Linear Thompson Sampling Revisited

Marc Abeille **Alessandro Lazaric**
Inria Lille - Nord Europe, Team SequeL

Abstract

We derive an alternative proof for the regret of Thompson sampling (TS) in the stochastic linear bandit setting. While we obtain a regret bound of order $\widetilde{O}(d^{3/2}\sqrt{T})$ as in previous results, the proof sheds new light on the functioning of the TS. We leverage on the structure of the problem to show how the regret is related to the sensitivity (i.e., the gradient) of the objective function and how selecting optimal arms associated to *optimistic* parameters does control it. Thus we show that TS can be seen as a generic randomized algorithm where the sampling distribution is designed to have a fixed probability of being optimistic, at the cost of an additional $\sqrt{d}$ regret factor compared to a UCB-like approach. Furthermore, we show that our proof can be readily applied to regularized linear optimization and generalized linear model problems.

1 Introduction

The multi-armed bandit (MAB) framework [Bubeck and Cesa-Bianchi, 2012] formalizes in a synthetic way the exploration-exploitation trade-off in sequential decision-making, where a learner needs to balance between exploiting current estimates to select actions maximizing the reward and exploring actions to improve the accuracy of its estimates. Two popular approaches have been developed to trade off exploration and exploitation: the *optimism in face of uncertainty* (OFU) principle (see e.g., Agrawal [1995], Auer et al. [2002]), which consists in choosing the optimal action according to upper-confidence bounds on the true values, and the Thompson Sampling (TS) strategy, which randomizes actions on the basis of their uncertainty. In this paper we mostly focus on this second approach.

TS is an general heuristic for decision-making problems characterized by some unknown parameters. The first version of this Bayesian heuristic dates back to Thompson [1933], but it has been rediscovered several times and successfully applied to address the exploration-exploitation trade-off in a wide range of problems (see e.g., Strens 2000, Chapelle and Li 2011, Russo and Van Roy 2014). The basic idea is to assume a *prior* distribution over the unknown parameters and to use the Bayes rule to update it using the samples obtained over time. More precisely, at each time step the learner gathers information by executing the optimal action corresponding to a random parameter sampled from the current posterior distribution.

Related literature. While the Bayesian perspective of TS provides a convenient tool to derive the sampling distribution, the algorithm is still valid under a frequentist approach, i.e., when the true parameter is not a random variable but a fixed parameter. As a result, the regret of TS (i.e., the difference between the rewards collected by the algorithm and the optimal action) has been analyzed both in the Bayesian and in the frequentist setting. In MAB, TS has been shown to achieve optimal performance in the frequentist setting (see e.g., May et al. 2012, Agrawal and Goyal 2012b, Kaufmann et al. 2012, Korda et al. 2013) and the dependency of the regret on its prior has been studied in the Bayesian case by Bubeck and Liu [2013]. In more general cases, such as the (generalized) linear bandit and reinforcement learning settings, most of the literature focused on the analysis of the Bayesian regret (see e.g., Russo and Van Roy [2014], Osband and Van Roy [2015], Russo and Van Roy [2016]). Notable exceptions are the analysis of TS in finite MDPs by Gopalan and Mannor [2015] and the study in linear contextual bandit (LB) by Agrawal and Goyal [2012b]. In this paper, we focus on LB and draw novel insights on the functioning of TS in this setting. In LB the value of an arm is obtained as the inner product between an arm feature vector x and an unknown global parameter $\theta^\star$. As opposed to the OFU approach, the main technical difficulty in analyzing TS lies in controlling the deviation in performance due to the randomness of the algorithm. Agrawal and Goyal [2012b] leverage on the MAB line of proof (as in Agrawal and Goyal

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

[2012a]) classifying arms as saturated and unsaturated depending on wether their standard deviation is smaller or bigger than their gap to the optimal arm.[1] While for unsaturated arms the regret is related to their standard deviation that decreases over time, they prove that TS has a small (but constant) probability to select saturated arms and thus it achieves a regret $\widetilde{O}\big(d^{3/2}\sqrt{T}\big)$.

Contributions. The major contributions of this paper are: **1)** Following the intuition of Agrawal and Goyal [2012b], we show that the TS does not need to sample from an actual Bayesian posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret. In particular, we show that the distribution should *over-sample* w.r.t. the standard least-squares confidence ellipsoid by a factor $\sqrt{d}$ to guarantee a constant probability of being optimistic. **2)** We provide an alternative proof of TS achieving the same result as Agrawal and Goyal [2012b]. One of our major finding is that, leveraging on the properties of support functions from convex geometry, we are able to prove that the regret is related to the gradient of the objective function, that is ultimately controlled by the norm of the optimal arms associated to any optimistic parameter θ. This provides a novel insight on the fact that whenever an optimistic parameter θ_t is chosen, not only its instantaneous regret is small but the corresponding optimal arm $x_t = \arg\max_x x^\mathsf{T}\theta_t$ represents a *useful exploration* step that improves the accuracy of the estimation of $\theta^\star$ over dimensions which are relevant to reduce regret in any subsequent non-optimistic step. This approach allows us to avoid the introduction of saturated/unsaturated arms and it illustrates why any TS-like algorithm (not necessarily Bayesian) with a constant probability of being optimistic has a bounded regret. **3**) Finally, we show how our proof can be easily adapted to regularized linear optimization (with arbitrary penalty) and to the generalized linear model (GLM), for which we derive the first frequentist regret bound for TS, which was first suggested by Agrawal and Goyal [2012b] as a venue to explore.

2 Preliminaries

The setting. We consider the stochastic linear bandit model. Let $\mathcal{X} \subset \mathbb{R}^d$ be an arbitrary (finite or infinite) set of arms. When an arm $x \in \mathcal{X}$ is pulled, a reward is generated as $r(x) = x^\mathsf{T}\theta^\star + \xi$, where $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown parameter and ξ is a zero-mean noise. An arm $x \in \mathcal{X}$ is evaluated according to its expected reward $x^\mathsf{T}\theta^\star$ and for any $\theta \in \mathbb{R}^d$ we denote the optimal arm and its value by

$$x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta, \qquad J(\theta) = \sup_{x\in\mathcal{X}} x^\mathsf{T}\theta. \quad (1)$$

[1]Here we refer to the definition introduced in the *arXiv* paper, which slightly differs from the original ICML paper.

