Linear Bandits in High Dimension and Recommendation Systems

Yash Deshpande* and Andrea Montanari†

December 29, 2012

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, propose a policy and prove upper and lower bounds on the cumulative "reward" that coincide up to constants in the data poor (high-dimensional) regime. Prior work on linear bandits has focused on the data rich (low-dimensional) regime and used cumulative "risk" as the figure of merit. For this data rich regime, we provide a simple modification for our policy that achieves near-optimal risk performance under more restrictive assumptions on the geometry of the problem. We test (a variation of) the scheme used for establishing achievability on the Netflix and MovieLens datasets and obtain good agreement with the qualitative predictions of the theory we develop.

1 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 [SPUP02, ZH08, SW06], rigorous and mathematical work has largely focused on the second objective [SJ03, SRJ05, CR09, Gro09, CT10, KMO10a, KMO10b, KLT11]. In this paper we address the first objective, building on recent work on linearly parametrized bandits [DHK08, RT10, AYPS11].

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.

*Y. Deshpande is with the Department of Electrical Engineering, Stanford University

†A. Montanari is with the Departments of Electrical Engineering and Statistics, Stanford University

A large body of literature has developed statistical methods to predict the feedback that 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 uses '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 resulting matrix of ratings $y = (y_{i,u})$ is well-approximated by a rank-p matrix.

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 methods 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 $\{x_t\}_{t \geq 1}$, $x_t \in \mathcal{X}_p$ wherein 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 $\mathcal{F}_t \equiv \sigma(\{y_\ell, x_\ell\}_{\ell=1}^{t})$). 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 [RT10]. The theory developed in that work, however, has two limitations that are important in the context of recommendation systems. First, the main objective of [RT10] 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 p of the latent feature vector are about 20 to 50 [BK07, Kor08, KBV09]. If the vector x_i include explicitly constructed features, p can easily become easily much larger. As a consequence, existing theory requires at least $t \gtrsim 100$ ratings, which is unrealistic for many recommendation systems and a large number of users.

Second, the policies that have been analyzed in [RT10] are based on an alternation of pure exploration and pure exploitation. In exploration phases, recommendations are completely independent of the user profile. This is somewhat unrealistic (and potentially harmful) in practice because it would translate into a poor user experience. Consequently, we postulate the following desirable properties for a "good" policy:

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 $r^{\text{opt}} \equiv \sup_{x \in \mathcal{X}_p} \langle x, \theta \rangle$, then the 'regret' $\sum_{\ell=1}^{t} (r^{\text{opt}} - r_\ell)$ 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 aim, in this paper, to address the first objection in a fairly general setting. In particular, when t is small, say a constant times p, we provide matching upper and lower bounds for the cumulative reward under certain mild assumptions on the set of arms $\mathcal{X}_p$. Under more restrictive assumptions on the set of arms $\mathcal{X}_p$, our policy can be extended to achieve near optimal regret as well. Although we will not prove a formal result of the type of Point 3, our policy is an excellent candidate in that respect.

The paper is organized as follows : in Section 2 we formally state our main results. In Section 3 we discuss further related work. Some explication on the assumptions we make on the set of arms $\mathcal{X}_p$ is provided in Section 4. In Section 5 we present numerical simulations of our policy on synthetic as well as realistic data from the Netflix and MovieLens datasets. We also compare our results with prior work, and in particular with the policy of [RT10]. Finally, proofs are given in Sections 6 and 7.

2 Main results

We denote by $\mathsf{Ball}(x;\rho)$ the Euclidean ball in $\mathbb{R}^p$ with radius ρ and center $x \in \mathbb{R}^p$. If x is the origin, we omit this argument and write $\mathsf{Ball}(\rho)$. Also, we denote the identity matrix as I_p.

Our achievability results are based on the following assumption on the set of arms $\mathcal{X}_p$.

Assumption 1. *Assume, without loss of generality, $\mathcal{X}_p \in \mathsf{Ball}(1)$. We further assume that there exists a subset of arms $\mathcal{X}'_p \subseteq \mathcal{X}_p$ such that:*

1. *For each $x \in \mathcal{X}'_p$ there exists a distribution $\mathbb{P}_x(z)$ supported on $\mathcal{X}_p$ with $\mathbb{E}_x(z) = x$ and $\mathbb{E}_x(zz^{\mathsf{T}}) \succeq (\gamma/p)\mathrm{I}_p$, for a constant $\gamma > 0$. Here $\mathbb{E}_x(\cdot)$ denotes expectation with respect to $\mathbb{P}_x$.*
2. *For all $\theta \in \mathbb{R}^p$, $\sup_{x\in\mathcal{X}'_p}\langle x, \theta\rangle \geq \kappa\|\theta\|_2$ for some $\kappa > 0$.*

Examples of sets satisfying Assumption 1 and further discussion of its geometrical meaning are deferred to Section 4. Intuitively, it requires that $\mathcal{X}_p$ is 'well spread-out' in the unit ball $\mathsf{Ball}(1)$.

Following [RT10] we will also assume $\theta \in \mathbb{R}^p$ to be drawn from a Gaussian prior $\mathsf{N}(0, \mathrm{I}_p/p)$. 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 $\langle x_1, \theta\rangle$, where x_1 is necessarily independent of θ, 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 explicitly computable and apply to any value of Δ. However they are constant-optimal for Δ bounded away from zero.

Let $\hat{\theta}_t$ be the posterior mean estimate of θ at time t, namely

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

A greedy policy would select the arm $x \in \mathcal{X}_p$ that maximizes the expected one-step reward $\langle x, \hat{\theta}_t\rangle$. As for the classical multiarmed bandit problem, we would like to combine this approach with random exploration of alternative arms. We will refer to our strategy as SmoothExplore since it combines exploration and exploitation in a continuous manner. This policy is summarized in Table 1.

Algorithm 1 SmoothExplore

```
initialize ℓ = 1, θ̂_1 = 0, θ̂_1/‖θ̂_1‖ = e_1, Σ_1 = I_p/p.
repeat
    Compute: x̃_ℓ = arg max_{x∈X'_p} ⟨θ̂_ℓ, x⟩.
    Play: x_ℓ ~ P_{x̃_ℓ}(·), observe y_t = ⟨x_t, θ⟩ + z_t.
    Update: ℓ ← ℓ + 1, θ̂_ℓ = arg min_{θ∈R^p} (1/(2σ^2)) Σ_{i=1}^{ℓ−1} (y_i − ⟨x_i, θ⟩)^2 + (1/(2p))‖θ‖^2.
until ℓ > t
```

The policy SmoothExplore uses a fixed mixture of exploration and exploitation as prescribed by the probability kernel $\mathbb{P}_x(\,\cdot\,)$. As formalized below, this is constant optimal in the data poor high-dimensional regime hence on small time horizons.

While the focus of this paper is on the data poor regime, it is useful to discuss how the latter blends with the data rich regime that arises on long time horizons. This also clarifies where the boundary between short and long time horizons sits. Of course, one possibility would be to switch to a long-time-horizon policy such as the one of [RT10]. Alternatively, in the spirit of approximate monotonicity, we can try to progressively reduce the random exploration component as t increases. We will illustrate this point for the special case $\mathcal{X}_p \equiv \mathsf{Ball}(1)$. In that case, we introduce a special case of SmoothExplore , called BallExplore , cf. Table 2. The amount of random exploration at time t is gauged by a parameter β_t that decreases from $\beta_1 = \Theta(1)$ to $\beta_t \to 0$ as $t \to \infty$.

Note that, for $t \le p\Delta$, β_t is kept constant with $\beta_t = \sqrt{2/3}$. In this regime BallExplore corresponds to SmoothExplore with the choice $\mathcal{X}'_p = \partial\mathsf{Ball}(1/\sqrt{3})$ (here and below ∂S denotes the boundary of a set S). It is not hard to check that this choice of $\mathcal{X}'_p$ satisfies Assumption 1 with $\kappa = 1/\sqrt{3}$ and $\gamma = 2/3$. For further discussion on this point, we refer the reader to Section 4.

