# Ensemble sampling for linear bandits: small ensembles suffice

**David Janz** DAVID.JANZ93@GMAIL.COM
*University of Alberta*

**Alexander E. Litvak** ALITVAK@UALBERTA.CA
*University of Alberta*

**Csaba Szepesvári** SZEPESVA@UALBERTA.CA
*University of Alberta*

## Abstract

We provide the first useful, rigorous analysis of ensemble sampling for the standard linear bandit setting. In particular, we show that for a $d$-dimensional linear bandit with an interaction horizon $T$, ensemble sampling with an ensemble of size $m$ on the order of $d \log T$ matches the standard regret bound for Thompson sampling up to a multiplicative factor of order $m\sqrt{\log T}$. Ours is the first result in any structured setting not to require the size of the ensemble to scale linearly with $T$ for near $\sqrt{T}$ order regret—which defeats the purpose of ensemble sampling—and the first that does not require a finite arm set.

## 1 Introduction

Ensemble sampling, as christened by Lu and Van Roy (2017), is a family of randomised algorithms for balancing exploration-and-exploitation in sequential decision making. The premise of the approach is that an ensemble of perturbed models of the value of the available decisions (actions, arms) is maintained, and the decision taken at each step of intersection is that which is optimal with respect to a randomly selected ensemble element (model).

Ensemble sampling can be seen as an approximation to the classic Thompson sampling algorithm (Thompson, 1933), also known as posterior sampling. Whereas Thompson sampling maintains a posterior distribution over models, and samples a new model from this distribution at each step, ensemble sampling can be thought to approximate this distribution with a finite, unweighted ensemble, which is updated incrementally—and in randomly selecting a model from this ensemble, ensemble sampling can be thought of as Thompson sampling that periodically reuses previously sampled models.

The advantage of ensemble sampling over Thompson sampling whenever incrementally updating the ensemble is cheap, but computing a posterior distribution and sampling from it is expensive. A classic example of this setting is in deep reinforcement learning, where the models—neural networks—are large, but trained incrementally. Here, ensemble sampling is used directly under the names of Bootstrapped DQN (Osband et al., 2016) and Ensemble+ (Osband et al., 2018), and as part of other reinforcement learning algorithms (say, in Dimakopoulou and Van Roy, 2018; Curi et al., 2020). Ensemble sampling has also been applied to online recommendation (Lu et al., 2018; Hao et al., 2020; Zhu and Van Roy, 2021), in behavioural sciences (Eckles and Kaptein, 2019) and marketing (Yang et al., 2020).

However, despite the practicality and seemingly simple nature of the ensemble sampling algorithm, we have no theoretical explanation for its performance. Here, the issue is that

the dependencies introduced by reusing models between time steps significantly complicate the analysis. Indeed, Qin et al. (2022) state that:

> *A lot of work has attempted to analyze ensemble sampling, but none of them has been successful.*

Our contribution is the first successful analysis of ensemble sampling. Our analysis follows broadly that of Thompson sampling given by Abeille and Lazaric (2017), but does not recover quite the same regret bound. We leave eliminating the slack (or showing that it cannot be done) for future work. While a little technical in places, our analysis is conceptually simple, and can, with a bit of effort, be extended beyond the linear setting. In particular, immediate extensions include generalised linear bandits (Filippi et al., 2010), kernelised bandits/Gaussian-process-based Bayesian optimisation (Srinivas et al., 2010), and deep learning—with the latter via the usual neural tangent kernel approach (Jacot et al., 2018).

## 2 Problem setting, formalism and notation

We now introduce, in turn, some general notation, the linear stochastic bandit setting that we consider, the relevant ridge regression estimates and their properties, and the probabilistic formalism which we shall adopt—the last of these is particularly important, for much of the difficulty in analysing ensemble sampling lies in having to work with conditional expectations.

**General notation** We denote by $\mathbb{N}^+$ the set of positive natural numbers and for $m \in \mathbb{N}^+$, we write $[m] = \{1, \dots, m\}$. For a vectors $v, u$ in $\mathbb{R}^\ell$, we denote by $\|v\|_2$ the canonical Euclidean norm of $u$ and by $\langle v, u\rangle$ the canonical Euclidean inner product between $v$ and $u$. $B_2^\ell$ denotes the closed canonical Euclidean unit ball in $\mathbb{R}^\ell$. $I_\ell$ denotes the identity matrix in $\mathbb{R}^{\ell\times\ell}$, and $0_\ell$ zero element of $\mathbb{R}^\ell$. For a matrix $M \in \mathbb{R}^{\ell\times k}$, $\|M\|$ denotes its operator norm from $(\mathbb{R}^k, \|\cdot\|_2)$ to $(\mathbb{R}^\ell, \|\cdot\|_2)$; whenever $k = \ell$ and $M$ is positive definite, $\|v\|_M$ denotes the $M$-weighted canonical Euclidean norm, given by $\|v\|_M^2 = \langle v, Mv\rangle$. For positive semidefinite matrices $A, B$ of matching dimensions, $A \preceq B$ denotes the usual semidefinite order.

**Problem setting** We consider the standard stochastic linear bandit setting. At each step $t \in [T]$, for a horizon length $T \in \mathbb{N}^+$, a learner selects an action $X_t$ from an arm set $\mathcal{X}$, a closed subset of the $d$-dimensional Euclidean unit ball $B_2^d$, and receives a random reward $Y_t \in \mathbb{R}$ of the form

$$Y_t = \langle X_t, \theta^\star\rangle + Z_t, \tag{1}$$

where $\theta^\star \in B_2^d$ is an unknown weight vector and $Z_t$ is a zero-mean 1-sub-Gaussian random variable independent on the past (see 'probabilistic formalism' for definition). The aim of the learner is to minimise its regret over the horizon, the quantity

$$R(T) = \max_{x\in\mathcal{X}} \sum_{t=1}^{T} \langle x - X_t, \theta^\star\rangle, \tag{2}$$

while ours will be to show a high probability bound on $R(T)$ that holds uniformly over $\theta^\star \in B_2^d$ when the learner uses the ensemble sampling algorithm, detailed shortly.

**Ridge regression** Algorithms we consider will estimate $\theta^\star$ using ridge regression. For a regularisation parameter $\lambda > 0$, ridge regression gives the estimate

$$\hat{\theta}_t = V_t^{-1} \sum_{i=1}^{t} X_i Y_i \quad \text{where} \quad V_t = V_0 + \sum_{i=1}^{t} X_i X_i^\mathsf{T} \quad \text{and} \quad V_0 = \lambda I, \tag{3}$$

and where we take $\hat{\theta}_0 = 0_d$. Importantly, the ridge regression gives a *good* estimator for $\theta^\star$, in that under our assumptions, for any $\delta > 0$, with probability $1 - \delta$, for all $t \in \mathbb{N}^+$, $\theta^\star \in \mathcal{C}_t(\delta)$, where

$$\mathcal{C}_t(\delta) = \hat{\theta}_t + \beta_t V_t^{-1/2} B_2^d \quad \text{with} \quad \beta_t = \sqrt{\lambda} + \sqrt{2\log(1/\delta) + \log(\det(V_t)/\lambda^d)}. \tag{4}$$

The above confidence sets were introduced to the bandit literature by Abbasi-Yadkori et al. (2011), and their construction relies on the method of Peña et al. (2009) and de la Pena et al. (2004) (see Chapter 20 of Lattimore and Szepesvári (2020) for an overview of this construction and result). We will make generous use of the related map

$$\psi_t(u) = \hat{\theta}_t + \beta_t V_t^{-1/2} u \quad \text{for} \quad u \in \mathbb{R}^d, \quad \text{observing in particular that} \quad \psi_t(B_2^d) = \mathcal{C}_t(\delta). \tag{5}$$

**Probabilistic formalism** We let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{N}}, \mathbb{P})$ be a complete filtered probability space. We assume the following measurability:

> We take each $\mathcal{F}_t$ such that $X_t$, $Y_t$ are $\mathcal{F}_t$-measurable—this makes $V_t$, $\beta_t$ and $\hat{\theta}_t$ likewise $\mathcal{F}_t$-measurable. Each $\mathcal{F}_t$ will shortly be expanded to contain the randomisation used by the algorithms.

We will use the shorthands $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mathbb{P}_t(A) = \mathbb{E}_t \mathbf{1}[A]$ for $A \in \mathcal{F}$, where $\mathbf{1}[A]$ denotes the characteristic function of $A$, with inequalities of random variables here and henceforth understood to hold in an almost sure sense.

For an index set $I$ of the form $\{t \in \mathbb{N} \colon i \geq t_0\}$ for some $t_0 \in \mathbb{N}$, we say a random sequence $(\xi_t)_{t\in I}$ is *adapted* if each $\xi_t$ is $\mathcal{F}_t$-measurable. For adapted real-valued sequence $(\xi_t)_{t\in\mathbb{N}^+}$ and non-negative adapted sequence $(\sigma_t)_{t\in\mathbb{N}}$, we say that each $\xi_{t+1}$ is $\mathcal{F}_t$-*conditionally* $\sigma_t$-*sub-Gaussian* if, for each $t \in \mathbb{N}$,

$$\mathbb{E}_t \exp(s\xi_{t+1}) \leq \exp(s^2\sigma_t^2/2) \quad \text{holds for all} \quad s \in \mathbb{R}. \tag{6}$$

If for some constant $c > 0$, the above holds unconditionally (that is, with $\mathbb{E}$ in place of $\mathbb{E}_t$) with $\sigma_t = c$ for all $t \in \mathbb{N}$, we say that each $\xi_t$ is $c$-*subGaussian*.

Finally, for a suitable set $A$, we write $\mathcal{U}(A)$ for the uniform probability measure on $A$. For $m \in \mathbb{N}^+$, we write $\Xi_1, \ldots, \Xi_m \sim \mathcal{U}(A)^{\otimes m}$ to denote that $\Xi_1, \ldots, \Xi_m$ are independent random variables each with law $\mathcal{U}(A)$.

## 3 Thompson and ensemble sampling algorithms

We now outline versions of Thompson sampling and ensemble sampling for the linear bandit problem. Our exposition is designed to draw out the similarities of the two methods, and sacrifices generality for simplicity. For a well-rounded and motivated introduction, see Chapter 36 of Lattimore and Szepesvári (2020) and, in particular, the notes and bibliographic remarks therein. Likewise, for introductions to ensemble sampling, see Lu and Van Roy (2017) and Osband et al. (2019), with the latter in the context of reinforcement learning.