Then $x^\star = x^\star(\theta^\star)$ is the optimal arm for $\theta^\star$ and $J(\theta^\star)$ is its optimal value. At each step t, the learner selects an arm $x_t \in \mathcal{X}$ based on the past observations (and possibly additional randomization), it observes the reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$, and it suffers a *regret* equal to the difference in expected reward between the optimal arm $x^\star$ and the arm x_t. All the information observed up to time t is encoded in the filtration $\mathcal{F}_t^x = (\mathcal{F}_1, \sigma(x_1, r_2, \dots, r_t, x_t))$, where $\mathcal{F}_1$ contains any prior knowledge (e.g., the bound S). The objective of the learner is to minimize the *cumulative regret* up to step T, i.e., $R(T) = \sum_{t=1}^T \big(x^{\star,\mathsf{T}}\theta^\star - x_t^\mathsf{T}\theta^\star\big)$.

We introduce general assumptions on the structure of the problem and on the noise ξ_{t+1}.

Assumption 1 (Arm set)**.** *The arm set $\mathcal{X}$ is a bounded closed (and hence compact) subset of $\mathbb{R}^d$ such that $\|x\| \le X$ for all $x \in \mathcal{X}$. We also assume $X = 1$.*

Assumption 2 (Bandit parameter)**.** *There exists $S \in \mathbb{R}^+$ such that $\|\theta^\star\| \le S$ and S is known.*

Assumption 3 (Noise)**.** *The noise process $\{\xi_t\}_t$ is a martingale difference sequence given $\mathcal{F}_t^x$ and it is conditionally R-subgaussian for some constant $R \ge 0$,*

$$\begin{aligned} &\forall t \ge 1,\ \mathbb{E}\left[\xi_{t+1}|\mathcal{F}_t^x\right] = 0, \\ &\forall \alpha \in \mathbb{R},\ \mathbb{E}\left[e^{\alpha\xi_{t+1}} \mid \mathcal{F}_t^x\right] \le \exp\big(\alpha^2 R^2/2\big). \end{aligned} \quad (2)$$

Technical tools. Let $(x_1, \dots, x_t) \in \mathcal{X}^t$ be a sequence of arms and $(r_2, \dots, r_{t+1})$ be the corresponding rewards, then $\theta^\star$ can be estimated by regularized least-squares (RLS). For any regularization parameter $\lambda \in \mathbb{R}^+$, the design matrix and the RLS estimate are defined as

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\mathsf{T}, \qquad \widehat{\theta}_t = V_t^{-1}\sum_{s=1}^{t-1} x_s r_{s+1}. \quad (3)$$

For any positive semi-definite matrix A, the weighted 2-norm $\|.\|_A$ is defined by $\|x\|_A^2 = x^\mathsf{T}Ax$. We recall an important concentration inequality for RLS estimates.

Proposition 1 (Thm. 2 in Abbasi-Yadkori et al. [2011a])**.** *For any $\delta \in (0,1)$, under Asm. 1,2, and 3, for any $\mathcal{F}_t^x$-adapted sequence $(x_1, \dots, x_t)$, the RLS estimator $\widehat{\theta}_t$ is such that for any fixed $t \ge 1$,*

$$\begin{aligned} &\|\widehat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta), \\ &\forall x \in \mathbb{R}^d,\ |x^\mathsf{T}(\widehat{\theta}_t - \theta^\star)| \le \|x\|_{V_t^{-1}}\beta_t(\delta), \end{aligned} \quad (4)$$

w.p. $1-\delta$ (w.r.t. the noise $\{\xi_t\}_t$ and any source of randomization in the choice of the arms), where

$$\beta_t(\delta) = R\sqrt{2\log\frac{(\lambda+t)^{d/2}\lambda^{-d/2}}{\delta}} + \sqrt{\lambda}S. \quad (5)$$

At step t, we define the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}} = \big\{\theta \in \mathbb{R}^d \mid ||\theta - \widehat{\theta}_t||_{V_t} \le \beta_t(\delta')\big\}$ centered in $\widehat{\theta}_t$ with orientation defined

Input: $\hat{\theta}_1$, $V_1 = \lambda I$, δ, T
Set $\delta' = \delta/(4T)$
for $t = \{1, \dots, T\}$ **do**
  Sample $\eta_t \sim \mathcal{D}^{\mathrm{TS}}$
  Compute parameter

$$\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta') V_t^{-1/2} \eta_t$$

  Compute optimal arm

$$x_t = x^\star(\widetilde{\theta}_t) = \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \widetilde{\theta}_t$$

  Pull arm x_t and observe reward r_{t+1}
  Compute V_{t+1} and $\widehat{\theta}_{t+1}$ using Eq. 3
end for

Figure 1: Thompson sampling algorithm.

by V_t and radius $\beta_t(\delta')$, where $\delta' = \delta/4T$.From Eq. 4 we have that $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$ with high probability. Finally, we report a standard result of RLS that, together with Prop. 1, shows that the prediction error on the x_ts used to construct the estimator $\widehat{\theta}_t$ is cumulatively small.

Proposition 2. *Let $\lambda \geq 1$, for any arbitrary sequence $(x_1, x_2, \dots, x_t) \in \mathcal{X}^t$ let V_{t+1} be the corresponding design matrix (Eq. 3), then*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \leq 2 \log \frac{\det(V_{t+1})}{\det(\lambda I)} \leq 2d \log\left(1 + \frac{t}{\lambda}\right). \quad (6)$$

This result plays a central role in most of the proofs for linear bandit, since the regret is usually related to $||x_s||_{V_s^{-1}}$ and Prop. 2 is used to bound its cumulative sum. While Agrawal and Goyal [2012b] achieve this by dividing arms in saturated and unsaturated, we follow a different path that leverages on the core features of the problem (structure of $J(\theta)$) and of TS (probability of being optimistic).