Algorithm 2 BallExplore

initialize $\ell = 1$, $\hat{\theta}_1 = 0$, $\hat{\theta}_1/\|\hat{\theta}_1\| = e_1$, $\Sigma_1 = \mathrm{I}_p/p$, $\mathsf{P}_1^{\perp} = \mathrm{I}_p - e_1 e_1^{\mathsf{T}}$.
repeat
  Compute: $\tilde{x}_\ell = \arg\max_{x\in\mathsf{Ball}(1)} \langle \hat{\theta}_\ell, x\rangle = \hat{\theta}_\ell/\|\hat{\theta}_\ell\|$, $\beta_\ell = \sqrt{2/3}\,\min(p\Delta/\ell, 1)^{1/4}$.
  Play: $x_t = \sqrt{1-\beta_\ell^2}\,\tilde{x}_\ell + \beta_\ell \mathsf{P}_\ell^{\perp} u_\ell$, where u_ℓ is a uniformly sampled unit vector, independent of the past.
  Observe: $y_t = \langle x_t, \theta\rangle + z_t$.
  Update: $\ell \leftarrow \ell + 1$, $\hat{\theta}_\ell = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2\sigma^2}\sum_{i=1}^{\ell-1}(y_i - \langle x_i, \theta\rangle)^2 + \frac{1}{2p}\|\theta\|^2$, $\mathsf{P}_\ell^{\perp} = \mathrm{I}_p - \hat{\theta}_\ell\hat{\theta}_\ell^{\mathsf{T}}/\|\hat{\theta}_\ell\|^2$.
until $\ell > t$

Our main result characterizes the cumulative reward

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

Theorem 1. *Consider the linear bandits problem with $\theta \sim \mathsf{N}(0, \mathrm{I}_{p\times p}/p)$, $x_t \in \mathcal{X}_p \subseteq \mathsf{Ball}(1)$ satisfying Assumption 1, and $p\sigma^2 = \Delta$. Further assume that $p \ge 2$ and $p\Delta \ge 2$.*

Then there exists a constant $C_1 = C_1(\kappa, \gamma, \Delta)$ bounded for κ, γ and Δ bounded away from zero, such that SmoothExplore *achieves, for $1 < t \le p\Delta$, cumulative reward*

$$R_t \ge C_1\, t^{3/2} p^{-1/2}.$$

Further, the cumulative reward of any *strategy is bounded for $1 \le t \le p\Delta$ as:*

$$R_t \le C_2\, t^{3/2} p^{-1/2}\,.$$

We may take the constants $C_1(\kappa, \gamma, \Delta)$ and $C_2(\Delta)$ to be:

$$C_1 = \frac{\kappa\sqrt{\Delta}\,C(\gamma,\Delta)}{24\,\alpha(\gamma,\Delta)}, \qquad C_2 = \frac{2}{3\sqrt{\Delta}},$$

$$\text{where} \qquad C(\gamma,\Delta) = \frac{\gamma}{4(\Delta+1)}, \qquad \alpha(\gamma,\Delta) = 1 + \left[3\log\left(\frac{96}{\Delta\, C(\gamma,\Delta)}\right)\right]^{1/2}.$$

In the special case where $\mathcal{X}_p = \mathsf{Ball}(1)$, we have the following result demonstrating that BallExplore has near-optimal performance in the long time horizon as well.

Theorem 2. *Consider the linear bandits problem with $\theta \sim \mathsf{N}(0, \mathrm{I}_{p\times p}/p)$ with the set of arms $\mathcal{X}_p$ is the unit ball, i.e.* $\mathsf{Ball}(1)$. *Assume, $p \ge 2$ and $p\Delta \ge 2$. Then* BallExplore *achieves for all $t > p\Delta$:*

$$R_t \ge r^{\mathrm{opt}}\, t - C_3\,(pt)^{1/2+\omega(p)}\,.$$

where:

$$\omega(p) = 1/(2(p+2)), \qquad C_3(\Delta) = 70\left(\frac{\Delta+1}{\sqrt{\Delta}}\right).$$

For $t > p\Delta$, we can obtain a matching upper bound by a simple modification of the arguments in [RT10].

Theorem 3 (Rusmevichientong and Tsitsiklis)**.** *Under the described model, the cumulative reward of any policy is bounded as follows*

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

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 our upper bound on the risk of BallExplore 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 for moderate to large values p (see Section 5 for a demonstration). It is an open problem to establish the exact asymptotic scaling[1] of BallExplore .

In the high-dimensional, data poor regime $t \le p\Delta$, the number of observations is smaller than the number of 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 t steps.

In order to understand this behavior, it is convenient to consider the noiseless case $\sigma = 0$. This is a somewhat degenerate case that, although not covered by the above theorem, yields useful intuition. In the noiseless case, acquiring t observations $y_1, \dots y_t$ is equivalent to learning the projection of θ on the 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 $\hat\theta_t$ (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.

3 Related work

Auer in [Aue02] first considered a model similar to ours, wherein the parameter θ and noise z_t are bounded almost surely. This work assumes $\mathcal{X}_p$ finite and introduces an algorithm based on upper confidence bounds. Dani et al. [DHK08] extended the policy of [Aue02] for arbitrary compact decision sets $\mathcal{X}_p$. For finite sets, [DHK08] prove an upper bound on the regret that is logarithmic in its cardinality $|\mathcal{X}_p|$, while for continuous sets they prove an upper bound of $O(\sqrt{pt}\log^{3/2} t)$. This result was further improved by logarithmic factors in [AYPS11]. 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 3 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.

Rusmevichientong and Tsitsiklis [RT10] 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

[1]Simulations suggest that the upper bound $(pt)^{1/2+\omega(p)}$ might be tight.

order optimal bounds for the regret, it suffers from the same shortcomings as confidence-ellipsoid based algorithms. In fact, [RT10] 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.

4 On Assumption 1

The geometry of the set of arms $\mathcal{X}_p$ is an important factor in the in the performance of any policy. For instance, [RT10], [DHK08] and [AYPS11] provide "problem-dependent" bounds on the regret incurred in terms of the difference between the reward of the optimal arm and the next-optimal arm. This characterization is reasonable in the long time horizon: if the posterior estimate $\hat{\theta}_t$ of the feature vector θ *coincided* with θ itself, only the optimal arm would matter. Since the posterior estimate converges to θ in the limit of large t, the local geometry of $\mathcal{X}_p$ around the optimal arm dictates the asymptotic behavior of the regret.

In the high-dimensional, short-time regime, the global geometry of $\mathcal{X}_p$ plays instead a crucial role. This is quantified in our results through the parameters κ and γ appearing in Assumption 1. Roughly speaking, this amounts to requiring that $\mathcal{X}_p$ is 'spread out' in the unit ball. It is useful to discuss this intuition in a more precise manner. For the proofs of statements in this section we refer to Appendix A.

A simple case is the one in which the arm set contains a ball.

Lemma 4.1. *If* $\mathsf{Ball}(\rho) \subseteq \mathcal{X}_p \subseteq \mathsf{Ball}(1)$*, then* $\mathcal{X}_p$ *satisfies Assumption 1 with* $\kappa = \rho/\sqrt{3}$*,* $\gamma = 2\rho^2/3$*.*

The last lemma does not cover the interesting case in which $\mathcal{X}_p$ is finite. The next result shows however that, for Assumption 1.2 to hold it is sufficient that the closure of the convex hull of $\mathcal{X}'_p$, denoted by $\overline{\mathrm{conv}}(\mathcal{X}'_p)$, contains a ball.

Proposition 4.2. *Assumption 1.2 holds if and only if* $\mathsf{Ball}(\kappa) \subseteq \overline{\mathrm{conv}}(\mathcal{X}'_p)$*.*

In other words, Assumption 1.2 is satisfied if $\mathcal{X}'_p$ is 'spread out' in all directions around the origin.

Finally, we consider a concrete example with $\mathcal{X}_p$ finite. Let $x_1, x_2, \dots, x_M$ to be i.i.d. uniformly random in $\mathsf{Ball}(1)$. We then refer to the set of arms $\mathcal{X}_p \equiv \{x_1, x_2, \dots, x_M\}$ as to a *uniform cloud.*

Proposition 4.3. *A uniform cloud* $\mathcal{X}_p$ *in dimension* $p \geq 20$ *satisfies Assumption 1 with* $M = 8^p$*,* $\kappa = 1/4$ *and* $\gamma = 1/32$ *with probability larger than* $1 - 2\exp(-p)$*.*

5 Numerical results

We will mainly compare our results with those of [RT10] 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 (exploration) 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 $\theta \sim \mathsf{N}(0, \mathrm{I}_p)$ randomly for $p = 30$, and produced observations y_t, $t \in \{1, 2, 3, \dots\}$ according to the general model (1) with $\Delta = p\sigma^2 = 1$ and arm set $\mathcal{X}_p = \mathsf{Ball}(1)$. The curves presented here are averages over $n = 5000$ realizations and statistical fluctuations are negligible.

The left frame illustrates the performance of SmoothExplore in the data poor (high-dimensional) regime $t \lesssim 2\,p\Delta$. We compare the cumulative reward R_t as achieved in simulations, with that of the phased policy of [RT10] and with the theoretical upper bound of Theorem 1 (and Theorem 3 for $t > p\Delta$). In the right frame we consider instead the data rich (low-dimensional) regime $t \gg p\Delta$. In this case it is more

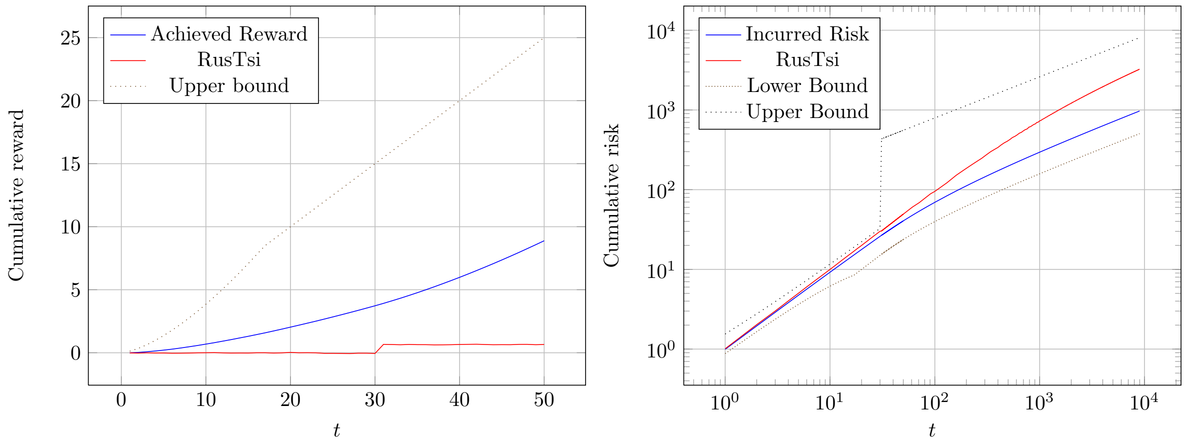

Figure 1: Left frame: Cumulative reward R_t in the data poor regime $t \lesssim 2\,p\Delta$ (here $p = 30$, $\Delta = 1$) as obtained through numerical simulations over synthetic data, together with analytical upper bound. Right frame: Cumulative risk in the data rich regime $t \gg p\Delta$ (again, $p = 30$, $\Delta = 1$).

convenient to plot the cumulative risk $tr^{\rm opt} - R_t$. We plot the curves corresponding to the ones in the left frame, as well as the upper bound (lower bound on the reward) from Theorems 1 and 2.