### 3.1 Thompson sampling

Our linear Thompson sampling algorithm, presented in Algorithm 1, is extremely simple: at each step $t \in [T]$, it picks an arm $X_t$ that is optimal according to an estimate $\theta_t$ sampled uniformly on $\psi_{t-1}(\sqrt{d}B_2^d) = \sqrt{d}\mathcal{C}_{t-1}(\delta)$, a $\sqrt{d}$-inflation of the ridge regression confidence set (with $\psi_{t-1}$ as defined in Eq. (5)). Here, we treat the confidence parameter $\delta \in (0, 1]$ as fixed implicitly.

**Algorithm 1** Linear Thompson sampling

**for** $t \in \mathbb{N}^+$ **do**
  Sample $U_t \sim \mathcal{U}(\sqrt{d}B_2^d)$ and compute $\theta_t = \psi_{t-1}(U_t)$
  Compute some $X_t \in \arg\max_{x \in \mathcal{X}} \langle x, \theta_t \rangle$, play arm $X_t$ and receive reward $Y_t$

The algorithm works by balancing exploitation and exploration. The perturbed models $\theta_t$ are not too far from the estimate $\hat{\theta}_{t-1}$, no more than $\sqrt{d}$-times the confidence width. Yet this inflation of $\sqrt{d}$ allows Thompson sampling to explore sufficiently, and in particular, to occasionally try models that are *optimistic* for the true parameter $\theta_\star$, in that their predicted value is higher than the true value of the optimal arm. These two properties are vital to the usual regret guarantees for Thompson sampling, as established by Agrawal and Goyal (2012, 2013) and Abeille and Lazaric (2017).

**Remark 1.** *Our use of the uniform distribution to generate perturbed parameters is purely for the sake of a clean exposition. After all, the usual analysis for the Gaussian (or sub-Gaussian) case begins by restricting to a high-probability event where each $\theta_t$ lands within some scaled version of the corresponding $\sqrt{d}\mathcal{C}_t(\delta)$ (as in Abeille and Lazaric, 2017).*

We call the sequence of uniform random variables $U_1, U_2, \ldots$ the *random noises* used by Thompson sampling. For the purposes of our analysis, we will assume the following:

> The sequence $U_1, U_2, \ldots$ is adapted, and each element of this sequence has been sampled independently before any interaction begins.

### 3.2 Ensemble sampling

We begin with a little formalism around the random quantities that feature in the upcoming Algorithm 2, ensemble sampling:

> Our ensemble sampling algorithm uses at each step $t \in \mathbb{N}^+$ the random variables $U_t^1, \ldots, U_t^m$ and $\xi_t, J_t$. We take each $\mathcal{F}_t$ to be such that these are $\mathcal{F}_t$-measurable, and assume that all these random variables for all $t \in \mathbb{N}^+$ are sampled independently of one another before any interaction begins.

With that in place, let us examine Algorithm 2. The algorithm is a lot simpler than it might seem. In particular, observe that if we fit a ridge regression estimate on the *fake data* $(X_1, U_1^j), \ldots, (X_{t-1}, U_{t-1}^j)$, we get the estimate $\tilde{\theta}_{t-1}^j = V_{t-1}^{-1} S_{t-1}^j$. We fit $m$ such estimates on the $m$ independent streams of targets $(U_t^j : t \in \mathbb{N}^+)_{j \in [m]}$, and select $X_t$ as optimal with respect to

$$\theta_t = \hat{\theta}_{t-1} + r_0 \xi_t \tilde{\theta}_{t-1}^{J_t}.$$

That is to say, ensemble sampling acts optimally with respect to the estimate $\hat{\theta}_{t-1}$ perturbed additively by $r_0 \xi_t \tilde{\theta}_{t-1}^{J_t}$, which is the parameter of one of the $m$-many ensemble elements, chosen uniformly at random (by the index $J_t \in [m]$), symmetrised (by $\xi_t \in \{-1, 1\}$) and rescaled (by $r_0 > 0$), where each ensemble element is, effectively, a random estimate of $0_d$.

**Algorithm 2** Linear ensemble sampling

**Input** noise scale $r_0 > 0$, ensemble size $m \in \mathbb{N}^+$
Sample $(X_0^j)_{j \in [m]} \sim \mathcal{U}(\sqrt{d} S^{d-1})^{\otimes m}$ and let $S_0^j = \lambda X_0^j$ for each $j \in [m]$
**for** $t \in \mathbb{N}^+$ **do**
  Sample $(\xi_t, J_t) \sim \mathcal{U}(\{\pm 1\} \times [m])$ and let $\theta_t = \psi_{t-1}(r_0 \xi_t V_{t-1}^{-1/2} S_{t-1}^{J_t})$
  Compute some $X_t \in \arg\max_{x \in \mathcal{X}} \langle x, \theta_t \rangle$, play arm $X_t$ and observe reward $Y_t$
  Sample $(U_t^j)_{j \in [m]} \sim \mathcal{U}([-1, 1])^{\otimes m}$ and let $S_t^j = S_{t-1}^j + X_t U_t^j$ for each $j \in [m]$

The reader may well suspect that the random variables $r_0 \xi_1 V_0^{-1/2} S_0^{J_1}, r_0 \xi_2 V_1^{-1/2} S_1^{J_2}, \ldots$ will serve the same function as the noises $U_1, U_2, \ldots$ used within Thompson sampling. Indeed, our approach will be to show a regret bound for randomised algorithms where each $\theta_t = \psi_{t-1}(\Xi_t)$ for any sequence of noises $\Xi_1, \Xi_2, \ldots$ that satisfy certain properties, and then show that those used by Thompson sampling and ensemble sampling do just that.

First, however, a couple remarks.

**Remark 2.** *The random sequence of targets used to fit the ensembles in our ensemble sampling algorithm is based on uniform random variables, as opposed to Gaussian random variables, as in the prior literature. Like in the case of Thompson sampling (see Remark 1), this serves only to simplify the proof. In this case, the simplification is quite significant. In the upcoming Remark 13, we point out where this specific form of the targets was used, and sketch how to make our proof go through with suitable sub-Gaussian targets.*

**Remark 3.** *The symmetrisation of the noises by the Rademacher random variables $\xi_1, \xi_2, \ldots$ does not feature within the previous formulations of ensemble sampling. This symmetrisation again makes the proof much more convenient—we point out in Remark 10 where and how it is used. While the result almost certainly goes through without this symmetrisation, the proof would become significantly more complex. We will not attempt it.*

**Remark 4.** *In the linear setting, ensemble sampling is less computationally efficient than Thompson sampling: incrementally updating the $m + 1$ ridge regression estimators, for the $m$ that our upcoming regret analysis necessities (and which is likely not improvable) is more expensive than producing a single sample from the posterior of, say, a conjugate Gaussian linear model—the classic instantiation of Thompson sampling. Obtaining a posterior sample directly, however, uses $d^2$ memory, whereas ensemble sampling requires only order $d$ memory. Memory cost may be of particular importance when the linear model is the linearisation of a neural network, as it often is in the literature (Antorán et al., 2022; Ash et al., 2022; Mackay, 1992), where a $d^2$ memory requirement is simply prohibitive. Either way, this is only an aside from the perspective of this work: our aim is a regret bound; we leave the relative advantages of the methods for others to settle.*

## 4 A general regret bound for optimistic randomised algorithms

Our analysis of ensemble sampling—and randomised algorithms more generally—will rely on the usual principle of *optimism*. To make this precise, consider some fixed instance parameters $\theta^\star \in \mathbb{R}^d$. Then, writing $J(\theta) = \max_{x\in\mathcal{X}}\langle x, \theta\rangle$, we call

$$\Theta^{\mathrm{OPT}} = \{\theta \in \mathbb{R}^d \colon J(\theta) \geq J(\theta^\star)\} \tag{7}$$

the set of parameters *optimistic* for $\theta^\star$. With this in place, our regret bound, a generalisation of that given for Thompson sampling by Abeille and Lazaric (2017), follows.

**Theorem 1.** *Fix $T \in \mathbb{N}^+ \cup \{+\infty\}$ and $\delta \in (0,1]$. Suppose $X_1, \ldots, X_T$ are such that for each $t \in [T]$,*

$$X_t \in \arg\max_{x\in\mathcal{X}}\langle x, \theta_t\rangle \tag{8}$$

*for some adapted sequence $(\theta_t)_{t\in\mathbb{N}}$. Let $(b_t)_{t\in\mathbb{N}}$ be an adapted non-negative sequence and let*

$$\Theta_t = \psi_t(b_t B_2^d) \quad \textit{for each} \quad t \in \mathbb{N}. \tag{9}$$

*Suppose that*

$$\mathcal{E} = \bigcap_{t=1}^{T} \left\{\theta^\star, \theta_t \in \Theta_{t-1}\right\} \quad \textit{satisfies} \quad \mathbb{P}(\mathcal{E}) \geq 1-\delta. \tag{10}$$

*Also, let*

$$p_{t-1} = \mathbb{P}(\theta_t \in \Theta^{OPT} \cap \Theta_{t-1} \mid \mathcal{F}_{t-1}) \quad \textit{for each} \quad t \in \mathbb{N}^+. \tag{11}$$

*Then, the probability that there exists a $\tau \in [T]$ such that*

$$R(\tau) > 2\sqrt{2}\max_{i\in[\tau]}\frac{b_{i-1}}{p_{i-1}}\beta_{i-1}\left(\sqrt{d\tau\log\left(1+\frac{\tau}{d\lambda}\right)} + \sqrt{\frac{(\tau+1)}{\lambda}\log\left(\frac{\sqrt{4\tau/\lambda+1}}{\delta}\right)}\right) \tag{12}$$

*does not exceed $2\delta$.*

We defer the proof of Theorem 1 to Appendix A.

Evidently, the key to establishing a regret bound for a randomised algorithm using the above theorem is to control the ratios $b_0/p_0, \ldots, b_{T-1}/p_{T-1}$. As a warm-up for our analysis of ensemble sampling, we now briefly state and prove such a bound for Thompson sampling.

**Claim 2.** *For Algorithm 1, Thompson sampling,*

$$\frac{b_{t-1}}{p_{t-1}} \leq 16\sqrt{3d\pi} \quad \textit{for all} \quad t \in \mathbb{N}^+. \tag{13}$$

**Corollary 1.** *Fix $\delta \in (0,1]$. A learner using Algorithm 1, Thompson sampling, incurs regret that is, with probability $1-\delta$, bounded as*

$$R(\tau) = O\big(\sqrt{d}(d\log\tau + \sqrt{d\log\tau\log 1/\delta} + \log 1/\delta)\sqrt{\tau}\big) \quad \textit{for all} \quad \tau \in \mathbb{N}^+.$$

**Remark 5.** *The above corollary recovers the same regret bound for linear Thompson sampling as established in Agrawal and Goyal (2013) and Abeille and Lazaric (2017). It might look tighter in terms of the logarithmic $\tau$ factors—that is as we present it for uniform rather than Gaussian noises, which yield a tighter result.*

To prove the aforementioned claim, we will need the following technical lemma, given as proposition 5 in Abeille and Lazaric (2017) (we provide a much cleaner proof in Appendix B).