3 Linear Thompson Sampling

Agrawal and Goyal [2012b] define TS for linear bandit as a Bayesian algorithm where a Gaussian prior over $\theta^\star$ is updated according to the observed rewards, a random sample is drawn from the posterior, and the corresponding optimal arm is selected at each step.

As hinted by Agrawal and Goyal [2012b], we show that TS can be defined as a generic randomized algorithm constructed on the RLS-estimate rather than an algorithm sampling from a Bayesian posterior (see Fig. 1). At any step t, given RLS-estimate $\widehat{\theta}_t$ and the design matrix V_t, TS samples a *perturbed* parameter $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta') V_t^{-1/2} \eta_t, \quad (7)$$

where η_t is a random sample drawn i.i.d. from a suitable multivariate distribution $\mathcal{D}^{\mathrm{TS}}$, which does not need to be associated with an actual posterior over $\theta^\star$. Then the optimal arm $x_t = x^\star(\widetilde{\theta}_t)$ is chosen, a reward r_{t+1} is observed and V_t and $\widehat{\theta}_t$ are updated according to Eq. 3. Notice that the resulting distribution on θ_t is obtained rotating η_t by the design matrix V_t and scaling it by $\beta_t(\delta)$. The computational complexity of TS is determined by the linear optimization problem solved when computing $x^\star(\widetilde{\theta}_t)$ and by the sampling process from $\mathcal{D}^{\mathrm{TS}}$. This is in contrast with OFUL [Abbasi-Yadkori et al., 2011a], which requires solving a bilinear optimization problem (i.e., $\arg\max_\theta \max_x x^\mathsf{T}\theta$).

The key aspect to ensure small regret is that the perturbation η_t is distributed so that TS explores *enough* but *not too much*. This translates into the following conditions on $\mathcal{D}^{\mathrm{TS}}$.

Definition 1. *$\mathcal{D}^{\mathrm{TS}}$ is a multivariate distribution on $\mathbb{R}^d$ absolutely continuous with respect to the Lebesgue measure which satisfies the following properties:*

1. *(anti-concentration) there exists a strictly positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\| = 1$,*

$$\mathbb{P}_{\eta \sim \mathcal{D}^{\mathrm{TS}}}\left(u^\mathsf{T}\eta \geq 1\right) \geq p,$$

2. *(concentration) there exists c, c' positive constants such that $\forall \delta \in (0,1)$*

$$\mathbb{P}_{\eta \sim \mathcal{D}^{\mathrm{TS}}}\left(\|\eta\| \leq \sqrt{cd \log \frac{c'd}{\delta}}\right) \geq 1 - \delta.$$

Once interpreted in the construction of $\widetilde{\theta}_t$, the definition of $\mathcal{D}^{\mathrm{TS}}$ basically requires TS to explore far enough from $\widehat{\theta}_t$ (anti-concentration) but not too much (concentration). This implies that TS performs "useful" exploration with enough frequency (notably it performs optimistic steps), but without selecting arms with too large regret. Let $\gamma_t(\delta) = \beta_t(\delta')\sqrt{cd \log(c'd/\delta)}$, then we introduce the high-probability ellipsoid $\mathcal{E}_t^{\mathrm{TS}} = \{\theta \in \mathbb{R}^d \mid \|\theta - \widehat{\theta}_t\|_{V_t} \leq \gamma_t(\delta')\}$. The difference between $\mathcal{E}_t^{\mathrm{RLS}}$ and $\mathcal{E}_t^{\mathrm{TS}}$ lies in the additional factor $\sqrt{d}$ in the definition of $\gamma_t(\delta)$ and it is crucial for both concentration and anti-concentration to hold at the same time. In Sect. 5 we prove that any distribution satisfying the conditions in Def. 1 introduces the right amount of randomness to achieve the desired regret without actually satisfying any Bayesian assumption. Def. 1 includes the Gaussian prior used by Agrawal and Goyal [2012b], but also other types of distributions such as the uniform on the unit ball $\mathcal{B}_d(0, \sqrt{d})$ or distributions concentrated on the boundary of $\mathcal{E}_t^{\mathrm{TS}}$ (refer to App. A for exact values of c, c', and p for uniform and Gaussian distributions).

4 Sketch of the proof

In this section we report a sketch of the proof providing a geometric intuition on the behavior of TS and how its actions (i.e., the sampled $\widetilde{\theta}_t$ and the corresponding

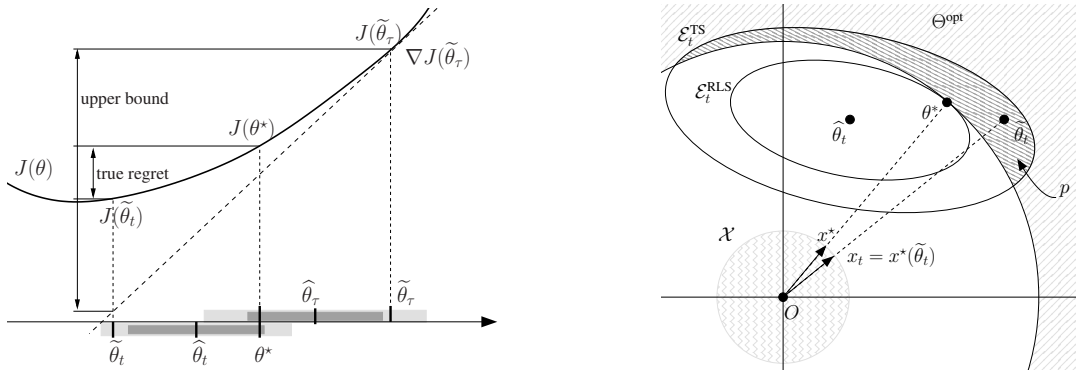

Figure 2: Illustration of the steps **2)** and **3)** of the proof in $\mathbb{R}^1$ and $\mathbb{R}^2$. *Left:* The regret at step t could be bounded by the gradient of the function J at a previous optimistic $\widetilde{\theta}_\tau$ times the distance between $\widetilde{\theta}_\tau$ and the current $\widetilde{\theta}_t$. Notice that $\theta^\star$ is always included in $\mathcal{E}_t^{\mathrm{RLS}}$ (in dark gray) and thus $\widetilde{\theta}$s sampled from $\mathcal{E}_t^{\mathrm{TS}}$ (in light gray) are never too far. *Right:* TS has a constant probability of being optimistic thanks to the over-sampling of $\mathcal{D}^{\mathrm{TS}}$.