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 multiplicative factor larger than 2 only for $t \geq 2^{2(p+2)} \approx 2 \cdot 10^{19}$.

The above set of numerical experiments used $\mathcal{X}_p = \mathsf{Ball}(1)$. For applications to recommendation systems, $\mathcal{X}_p$ is in correspondence with a certain catalogue of achievable products or contents. In particular, $\mathcal{X}_p$ is expected to be finite. It is therefore important to check how does SmoothExplore perform for a realistic sets of arms. We plot results obtained with the Netflix Prize dataset and the MovieLens 1M dataset in Figure 2. Here the feature vectors x_i's for movies are obtained using the matrix completion algorithm of [KMO10b]. 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). Similar to 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 $y_{i,u} = \widetilde{y}_{i,u} - a_u$.

We implement a slightly modified version of SmoothExplore for these simulations. At each time we compute the ridge regression estimate of the user feature vector $\hat{\theta}_t$ as before 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 $\tilde{x}_t$ and radius β_t. We list all the movies whose feature vectors fall in this ball, and recommend a uniformly randomly chosen one in this list.

Classical bandit theory implies the reward behavior is of the 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 $t \lesssim 2p$. The description appears 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. In particular, the fact that the "noise" encountered in practice is manifestly non-Gaussian does not affect the qualitative predictions of Theorem 1.

A full validation of our approach would require an actual interactive realization of a recommendation system [DM13]. Unfortunately, such validation cannot be provided by existing datasets, such as the ones

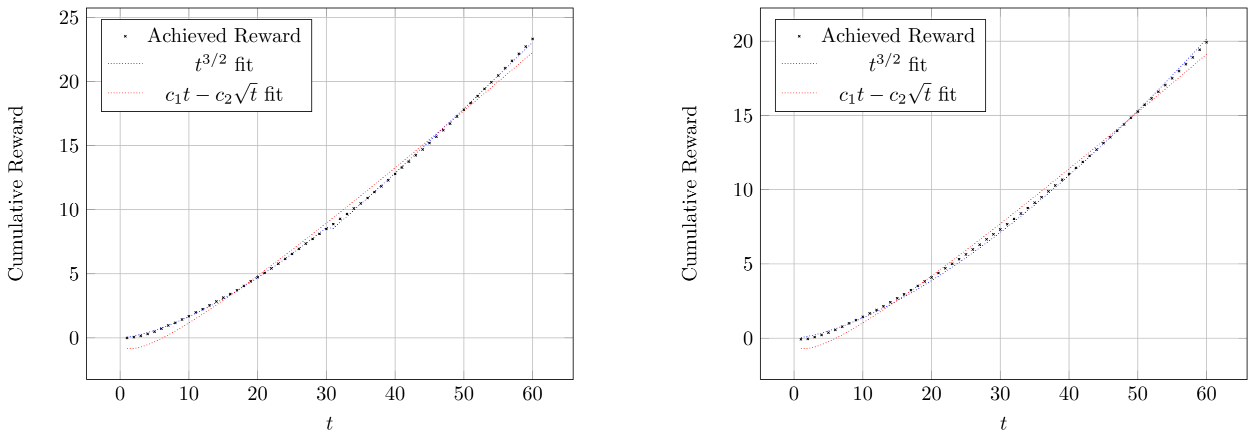

Figure 2: Results using the Netflix (left frame) and MovieLens 1M (right frame) datasets. SmoothExplore is effective in learning the user's preferences and is well described by the predicted behavior of Theorem 1.

used here. A naive approach would be to use the actual ratings as the feedback y_{iu}, but this suffers from many shortcomings. First of all, each user rates a sparse subset (of the order of 100 movies) of the whole database of movies, and hence any policy to be tested would be heavily constrained and distorted. Second, the set of rated movies is a biased subset (since it is selected by the user itself).

6 Proof of Theorem 1

We begin with some useful notation. Define the σ-algebra $\mathcal{F}_t \equiv \sigma(\{y_\ell, x_\ell\}_{\ell=1}^{t})$. Also let $\mathcal{G}_t \equiv \sigma(\{y_\ell\}_{\ell=1}^{t-1}, \{x_\ell\}_{\ell=1}^{t})$. We let $\hat{\theta}_t$ and Σ_t denote the posterior mean and covariance of θ given $t-1$ observations. Since θ is Gaussian and the observations are linear, it is a standard result that these can be computed as:

$$\Sigma_t \equiv \mathrm{Cov}(\theta|\mathcal{F}_{t-1}) = \left(p\mathrm{I}_p + \frac{1}{\sigma^2}\sum_{\ell=1}^{t-1} x_\ell x_\ell^{\mathsf{T}}\right)^{-1}$$

$$\hat{\theta}_t \equiv \mathbb{E}(\theta|\mathcal{F}_{t-1}) = \Sigma_t\left(\sum_{\ell=1}^{t-1}\frac{y_\ell}{\sigma^2}x_\ell\right).$$

Note that since θ is Gaussian and the measurements are linear the posterior mean coincides with the maximum likelihood estimate for θ. This ensures our notation is consistent.

6.1 Upper bound on reward

At time ℓ, the expected reward $r_\ell = \mathbb{E}(\langle x_\ell, \theta\rangle) \le \mathbb{E}(\|\hat{\theta}_\ell\|) \le \left[\mathbb{E}(\|\hat{\theta}_\ell\|^2)\right]^{1/2}$, where the first inequality follows from Cauchy-Schwarz, that $\hat{\theta}_\ell$ is unbiased and that $\|x_\ell\| \le 1$. Since $1 = \mathbb{E}(\|\theta\|^2) = \mathbb{E}(\|\hat{\theta}_\ell\|^2) + \mathbb{E}(\mathsf{Tr}\Sigma_\ell)$:

$$r_\ell^2 \le 1 - \mathbb{E}\left(\mathsf{Tr}(\Sigma_\ell)\right). \tag{3}$$

We have, applying Jensen's inequality and further simplification:

$$
\begin{aligned}
\mathbb{E}\mathsf{Tr}(\Sigma_\ell) &\geq p^2/\mathbb{E}(\mathsf{Tr}(\Sigma_\ell^{-1})) \\
&= p^2/\mathbb{E}\mathsf{Tr}\left(p\mathrm{I}_p + \frac{1}{\sigma^2}\sum_{j=1}^{\ell-1} x_j x_j^{\mathsf{T}}\right) \\
&\geq \left(1 + \frac{\ell-1}{p^2\sigma^2}\right)^{-1}.
\end{aligned}
$$

Using this to bound the right hand side of Eq. (3)

$$
\begin{aligned}
r_\ell^2 &\leq 1 - \frac{1}{1 + (\ell-1)/(p\sigma)^2} \\
&= \frac{(\ell-1)/p}{(\ell-1)/p + p\sigma^2} \\
&\leq \frac{1}{p\sigma^2}\frac{\ell-1}{p}.
\end{aligned}
$$

The cumulative reward can then be bounded as follows:

$$
\begin{aligned}
\sum_{\ell=1}^{t} r_\ell &\leq \frac{1}{\sqrt{p\sigma^2}}\sum_{\ell=1}^{t}\sqrt{\frac{\ell-1}{p}} \\
&\leq \frac{2}{3\sqrt{p\sigma^2}} t^{3/2} p^{-1/2} \\
&= C_2(\Delta) t^{3/2} p^{-1/2}.
\end{aligned}
$$

Here we define $C_2(\Delta) \equiv 2/3\sqrt{\Delta}$.

6.2 Lower bound on reward

We compute the expected reward earned by SmoothExplore at time t as:

$$
\begin{aligned}
r_t &= \mathbb{E}(\langle x_t, \theta\rangle) \\
&= \mathbb{E}(\mathbb{E}(\langle x_t, \theta\rangle | \mathcal{G}_{t-1})) \\
&= \mathbb{E}(\mathbb{E}(\langle x_t, \hat{\theta}_t\rangle | \mathcal{G}_{t-1})) \\
&= \mathbb{E}(\langle \tilde{x}_t, \hat{\theta}_t\rangle) \\
&\geq \kappa\, \mathbb{E}(\|\hat{\theta}_t\|).
\end{aligned}
\tag{4}
$$

The following lemma guarantees that $\|\hat{\theta}_t\|$ is $\Omega(\sqrt{t})$.