**Lemma 3.** *Fix $t \in \mathbb{N}$. Then, for any measure $Q$ over $\mathbb{R}^d$ and $b > 0$,*

$$Q(\Theta^{OPT} \cap \psi_t(bB_2^d)) \geq \inf_{u \in S^{d-1}} Q(\psi_t(H_u \cap bB_2^d)), \tag{14}$$

*where $H_u$ denotes the closed halfspace $\{v \in \mathbb{R}^d \colon \langle v, u \rangle \geq 1\}$.*

**Proof of Claim 2** Since each $U_t$ is in $\sqrt{d}B_2^d$, taking $b_{t-1} = \sqrt{d}$ for all $t \in \mathbb{N}^+$, leads to $\theta_t \in \Theta_{t-1}$ almost surely for all $t \in \mathbb{N}^+$. Also, since $b_t \geq 1$, $\mathcal{C}_{t-1} \subset \Theta_{t-1}$ for each $t \in \mathbb{N}^+$, and thus $\mathcal{E}$ holds with the prescribed probability. Now apply Lemma 3 with $Q(A) = \mathbb{P}_{t-1}(\theta_t \in A)$, and note that the right hand side of the inequality therein is the probability that $U_t$ is within a spherical cap of the form $H_u \cap \sqrt{d}B_2^d$ for some $u \in S^{d-1}$. By the rotational invariance of $U_t$, we may consider just $u = 1$. This probability is then just the ratio of the volume of this spherical cap to the volume of the the ball $\sqrt{d}B_2^d$. A simple geometric argument shows that this is lower bounded by $1/(16\sqrt{3\pi})$, independently of $d$. ■

We thus have a generic way of obtaining high probability regret bounds for randomised algorithms that recovers the usual result for Thompson sampling. What has changed from the result of Abeille and Lazaric (2017)?

1. We removed the assumption that the distribution of each $\theta_t$ is absolutely continuous with respect to the Lebesgue measure.

2. We allow the *probabilities of optimism* $p_0, p_1, \ldots$ to be an adapted sequence of random variables, rather than asking for the probability of optimism to be lower bounded by some fixed real number $p \in (0, 1]$, with high probability, a priori.

The above two changes are vital for ensemble sampling, where the distribution of $\theta_t$, conditioned on $\mathcal{F}_{t-1}$, is finitely supported, and where we have to deal with dependencies between time-steps. While at it, we also made the result anytime—recall that the regret bound for Thompson sampling in Corollary 1 holds uniformly over $\tau \in \mathbb{N}^+$.

## 5 Analysis of ensemble sampling

Our advertised result is captured by the following claim and its corollary.

**Claim 4.** *Fix $\delta \in (0, 1]$. Take $r_0 = 7$, $\lambda \geq 5$ and $m \geq 400 \log(2NT/\delta)$ for $N = (134\sqrt{1 + T/\lambda})^d$. Then, for Algorithm 2, linear ensemble sampling, we have that*

$$\frac{b_{t-1}}{p_{t-1}} \leq 20\sqrt{2}m^{3/2} \quad \textit{for all} \quad t \in [T]. \tag{15}$$

**Corollary 2.** *Fix $\delta \in (0, 1]$ and $T \in \mathbb{N}^+$. Take $r_0 = 7$, $\lambda \geq 5$ and $m = O(d \log T/\delta)$. Then the regret incurred by a learner using Algorithm 2 with those parameters is, with probability at least $1 - \delta$, bounded as*

$$R(T) = O\big((d \log T/\delta)^{3/2}(d \log \tau + \sqrt{d \log \tau \log 1/\delta} + \log 1/\delta)\sqrt{\tau}\big) \quad \textit{for all} \quad \tau \in [T]\,.$$

With regard to the assumptions within the above result, we have the following remarks:

**Remark 6.** *It suffices that $\lambda > 1$, but then terms dependent on $\lambda$ appear, exploding as $\lambda \downarrow 1$.*

**Remark 7.** *Our ensemble sampling algorithm requires the ensemble size $m$ to be fixed in advance, and as $m$ depends (logarithmically) on the horizon $T$, the method only provides guarantees for a fixed, finite horizon $T$, and its regret has direct dependence on $T$. One could envisage online schemes for constructing new ensemble elements as needed—this would, however, likely require us to store past observations, and the method would no longer be a streaming algorithm.*

Recall also Remarks 2 and 3 regarding the relationship between our ensemble sampling algorithm and that of, say, Lu et al. (2018) and Qin et al. (2022). The following two remarks compare our result to the aforementioned prior work.

**Remark 8.** *The seminal work of Lu and Van Roy (2017) makes strong claims on the frequentist regret of linear ensemble sampling. Their argument is, however, flawed.*[1]

**Remark 9.** *The only correct result on the regret of linear ensemble sampling is by Qin et al. (2022), where for a $d$ dimensional linear bandit with an arm set $\mathcal{X}$ of cardinality $K$, they bound the Bayesian regret incurred as*

$$BR(T) \leq \sqrt{dT \log K} + T\sqrt{\frac{K \log(Tm)}{m}}(d \wedge \log K)\,,$$

*where Bayesian regret here denotes that averaged over $\theta_\star \sim \mathcal{N}(0, I_d)$. Observe that this bound necessities an ensemble size linear in $T$ in order to recover Bayesian regret that scales as $\sqrt{T}$ (up to constant and polylogarithmic factors), which largely defeats the purpose of ensemble sampling. Furthermore, the ensemble size $m$ needs to scale linearly with $K$ to get a $\log K$ overall dependence on $K$. If we want to tackle a bandit with $\mathcal{X} = B_2^d$, order $e^{d-1}$-many arms would be needed to discretise it, and so an ensemble size $m$ exponential in $d$.*

In light of the above remarks, our result is the only result for ensemble sampling that justifies its effectiveness.

### 5.1 Proof of Claim 4

To establish 4, we need to, for each $t \in [T]$, control properties of the $\mathcal{F}_{t-1}$-conditional distributions of $r_0 \xi_t V_{t-1}^{-1/2} S_{t-1}^{J_t}$. Observe that this is a uniform distribution supported on

$$\mathcal{S}_{t-1} = \left\{ \pm r_0 V_{t-1}^{-1/2} S_{t-1}^1, \ldots, \pm r_0 V_{t-1}^{-1/2} S_{t-1}^m \right\},$$

a set of $2m$-many elements. What we will show is that there exists a high probability event on which, at every $t \in [T]$, there exists at least one $w \in \mathcal{S}_{t-1}$ such that $\theta_t = \psi_{t-1}(w)$ is optimistic for $\theta_\star$, yielding that $p_{t-1} \geq 1/(2m)$, and that the set $\mathcal{S}_{t-1}$ is not too large—specifically, that for all $u \in \mathcal{S}_{t-1}$, $\|w\|$ is on the order of $\sqrt{m}$, and thus that $b_{t-1}$ on the order of $\sqrt{m}$ suffices.

1. As confirmed by the authors.

To formalise this argument, let $\Gamma_0, \Gamma_1, \ldots$ be the sequence of positive definite matrices in $\mathbb{R}^{d\times m}$ with columns $\Gamma_t^j = V_t^{-1/2} S_t^j,\ j \in [m]$. Consider the smallest and largest singular values of each $\Gamma_t$. These are

$$s_d(\Gamma_t) = \min_{u\in S^{d-1}} \|\Gamma_t^T u\|_2 = \min_{u\in S^{d-1}} \left(\sum_{j=1}^m \langle \Gamma_t^j, u\rangle^2\right)^{\frac{1}{2}} \quad \text{and} \quad s_1(\Gamma_t) = \|\Gamma_t^T\| = \max_{u\in S^{d-1}} \|\Gamma_t^T u\|_2\,,$$

and are bounded as follows:

**Theorem 5.** *For $\lambda \geq 5$, $m \geq 400\log(3+T) \vee 1750d$, $N = (134\sqrt{1+T/\lambda})^d$ and $r_0 = 7$,*

$$\mathbb{P}\left(\forall t \in [T],\ \sqrt{m} \leq s_d(r_0\Gamma_{t-1}) \leq s_1(r_0\Gamma_{t-1}) \leq 10\sqrt{m}\right) \geq 1 - NTe^{-\frac{m}{400}}. \tag{16}$$

Theorem 5 will be proven after Claim 4.

**Proof of Claim 4** Observe that our choice of $m$ satisfies $m \geq 400\log(3+T) \vee 1750d$ and, writing $\mathcal{E}'$ for the event of Theorem 5, yields $\mathbb{P}(\mathcal{E}') \geq 1 - \delta/2$. Let $\mathcal{E}^\star$ denote the event that $\{\forall t \geq 0, \theta^\star \in \mathcal{C}_t\}$. Then, choosing $(\beta_t)_{t\in\mathbb{N}}$ with $\delta/2$ in place of $\delta$, $\mathbb{P}(\mathcal{E}^\star) \geq 1 - \delta/2$. Thus $\mathbb{P}(\mathcal{E}' \cap \mathcal{E}^\star) \geq 1 - \delta$. Now, since on $\mathcal{E}'$,

$$\|r_0\xi_t\Gamma_{t-1}^{J_t}\|_2 \leq \max_j \|r_0\Gamma_{t-1}^j\|_2 \leq s_1(r_0\Gamma_{t-1}) \leq 10\sqrt{m}, \quad \forall t \in [T]\,, \tag{17}$$

taking $b_{t-1} = 10\sqrt{m}$ for all $t \in \mathbb{N}^+$, we have that for the event $\mathcal{E}$ of Theorem 1, $\mathcal{E}' \cap \mathcal{E}^\star \subset \mathcal{E}$, and so $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Therefore, we can apply Theorem 1.