x_t) influence the regret. For the sake of illustration, we consider the unit ball $\mathcal{X} = \{\|x\| \leq 1\}$, such that the optimal arm is just the projection of θ on the ball ($x^\star(\theta) = \theta/\|\theta\|$), and the optimal value is $J(\theta) = \theta^\mathsf{T}\theta/\|\theta\| = \|\theta\|$. We start by decomposing the regret using the definition of $J(\theta)$ as

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \Big(\big(x^{\star,\mathsf{T}}\theta^\star - x_t^\mathsf{T}\widetilde{\theta}_t\big) + \big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star\big) \Big) \\
&= \underbrace{\sum_{t=1}^{T} \big(J(\theta^\star) - J(\widetilde{\theta}_t)\big)}_{R^{\mathrm{TS}}(T)} + \underbrace{\sum_{t=1}^{T} \big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\theta^\star\big)}_{R^{\mathrm{RLS}}(T)},
\end{aligned}
$$

where R^{TS} depends on the randomization of TS and R^{RLS} mostly depends on the properties of RLS.

Bounding $R^{\mathbf{RLS}}(T)$. The decomposition $R^{\mathrm{RLS}}(T) = \sum_{t=1}^{T} \big(x_t^\mathsf{T}\widehat{\theta}_t - x_t^\mathsf{T}\theta^\star\big) + \sum_{t=1}^{T} \big(x_t^\mathsf{T}\widetilde{\theta}_t - x_t^\mathsf{T}\widehat{\theta}_t\big)$, shows that both RLS estimate $\widehat{\theta}_t$ and TS parameter $\widetilde{\theta}_t$ should concentrate appropriately. Since at each step t, $\widetilde{\theta}_t$ is sampled from $\mathcal{D}^{\mathrm{TS}}$, the second term is kept under control by construction, while the first sum deals with the prediction error of RLS. As opposed to R^{TS}, this error is not related to the exploration scheme and it is small for any sequence of arms. Intuitively, this is due to the fact that the RLS estimate is the minimizer of the regularized cumulative squared error $\widehat{\theta}_{T+1} = \arg\min_\theta \big(\sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\theta|^2 + \lambda\|\theta\|^2 \big)$, so that $x_t^\mathsf{T}\widehat{\theta}_{T+1}$ is an accurate prediction *on the arms observed so far*. The RLS minimizes the error in "hindsight" (i.e., after all rewards up to T) and therefore it also controls the *online* error $|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_{t+1}|^2$, since by induction

$$
\sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{T+1}|^2 + \lambda\|\widehat{\theta}_{T+1}\|^2 \geq
\sum_{t=1}^{T} |r_{t+1} - x_t^\mathsf{T}\widehat{\theta}_{t+1}|^2 + \lambda\|\widehat{\theta}_1\|^2.
$$

Having a small *online* error also implies a small *prediction* error $|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_t|^2$. In fact, using a recursive version of Eq. 3, we have $\hat{\theta}_{t+1} = \hat{\theta}_t + V_t^{-1}x_t(1 + \|x_t\|^2_{V_t^{-1}})^{-1}(r_{+1} - x_t^\mathsf{T}\hat{\theta}_t)$, which, together with $\|x_t\|^2_{V_t^{-1}} \leq 1/\lambda$, leads to $|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_{t+1}| \geq \frac{\lambda}{1+\lambda}|r_{t+1} - x_t^\mathsf{T}\hat{\theta}_t|$. Since the cumulative prediction error is small, then the associated regret $\sum_{t=1}^{T} |x_t^\mathsf{T}\widehat{\theta}_t - x_t^\mathsf{T}\theta^\star|$ is also small. This result can be seen as an intrinsic *on-policy* error guarantee of RLS. Nonetheless, notice that while RLS minimizes the prediction error for any sequence of arms, this does not imply the consistency of the estimator. For instance, when the same arm x is repeatedly played, the unknown parameter $\theta^\star$ is well-estimated in the direction of x (thus making $R^{\mathrm{RLS}}(T)$ small) but it is poorly estimated in any other directions. This shows the need for a careful exploration strategy to recover consistency and hence a sub-linear regret.

Bounding $R^{\mathbf{TS}}(T)$. We denote by $R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t)$ each term in $R^{\mathrm{TS}}(T)$. For optimistic algorithms this term is bounded by 0 at any step since w.h.p. $J(\widetilde{\theta}_t) \geq J(\theta^*)$ by construction. In the Bayesian regret analysis of TS, this term is equal to 0 by assumption that θ^* is drawn from the same prior as $\widetilde{\theta}_t$. On the other hand, in the frequentist analysis, we have to control the deviations caused by the random sampling of $\widetilde{\theta}_t$. This is achieved by showing that the arms selected by TS provide "useful" information about $\theta^\star$ and contribute to keep the regret small. We follow three steps: **1)** we show that the regret is related to the sensitivity of J w.r.t. the errors in estimating $\theta^\star$ and we bound the regret with the gradient of $J(\theta)$ at any *optimistic* θ; **2)** we show how the gradient in a point θ is intrinsically related to its corresponding optimal arm $x^\star(\theta)$; **3)** since we prove that TS is frequently optimistic, then we can finally link $x^\star(\theta)$ to $x_t = x^\star(\widetilde{\theta}_t)$ and Prop. 2 allows us to finally bound the overall regret.