Lemma 6.1. *Under the conditions of Theorem 1 we have, for all $t > 1$:*

$$
\mathbb{E}\|\hat{\theta}_t\| \geq C'(\gamma, \Delta) t^{1/2} p^{-1/2}.
$$

Here:

$$
\begin{aligned}
C'(\gamma, \Delta) &= \frac{1}{2}\frac{C(\gamma,\Delta)}{\alpha(\gamma,\Delta)}\sqrt{\frac{\Delta}{8}}, \\
\textit{where} \quad C(\gamma, \Delta) &= \frac{\gamma}{4(\Delta+1)} \\
\alpha(\gamma, \Delta) &= 1 + \left[3\log\left(\frac{96}{\Delta C(\gamma,\Delta)}\right)\right]^{1/2}.
\end{aligned}
$$

Using this lemma we proceed by bounding the right side of Eq. (4):

$$r_t \geq \kappa C' \sqrt{\frac{t-1}{p}}.$$

Computing cumulative reward R_t we have:

$$\begin{aligned}
R_t &= \sum_{\ell=1}^{t} r_\ell \\
&\geq \sum_{\ell=1}^{t} \kappa C' \sqrt{\frac{\ell-1}{p}} \\
&\geq \kappa C' \int_0^{t-1} \sqrt{\frac{\nu}{p}} \mathrm{d}\nu \\
&\geq \frac{2}{3} \kappa C' (t-1)^{3/2} p^{-1/2} \\
&\geq \frac{\kappa C'}{3\sqrt{2}} t^{3/2} p^{-1/2}.
\end{aligned}$$

Thus, letting $C_1(\kappa, \gamma, \Delta) = \kappa C'(\gamma, \Delta)/3\sqrt{2}$, we have the required result.

6.3 Proof of Lemma 6.1

In order to prove that $\mathbb{E}(\|\hat{\theta}_t\|) = \Omega(\sqrt{t})$, we will first show that $\mathbb{E}(\|\hat{\theta}_t\|^2)$ is $\Omega(t)$. Then we prove that $\|\hat{\theta}_t\|$ is sub-gaussian, and use this to arrive at the required result.

Lemma 6.2 (Growth of Second Moment)**.** *Under the conditions of Theorem 1:*

$$\mathbb{E}\|\hat{\theta}_t\|^2 \geq C(\Delta, \gamma) \frac{t-1}{p},$$

where

$$C(\Delta, \gamma) = \frac{\gamma}{4(\Delta + 1)}.$$

Proof. We rewrite $\hat{\theta}_t$ using the following inductive form:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \Sigma_{t+1} \left(\frac{1}{\sigma^2} x_t x_t^{\mathsf{T}} \right) v_t + \Sigma_{t+1} \frac{z_t}{\sigma^2} x_t. \tag{5}$$

Here $v_t \equiv \theta - \hat{\theta}_t$ is a random zero mean vector. Conditional on $\mathcal{F}_{t-1}$, v_t is distributed as $\mathsf{N}(0, \Sigma_t)$ and is independent of x_t and z_t. Recall that the σ-algebra $\mathcal{G}_t = \sigma(\{y_\ell\}_{\ell=1}^{t-1}, \{x_\ell\}_{\ell=1}^{t}) \supseteq \mathcal{F}_{t-1}$. Then we have:

$$\mathbb{E}(\|\hat{\theta}_{t+1}\|^2 | \mathcal{G}_t) = \|\hat{\theta}_t\|^2 + \frac{1}{\sigma^4} \mathbb{E}\left[v_t^{\mathsf{T}} \left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right)^{\mathsf{T}} \left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right) v_t \middle| \mathcal{G}_t \right] + \frac{1}{\sigma^4} \mathbb{E}\left[z_t^2 | \mathcal{G}_t \right] \left(\Sigma_{t+1} x_t \right)^{\mathsf{T}} \left(\Sigma_{t+1} x_t \right). \tag{6}$$

The cross terms cancel since v_t and z_t conditionally on $\mathcal{G}_t$ are independent and zero mean. The expectation in the second term can be reduced as follows:

$$\begin{aligned}
\mathbb{E}\left[v_t^{\mathsf{T}} \left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right)^{\mathsf{T}} \left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right) v_t \middle| \mathcal{G}_t \right] &= \mathsf{Tr}\left[\left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right) \Sigma_t \left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right)^{\mathsf{T}} \right] \\
&= \mathsf{Tr}\left[\left(\Sigma_{t+1} x_t x_t^{\mathsf{T}} \right) \Sigma_t \left(x_t x_t^{\mathsf{T}} \Sigma_{t+1} \right) \right] \\
&= (x_t^{\mathsf{T}} \Sigma_t x_t) \mathsf{Tr}\left[\Sigma_{t+1} x_t x_t^{\mathsf{T}} \Sigma_{t+1} \right] \\
&= \left(x_t^{\mathsf{T}} \Sigma_t x_t \right) \left(x_t^{\mathsf{T}} \Sigma_{t+1}^2 x_t \right).
\end{aligned}$$

The third term can be seen to be:

$$\mathbb{E}\left[z_t^2|\mathcal{G}_t\right](\Sigma_{t+1}x_t)^{\mathsf{T}}(\Sigma_{t+1}x_t) = \sigma^2 x_t^{\mathsf{T}}\Sigma_{t+1}^2 x_t.$$

Thus we have, continuing Eq. (6):

$$\mathbb{E}(\|\hat{\theta}_{t+1}\|^2|\mathcal{G}_t) = \|\hat{\theta}_t\|^2 + \frac{1}{\sigma^4}\left(\sigma^2 + x_t^{\mathsf{T}}\Sigma_t x_t\right)\left(x_t^{\mathsf{T}}\Sigma_{t+1}^2 x_t\right). \tag{7}$$

Since $\Sigma_{t+1} = \left(\Sigma_t^{-1} + \frac{1}{\sigma^2}x_t x_t^{\mathsf{T}}\right)^{-1} = \Sigma_t - \Sigma_t x_t x_t^{\mathsf{T}}\Sigma_t/(\sigma^2 + x_t^{\mathsf{T}}\Sigma_t x_t)$, some calculation yields that:

$$x_t^{\mathsf{T}}\Sigma_{t+1}^2 x_t = \frac{\sigma^4\left(x_t^{\mathsf{T}}\Sigma_t^2 x_t\right)}{\left(\sigma^2 + x_t^{\mathsf{T}}\Sigma_t x_t\right)^2}.$$

Thus Eq. (7) reduces to

$$\mathbb{E}(\|\hat{\theta}_{t+1}\|^2|\mathcal{G}_t) = \|\hat{\theta}_t\|^2 + \frac{x_t^{\mathsf{T}}\Sigma_t^2 x_t}{\sigma^2 + x_t^{\mathsf{T}}\Sigma_t x_t}. \tag{8}$$

We now bound the additive term in Eq. (8). We know that $\Sigma_t \preceq \mathrm{I}/p$ (the prior covariance), thus $x_t^{\mathsf{T}}\Sigma_t x_t \le 1/p$ since $x_t \in \mathcal{X}_p \subseteq \mathsf{Ball}(1)$. Hence the denominator in Eq. (8) is upper bounded by $\sigma^2 + 1/p$. To bound the numerator:

$$\begin{aligned}
\mathbb{E}[x_t^{\mathsf{T}}\Sigma_t^2 x_t|\mathcal{F}_{t-1}] &= \mathbb{E}[\mathsf{Tr}(\Sigma_t^2 x_t x_t^{\mathsf{T}})|\mathcal{F}_{t-1}] \\
&= \mathsf{Tr}[\Sigma_t^2\mathbb{E}(x_t x_t^{\mathsf{T}}|\mathcal{F}_{t-1})] \\
&\ge \frac{\gamma}{p}\mathsf{Tr}(\Sigma_t^2),
\end{aligned}$$

since $\mathbb{E}_{\tilde{x}_t}(x_t x_t^{\mathsf{T}}) \succeq (\gamma/p)\mathrm{I}_p$ by Assumption 1. Using this in Eq. (8), we take expectations to get:

$$\mathbb{E}(\|\hat{\theta}_{t+1}\|^2) \ge \mathbb{E}(\|\hat{\theta}_t\|^2) + \frac{\gamma}{\Delta+1}\mathbb{E}[\mathsf{Tr}(\Sigma_t^2)]. \tag{9}$$