It remains to lower bound $p_0, \ldots, p_{T-1}$. Using Lemma 3 with $Q(A) = \mathbb{P}_{t-1}(\theta_t \in A)$, we have that for all $t \in \mathbb{N}^+$

$$\begin{aligned} p_{t-1} &\geq \inf_{u\in S^{d-1}} \mathbb{P}_{t-1}(\psi_{t-1}(r_0\xi_t\Gamma_{t-1}^{J_t}) \in \psi_{t-1}(H_u \cap b_{t-1}B_2^d)) \\ &= \inf_{u\in S^{d-1}} \mathbb{P}_{t-1}(r_0\xi_t\Gamma_{t-1}^{J_t} \in H_u \cap b_{t-1}B_2^d)\,. \end{aligned}$$

Now, since we assumed $\mathcal{E}'$ holds, for all $t \in [T]$, we have the bound

$$1 \leq \frac{s_d^2(r_0\Gamma_{t-1})}{m} = \min_{u\in S^{d-1}} \frac{1}{m}\sum_{j=1}^m \langle r_0\Gamma_{t-1}^j, u\rangle^2 \leq \min_{u\in S^{d-1}} \max_j \langle r_0\Gamma_{t-1}^j, u\rangle^2\,. \tag{18}$$

Thus, for any $u \in S^{d-1}$ there exists a pair $(s,j) \in \{\pm 1\}\times[m]$ such that $r_0 s\Gamma_{t-1}^j \in H_u \cap b_{t-1}B_2^d$, and therefore

$$\inf_{u\in S^{d-1}} \mathbb{P}_{t-1}(r_0\xi_t\Gamma_{t-1}^{J_t} \in H_u \cap b_{t-1}B_2^d) = \inf_{u\in S^{d-1}} \frac{1}{2m}\sum_{(s,j)} \mathbf{1}[r_0 s\Gamma_{t-1}^j \in H_u \cap b_{t-1}B_2^d] \geq \frac{1}{2m}\,,$$

where the summation runs over all $(s,j) \in \{\pm 1\}\times[m]$.

This establishes the claim (with the $\sqrt{2}$ factor present in the claimed ratio there to account for using $\delta/2$ in place of $\delta$ in the definition of each $\beta_t$). ■

**Remark 10** (On symmetrisation)**.** *In proving Claim 4, we use the symmetrisation by the Rademacher random variable $\xi_t$ in order to move from statements on minimum singular values, which are used to show the existence of at least one $r_0\Gamma_{t-1}^j$ in a given symmetrised half-space $H_u \cup H_{-u}$, to the probability that $r_0\xi_t\Gamma_{t-1}^j$ is in either of the half spaces, $H_u$ or $H_{-u}$. Else, for any $u$, Eq. (18), the middle sum would need to consider only the $\Gamma_{t-1}^j$ such that $\langle \Gamma_{t-1}^j, u\rangle \geq 0$, breaking the correspondence with the minimum singular value $\sigma_d(\Gamma_{t-1})$.*

**Remark 11** (Can we improve the bound?)**.** *Our argument is that, for any $u \in S^{d-1}$, we lower bound the maximum $\max_j \langle r_0\Gamma_{t-1}^j, u\rangle^2$ by the average $\frac{1}{m}\sum_{j=1}^m \langle r_0\Gamma_{t-1}^j, u\rangle^2$, which we show exceeds 1. Lower bounding the maximum gets us the $1/(2m)$ lower bound for $p_{t-1}$, by showing the existence of at least one element of $\mathcal{S}_{t-1}$ in $H_u$. To lower bound $p_{t-1}$ by a constant, we would want to show that a constant proportion of the elements of $\mathcal{S}_{t-1}$ lies in $H_u$, or, equivalently, lower bound the $\gamma m$-order statistic of $\langle r_0\Gamma_{t-1}^1, u\rangle^2, \ldots, \langle r_0\Gamma_{t-1}^m, u\rangle^2$ for some constant $\gamma \in (0,1)$. While order statistics are relatively well-studied for independent random variables (see, for example, Litvak and Tikhomirov (2018) and Gordon et al. (2012) and the references therein), order statistics of singular values corresponding to sequences of random matrices with the kind of dependencies inherent to our problem are virgin territory.*

### 5.2 Proving Theorem 5, bound on singular values

Theorem 5 for $t = 0$ follows by classical results on sub-Gaussian matrices with independent rows. Indeed, we show the following in Appendix D.

**Lemma 6.** *Whenever $m \geq 1750d$, $\mathbb{P}\left(\frac{1}{2}\sqrt{m} \leq s_d(\Gamma_0) \leq s_1(\Gamma_0) \leq \frac{3}{2}\sqrt{m}\right) \geq 1 - e^{-\frac{m}{400}}$.*

To extend the result to $t > 0$, we will consider the processes $R^j(u)$ and $R(u)$ defined for $u \in \mathbb{R}^d$ by

$$R_t^j(u) = \frac{\langle u, S_t^j\rangle^2}{\|u\|_{V_t}^2} \quad \text{and} \quad R_t(u) = \frac{1}{m}\sum_j R_t^j(u). \tag{19}$$

Note that for $v = V^{1/2}u \neq 0$ one has $R_t^j(u) = \langle v, \Gamma_t^j\rangle^2/\|v\|^2$. Since $V_t$ is positive-definite (and hence a bijection) we observe the following relations.

**Claim 7.** *For all $t \geq 0$, $j \leq m$,*

$$\sup_{u\neq 0} R_t^j(u) = \sup_{v\neq 0} \frac{\langle v, \Gamma_t^j\rangle^2}{\|v\|^2} = \|\Gamma_t^j\|^2 \quad \textit{and} \quad \inf_{u\neq 0} R_t(u) = \inf_{v\neq 0} \frac{\|\Gamma_t^T v\|^2}{m\|v\|^2} = \frac{s_d^2(\Gamma_t)}{m}. \tag{20}$$

With that, Theorem 5 will follow from the following bounds on $R_t(u)$ for a fixed $u \in S^{d-1}$, together with a covering argument.

**Lemma 8.** *Fix $u \in S^{d-1}$ and $\lambda \geq 5$. Suppose that $\frac{1}{2} \leq R_0(u) \leq \frac{3}{2}$ and that $m \geq 400\log(3+2T)$. Then,*

$$\mathbb{P}\left\{\forall t \in [T],\ \frac{9}{100} \leq R_t(u) \leq \frac{5}{3}\right\} \geq 1 - Te^{-\frac{m}{400}}. \tag{21}$$

The above lemma will be proven after Theorem 5. We will, of course, need the following well known bound on epsilon-nets (see, for example, Lemma 4.10 in (Pisier, 1999)).

**Lemma 9.** *For all $\delta$ in $(0,1]$, there exists an $\delta$-net $\mathcal{N}$ of $S^{d-1}$ with $|\mathcal{N}| \leq \left(1+\frac{2}{\delta}\right)^d$.*

**Proof of Theorem 5** For some $\delta \in (0,1)$, the value of which shall be determined shortly, let $\mathcal{N}_\delta$ be a $\delta$-net of $S^{d-1}$. Consider the event of Lemma 6: on that event, and with $m \geq 400\log(3+T) \vee 1750d$, the conditions of Lemma 8 are satisfied for all $u \in S^{d-1}$. Let

$$\mathcal{E}_\delta = \left\{ \forall v \in \mathcal{N}_\delta,\ \forall t \in [T],\ \frac{9}{100} \leq R_t(v) \leq \frac{5}{3} \right\}. \tag{22}$$

By a union bound over the aforementioned events, $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - (|\mathcal{N}_\delta|+1)Te^{-\frac{m}{400}}$. We will now use a covering argument to show that for $\delta$ sufficiently small, $\mathcal{E}_\delta$ is a subset of the event given in the theorem.

For this, note that for every $u \neq 0$ and $z = V_t^{1/2}u$,

$$R_t(u) = \frac{1}{m}\sum_{j=1}^m R_t^j = \frac{1}{m}\sum_{j=1}^m \frac{\langle z, \Gamma_t^j\rangle^2}{\|z\|^2} = \frac{\|\Gamma_t z\|^2}{m\|z\|^2}, \tag{23}$$

and that for all non-negative $a, b, A, B$ with $b \geq a > 0$,

$$\left|\frac{A^2}{a^2} - \frac{B^2}{b^2}\right| = \left|\frac{A^2(b^2-a^2) + (A^2-B^2)a^2}{a^2b^2}\right| \leq \frac{2A^2}{a^2}\frac{|b-a|}{b} + \frac{|A-B|(A+B)}{b^2}. \tag{24}$$

Let $u \in S^{d-1}$, $v \in \mathcal{N}$ be such that $\|u-v\| \leq \delta$ and $z = V_t^{1/2}u$, $w = V_t^{1/2}v$. Denote $A = \|\Gamma_t z\|$, $B = \|\Gamma_t w\|$, $a = \|z\|$, $b = \|w\|$. Assume without loss of generality that $b \geq a$. Since $v \in S^{d-1}$, $b \geq \sqrt{\lambda}$. Then,

$$2\frac{A^2}{a^2}\frac{|b-a|}{b} \leq \frac{2\|\Gamma_t\|^2\|z-w\|}{\sqrt{\lambda}} \leq 2\|\Gamma_t\|^2\frac{\|V_t^{1/2}\|}{\sqrt{\lambda}}\,\delta \tag{25}$$

and likewise

$$\frac{|A-B|(A+B)}{b^2} \leq \frac{2\|\Gamma_t\|\|\Gamma_t(z-w)\|}{\sqrt{\lambda}} \leq 2\|\Gamma_t\|^2\frac{\|V_t^{1/2}\|}{\sqrt{\lambda}}\,\delta, \tag{26}$$

and so we choose $\delta = \sqrt{\lambda}/(132\|V_t^{1/2}\|)$, such that

$$|R_t(u) - R_t(v)| \leq \frac{4\|\Gamma_t\|^2\|V_t^{1/2}\|}{m\sqrt{\lambda}}\delta \leq \frac{\|\Gamma_t\|^2}{33m}. \tag{27}$$

Then by Claim 7, on $\mathcal{E}_\delta$, for our choice of $\delta$,

$$\|\Gamma_t\|^2 = m\sup_{u\neq 0} R_t(u) \leq m\sup_{v\in\mathcal{N}_\delta} R_t(u) + \frac{\|\Gamma_t\|^2}{51} \quad \text{and so} \quad \|\Gamma_t\|^2 \leq \frac{55}{32}m, \tag{28}$$

and so, by the same argument, on $\mathcal{E}_\delta$, we have that

$$s_d^2(\Gamma_t) \geq m\inf_{v\in\mathcal{N}} R_t(u) - \frac{\|\Gamma_t\|^2}{33} \geq \frac{11}{400}m. \tag{29}$$

Now examine the event in the statement of the theorem: clearly, $\mathcal{E}_\delta$ is contained within. And since $\|V_t\| \leq t+\lambda$, by Lemma 9, $|\mathcal{N}_\delta|+1 \leq N$. ■

### 5.3 Proof of Lemma 8

Finally, we prove Lemma 8. We will need the following de la Peña-type concentration result, established in Appendix C.