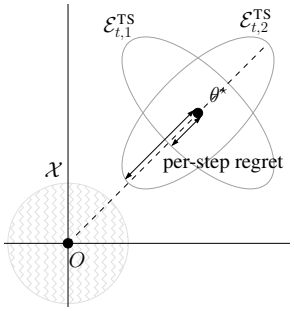

Figure 3: While $\mathcal{E}_{t,1}^{\mathrm{TS}}$ and $\mathcal{E}_{t,2}^{\mathrm{TS}}$ have an equivalent accurate estimation of $\theta^\star$, $\mathcal{E}_{t,1}^{\mathrm{TS}}$ has smaller regret than $\mathcal{E}_{t,2}^{\mathrm{TS}}$.

Step 1 (regret and sensitivity of J). We first show why the exploration of TS should be *well adapted* to $J(\theta)$. Using the definition of $J(\theta) = \|\theta\|$ we have

$$R_t^{\mathrm{TS}} = J(\theta^\star) - J(\widetilde{\theta}_t) = \|\theta^\star\| - \|\widetilde{\theta}_t\| \le \|\theta^\star - \widetilde{\theta}_t\| \le \frac{\|\theta^\star - \widetilde{\theta}_t\|_{V_t}}{\sqrt{\lambda_{\min,t}}},$$

where $\lambda_{\min,t}$ is the smallest eigenvalue of V_t. This bound shows that it is sufficient to estimate $\theta^\star$ accurately over all its components (i.e., $\lambda_{\min,t}$ tends to infinity) to obtain a no-regret algorithm. Nonetheless, the desired regret bound of $O(\sqrt{T})$ is obtained only if $\lambda_{\min,t}$ increases as $O(t)$. While this could be achieved by a fully explorative algorithm (e.g., a round robin over the canonic vectors e_i reduces the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ to a ball of radius $\lambda_{\min,t}$), it would severely increase the second term of $R^{\mathrm{RLS}}(T)$ and cause an overall linear regret[2]. Fortunately, inspecting the definition of R_t^{TS} reveals that not all components of $\theta^\star$ must be equally well estimated. In fact, we have w.h.p. that

$$R_t^{\mathrm{TS}} \le \sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \big(J(\theta) - J(\theta')\big).$$

This shows that R_t^{TS} is determined by the *diameter* of ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ w.r.t. J, which suggests that the estimation of $\theta^\star$ should be more accurate on the dimensions on which J is more sensitive. In the case of $\mathcal{X}$ unit ball, the most sensitive direction of J is $\theta^\star/\|\theta^\star\|$ itself and Fig. 3 illustrates two opposite cases where the accuracy in the estimation of $\theta^\star$ is the same (i.e., V_t has the same eigenvalues) but the regret may be very different. Let $\Theta^{\mathrm{opt}} = \{\theta : J(\theta) \ge J(\theta^\star)\}$ be the set of optimistic parameters. In our example $J(\theta) = \|\theta\|$ is convex thus we can make explicit the dependency of the regret on the sensitivity of J through its gradient evaluated at any $\theta \in \Theta^{\mathrm{opt}}$ as (see Prop. 3 for the general case)

$$R_t^{\mathrm{TS}} \le \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta) - J(\theta') \le \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \nabla J(\theta)^{\mathsf{T}}(\theta - \theta'),$$

which shows that the regret of non-optimistic $\widetilde{\theta}_t$ is bounded by the gradient of $J(\theta)$ at any optimistic θ and its distance to any other point in the TS ellipsoid.

Step 2 (sensitivity of J and optimal arm). According to Prop. 1, the difference $\theta - \theta'$ in the previous inequality is well controlled whenever θ belongs to the ellipsoid, while the first term cannot be immediately controlled by the algorithm. Nonetheless, we notice that since $J(\theta) = \|\theta\|$, then $\nabla J(\theta) = \theta/\|\theta\| = x^\star(\theta)$ (see Lem. 2 for the general case). This shows how selecting the optimal arm associated to an optimistic θ is equivalent to controlling the gradient of J, which results in

$$R_t^{\mathrm{TS}} \le \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} x^\star(\theta)^{\mathsf{T}}(\theta - \theta').$$

From Prop. 2, we could conclude that the regret would be cumulatively small if $x^\star(\theta)$ corresponded to the arms chosen by the TS ($x_t = x^\star(\widetilde{\theta}_t)$). As a result, we need a θ **1)** that is optimistic (i.e., $\theta \in \Theta^{\mathrm{opt}}$), **2)** it belongs or is close to the ellipsoid $\mathcal{E}_t^{\mathrm{TS}}$ and **3)** it is used to select an arm x_t. The first two requirements are at the core of the choice of the TS distribution in Def. 1 where the anticoncentration property guarantees enough probability to be optimistic, while the concentration property implies that $\widetilde{\theta}$s are within a small ellipsoid. Let $\tau < t$ be any step when TS selects $\widetilde{\theta}_\tau \in \Theta^{\mathrm{opt}}$ with corresponding arm $x_\tau = x^\star(\widetilde{\theta}_\tau)$, then we have (see an illustration of this bound in Fig. 2 in the 1-d case)

$$R_t^{\mathrm{TS}} \le \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} x_\tau^{\mathsf{T}}(\widetilde{\theta}_\tau - \theta') \le \|x_\tau\|_{V_\tau^{-1}} \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta}_\tau - \theta'\|_{V_\tau}.$$

Introducing $\theta^\star$ and using the fact that the design matrices forms a non-decreasing sequence (e.g. $V_\tau \le V_t$), we decompose

$$\begin{aligned}\sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta}_\tau - \theta'\|_{V_\tau} &\le \|\widetilde{\theta}_\tau - \theta^\star\|_{V_\tau} + \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta^\star - \theta'\|_{V_\tau} \\ &\le \|\widetilde{\theta}_\tau - \theta^\star\|_{V_\tau} + \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta^\star - \theta'\|_{V_t}\end{aligned}$$