Considering the second term in Eq. (9):

$$\begin{aligned}
\mathbb{E}[\mathsf{Tr}(\Sigma_t^2)] &\ge p\,\mathbb{E}[\det(\Sigma_t^2)^{1/p}] \\
&= p\,\mathbb{E}\left[\left(\prod_{j=1}^{p}\frac{1}{p+\lambda_j/\sigma^2}\right)^{2/p}\right],
\end{aligned}$$

where λ_j is the j^{th} eigenvalue of $\sum_{\ell=1}^{t-1} x_\ell x_\ell^{\mathsf{T}}$. Continuing the chain of inequalities:

$$\begin{aligned}
\mathbb{E}[\mathsf{Tr}(\Sigma_t^2)] &\ge \frac{1}{p}\,\mathbb{E}\left[\prod_{j=1}^{p}\left(1+\frac{\lambda_j}{\Delta}\right)^{-2/p}\right] \\
&\ge \frac{1}{p}\,\mathbb{E}\left[\prod_{j=1}^{p}\exp\left(-\frac{2\lambda_j}{p\Delta}\right)\right] \\
&= \frac{1}{p}\,\mathbb{E}\left[\exp\left\{-\frac{2}{p\Delta}\mathsf{Tr}\Big(\sum_{\ell=1}^{t} x_\ell x_\ell^{\mathsf{T}}\Big)\right\}\right] \\
&\ge \frac{1}{p}\exp\left\{-\frac{2(t-1)}{p\Delta}\right\},
\end{aligned}$$

where the last inequality follows from the fact that $x_\ell \in \mathsf{Ball}(1)$ for each ℓ. Combining this with Eq. (9) gives:

$$\mathbb{E}(\|\hat{\theta}_{t+1}\|^2 \geq \mathbb{E}(\|\hat{\theta}_t\|^2) + \frac{\gamma}{\Delta+1}\frac{1}{p}\exp\left\{-\frac{2(t-1)}{p\Delta}\right\}. \tag{10}$$

Summing over t this implies:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}_t\|^2] &\geq \frac{\gamma}{p(\Delta+1)}\frac{1-\exp\{2(t-1)/p\Delta\}}{1-\exp\{-2/p\Delta\}} \\
&\geq \frac{\gamma\Delta}{2(\Delta+1)}(1-\exp\{-2(t-1)/p\Delta\}) \\
&\geq \frac{\gamma}{2(\Delta+1)}\left(1-\exp\{-2(p\Delta-1)/p\Delta\}\right)\left(\frac{t-1}{p}\right).
\end{aligned}$$

The last inequality follows from fact that $1-\exp(-z)$ is concave in z. Using $p\Delta \geq 2$, we obtain:

$$\begin{aligned}
\mathbb{E}(\|\hat{\theta}_t\|^2) &\geq \frac{\gamma(1-e^{-1})}{2(\Delta+1)}\frac{t-1}{p} \\
&\geq \frac{\gamma}{4(\Delta+1)}\frac{t-1}{p}.
\end{aligned}$$

□

Lemma 6.3 (Sub-Gaussianity of $\|\hat{\theta}_t\|$)**.** *Under the conditions of Theorem 1*

$$\mathbb{P}\left(\|\hat{\theta}_t\| \geq \sqrt{\frac{8(t-1)}{p\Delta}}\nu\right) \leq e^{-(\nu-1)^2/3}.$$

Proof. Note that $\hat{\theta}_t$ is a (vector-valued) martingale. The associated difference sequence given by (cf. Eq. (5))

$$\xi_t = \frac{\langle v_t, x_t\rangle + z_t}{\sigma^2}\Sigma_{t+1}x_t.$$

Note that $\hat{\theta}_t = \sum_{\ell=1}^{t-1}\xi_\ell$. We have that $\mathbb{E}(\xi_t|\mathcal{F}_{t-1}) = 0$. Then conditionally on $\mathcal{G}_t$, $\|\xi_t\| = |w_t|\frac{\Sigma_{t+1}x_t}{\|\Sigma_{t+1}x_t\|}$, where $w_t \equiv \frac{\langle v_t, x_t\rangle + z_t}{\sigma^2}\|\Sigma_{t+1}x_t\|$ is Gaussian with variance given by:

$$\begin{aligned}
\mathrm{Var}(w_t|\mathcal{G}_t) &= \frac{\sigma^2 + x_t^\mathsf{T}\Sigma_t x_t}{\sigma^4}x_t^\mathsf{T}\Sigma_{t+1}^2 x_t \\
&= \frac{x_t^\mathsf{T}\Sigma_t^2 x_t}{\sigma^2 + x_t^\mathsf{T}\Sigma_t x_t} \\
&\leq \frac{1}{p\Delta},
\end{aligned}$$

since $0 \preceq \Sigma_t \preceq \mathrm{I}/p$ and $\|x_t\| \leq 1$. Thus, we have the following "light-tail" condition on ξ_t:

$$\mathbb{E}(e^{\lambda\|\xi_t\|^2}|\mathcal{G}_t) \leq \left(1-\frac{2\lambda}{p\Delta}\right)^{-1/2}.$$

Using $\lambda = p\Delta/4$, we obtain:

$$\mathbb{E}(e^{p\Delta\|\xi_t\|^2/4}|\mathcal{G}_t) \leq \sqrt{2} \leq e.$$

Now using Theorem 2.1 in [JN08] we obtain that:

$$\mathbb{P}\left(\|\hat{\theta}_t\| \geq \sqrt{\frac{8(t-1)}{p\Delta}}(1+\nu)\right) \leq e^{-\nu^2/3},$$

which implies the lemma. $\square$

We can now prove Lemma 6.1. We have:

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_t\|^2] &= \mathbb{E}\big[\|\hat{\theta}_t\|^2 \mathbb{I}_{\|\hat{\theta}_t\|^2 \leq a}\big] + \mathbb{E}\big[\|\hat{\theta}_t\|^2 \mathbb{I}_{\|\hat{\theta}_t\|^2 \geq a}\big] \\ &\leq \sqrt{a}\mathbb{E}[\|\hat{\theta}_t\|] + \int_a^\infty \mathbb{P}(\|\hat{\theta}_t\|^2 \geq y)\mathrm{d}y. \end{aligned} \tag{11}$$

Here we use the fact that $\|\hat{\theta}_t\|$ is a positive random variable. Employing Lemma 6.3 to bound the second term:

$$\begin{aligned}\int_a^\infty \mathbb{P}(\|\hat{\theta}_t\|^2 \geq y)\mathrm{d}y &\leq \frac{8(t-1)}{p\Delta}\int_\alpha^\infty 2\nu e^{-(\nu-1)^2/3}\mathrm{d}\nu, \\ &= \frac{8(t-1)}{p\Delta}\left(\int_\alpha^\infty 2(\nu-1)e^{-(\nu-1)^2/3}\mathrm{d}\nu + 2\int_\alpha^\infty e^{-(\nu-1)^2/3}\mathrm{d}\nu\right) \\ &\leq \frac{8(t-1)}{p\Delta}\frac{3\alpha}{\alpha-1}e^{-(\alpha-1)^2/3}, \end{aligned}$$

where we define $\alpha = \sqrt{ap\Delta/8(t-1)}$. Using this and the result of Lemma 6.2 in Eq. (11)

$$\begin{aligned}\mathbb{E}\|\hat{\theta}_t\| &\geq \left(\frac{C(\gamma,\Delta)}{\alpha}\sqrt{\frac{\Delta}{8}} - \frac{3\sqrt{8}}{(\alpha-1)\sqrt{\Delta}}e^{-(\alpha-1)^2/3}\right)\sqrt{\frac{t-1}{p}} \\ &\geq \left(\frac{C(\gamma,\Delta)}{\alpha}\sqrt{\frac{\Delta}{8}} - \frac{6}{\alpha}\sqrt{\frac{8}{\Delta}}e^{-(\alpha-1)^2/3}\right)\sqrt{\frac{t-1}{p}}, \end{aligned}$$

where the last inequality holds when $\alpha \geq 2$. Using $\alpha(\gamma,\Delta) = 1 + [3\log(96/\Delta C(\gamma,\Delta))]^{1/2} > 2$, the second term in leading constant is half that of the first, and we get the desired result.

7 Proof of Theorem 2

We now consider the large time horizon of $t > p\Delta$ for strategy BALLEXPLORE , assuming the special case $\mathcal{X}_p = \mathsf{Ball}(1)$. Throughout, we will adopt the notation $\bar{\beta}_t^2 = 1 - \beta_t^2$. To begin, we bound the mean squared error in estimating θ using the following

Lemma 7.1 (Upper bound on Squared Error)**.** *Under the conditions of Theorem 2 we have* $\forall\, t \geq p\Delta + 1$*:*

$$\mathbb{E}(\mathsf{Tr}(\Sigma_t)) \leq C_4(\Delta)\sqrt{\frac{p}{t}},$$

where $C_4(\Delta) \equiv 3(\Delta+1)/\sqrt{\Delta}$.