**Lemma 10.** *Let $(A_t)_{t\in\mathbb{N}^+}$ be an adapted real-valued sequence and $(\sigma_t)_{t\in\mathbb{N}}$ an nonnegative, adapted sequence. Suppose that for some fixed $m > 0$, each $A_{t+1}$ is $\mathcal{F}_t$-conditionally $\sigma_t/\sqrt{m}$-sub-Gaussian. Then, for any $n \in \mathbb{N}^+$, and all $\alpha > 0$ satisfying $\alpha^2 m \geq 2\log\left(1 + \sum_{i=1}^n \sigma_i^2\right)$,*

$$\mathbb{P}\left\{\exists \tau \in [n]\colon \left|\sum_{i=1}^{\tau} A_i\right| > \alpha\left(\sum_{i=1}^{\tau} \sigma_i^2 + 1\right)\right\} \leq e^{-\alpha^2 m/4}. \tag{30}$$

Since we now consider a fixed $u \in S^{d-1}$, we will write $R_t^j := R_t^j(u)$ and $R_t := R_t(u)$. Let

$$D_t = \mathbb{E}_t R_{t+1} - R_t \quad \text{and} \quad W_{t+1} = R_{t+1} - E_t R_{t+1} \tag{31}$$

be respectively the drift and the noise of the process $(R_t)_{t\in\mathbb{N}}$. Also let

$$Q_t = \langle u, X_{t+1}\rangle^2 / \|u\|^2_{V_{t+1}} \quad \text{and} \quad \sigma_t^2 = 2Q_t^2 + Q_t R_t. \tag{32}$$

In Appendix E, we verify that the above defined quantities satisfy the following claims:

**Claim 11.** *$D_t = (\frac{2}{3} - R_t)Q_t$ for all $t \in \mathbb{N}$.*

**Claim 12.** *Each $W_{t+1}$ is conditionally $\sigma_t$-subGaussian.*

**Claim 13.** *For any $0 \leq \tau \leq t < T$, we have that*

$$\sum_{i=\tau}^{t} \sigma_i^2 + 1 \leq 3 + \sum_{i=\tau}^{t} Q_i R_i, \tag{33}$$

*and if, furthermore, $R_0 \leq 2$, we also have the bound*

$$\sum_{i=\tau}^{t} \sigma_i^2 + 1 \leq (3 + 2T)^2. \tag{34}$$

**Proof of Lemma 8** Let $(\tau, t)$ be a pair of time-steps satisfying $0 \leq \tau \leq t < T$. By Lemma 10, Claim 12 and Claim 13, for any $\alpha > 0$ such that $\alpha^2 m \geq 4\log(3+T)$, the event

$$\mathcal{E}_\tau(\alpha) = \left\{\exists t \geq \tau\colon \left|\sum_{i=\tau}^{t} W_{i+1}\right| > \alpha\left(3 + \sum_{i=\tau}^{t} Q_i R_i\right)\right\} \quad \text{satisfies} \quad \mathbb{P}(\mathcal{E}_\tau(\alpha)) \leq e^{-\alpha^2 m/4}. \tag{35}$$

Now, we decompose $R_{t+1}$ as

$$R_{t+1} = R_{t+1} - \mathbb{E}_t R_{t+1} + \mathbb{E}_t R_{t+1} - R_t + R_t = W_{t+1} + D_t + R_t, \tag{36}$$

which unrolled back to $\tau$, together with Claim 11, gives

$$R_{t+1} = \sum_{i=\tau}^{t} W_{i+1} + \sum_{i=\tau}^{t} \left(\frac{2}{3} - R_i\right) Q_i + R_\tau. \tag{37}$$

Observe from the above that $R_0, R_1, \dots$ is a process that drifts towards $\frac{2}{3}$, with the strength of the drift proportional to the current level of deviation. We will now argue that, if $R_t$ drifts sufficiently far from $\frac{2}{3}$, the drift will overwhelm the effect of the noises $(W_{t+1})$.

*Lower bound.* Let $0 \le \tau \le s < T$ be such that $R_\tau \ge \frac{1}{2} > R_{t+1}$ for all $t \in \{\tau, \dots, s\}$ and $s$ is maximal (since we have assumed $R_0 \ge \frac{1}{2}$, if no such $\tau$ exists, we are done). Then, for any $t \in \{\tau, \dots, s\}$, if the complement of $\mathcal{E}_\tau(\alpha)$ holds for some $\alpha \le 1/10$, then

$$R_{t+1} \ge \sum_{i=\tau+1}^{t} \left((1-\alpha)R_i - \frac{2}{3}\right)Q_i + (1-(1+\alpha)Q_\tau)R_\tau + \frac{2}{3}Q_\tau - 3\alpha \ge \frac{9}{100}. \tag{38}$$

where we used that $((1-\alpha)R_i - 1)Q_i \ge 0$ for all $i \in \{\tau+1, \dots, t\}$ for our choice of $\alpha$ and that $Q_\tau \le \frac{1}{\lambda} \le \frac{1}{5}$. The lower bound thus holds on the complement of $\mathcal{E}_0(1/10) \cup \dots \cup \mathcal{E}_{T-1}(1/10)$, the probability of which is no less than $1 - Te^{-\frac{m}{400}}$.

*Upper bound.* The upper bound follows near-verbatim, taking $\tau$ with $R_\tau \le \frac{3}{2} < R_{\tau+1}$. ■

**Remark 12.** *The lower bound of Lemma 8 was, of course, the difficult direction. Indeed, the upper bound follows rather easily from standard bounds, say Theorem 20.4 in Lattimore and Szepesvári (2020)—the same de la Peña-style result used to establish the confidence sets used here for ridge regression.*

**Remark 13** (On the use of uniform noise)**.** *The proof of Lemma 8 was where we used that the targets $(U_t^j)$ are uniform—or, in particular, that they are bounded random variables—for each $W_{t+1}$ features $(U_t^j)^2$ terms, and might otherwise be only sub-exponential. Of course, in that case, we would simply use a truncation argument: pick some truncation level $a > 0$, set $W'_{t+1} = W_{t+1} \wedge a$ for each $t \in \mathbb{N}^+$ and work with the process given by the recursion $R'_{t+1} = W'_{t+1} + D_t + R'_t$. Then, $R_t \ge R'_t$ for all $t \in \mathbb{N}$, and the truncated noises $(W'_{t+1})$ are once again sub-Gaussian, so our approach to lower bounding $R_t$ would also work for $R'_t$. We would then establish the upper bound as in Remark 12, observing that the resulted cited therein does not require the targets to be bounded.*

## 6 Discussion

We showed that linear ensemble sampling genuinely works. Per Remark 9 and Remark 8, ours is the first theoretical result for linear ensemble sampling to carry any real weight. As discussed in Remarks 2, 3, 10 and 13, while the algorithm we study varies from that presented in Lu et al. (2018) and Qin et al. (2022), the differences are largely cosmetic. Our result might not be tight. We discuss why in Remark 11—in short, getting a tighter regret bound, if possible, might not be easy. Improving the regret bound presented here for ensemble sampling might first require developing a better understanding of Thompson sampling itself. We also do not envisage the size of the ensemble $m$ being improvable by more than absolute constants. On a more positive note, there should be little challenge in extending our result to the usual non-linear settings: generalised linear models, kernels, and neural networks, via the neural tangent kernel.

## References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2011. Cited on page 3.

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017. Cited on pages 2, 4, 6, 7.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, 2012. Cited on page 4.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, 2013. Cited on pages 4, 6.

J. Antorán, S. Padhy, R. Barbano, E. Nalisnick, D. Janz, and J. M. Hernández-Lobato. Sampling-based inference for large linear models, with application to linearised Laplace. In *International Conference on Learning Representations*, 2022. Cited on page 5.

J. T. Ash, C. Zhang, S. Goel, A. Krishnamurthy, and S. Kakade. Anti-concentrated confidence bonuses for scalable exploration. In *ICLR*, 2022. Cited on page 5.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. Cited on pages 16, 20.

S. Curi, F. Berkenkamp, and A. Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 2020. Cited on page 1.

V. H. de la Pena, M. J. Klass, and T. Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 2004. Cited on pages 3, 19.

M. Dimakopoulou and B. Van Roy. Coordinated exploration in concurrent reinforcement learning. In *International Conference on Machine Learning*, 2018. Cited on page 1.

D. Eckles and M. Kaptein. Bootstrap thompson sampling and sequential decision problems in the behavioral sciences. *Sage Open*, 9(2):2158244019851675, 2019. Cited on page 1.

K. W. Fang. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 1990. Cited on page 20.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *NeurIPS*, 2010. Cited on page 2.

Y. Gordon, A. E. Litvak, C. Schütt, and E. Werner. Uniform estimates for order statistics and Orlicz functions. *Positivity*, 16(1):1–28, 2012. Cited on page 10.

B. Hao, J. Zhou, Z. Wen, and W. W. Sun. Low-rank tensor bandits. *arXiv preprint arXiv:2007.15788*, 2020. Cited on page 1.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 2018. Cited on page 2.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. Cited on pages 3, 13, 16, 19.

A. E. Litvak and K. Tikhomirov. Order statistics of vectors with dependent coordinates, and the Karhunen-Loève basis. *The Annals of Applied Probability*, 28(4):2083–2104, 2018. Cited on page 10.

X. Lu and B. Van Roy. Ensemble sampling. *Advances in Neural Information Processing Systems*, 2017. Cited on pages 1, 3, 8.

X. Lu, Z. Wen, and B. Kveton. Efficient online recommendation via low-rank ensemble sampling. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018. Cited on pages 1, 8, 13.

D. J. C. Mackay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992. Cited on page 5.

I. Osband, J. Aslanides, and A. Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2018. Cited on page 1.

I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. *Advances in Neural Information Processing Systems*, 2016. Cited on page 1.

I. Osband, B. Van Roy, D. J. Russo, Z. Wen, et al. Deep Exploration via Randomized Value Functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. Cited on page 3.

V. H. Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009. Cited on page 3.

G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999. Cited on page 10.

C. Qin, Z. Wen, X. Lu, and B. Van Roy. An analysis of ensemble sampling. In *Advances in Neural Information Processing Systems*, 2022. Cited on pages 2, 8, 13.

M. Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023. Cited on page 20.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In 2010. Cited on page 2.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933. Cited on page 1.

R. Vershynin. *High-dimensional probability: an introduction with applications in data science*, volume 47. Cambridge University Press, 2018. Cited on page 20.