Since by Prop. 1 $\theta^\star$ is contained in all confidence ellipsoids with high probability, then

$$\begin{aligned}R_t^{\mathrm{TS}} &\le \big(\beta_\tau(\delta') + \gamma_\tau(\delta') + \beta_t(\delta') + \gamma_t(\delta')\big)\|x_\tau\|_{V_\tau^{-1}} \\ &\le \big(2\beta_T(\delta') + 2\gamma_T(\delta')\big)\|x_\tau\|_{V_\tau^{-1}}.\end{aligned}$$

Let K be the number of times $\widetilde{\theta}_t \in \Theta^{\mathrm{opt}}$, t_k the corresponding steps, and $\nu_k = t_k - t_{k-1}$, then the final regret can be written as

$$R^{\mathrm{TS}}(T) \le 2\big(\beta_T(\delta') + \gamma_T(\delta')\big) \sum_{k=1}^{K} \nu_k \|x_{t_k}\|_{V_{t_k}^{-1}}.$$

Step 3 (optimism). This bound shows the importance that TS is optimistic with high frequency. In fact,

[2]This happens because x_t would be optimal w.r.t. a $\widetilde{\theta}_t$, which is *not* in the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$.

whenever $\widetilde{\theta}_t$ is in Θ^{opt}, not only the corresponding instantaneous regret R_t^{TS} is upper-bounded by 0, but the exploration performed by playing arm $x^\star(\widetilde{\theta}_t)$ has also a positive impact in controlling the regret for any subsequent non-optimistic step. Consider the extreme case when TS is never optimistic, then $K=1$, $\nu_1 = T$ and $R^{\mathrm{TS}}(T) = O(T)$. On the other hand, if TS is optimistic with a constant frequency, then we can easily show that $R^{\mathrm{TS}}(T)$ is bounded by $\widetilde{O}(\sqrt{T})$. Consider the case where an optimistic θ is chosen with probability p. Since $\mathbb{E}[\nu_k] = 1/p$, we can prove that w.h.p. $R^{\mathrm{TS}}(T) \leq \widetilde{O}(1/p\sqrt{T})$ by Cauchy-Schwarz and Prop. 2 applied to $\sum_{k=1}^{K} \|x_{t_k}\|^2_{V_{t_k}^{-1}}$, where $K \approx T$. Unfortunately, sampling $\widetilde{\theta}_t$ from the RLS ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$ may have a very small probability of being optimistic (see e.g., Fig. 2, where sampling uniformly in $\mathcal{E}_t^{\mathrm{RLS}}$ has zero probability to return a $\widetilde{\theta}_t \in \Theta^{\mathrm{opt}}$). For this reason, TS is required to draw $\widetilde{\theta}_t$ from a distribution *over-sampling* by a factor $\sqrt{d}$ w.r.t. $\mathcal{E}_t^{\mathrm{RLS}}$ as in the definition of $\mathcal{D}^{\mathrm{TS}}$. This guarantees a fixed probability p of being optimistic (see Lem. 3) and the final desired regret.

5 Formal Proof

In this section we report the main steps of the regret analysis, while we postpone technical lemmas to the supplementary material. We prove the following result.

Theorem 1. *Under assumptions 1,2,3, the regret of* TS *is bounded w.p.* $1-\delta$ *as* ($\delta' = \frac{\delta}{4T}$)

$$R(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')(1+4/p)\big)\sqrt{2Td\log\big(1+\frac{T}{\lambda}\big)} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}. \tag{8}$$

As anticipated in introduction, this bound is of order $\widetilde{O}(d^{3/2}\sqrt{T})$ and it entirely matches the result of Agrawal and Goyal [2012b]. The analysis of the regret requires extra care in the definition of the filtrations. While in analyzing R^{RLS} we consider all the knowledge up to step t (i.e., including the sampled parameter $\widetilde{\theta}_t$), in R^{TS} we need to study the randomness of $\widetilde{\theta}_t$ conditional on all the information before sampling η_t. We introduce an additional filtration besides $\mathcal{F}_t^x$.

Definition 2. *We define the filtration $\mathcal{F}_t$ as the accumulated information up to time t before the sampling procedure, i.e., $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, r_2, x_2, \dots, x_{t-1}, r_{t-1}))$.*

Notice that $\widehat{\theta}_t$ and V_t^{-1} are both $\mathcal{F}_t$ and $\mathcal{F}_t^x$ adapted, while $\widetilde{\theta}_t$ is a random variable w.r.t. $\mathcal{F}_t$ and it is fixed when considering $\mathcal{F}_t^x$. Hence we have $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_2^x \subset \mathcal{F}_3 \subset \mathcal{F}_3^x, \dots$. We are now ready to introduce the high-probability events we use in the rest of the proof.

Definition 3. *Let $\delta \in (0,1)$ and $\delta' = \delta/(4T)$ and $t \in [1,T]$. We define $\widehat{E}_t$ as the event where the RLS estimate concentrates around $\theta^\star$ for all steps $s \leq t$, i.e., $\widehat{E}_t = \big\{\forall s \leq t,\ \|\widehat{\theta}_s - \theta^\star\|_{V_s} \leq \beta_s(\delta')\big\}$. We also define $\widetilde{E}_t$ as the event where the sampled parameter $\widetilde{\theta}_s$ concentrates around $\widehat{\theta}_s$ for all steps $s \leq t$, i.e., $\widetilde{E}_t = \big\{\forall s \leq t,\ \|\widetilde{\theta}_s - \widehat{\theta}_s\|_{V_s} \leq \gamma_s(\delta')\big\}$.*

Then we have that $\widehat{E} := \widehat{E}_T \subset \cdots \subset \widehat{E}_1$, $\widetilde{E} := \widetilde{E}_T \subset \cdots \subset \widetilde{E}_1$ and we use $E_t = \widehat{E}_t \cap \widetilde{E}_t$ and $E = \widehat{E} \cap \widetilde{E}$.

Lemma 1. *Under Asm. 2, 3 we have* $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \frac{\delta}{2}$.