Proof. As $\Sigma_t = (\Sigma_{t-1}^{-1} + \frac{1}{\sigma^2}x_t x_t^{\mathsf{T}})^{-1}$, we use the inversion lemma to get:

$$\begin{aligned}\mathsf{Tr}(\Sigma_t) &= \mathsf{Tr}(\Sigma_{t-1}) - \frac{x_t^{\mathsf{T}}\Sigma_{t-1}^2 x_t}{\sigma^2 + x_t^{\mathsf{T}}\Sigma_{t-1}x_t} \\ &\leq \mathsf{Tr}(\Sigma_{t-1}) - \frac{p}{(\Delta+1)}x_t^{\mathsf{T}}\Sigma_{t-1}^2 x_t, \end{aligned}$$

where the inequality follows from $\Sigma_{t-1} \preceq \mathrm{I}_p/p$ and $\|x_t\|^2 \le 1$ for each ℓ. Using $x_t = \bar{\beta}_t \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|} + \beta_t \mathsf{P}_t^{\perp} u_t$ and taking expectations on either side, we obtain:

$$\begin{aligned}\mathbb{E}(\mathsf{Tr}(\Sigma_t)) &\le \mathbb{E}(\mathsf{Tr}(\Sigma_{t-1})) - \frac{p}{\Delta+1}\left[\left(\bar{\beta}_t^2 - \frac{\beta_t^2}{p}\right)\frac{\hat{\theta}_t^{\mathsf{T}}\Sigma_{t-1}^2\hat{\theta}_t}{\|\hat{\theta}_t\|^2} + \frac{\beta_t^2}{p}\mathbb{E}(\mathsf{Tr}(\Sigma_{t-1}^2))\right] \\ &\le \mathbb{E}(\mathsf{Tr}(\Sigma_{t-1})) - \frac{\beta_t^2}{\Delta+1}\mathbb{E}(\mathsf{Tr}(\Sigma_{t-1}^2)),\end{aligned}$$

where we used $\bar{\beta}_t^2 - \beta_t^2/p \ge 0$. This follows because $\beta_t^2 \le 2/3 \le p/(p+1)$ when $t \ge p\Delta$ and $p \ge 2$. Employing Cauchy-Schwartz twice and using substituting for β_t^2 we get the following recursion in $\mathbb{E}(\mathsf{Tr}(\Sigma_t))$:

$$\mathbb{E}(\mathsf{Tr}(\Sigma_t)) \le \mathbb{E}(\mathsf{Tr}(\Sigma_{t-1})) - \frac{2\sqrt{\Delta}}{3(\Delta+1)}\frac{1}{\sqrt{pt}}[\mathbb{E}(\mathsf{Tr}(\Sigma_{t-1}))]^2. \tag{12}$$

The function $f(z) = z - z^2/b$ is increasing z when $z \in (0, b/2)$. For the recursion above:

$$\begin{aligned} b = b(t) &= \frac{3}{2}\sqrt{\frac{pt}{\Delta}}(\Delta+1) \\ &> p(\Delta+1) \\ &\ge 4, \end{aligned}$$

since $p\Delta \ge 2$ and $p \ge 2$. Also, we know that $\Sigma_t \preceq \mathrm{I}_p/p$ and hence $\mathsf{Tr}(\Sigma_t) \le 1$ with probability 1 and that $\mathbb{E}(\mathsf{Tr}(\Sigma_t))$ is decreasing in t. Thus the right hand side of the recursion is increasing in its argument. A standard induction argument then implies that $\mathbb{E}(\mathsf{Tr}(\Sigma_t))$ is bounded pointwise by the solution to the following equation:

$$y(t) = y(t_0) - c\int_{t_0}^{t}\frac{y^2(s)}{\sqrt{s}}\mathrm{d}s,$$

with the initial condition $t_0 = p\Delta,\ y(t_0) = 1$, where $c = 2\sqrt{\Delta}/3(\Delta+1)\sqrt{p}$. The solution is explicitly computed to yield:

$$\mathbb{E}(\mathsf{Tr}(\Sigma_t)) \le \left[1 + \frac{c'}{2}\left(\sqrt{\frac{t}{p}} - \sqrt{\Delta}\right)\right]^{-1},$$

where $c' = c\sqrt{p} = 2\sqrt{\Delta}/3(\Delta+1)$. Since the constant term is always positive, we can remove it and obtain the required result.

□

We can now prove the following result:

Lemma 7.2. *For all $t > p\Delta$, under the conditions of Theorem 2:*

$$\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|} - \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right] \le 12(\Delta+1)\sqrt{\frac{e}{\Delta}}\left(\frac{p}{t}\right)^{1/2-1/2(p+2)}.$$

Proof. Using the linearity of expectation:

$$\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|} - \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right] \;\le\; \mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|} - \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\mathbb{I}(\|\theta\| < \varepsilon)\right] + \mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|} - \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\mathbb{I}(\|\theta\| \ge \varepsilon)\right]. \tag{13}$$

We bound the first term as follows:

$$\begin{aligned}
\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\mathbb{I}(\|\theta\|<\varepsilon)\right] &\leq \mathbb{E}\left[\|\theta\|\left\|\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right\|\mathbb{I}(\|\theta\|\leq\varepsilon)\right] \\
&\leq 2\varepsilon\mathbb{P}(\|\theta\|\leq\varepsilon) \\
&\leq 2\varepsilon^{p+1}e^{p/2}.
\end{aligned}$$

The first inequality is Cauchy-Schwartz, the second follows from bounds on the norm of either vectors while the third is a standard Chernoff bound computation using the fact that $\theta \sim \mathsf{N}(0, \mathrm{I}_p/p)$. The second term can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\mathbb{I}(\|\theta\|\geq\varepsilon)\right] &\leq \mathbb{E}\left(\frac{2\|\theta-\hat{\theta}_t\|^2}{\|\theta\|}\mathbb{I}(\|\theta\|\geq\varepsilon)\right) \\
&\leq \frac{2}{\varepsilon}\mathbb{E}(\|\theta-\hat{\theta}_t\|^2) \\
&\leq \frac{2}{\varepsilon}\mathbb{E}\left(\mathsf{Tr}\Sigma_t\right).
\end{aligned}$$

The first inequality follows from Lemmas 3.5 and 3.6 of [RT10], the second follows from the fact that $\|\theta-\hat{\theta}_t\|^2$ is nonnegative and the indicator is used. Combining the bounds above and Lemma 7.1 we get:

$$\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right] \leq 2\varepsilon^{p+1}e^{p/2}+\frac{2C_4(\Delta)}{\varepsilon}\sqrt{\frac{p}{t}}.$$

Optimizing over ε we obtain:

$$\begin{aligned}
\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right] &\leq 4\left(C_4(\Delta)e^{1/2}\sqrt{\frac{p}{t}}\right)^{1-1/(p+2)} \\
&\leq 4\,e^{1/2}C_4(\Delta)\left(\frac{p}{t}\right)^{1/2-1/2(p+2)}.
\end{aligned}$$

□

We using Lemma 7.2 we can now prove Theorem 2 for the large time horizon. Let ρ_t denote the expected regret incurred by SmoothExplore at time $t > p\Delta$. By definition, we write it as:

$$\begin{aligned}
\rho_t &= \mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\bar{\beta}_t\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}-\beta_t\mathsf{P}_t^{\perp}u_t\right)\right] \\
&= \mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\bar{\beta}_t\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right],
\end{aligned}$$

as u_t is zero mean conditioned on past observations. We split the first term in two components to get:

$$\rho_t \leq (1-\bar{\beta}_t)\mathbb{E}\|\theta\| + \bar{\beta}_t\mathbb{E}\left[\theta^{\mathsf{T}}\left(\frac{\theta}{\|\theta\|}-\frac{\hat{\theta}_t}{\|\hat{\theta}_t\|}\right)\right]$$

We know that $0 \leq 1-\bar{\beta}_t \leq \beta_t^2 = \sqrt{4p\Delta/9t}$. We use this and the result of Lemma 7.2 to bound the right hand side above as:

$$\rho_t \leq \frac{2}{3}\left(\frac{p\Delta}{t}\right)^{1/2} + 12(\Delta+1)\sqrt{\frac{e}{\Delta}}\left(\frac{p}{t}\right)^{1/2-\omega(p)},$$

where we define $\omega(p) \equiv 1/(2(p+2))$. Summing over the relevant interval and bounding by the corresponding integrals, we obtain:

$$\sum_{\ell=p\Delta+1}^{t} \rho_\ell \le \frac{4\sqrt{\Delta}}{3}(pt)^{1/2} + 24(\Delta+1)\sqrt{\frac{e}{\Delta}}(pt)^{1/2+\omega(p)}$$
$$\le C_3(\Delta)(pt)^{1/2+\omega(p)},$$

where $C_3(\Delta) = 4\sqrt{\Delta}/3 + 24(\Delta+1)\sqrt{e}/\sqrt{\Delta}$ and $\omega(p) = 1/2(p+2)$. We can use $C_3(\Delta) \equiv 70(\Delta+1)/\sqrt{\Delta}$ for simplicity.

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.

A Properties of the set of arms

A.1 Proof of Lemma 4.1

We let $\mathcal{X}'_p = \partial\mathsf{Ball}(\rho/\sqrt{3})$, where ∂S denotes the boundary of a set S. For each $x \in \mathcal{X}'_p$ denote the projection orthogonal to it by $\mathsf{P}_x^\perp$. We use the distribution $\mathbb{P}_x(z)$ induced by:

$$z = x + \sqrt{\frac{2}{3}}\rho\,\mathsf{P}_x^\perp u,$$

where u is chosen uniformly at random on the unit sphere. This distribution is in fact supported on $\mathsf{Ball}(\rho) \subseteq \mathcal{X}_p$. Also, we have, for all $x \in \mathcal{X}'_p$, $\mathbb{E}_x(z) = x$. Computing the second moment:

$$\begin{aligned}\mathbb{E}_x(zz^\mathsf{T}) &= \mathbb{E}\left(xx^\mathsf{T} + \frac{2\rho^2}{3}\mathsf{P}_x^\perp uu^\mathsf{T}\mathsf{P}_x^\perp\right)\\ &= xx^\mathsf{T} + \frac{2\rho^2}{3p}\mathsf{P}_x^\perp\\ &= \frac{2\rho^2}{3p}\mathrm{I}_p + \left(1-\frac{2}{p}\right)xx^\mathsf{T}\\ &\succeq \frac{2\rho^2}{3p}\mathrm{I}_p,\end{aligned}$$

where in the first equality we used linearity of expectation, and that the projection mapping is idempotent. This yields $\gamma = 2\rho^2/3$. Since $\mathcal{X}'_p = \partial\mathsf{Ball}(\rho/\sqrt{3})$ we obtain $\kappa = \inf_{\{\theta:\|\theta\|=1\}}\sup_{\{x\in\mathcal{X}'_p\}}\langle\theta, x\rangle = \rho/\sqrt{3}$. Thus this construction satisfies Assumption 1. Note the fact that BallExplore is a special case of SmoothExplore follows from the fact that we can use $\rho = 1$ above when $\mathcal{X}_p = \mathsf{Ball}(1)$.