J. Ville. *Etude critique de la notion de collectif*. PhD thesis, Faculté des sciences de Paris, 1939. Cited on page 19.

J. Yang, D. Eckles, P. Dhillon, and S. Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020. Cited on page 1.

Z. Zhu and B. Van Roy. Deep Exploration for Recommendation Systems. *arXiv preprint arXiv:2109.12509*, 2021. Cited on page 1.

## Appendix A. Proof of regret bound

We will use the following concentration inequality, a simple consequence of Exercise 20.8 in Lattimore and Szepesvári (2020) and Hoeffding's lemma (Lemma 2.2, Boucheron et al. (2013)).

**Lemma 14.** *Fix $0 < \delta \leq 1$. Let $(\xi_t)_{t\in\mathbb{N}^+}$ be a real-valued martingale difference sequence satisfying $|\xi_t| \leq c$ almost surely for each $t \in \mathbb{N}^+$ and some $c > 0$. Then,*

$$\mathbb{P}\left(\exists\tau\colon \left(\sum_{t=1}^{\tau}\xi_t\right)^2 \geq 2c^2(\tau+1)\log\left(\frac{\sqrt{c^2\tau+1}}{\delta}\right)\right) \leq \delta. \tag{39}$$

We will also need the following classic result (Lemma 19.4 in Lattimore and Szepesvári (2020)).

**Lemma 15** (Elliptical potential lemma)**.** *Let $(x_t)_{t\in\mathbb{N}^+}$ be a sequence of vectors in $B_2^d$, let $V_0 = \lambda I$ for some $\lambda > 0$ and $V_t = V_0 + \sum_{i=1}^{t} x_i x_i^\mathsf{T}$ for each $t \in \mathbb{N}^+$. Then, for all $\tau \in \mathbb{N}^+$,*

$$\sum_{t=1}^{\tau} \|x_t\|^2_{V_{t-1}^{-1}} \leq 2d\log\left(1+\frac{\tau}{\lambda d}\right). \tag{40}$$

**Claim 16.** *For any $t \in \mathbb{N}^+$, $X_t$ is a subgradient of $J$ at $\theta_t$.*

**Proof** Fix $t \in \mathbb{N}^+$. For any $\theta \in \mathbb{R}^d$,

$$J(\theta_t) + \langle X_t, \theta - \theta_t\rangle = \langle X_t, \theta_t\rangle + \langle X_t, \theta - \theta_t\rangle = \langle X_t, \theta\rangle \leq \max_{x\in\mathcal{X}}\langle x, \theta\rangle = J(\theta), \tag{41}$$

which is the defining inequality for a subgradient. ■

**Proof of regret bound, Theorem 1** For any $\tau \in [T]$, the regret is split into two parts, which we will control separately:

$$R(\tau) = \sum_{t=1}^{\tau}\left(J(\theta^\star) - J(\theta_t)\right) + \sum_{t=1}^{\tau}\left(J(\theta_t) - \langle X_t, \theta^\star\rangle\right). \tag{42}$$

Fix an index $t \in [T]$ and consider $J(\theta_t) - \langle X_t, \theta^\star \rangle$. We have that, on $\mathcal{E}$,

$$J(\theta_t) - \langle X_t, \theta^\star \rangle = \langle X_t, \theta_t - \theta^\star \rangle \leq \|X_t\|_{V_{t-1}^{-1}} \|\theta_t - \theta^\star\|_{V_{t-1}} \leq \gamma_{t-1} \|X_t\|_{V_t^{-1}}. \tag{43}$$

where the first inequality is by Cauchy-Schwartz and second uses that on $\mathcal{E}$, we have $\theta_t, \theta^\star \in \Theta_{t-1}$, and the definition of $\gamma_{t-1}$.

Now consider $J(\theta^\star) - J(\theta_t)$, again for a fixed index $t \in [T]$. Let $\theta^-$ be a minimiser $J$ over $\Theta_{t-1}$ (which is well defined, since $J$ is continuous and $\Theta_{t-1}$ closed) and let $\theta^+$ be any element of $\Theta^{\mathrm{OPT}}$. Then, on $\mathcal{E}$, since $\theta^\star, \theta_t \in \Theta_{t-1}$,

$$J(\theta^\star) - J(\theta_t) \leq J(\theta^\star) - J(\theta^-) \leq J(\theta^+) - J(\theta^-). \tag{44}$$

Moreover, likewise for any probability measure $Q$ over $\Theta^{\mathrm{OPT}}$, we have

$$J(\theta^\star) - J(\theta_t) \leq \int J(\theta^+) - J(\theta^-) \, \mathrm{d}Q(\theta^+). \tag{45}$$

Writing $\Theta_{t-1}^{\mathrm{OPT}} = \Theta^{\mathrm{OPT}} \cap \Theta_{t-1}$, we choose $Q = Q_{t-1}$ for the integral above given by

$$Q_{t-1} = \begin{cases} \mathbb{P}(\theta_t \in \cdot \cap \Theta_{t-1}^{\mathrm{OPT}} \mid \mathcal{F}_{t-1}) / p_{t-1} \,, & p_{t-1} > 0 \,; \\ \text{any arbitrary probability measure}, & \text{otherwise} \,. \end{cases} \tag{46}$$

Then, by definition of $Q_{t-1}$ and since $\theta^-$ is $\mathcal{F}_{t-1}$-measurable, we get

$$J(\theta^\star) - J(\theta_t) \leq \mathbb{E}_t[(J(\theta_t) - J(\theta^-)) \mathbf{1}[\theta_t \in \Theta_{t-1}^{\mathrm{OPT}}] \mid \mathcal{F}_{t-1}] / p_{t-1}, \tag{47}$$

where for $p_{t-1} = 0$ we take the upper bound to be positive infinity. Observing that $X_t$ is a subgradient of $J$ at $\theta_t$ (Claim 16 and Eq. (41)) and applying Cauchy-Schwartz, we have that

$$J(\theta_t) - J(\theta^-) \leq \langle X_t, \theta_t - \theta^- \rangle \leq \|X_t\|_{V_{t-1}^{-1}} \|\theta^- - \theta_t\|_{V_{t-1}}, \tag{48}$$

Moreover, recalling that $\theta^- \in \Theta_{t-1}$, that $\Theta_{t-1}^{\mathrm{OPT}} \subset \Theta_{t-1}$ and by definition of $\gamma_{t-1}$,

$$\|\theta^- - \theta_t\|_{V_{t-1}} \mathbf{1}[\theta_t \in \Theta_{t-1}^{\mathrm{OPT}}] \leq \gamma_{t-1}. \tag{49}$$

So, since $\gamma_{t-1}$ is, by assumption, $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[(J(\theta_t) - J(\theta^-)) \mathbf{1}[\theta_t \in \Theta_{t-1}^{\mathrm{OPT}}] \mid \mathcal{F}_{t-1}] / p_{t-1} \leq \frac{\gamma_{t-1}}{p_{t-1}} \mathbb{E}[\|X_t\|_{V_{t-1}^{-1}} \mid \mathcal{F}_{t-1}]. \tag{50}$$

Chaining the above inequalities and writing $\Delta_t = \mathbb{E}[\|X_t\|_{V_{t-1}^{-1}} \mid \mathcal{F}_{t-1}] - \|X_t\|_{V_{t-1}^{-1}}$, we have

$$J(\theta^\star) - J(\theta_t) \leq \frac{\gamma_{t-1}}{p_{t-1}} \mathbb{E}[\|X_t\|_{V_{t-1}^{-1}} \mid \mathcal{F}_{t-1}] = \frac{\gamma_{t-1}}{p_{t-1}} \left( \|X_t\|_{V_{t-1}^{-1}} + \Delta_t \right), \tag{51}$$

Combining Eqs. (43) and (51) with the regret decomposition in Eq. (42), for any $\tau \in [T]$,

$$R(\tau) \leq \sum_{t=1}^{T} \left( \left( \gamma_{t-1} + \frac{\gamma_{t-1}}{p_{t-1}} \right) \|X_t\|_{V_{t-1}^{-1}} + \frac{\gamma_{t-1}}{p_{t-1}} \Delta_t \right) \leq \max_{i \in [\tau]} \frac{\gamma_{i-1}}{p_{i-1}} \left( 2 \sum_{t=1}^{\tau} \|X_t\|_{V_{t-1}^{-1}} + \sum_{t=1}^{\tau} \Delta_t \right). \tag{52}$$

Now, by Cauchy-Schwartz and the elliptical potential lemma (Lemma 15), for any $\tau \in \mathbb{N}^+$,

$$\sum_{t=1}^{\tau} \|X_t\|_{V_{t-1}^{-1}} \leq \left(\tau \sum_{t=1}^{\tau} \|X_t\|^2_{V_{t-1}^{-1}}\right)^{\frac{1}{2}} \leq \sqrt{2\tau d \log\left(1 + \frac{\tau}{d\lambda}\right)}. \tag{53}$$

To deal with the second sum, observe that since for all $t \in \mathbb{N}^+$, $V_{t-1} \succeq \lambda I$ and $X_t \in B_2^d$,

$$\|X_t\|^2_{V_{t-1}^{-1}} = \langle X_t, V_{t-1}^{-1} X_t \rangle \leq \|X_t\|_2^2/\lambda \leq 1/\lambda \quad \text{and so} \quad |\Delta_t| \leq 2/\sqrt{\lambda} \quad \text{for all } t \in \mathbb{N}^+. \tag{54}$$

Also, observe that $(\Delta_t)_{t\in\mathbb{N}}$ is a martingale. Thus, we can apply Lemma 14 with $c = 2/\sqrt{\lambda}$, obtaining

$$\mathbb{P}\left(\exists \tau \in \mathbb{N}^+ \colon \sum_{t=1}^{\tau} \Delta_t \geq 2\sqrt{\frac{2(\tau+1)}{\lambda} \log\left(\frac{\sqrt{4\tau/\lambda + 1}}{\delta}\right)}\right) \leq \delta. \tag{55}$$

Combined with Eq. (52), the bounds on the two sums, Eq. (53) and Eq. (55), together with a union bound, yield the claim. ■

## Appendix B. Generic optimism with elliptical confidence sets

**Lemma 17.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex function and let $u$ be its maximizer over the unit ball. Then, for any $v \in H_u \doteq \{v \in \mathbb{R}^d \colon \langle v, u\rangle \geq 1\}$, we have $F(v) \geq F(u)$.*