Conditioned on $\mathcal{F}_t$ and event $\widehat{E}_t$, we have $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$, while on event $\widetilde{E}_t$ we have $\widetilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$, then we directly bound the regret as

$$\begin{aligned} R(T) &\leq \sum_{t=1}^{T} \big(J(\theta^\star) - J(\widetilde{\theta}_t)\big)\mathbb{1}\{E_t\} + \sum_{t=1}^{T} \big(x_t^T\widetilde{\theta}_t - x_t^T\theta^\star\big)\mathbb{1}\{E_t\} \\ &\leq \sum_{t=1}^{T} R_t^{\mathrm{TS}}\mathbb{1}\{E_t\} + \sum_{t=1}^{T} R_t^{\mathrm{RLS}}\mathbb{1}\{E_t\}, \end{aligned}$$

w.p. $1-\delta/2$. In the interest of space we only report the formal proof to bound R_t^{TS}, while the bound on $R^{\mathrm{RLS}}(T)$ and the overall regret is postponed to App. D.

Similar to the sketch in Sect. 4, the proof follows three steps: **1)** we use the convexity of J to upper-bound the regret by its expectation conditioned on being optimistic and to relate it to the gradient of J, **2)** we relate the gradient of J to the arms chosen by TS over time, **3)** we show that despite the randomization, TS has a constant probability of being optimistic.

Step 1 (Regret and gradient of $J(\theta)$). On event E_t, $\widetilde{\theta}_t$ belongs to $\mathcal{E}_t^{\mathrm{TS}}$ and thus

$$R_t^{\mathrm{TS}}\mathbb{1}\{E_t\} \leq \big(J(\theta^\star) - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta)\big)\mathbb{1}\{\widehat{E}_t\}.$$

Recalling that Θ^{opt} is the set of all optimistic θs, we can bound the previous expression by the expectation over any random choice of $\widetilde{\theta}$ in $\Theta_t^{\mathrm{opt}} := \Theta^{\mathrm{opt}} \cap \mathcal{E}_t^{\mathrm{TS}}$ where we restrict the optimistic set to the high-probability sampling ellipsoid, that is

$$R_t^{\mathrm{TS}} \leq \mathbb{E}\Big[\big(J(\widetilde{\theta}) - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} J(\theta)\big)\mathbb{1}\{\widehat{E}_t\}\Big|\mathcal{F}_t, \widetilde{\theta} \in \Theta_t^{\mathrm{opt}}\Big],$$

where $\widetilde{\theta} = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$ with $\eta \sim \mathcal{D}^{\mathrm{TS}}$ is the TS sampling distribution. We now rely on the following characterization of $J(\theta)$ (see App. C).

Proposition 3. *For any set of arm $\mathcal{X}$ satisfying Asm. 1, $J(\theta) = \sup_x x^\mathsf{T}\theta$ has the following properties: **1)** J is real-valued as the supremum is attained in $\mathcal{X}$, **2)** J is convex on $\mathbb{R}^d$, **3)** J is continuous with continuous first derivative except for a zero-measure set w.r.t. the Lebesgue's measure.*

These properties follow from the fact that J is the *support function* of $\mathcal{X}$ and it shows that J is convex for any arm set $\mathcal{X}$. As a result, we can directly relate R_t^{TS} to the gradient of J as

$$R_t^{\mathrm{TS}} \le \mathbb{E}\Big[\sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}} \nabla J(\widetilde{\theta})^{\mathsf{T}}(\widetilde{\theta}-\theta)\mathbb{1}\{\widehat{E}_t\}\Big|\mathcal{F}_t,\widetilde{\theta}\in\Theta_t^{\mathrm{opt}}\Big]$$
$$\le \mathbb{E}\Big[\|\nabla J(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}}\|\widetilde{\theta}-\theta\|_{V_t}\Big|\mathcal{F}_t,\widetilde{\theta}\in\Theta_t^{\mathrm{opt}},\widehat{E}_t\Big]\mathbb{P}(\widehat{E}_t)$$
$$\le 2\gamma_t(\delta')\mathbb{E}\Big[\|\nabla J(\widetilde{\theta})\|_{V_t^{-1}}\Big|\mathcal{F}_t,\widetilde{\theta}\in\Theta_t^{\mathrm{opt}},\widehat{E}_t\Big]\mathbb{P}(\widehat{E}_t)$$

where we use Cauchy-Schwarz, we "push" the event $\widehat{E}_t$ into the conditioning and we use the fact that $\widetilde{\theta}\in\mathcal{E}_t^{\mathrm{TS}}$.

Step 2 (From gradient of $J(\theta)$ to optimal arm $x^\star(\theta)$). In the sketch of the proof there was a direct relationship between $\nabla J(\theta)$ and the optimal arm corresponding to θ by direct construction. In the next lemma, we show that this connection is true for any arm set $\mathcal{X}$ (proof in App. C).

Lemma 2. *Under Asm. 1, for any $\theta\in\mathbb{R}^d$, we have $\nabla J(\theta)=x^\star(\theta)$ except for a zero-measure set w.r.t. the Lebesgue's measure.*

This property strongly connects the exploration of TS to the actual regret. In fact, together with Prop. 2, it implies that selecting the optimal arm associated with any optimistic θ is equivalent to reducing the gradient of J and ultimately the regret R_t^{TS}. This motivates the next step where we show that since TS is often optimistic, then the arm $x_t=x^\star(\widetilde{\theta}_t)$ contributes to the reduction of the regret.

Step 3 (Optimism). The optimism of TS is a direct consequence of the convexity of J and the fact that the distribution of η is oversampling by a factor $\sqrt{d}$ w.r.t. the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$ (proof in App. D).