A.2 Proof of Proposition 4.2

Throughout we will denote by $\mathrm{conv}(S)$ the convex hull of set S, and by $\overline{\mathrm{conv}}(S)$ its closure. Also, it is sufficient to consider Assumption 1.2 for $\|\theta\| = 1$.

It is immediate to see that $\mathsf{Ball}(\kappa) \subseteq \overline{\mathrm{conv}}(\mathcal{X}'_p)$ implies Assumption 1.2. Indeed

$$\begin{aligned}\sup\left\{\langle\theta, x\rangle :\ x \in \mathcal{X}'_p\right\} &= \sup\left\{\langle\theta, x\rangle :\ x \in \mathrm{conv}(\mathcal{X}'_p)\right\}\\ &= \max\left\{\langle\theta, x\rangle :\ x \in \overline{\mathrm{conv}}(\mathcal{X}'_p)\right\}\\ &\ge \max\left\{\langle\theta, x\rangle :\ x \in \mathsf{Ball}(\kappa)\right\} \ge \kappa\|\theta\|,\end{aligned}$$

where the last inequality follows by taking $x = \kappa\theta/\|\theta\|$.

In order to prove the converse, let

$$\kappa_0 \equiv \sup\Big\{\rho : \ \mathsf{Ball}(\rho) \in \mathrm{conv}(\mathcal{X}'_p)\Big\}.$$

We then have $\mathsf{Ball}(\kappa_0) \subseteq \overline{\mathrm{conv}}(\mathcal{X}'_p)$. Assume by contradiction that $\kappa_0 < \kappa$. Then there exists at least one point x_0 on the boundary of $\overline{\mathrm{conv}}(\mathcal{X}'_p)$ such that $\|x_0\| = \kappa_0$ (else κ_0 would not be the supremum).

By the supporting hyperplane theorem, there exists a closed half space $\mathcal{H}$ in $\mathbb{R}^p$ such that $\overline{\mathrm{conv}}(\mathcal{X}'_p) \subseteq \mathcal{H}$ and x_0 is on the boundary $\partial\mathcal{H}$ of $\mathcal{H}$. It follows that $\mathsf{Ball}(\kappa_0) \subseteq \mathcal{H}$ has well, and therefore $\partial\mathcal{H}$ is tangent to the ball at x_0. Summarizing

$$\overline{\mathrm{conv}}(\mathcal{X}'_p) \subseteq \mathcal{H} \equiv \Big\{ x \in \mathbb{R}^p \, : \langle x, x_0\rangle \le \kappa_0\|x_0\| \Big\}. \tag{14}$$

By taking $\theta = x_0/\|x_0\|$, we then have, for any $x \in \overline{\mathrm{conv}}(\mathcal{X}'_p)$, $\langle \theta, x\rangle \le \kappa_0 < \kappa$, which is in contradiction with Assumption 1.2.

A.3 Proof of Proposition 4.3

A.3.1 Proof of condition 1

Choose $\mathcal{X}'_p = \mathcal{X}_p \cap \mathsf{Ball}(\rho)$. We first prove that $f(\theta) \equiv \max_{x\in\mathcal{X}'_p}\langle \theta, x\rangle$ is Lipschitz continuous with constant ρ. Then, employing an υ-net argument, we prove that this choice of $\mathcal{X}'_p$ satisfies Assumption 1.1 with high probability.

Let $f(\theta_i) = \langle \theta_i, x_i\rangle$ for $i = 1, 2$. Without loss of generality, assume $f(\theta_1) > f(\theta_2)$. We then have:

$$\begin{aligned}
|f(\theta_1) - f(\theta_2)| &= |\langle \theta_1, x_1\rangle - \langle \theta_2, x_2\rangle| \\
&= |\langle \theta_1, x_1\rangle - \langle \theta_2, x_1\rangle + \langle \theta_2, x_1\rangle - \langle \theta_2, x_2\rangle| \\
&\le |\langle \theta_1 - \theta_2, x_1\rangle| \\
&\le \|x_1\|\|\theta_1 - \theta_2\| \\
&\le \rho\|\theta_1 - \theta_2\|,
\end{aligned}$$

where the first inequality follows since x_2 maximizes $\langle \theta_2, x_2\rangle$, the second is Cauchy-Schwarz and the third from the fact that $x_1 \in \mathcal{X}_p \cap \mathsf{Ball}(\rho)$.

Since $f(\theta) = \|\theta\| f(\theta/\|\theta\|)$, it suffices to consider θ on the unit sphere S_p. Suppose Υ is an υ-net of the unit sphere, i.e. a maximal set of points that are separated from each other by at least υ. We can bound $|\Upsilon|$ by a volume packing argument: consider balls of radius $\upsilon/2$ around every point in Υ. Each of these is disjoint (by the property of an υ-net) and, by the triangle inequality, are all contained in a ball of radius $1 + \upsilon/2$. The latter has a volume of $(1 + 2\upsilon^{-1})^p$ times that of each of the smaller balls, thus yielding that $|\Upsilon| \le (1 + 2\upsilon^{-1})^p$.

Now, $|\mathcal{X}'_p|$ is binomial with mean $M\rho^p$ and variance $M\rho^p(1-\rho^p)$. Consider a single point $\theta \in \Upsilon$. Due to rotational invariance we may assume $\theta = e_1$, the first canonical basis vector. Conditional on the event $E_n = \{\omega : \ |\mathcal{X}'_p| = n\}$, the arms in $\mathcal{X}'_p$ are uniformly distributed in $\mathsf{Ball}(\rho)$. Thus we have (assuming $z > 0$):

$$\begin{aligned}
\mathbb{P}(\max_{x\in\mathcal{X}'_p}\langle x, e_1\rangle \le z\rho|E_n) &= \prod_{j=1}^{n}\mathbb{P}(\langle x_j, e_1\rangle \le z\rho|E_n) \\
&= \left(\mathbb{P}(\langle x_1, e_1\rangle \le z\rho|x_1 \in \mathsf{Ball}(\rho)\right)^n \qquad (15)\\
&= \left(\mathbb{P}(\langle x_1, e_1\rangle \le z\right)^n \qquad (16)\\
&, \qquad (17)
\end{aligned}$$

since the $\langle x_j, e_1\rangle$, $j \in \{1, \ldots, n\}$ are iid, and the conditional distribution of x_1 given $x \in \mathsf{Ball}(\rho)$ is the same as the unconditional distribution of $\rho\, x$. Let $Y_1 \cdots Y_p \sim \mathsf{N}(0, 1/2)$ be iid and $Z \sim \mathrm{Exp}(1)$ be independent of the

Y_i. Then by Theorem 1 of [BGMN05] $\langle x_1, e_1\rangle$ is distributed as $\rho Y_1/(\sum_{i=1}^p Y_i^2+Z)^{1/2}$. By a standard Chernoff argument, $\mathbb{P}(\sum_{i=2}^p Y_i^2 \geq 2(p-1)) \leq \exp\{-c(p-1)\}$ where $c = (\log 2 - 1)/2$. Also, $\mathbb{P}(Z \geq p) = \exp(-p)$ and $\mathbb{P}(Y_1^2 \geq p) \leq 2\exp(-p)$. This allows us the following bound:

$$\begin{aligned}\mathbb{P}(\langle x_1, e_1\rangle \leq z) &= \mathbb{P}\left(\frac{Y_1}{\sum_{i=1}^p Y_i^2 + Z} \leq z\right) \\ &\leq \nu(p) + (1-\nu(p))\mathbb{P}\left(\frac{Y_1}{\sqrt{4p-2}} \leq z \Big| Y_1^2 \leq p\right),\end{aligned}$$

where $\nu(p) \equiv 3\exp(-p) + \exp(-c(p-1))$. We further simplify to obtain:

$$\begin{aligned}\mathbb{P}(\langle x_1, e_1\rangle \leq z) &\leq 1 - (1-\nu(p))\mathbb{P}\left(\frac{Y_1}{\sqrt{4p-2}} \geq z \Big| Y_1^2 \leq p\right) \\ &\leq 1 - (1-\nu(p))\left(F_G(\sqrt{2p}) - F_G(z\sqrt{8p-4})\right),\end{aligned}$$

and $F_G(\cdot)$ denotes the Gaussian cumulative distribution function. Employing this in Eq. (17):

$$\begin{aligned}\mathbb{P}(\max_{x\in\mathcal{X}_p'}\langle x, e_1\rangle \leq z\rho | E_n) &\leq \left[1 - (1-\nu(p))\left(F_G(\sqrt{2p}) - F_G(z\sqrt{8p-4})\right)\right]^n \\ &\leq \exp\left[-n(1-\nu(p))(F_G(\sqrt{2p}) - F_G(z\sqrt{8p-4})\right].\end{aligned}$$