**Proof** For any $v \in \mathbb{R}^d$ with $\langle v, u\rangle > 1$, the ray from $v$ to $u$ enters the interior of the unit ball. Hence, for any such $v$, there exists a $z \in B_2^d$ and $\alpha \in (0,1)$ such that $u = \alpha z + (1-\alpha)v$. By convexity and maximality,

$$F(u) = F(\alpha z + (1-\alpha)v) \leq \alpha F(z) + (1-\alpha)F(v) \leq \alpha F(u) + (1-\alpha)F(v). \tag{56}$$

Hence $F(u) \leq F(v)$. Since any finite convex function on an open set is continuous, the result holds for any $v \in H_u$. ■

**Proof of Lemma 3** Write $F = J \circ \psi_t$; since $J$ is convex and $\psi_t$ is affine, $F$ is convex. Let $u^+$ be the maximiser of $F$ over $B_2^d$ and note that since $B_2^d$ is strictly convex, $u^+ \in \partial B_2^d = S^{d-1}$. By assumption, $\theta^\star \in \psi_t(B_2^d)$, and so $J(\theta^\star) \leq F(u^+)$. By Lemma 17, $F(u^+) \leq F(u')$ for any $u' \in H_{u^+}$. Thus $\psi_t(H_{u^+}) \subset \Theta^{\mathrm{OPT}}$. Moreover, by assumption, $\Theta_t \subset \psi_t(b_t B_2^d)$. These two inclusions yield

$$\Theta^{\mathrm{OPT}} \cap \Theta_t \supset \psi_t(H_{u^+}) \cap \psi_t(b_t B_2^d) \supset \psi_t(H_{u^+} \cap b_t B_2^d). \tag{57}$$

Thus, for any measure $Q$ on $\mathbb{R}^d$,

$$Q(\Theta^{\mathrm{OPT}} \cap \Theta_t) \geq Q(\psi_t(H_{u^+} \cap b_t B_2^d)) \geq \inf_{u \in S^{d-1}} Q(\psi_t(H_u \cap b_t B_2^d)).$$ ■

## Appendix C. Concentration result

Lemma 10 is effectively a corollary to the following de la Peña-type concentration result.

**Lemma 18.** *Let $(\mathcal{H}_i)_{i\in\mathbb{N}}$ be a filtration and $((A_i, B_i))_{i\in\mathbb{N}^+}$ be pairs of random variables such that each $A_i$ is $\mathcal{H}_{i-1}$-conditionally $B_i$-subGaussian. Then, for any $x, y > 0$,*

$$\mathbb{P}\left\{\exists\tau > 0\colon \left(\sum_{i=1}^{\tau} A_i\right)^2 \geq \left(\sum_{i=1}^{\tau} B_i^2 + y\right)\left(x + \log\left(1 + \frac{1}{y}\sum_{i=1}^{\tau} B_i^2\right)\right)\right\} \leq e^{-x/2}. \tag{58}$$

**Proof of Lemma 10** Consider the right hand side of the event in Eq. (58); choosing $y = 1/m$, $x = \alpha^2 m/2$, and substituting $B_i^2 = \sigma_i^2/m$, it is equal to

$$\frac{1}{m}\left(\sum_{i=1}^{\tau}\sigma_i^2 + 1\right)\left(\frac{\alpha^2 m}{2} + \log\left(1 + \sum_{i=1}^{\tau}\sigma_i^2\right)\right) \leq \alpha^2\left(\sum_{i=1}^{\tau}\sigma_i^2 + 1\right), \tag{59}$$

where the inequality follows by assumption on $\alpha$. We conclude by using the simple observation that for $z \geq 0$, $(z+1) \leq (z+1)^2$. ■

The result of Lemma 18 is implied immediately by Theorem 2.1 in de la Pena et al. (2004), but since a direct proof is brief, we include it.

**Proof of Lemma 18** For any $s \in \mathbb{R}$, define the random process $M_1(s), M_2(s), \ldots$ given by

$$M_n(s) = \exp\left(s\sum_{i=1}^{n} A_i - s^2/2\sum_{i=1}^{n} B_i^2\right) \quad \text{for all} \quad n \in \mathbb{N}^+. \tag{60}$$

Note that $(M_n(s))_{s\in\mathbb{N}^+}$ is a nonnegative supermartingale satisfying $\mathbb{E}M_1(s) \leq 1$. Indeed, for any $n \in \mathbb{N}^+$,

$$\mathbb{E}M_n(s) = \mathbb{E}M_{n-1}(s)\mathbb{E}[\exp(sA_n - s^2/2B_n^2) \mid \mathcal{H}_{n-1}] \leq \mathbb{E}M_{n-1}(s) \leq \ldots \leq \mathbb{E}M_1(s) \leq 1. \tag{61}$$

Let $\bar{M}_1, \bar{M}_2, \ldots$ be the process given by $\bar{M}_n = \int M_n d\mathcal{N}(0, y)$ for all $n \in \mathbb{N}^+$. Then, by Lemma 20.3 of Lattimore and Szepesvári (2020), $(\bar{M}_n)_{n\in\mathbb{N}^+}$ is again a nonnegative supermartingale. Evaluating the integral that defines each $\bar{M}_n$, we see that

$$\bar{M}_n = \sqrt{\frac{y}{\sum_{i=1}^{n} B_i^2 + y}}\exp\left(\frac{(\sum_{i=1}^{n} A_i)^2}{2(\sum_{i=1}^{n} B_i^2 + y)}\right) \quad \text{for all} \quad n \in \mathbb{N}^+. \tag{62}$$

Applying Ville's inequality to $(\bar{M}_n)$ (Ville, 1939), we have that

$$e^{-x/2} \geq e^{-x/2}\mathbb{E}\bar{M}_1 \geq \mathbb{P}\left(\sup_{n\in\mathbb{N}^+} \bar{M}_n \geq e^{x/2}\right) = \mathbb{P}\left(\exists n \in \mathbb{N}^+\colon \log\bar{M}_n \geq x/2\right), \tag{63}$$

which, after plugging in the expression for $\bar{M}_n$ and rearranging, is the stated inequality. ■

## Appendix D. Proof of initialisation result, Lemma 6

Lemma 6 is an immediate consequence of the following theorem (take $\delta = \sqrt{m}/20$).

**Theorem 19.** *Let $M \in \mathbb{R}^{m\times d}$, $m \geq d$, be a matrix with rows $M_1, \ldots, M_m$ distributed uniformly and independently on $\sqrt{d}S^{d-1}$. Then, for $C = 2(1+\sqrt{3})$, and for all $\delta > 0$,*

$$\mathbb{P}\{\sqrt{m} - C(\sqrt{3d} + \delta) \leq s_d(M) \leq s_1(M) \leq \sqrt{m} + C(\sqrt{3d} + \delta)\} \geq 1 - e^{-\delta^2}\,. \tag{64}$$

**Claim 20.** *Fix $x \in S^{d-1}$, let $U \sim \mathcal{U}(S^{d-1})$ and $U_x^2 = \langle U, x\rangle^2$. Then,*

$$\mathbb{E}\exp(s\,|U_x^2 - \mathbb{E}U_x^2|) \leq \exp\left(\frac{s^2\nu/2}{1-cs}\right) \quad \text{for all} \quad 0 < s < 1/c \tag{65}$$

*and some $\nu, c > 0$ that satisfy $\nu \leq 2/d^2$ and $c \leq 4/d$, and where $\mathbb{E}U_x^2 = 1/d$.*

**Proof** It is known that the thus defined $U_x^2$ has distribution Beta($\frac{1}{2}, \frac{d-1}{2}$) (see, for example, Theorem 1.5 and the discussion thereafter in Fang, 1990), which has the stated expectation. We thus need only look up moment generating function bounds for beta random variables. Skorski (2023) derives such in their proof of their Theorem 1, and our result follows by substituting in the parameters of our beta distribution, and bounding crudely. ■

**Proof of Theorem 19** For $x \in S^{d-1}$, consider $Z_x^2 = \frac{1}{m}\|Mx\|_2^2 = \frac{d}{m}\sum_{j=1}^m \langle M_j/\sqrt{d}, x\rangle^2$. Observe that each $M_j/\sqrt{d} \sim \mathcal{U}(S^{d-1})$. Using Claim 20 and that $M_1, \ldots, M_m$ are independent, we have that, for all $0 < sd/m < 1/c$,

$$\mathbb{E}\exp(s\,|Z_x^2 - 1|) = \prod_{j=1}^m \mathbb{E}\exp\left(\frac{sd}{m}|U_x^2 - \mathbb{E}U_x^2|\right) \leq \exp\left(\frac{s^2d^2\nu/(2m)}{1-csd/m}\right).$$

Examining section 2.4 of Boucheron et al. (2013), we see that $Z_x^2 - 1$ is what would be termed there sub-gamma with parameters $(d^2\nu/m,\, cd/m)$ on both tails. Thus, it satisfies the there-stated Bernstein-type bound for sub-gamma random variables that, combined with a union bound over the two tails, and the bounds $\nu \leq 2/d^2$ and $c \leq 4/d$ from Claim 20, gives that, for all $r > 0$,

$$\mathbb{P}(|Z_x^2 - 1| \geq \sqrt{4r/m} + 4r/m) \leq 2e^{-r}\,.$$

Now let $\mathcal{N}$ be a $\frac{1}{4}$-net of $S^{d-1}$. By the usual variational representation of norm argument, $\sup_{x\in S^{d-1}} |z_x^2 - 1| \leq 2\max_{x\in\mathcal{N}} |z_x^2 - 1|$ (see, e.g., exercise 4.4.3 in Vershynin, 2018). Also, by our bound on nets from Lemma 9, $|\mathcal{N}| \leq 9^d$. Thus, for any $r > 0$, the event

$$\mathcal{E}_r = \left\{\sup_{x\in S^{d-1}} |Z_x^2 - 1| \geq 4\sqrt{r/m} + 8r/m\right\} \quad \text{satisfies} \quad \mathbb{P}(\mathcal{E}_r) \leq 2|\mathcal{N}|e^{-r} \leq \exp(3d - r)\,. \tag{66}$$

Next, observe that since $Z_x > 0$, we have that $|Z_x^2 - 1| \geq |Z_x - 1| \vee |Z_x - 1|^2$. So,

$$|Z_x^2 - 1| \geq \lambda|Z_x - 1| + (1-\lambda)|Z_x - 1|^2 \quad \text{for all } \lambda \in [0,1]. \tag{67}$$

Using the above inequality with $\lambda = \sqrt{3} - 1$ shows that

$$|Z_x^2 - 1| \leq 4\sqrt{r/m} + 8r/m \implies |Z_x - 1| \leq 2(1+\sqrt{3})\sqrt{r/m}\,. \tag{68}$$

