theory developed for $\mathcal{X}_p$ equal to the unit ball seems to apply qualitatively.

V. CONCLUSION

The (essentially) order-optimal results of Theorem 1 can be extended in a straightforward manner to include more general decision sets $\mathcal{X}_p$ such as ellipsoids. Further directions of work include establishing similar guarantees under weaker conditions on the noise and user model. More importantly, our focus on the dichotomy between the data-poor and data-rich regimes highlights it as an important yet hitherto neglected feature of the bandit model. We believe our results establish that the linear bandits are a good framework for understanding interactivity in recommendation systems. It would be interesting to see if the model and insight could be extended to include features like adaptivity (where user preferences may change with time) and influence (where the recommendations may alter the user characteristics).

ACKNOWLEDGMENTS

This work was partially supported by the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211, and the AFOSR grant FA9550-10-1-0360.

REFERENCES

[1] A. Schein, A. Popescul, L. Ungar, and D. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 253–260.

[2] M. Zhang and N. Hurley, "Avoiding monotony: improving the diversity of recommendation lists," in *Proceedings of the 2008 ACM conference on Recommender Systems*, 2008.

[3] M. Slaney and W. White, "Measuring playlist diversity for recommendation systems," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, 2006.

[4] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *20th International Conference on Machine Learning*. AAAI Press, 2003, pp. 720–727.

[5] N. Srebro, J. D. M. Rennie, and T. S. Jaakola, "Maximum-margin matrix factorization," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 1329–1336.

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundation of computational mathematics*, vol. 9, no. 6, pp. 717–772, February 2009.

[7] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," 2009, arXiv:0910.1879.

[8] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, July 2010.

[9] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, pp. 2302–2329, 2011.

[10] V. Dani, T. Hayes, and S. Kakade, "Stochastic linear optimization under bandit feedback," in *COLT*, 2008, pp. 355–366.

[11] P. Rusmevichientong and J. Tsitsiklis, "Linearly parameterized bandits," *Math. Oper. Res.*, vol. 35, no. 2, pp. 395–411, 2010.

[12] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *NIPS*, 2011, pp. 2312–2320.

[13] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, p. 2002, 2002.