For $p \geq 6$, we have that $1 - \nu(p) \geq 1/2$ and $F_G(\sqrt{2p}) - F_G(\sqrt{8p-4}) \geq 3^{-p}/2$. Using this, substituting $z = 1/2$ and that $|\mathcal{X}_p'| \geq M\rho^p/2$ with probability at least $1 - \exp(-M\rho^p/8)$ we now have:

$$\mathbb{P}(\max_{x\in\mathcal{X}_p'}\langle x, e_1\rangle \leq \rho/2) \leq \exp(-M\rho^p 3^{-p}/4) + \exp(-M\rho^p/8)$$

We may now union bound over Υ using rotational invariance to obtain:

$$\mathbb{P}(\min_{\theta\in\Upsilon}\max_{x\in\mathcal{X}_p'}\langle x, \theta\rangle \leq \rho/2) \leq (1+2\upsilon^{-1})^p(\exp(-M\rho^p 3^{-p}/4) + \exp(-M\rho^p/8))$$

Using $\rho = 1/2$, $\upsilon = 1/2$, $M = 8^p$ and that $f(\theta)$ is Lipschitz, we then obtain:

$$\begin{aligned}\mathbb{P}(\min_{\|\theta\|=1}\max_{\mathcal{X}_p'}\langle x, \theta\rangle \leq 1/4) &\leq 5^p[\exp(-4^{p-1}/3^p) + \exp(-4^p/8)] \\ &\leq \exp(-p),\end{aligned}$$

when $p \geq 20$.

A.3.2 Proof of condition 2

Fix radii ρ and δ such that $\rho + \delta \leq 1$. We choose the $\mathcal{X}_p'$ to be $\mathcal{X}_p \cap \mathsf{Ball}(\rho)$. Consider a point x such that $\|x\| \leq 1 - \delta$. We consider the events E_i, D_i:

$$\begin{aligned}E_i &\equiv \{\nexists \text{ a distribution } \mathbb{P}_{x_i} \text{ satisfying Assumption 1.2}\} \\ D_i &\equiv \{x_i \in \mathsf{Ball}(\rho)\}\end{aligned}$$

We now bound $\mathbb{P}(E_i|D_i)$. Within a distance δ around x_i, there will be, in expectation, $M\delta^p$ arms (assuming the total number of points is $M+1$). Indeed the distribution of the number of arms within distance δ around x_i is binomial with mean $M\delta^p$ and variance $M\delta^p(1-\delta^p)$.

Conditional on the number of arms in $\mathsf{Ball}(\delta, x_i)$ being n, these arms are uniformly distributed in $\mathsf{Ball}(\delta, x_i)$ and are independent of x_i. We will use $\mathbb{P}_n$ to denote this conditional probability measure. Denote the arms within distance δ from x_i to be $v_1, v_2 \dots v_n$. Define $u_j \equiv v_j - x_j$, $\bar{u} \equiv (\sum_{j=1}^n u_j)/n$ for all j and $Q = (\sum_{j=1}^n u_j u_j^\mathsf{T})/n$. To construct the probability distribution $\mathbb{P}_{x_j}$, we let the weight w_j on the arm v_j to be:

$$w_j = \frac{1}{n}\left(\frac{1 - u_j^\mathsf{T} Q^{-1}\bar{u}}{1 - \bar{u}^\mathsf{T} Q^{-1}\bar{u}}\right)$$

It is easy to check that these weights yield the correct first moment, i.e. $\sum_{j=1}^n w_j v_j = x_i$. Before considering the second moment, we first show that Q concentrates around its mean. It is straightforward to compute that $\mathbb{E}(Q) = \mathbb{E}(u_1 u_1^\mathsf{T}) = \mu \mathrm{I}_p$, where $\mu = \delta^2/(p+2)$. By the matrix Chernoff bound [AW02, Tro12], there exist $c > 0$ such that:

$$\mathbb{P}_n(\|Q^{-1}\| \geq \frac{2}{\mu}) \leq p \exp(-cn\mu/\delta^2), \tag{18}$$

where $\|Q\|$ denotes the operator norm and the probability is over the distribution of the u_j. We further have, for all j:

$$w_j \geq \frac{1}{n}\left(1 - \|u_j\|\|Q^{-1}\bar{u}\|\right) \geq \frac{1}{n}\left(1 - \delta\|Q^{-1}\|\|\bar{u}\|\right). \tag{19}$$

Also, using Theorem 2.1 of [JN08] we obtain that:

$$\begin{aligned}\mathbb{P}_n(\|\bar{u}\| \geq \delta/n^{1/4}) &\leq \exp\left\{-\frac{(n^{1/4}-1)^2}{2}\right\} \\ &\leq \exp(-n^{1/2}/4),\end{aligned}$$

for $n \geq 16$. Combining this with Eq. (18) and continuing inequalities in Eq. (19), we obtain, for all j,:

$$w_j \geq \frac{1}{n}\left(1 - \frac{2(p+2)}{n^{1/4}}\right), \tag{20}$$

with probability at least $1 - \omega(n,p)$ where $\omega(n,p) = p\exp(-cn/2p) + \exp(-n^{1/2})$. We can now bound the second moment of $\mathbb{P}_x$:

$$\begin{aligned}\sum_{j=1}^n w_j v_j v_j^\mathsf{T} &= \sum_{j=1}^n w_j u_j u_j^\mathsf{T} + xx^\mathsf{T} \\ &\succeq \sum_{j=1}^n w_j u_j u_j^\mathsf{T} \\ &\succeq \left(\frac{1}{2} - \frac{p+2}{n^{1/4}}\right)\frac{\delta^2}{p+2}\mathrm{I}_p\end{aligned}$$

where the last inequality holds with probability at least $1 - \omega(n,p)$. Thus we can obtain $\gamma = \delta^2/8$ for $n \geq [4(p+2)]^4$.

In addition, a standard Chernoff bound argument yields that the number of arms in $\mathsf{Ball}(\delta, x_i)$ is at least $M\delta^p/2$ with probability at least $1 - \exp(-M\delta^p/8)$. With this, we can bound $\mathbb{P}(E_i|D_i)$:

$$\mathbb{P}(E_i|D_i) \leq \exp(-M\delta^p/8) + \omega(M\delta^p/2, p).$$

The event F that the uniform cloud does not satisfy Assumption 1.2 can now be decomposed as follows:

$$
\begin{aligned}
\mathbb{P}(F) &= \mathbb{P}\left(\bigcup_{i=1}^{M}(E_i \cap D_i)\right) \\
&\le \sum_{i=1}^{M+1} \mathbb{P}(E_i|D_i)\mathbb{P}(D_i) \\
&\le 2M\rho^p\{\exp(-M\delta^p/8) + \omega(M\delta^p/8, p)\}
\end{aligned}
$$

Choosing $\delta = \rho = 1/2$, with $M = 8^p$, we get that the uniform cloud satisfies Assumption 1.2 with $\gamma \ge \delta^2/8 = 1/32$ with probability at least $1 - 2 \cdot 4^p[\exp(-4^p/8) + \omega(4^p/8, p)] \ge 1 - \exp(-p)$ when $p \ge 10$. Summarizing the proofs of both conditions we have, choosing the number of points $M = 8^p$, the subset $\mathcal{X}'_p = \mathcal{X}_p \cap \mathsf{Ball}(1/2)$, we obtain constants $\kappa = 1/4$ and $\gamma = 1/32$ with probability at least $1 - 2\exp(-p)$, provided $p \ge 20$.

References

[Aue02] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:2002, 2002.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569 – 579, March 2002.

[AYPS11] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

[BGMN05] F. Barthe, O. Guédon, S. Mendelson, and A. Naor. A probabilistic approach to the geometry of the lpn-ball. *The Annals of Probability*, 33(2):480–513, 2005.

[BK07] R. M. Bell and Y. Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 43–52, Washington, DC, USA, 2007. IEEE Computer Society.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundation of computational mathematics*, 9(6):717–772, February 2009.

[CT10] E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[DHK08] V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

[DM13] Y. Deshpande and A. Montanari. Implementing an interactive recommendation system. In preparation, 2013.

[Gro09] D. Gross. Recovering low-rank matrices from few coefficients in any basis. arXiv:0910.1879, 2009.

[JN08] A. Juditsky and A. Nemirovski. Large Deviations of Vector-Valued Martingales in 2-Smooth Normed Spaces. arXiv:0809.0813, 2008.

[KBV09] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[KLT11] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39:2302–2329, 2011.

[KMO10a] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, June 2010.

[KMO10b] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, July 2010.

[Kor08] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, New York, NY, USA, 2008. ACM.

[RT10] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

[SJ03] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *20th International Conference on Machine Learning*, pages 720–727. AAAI Press, 2003.

[SPUP02] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. 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*, pages 253–260, 2002.

[SRJ05] N. Srebro, J. D. M. Rennie, and T. S. Jaakola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, 2005.

[SW06] 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.

[Tro12] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[ZH08] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM conference on Recommender Systems*, 2008.