Therefore, taking $\sqrt{r} = \sqrt{3d} + \delta$ gives us that, with probability at least $1 - e^{-\delta^2}$,

$$\sqrt{m} \sup_{x \in S^{d-1}} |Z_x - 1| \leq 2(1+\sqrt{3})(\sqrt{3d} + \delta)\,.$$

Seeing as

$$\sqrt{m} \inf_{x \in S^{d-1}} Z_x = s_d(M) \leq s_1(M) = \sqrt{m} \sup_{x \in S^{d-1}} Z_x$$

we have now proven the stated theorem. ■

## Appendix E. Proofs of claims

**Proof of Claim 11** Fix $u \in S^{d-1}$ and note that

$$R_{t+1}^j = \frac{\langle u, S_t^j + U_{t+1} X_{t+1}\rangle^2}{\|u\|_{V_{t+1}}^2} = \frac{\langle u, S_t^j\rangle^2 + (U_{t+1}^j)^2\langle u, X_{t+1}\rangle^2 + 2U_{t+1}^j\langle u, S_t^j\rangle\langle u, X_{t+1}\rangle}{\|u\|_{V_t}^2 + \langle u, X_{t+1}\rangle^2}. \tag{69}$$

Recall that $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_t]$, that $X_{t+1}$ and $S_t^j$ are $\mathcal{F}_t$-measurable and that $U_{t+1}^j$ is independent of $\mathcal{F}_t$. The latter of these gives $\mathbb{E}_t U_{t+1}^j = 0$ and $\mathbb{E}_t (U_{t+1}^j)^2 = \frac{2}{3}$. With that, we have that

$$\begin{aligned}
\mathbb{E}_t R_{t+1}^j - R_t^j &= \frac{\langle u, S_t^j\rangle^2 + \frac{2}{3}\langle u, X_{t+1}\rangle^2}{\|u\|_{V_t}^2 + \langle u, X_{t+1}\rangle^2} - \frac{\langle u, S_t^j\rangle^2}{\|u\|_{V_t}^2} && (70)\\
&= \frac{\frac{2}{3}\langle u, X_{t+1}\rangle^2 \|u\|_{V_t}^2 - \langle u, S_t^j\rangle^2\langle u, X_{t+1}\rangle^2}{\|u\|_{V_t}^2\left(\|u\|_{V_t}^2 + \langle u, X_{t+1}\rangle^2\right)} && (71)\\
&= \frac{\langle u, X_{t+1}\rangle^2}{\|u\|_{V_t}^2 + \langle u, X_{t+1}\rangle^2}\left(\frac{2}{3} - \frac{\langle u, S_t^j\rangle^2}{\|u\|_{V_t}^2}\right) && (72)\\
&= Q_t\left(\frac{2}{3} - R_t^j\right). && (73)
\end{aligned}$$

The statement follows by averaging over $j \in \{1, \ldots, m\}$. ■

**Proof of Claim 12** Subtracting Eq. (70) from Eq. (69) and averaging over $j \in \{1, \ldots, m\}$, we see that

$$W_{t+1} = R_{t+1} - \mathbb{E}_t R_{t+1} = \frac{Q_i}{m}\sum_{i=1}^m ((U_{i+1}^j)^2 - \frac{2}{3}) + \frac{1}{m}\sum_{i=1}^m U_{i+1}^j H_i^j \tag{74}$$

where $H_i^j = \langle u, X_{i+1}\rangle\langle u, S_i^j\rangle / \|u\|_{V_{i+1}}^2$. Note that $Q_i$ and $H_i$ are $\mathcal{F}_i$ measurable and that $U_{i+1}^1, \ldots, U_{i+1}^m$ are independent of $\mathcal{F}_i$ and one another, and their absolute values are bounded by 1. Thus, examining the two terms in the sum we see that:

- $\frac{Q_i}{m}\sum_{i=1}^m((U_{i+1}^j)^2-\frac{2}{3})$ is $\mathcal{F}_i$-conditionally $\frac{Q_i}{\sqrt{m}}$-sub-Gaussian.
- $\frac{1}{m}\sum_{i=1}^m U_{i+1}^j H_i^j$ is $\mathcal{F}_i$-conditionally $\frac{H_t}{\sqrt{2m}}$-sub-Gaussian, where

$$(H_t)^2 := \frac{1}{m}\sum_{j=1}^m (H_i^j)^2 = \frac{1}{m}\sum_{j=1}^m \frac{\langle u, X_{i+1}\rangle^2 \langle u, S_i^j\rangle^2}{\|u\|_{V_{i+1}}^4} = \frac{Q_i}{m}\sum_{j=1}^m \frac{\langle u, S_i^j\rangle^2}{\|u\|_{V_{i+1}}^2} \le Q_i R_i. \quad (75)$$

The result follows by recalling that if the sum of an $a$-sub-Gaussian random variable and a $b$-sub-Gaussian random variable is $\sqrt{2(a^2+b^2)}$-sub-Gaussian. ■

The proof of the final claim will require the following simple lemma.

**Lemma 21.** *Let $b_1, b_2, \ldots$ be a sequence of real numbers in $[0,1]$. Then, for any $\lambda > 0$ and $n \in \mathbb{N}^+$,*

$$\sum_{j=1}^n \frac{b_j}{\lambda + \sum_{i=1}^j b_i} \le \frac{2}{\lambda} + 2\log(\lambda + n - 1) \quad \textit{and} \quad \sum_{j=1}^n \left(\frac{b_j}{\lambda + \sum_{i=1}^j b_i}\right)^2 \le \frac{4(\lambda+1)}{\lambda^2}. \quad (76)$$

**Proof** Let $(n_k \colon k \ge 1)$ be a finite sequence of integers where each $n_k$ is the largest integer such that $\sum_{i=1}^{n_k} b_i \le k$, and we stop if $n_k = n$. Then for some $\ell \in \mathbb{N}$, we have a sequence $0 = n_0 < n_1 < \ldots < n_\ell = n$ such that for all $1 \le k \le \ell$, $k - 1 \le \sum_{i=1}^{n_k} b_i \le k$, and if $k < \ell$, then $k \le \sum_{i=1}^{n_k+1} b_i$. Hence $\sum_{i=n_k+1}^{n_{k+1}} b_i \le 2$. Therefore, for the first sum,

$$\begin{aligned}\sum_{j=1}^n \frac{b_j}{\lambda + \sum_{i=1}^j b_i} &= \sum_{k=0}^{n-1}\sum_{j=n_k+1}^{n_{k+1}} \frac{b_j}{\lambda + \sum_{i=1}^j b_i} \le \sum_{k=0}^{n-1}\sum_{j=n_k+1}^{n_{k+1}} \frac{b_j}{\lambda + \sum_{i=1}^{n_k+1} b_i} \\ &\le \sum_{k=0}^{n-1} \frac{1}{\lambda + k}\sum_{j=n_k+1}^{n_{k+1}} a_j \le 2\sum_{k=0}^{n-1}\frac{1}{\lambda+k} \le \frac{2}{\lambda} + 2\int_0^{n-1}\frac{1}{\lambda+x}\,\mathrm{d}x,\end{aligned}$$

which is in equal to the stated upper bound. For the second sum,

$$\begin{aligned}\sum_{j=1}^n \left(\frac{b_j}{\lambda + \sum_{i=1}^j b_i}\right)^2 &= \sum_{k=0}^{n-1}\sum_{j=n_k+1}^{n_{k+1}} \left(\frac{b_j}{\lambda + \sum_{i=1}^j b_i}\right)^2 \le \sum_{k=0}^{n-1}\sum_{j=n_k+1}^{n_{k+1}} \left(\frac{b_j}{\lambda + \sum_{i=1}^{n_k+1} b_i}\right)^2 \\ &\le \sum_{k=0}^{n-1}\frac{1}{(\lambda+k)^2}\left(\sum_{j=n_k+1}^{n_{k+1}} b_j\right)^2 \le \sum_{k=0}^{n-1}\frac{4}{(\lambda+k)^2} \le \frac{4}{\lambda^2} + \int_0^\infty \frac{4}{(\lambda+x)^2}\,\mathrm{d}x,\end{aligned}$$

which is again equal to the stated upper bound. ■

**Proof of Claim 13** Noting that since $\|u\| = 1$ and $\lambda \ge 5$, by Lemma 21,

$$\sum_{i=\tau}^t Q_i \le \sum_{i=0}^{T-1}\frac{\langle u, X_{i+1}\rangle^2}{\lambda + \sum_{j=0}^{i+1}\langle u, X_j\rangle^2} \le \frac{2}{5} + \log(4+T) \le \frac{17}{5} + T \quad (77)$$

and

$$\sum_{i=\tau}^{t} Q_i^2 \le \sum_{i=0}^{T-1} Q_i^2 = \sum_{i=0}^{T-1} \left( \frac{\langle u, X_{i+1} \rangle^2}{\lambda + \sum_{j=0}^{i+1} \langle u, X_j \rangle^2} \right)^2 \le \frac{5}{\lambda} \le 1. \tag{78}$$

Using these, we have

$$1 + \sum_{i=\tau}^{t} \sigma_i^2 = 1 + 2 \sum_{i=\tau}^{t} Q_i^2 + \sum_{i=\tau}^{t} R_i Q_i \le 3 + \sum_{i=\tau}^{t} R_i Q_i \le 3 + \left( \frac{17}{5} + T \right) \max_{\tau \le i \le t} R_i, \tag{79}$$

which establishes the first part of the claim. Now, since $(a+b)^2 \le 2a^2 + 2b^2$ and by symmetry,

$$R_i^j = \frac{\langle u, S_0^j + \sum_{i=1}^{i} U_\ell^j X_\ell \rangle^2}{\lambda + \sum_{\ell=1}^{i} \langle u, X_\ell \rangle^2} \le 2R_0^j + 2 \frac{\left( \sum_{\ell=1}^{i} \langle u, X_\ell \rangle \right)^2}{\lambda + \sum_{\ell=1}^{i} \langle u, X_\ell \rangle^2} \le 2R_0^j + 2 \max_{b \in [0,1]} \frac{(ib)^2}{\lambda + ib^2} \tag{80}$$

$$\le 2R_0^j + 2i. \tag{81}$$

By definition, $R_i = \frac{1}{m} \sum_{j=1}^{m} R_i^j$, and by assumption $R_0 \le 2$ and $i \le T - 1$, so $R_i \le 4 + 2i \le 2 + 2T$. And so,

$$3 + \left( \frac{17}{5} + T \right) \max_{\tau \le i \le t} R_i \le 3 + \left( \frac{17}{5} + T \right) (2 + 2T) \le (3 + 2T)^2, \tag{82}$$

which shows the second part of the claim. ■