Lemma 3. *Let $\Theta_t^{opt}:=\{\theta\in\mathbb{R}^d|J(\theta)\ge J(\theta^\star)\}\cap\mathcal{E}_t^{\mathrm{TS}}$ be the set of optimistic parameters, $\widetilde{\theta}_t=\widehat{\theta}_t+\beta_t(\delta')V_t^{-1/2}\eta$ with $\eta\sim\mathcal{D}^{\mathrm{TS}}$, then $\forall t\ge1$, $\mathbb{P}\big(\widetilde{\theta}_t\in\Theta_t^{opt}|\mathcal{F}_t,\widehat{E}_t\big)\ge p/2$.*

Let $f(\widetilde{\theta}_t)$ be an arbitrary non-negative function of $\widetilde{\theta}_t$, then we can write the full expectation as

$$\mathbb{E}[f(\widetilde{\theta}_t)|\mathcal{F}_t,\widehat{E}_t]\ge\mathbb{E}\big[f(\widetilde{\theta}_t)|\widetilde{\theta}_t\in\Theta_t^{\mathrm{opt}},\mathcal{F}_t,\widehat{E}_t\big]\mathbb{P}\big(\widetilde{\theta}_t\in\Theta_t^{\mathrm{opt}}\big)$$
$$\ge\mathbb{E}\big[f(\widetilde{\theta}_t)|\widetilde{\theta}_t\in\Theta_t^{\mathrm{opt}},\mathcal{F}_t,\widehat{E}_t\big]p/2.$$

Setting $f(\widetilde{\theta})=2\gamma_t(\delta')\|x^\star(\widetilde{\theta})\|_{V_t^{-1}}$ and reintegrating $\widehat{E}_t$, we have $R_t^{\mathrm{TS}}\le 4\gamma_t(\delta')/p\;\mathbb{E}\big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}}\mathbb{1}\{\widehat{E}_t\}|\mathcal{F}_t\big]$ where $2/p$ can be interpreted as the expected time between any two optimistic samples. Finally, we can use Azuma's inequality to obtain the final bound with probability at least $1-\delta/2$

$$R^{\mathrm{TS}}(T)\le\frac{4\gamma_T(\delta')}{p}\Big(\sum_{t=1}^T\|x_t\|_{V_t^{-1}}+\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}\Big),$$

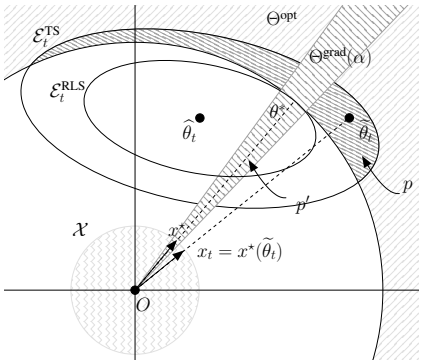

Figure 4: Illustration of the non-optimistic region that could contribute to reduce the regret.

where x_t is the optimal arm $x^\star(\widetilde{\theta}_t)$ selected by TS. The proof is concluded using Cauchy-Schwarz and Prop. 2 to bound $R^{\mathrm{TS}}(T)$ and Prop. 1 to bound $R^{\mathrm{RLS}}(T)$.

6 Discussion

We developed an alternative proof for TS in LB with novel insights on the core elements of the algorithm (*optimism*) and the structure of the problem (*support function $J(\theta)$*). There are a number of possible applications of our results and future directions of investigation.

Regularized linear optimization. Our proof holds for any arm set $\mathcal{X}$ and the corresponding constrained optimization problem $\max_{x\in\mathcal{X}}x^{\mathsf{T}}\theta^\star$. Similarly, we can apply it to any regularized linear optimization problem $\max_{x\in\mathbb{R}^d}f_{\mu,c}(x;\theta)$, with $f_{\mu,c}(x;\theta)=x^{\mathsf{T}}\theta^\star+\mu c(x)$, where μ is a constant and $c(x)$ is an arbitrary penalty function of x (e.g., norm-regularization). While there always exists a set of of constraints (corresponding to a set of arms $\mathcal{X}_{c,\mu,\theta^\star}$) such that the solution to the constrained and regularized problems coincide, such mapping is often unknown (e.g., $c(x)=\|x\|_1$) and thus TS cannot be run on $\mathcal{X}_{c,\mu,\theta^\star}$ but we need to directly deal with the regularized problem (i.e., sampling $\widetilde{\theta}_t$ and pulling arm $x_t=\arg\max_x f_{\mu,c}(x;\widetilde{\theta}_t)$). In this case, it can be seen that the three main steps of our proof still hold. In fact (see App. G), **1)** $J(\theta)$ is convex, **2)** the gradient of $J(\theta)$ corresponds to the optimal arm $x^*(\theta)$, **3)** Lemma 3 holds unchanged since it relies on the convexity of $J(\theta)$ and the TS distribution $\mathcal{D}^{\mathrm{TS}}$ is the same. As a result, the regret bound follows. On the other hand, the original proof by Agrawal and Goyal [2012b] could be less readily applied to this case. First notice that the mapping from μ and $c(x)$ to the constrained set $\mathcal{X}_{c,\mu,\theta^\star}$ requires the unknown parameter $\theta^\star$. This means that if we pass from the regularized problem to the constrained problem at each time step t, we would be working on a set $\mathcal{X}_{c,\mu,\widetilde{\theta}_t}$ which keeps changing over time. While Agrawal and Goyal [2012b] study the contextual bandit problem where $\mathcal{X}_t$ changes arbitrarily over time, in this case $\mathcal{X}_t$ would

change in response to $\widetilde{\theta}_t$ itself (i.e., it would not available in advance) and the analysis would bound the per-step regret $r_t = \max_{x \in \mathcal{X}_{c,\mu,\widetilde{\theta}_t}} x^\mathsf{T}\theta^\star - x_t^\mathsf{T}\theta$, which does not correspond to the desired regret on $f_{\mu,c}$ (the true optimal arm $x^\star(\theta^\star)$ may not even be in $\mathcal{X}_{c,\mu,\widetilde{\theta}_t}$). Alternatively, we need to formulate a suitable definition of saturated and unsaturated arms for $f_{\mu,c}(x;\theta)$, which does not seem trivial and it may require developing a more *ad-hoc* analysis.

Other extensions. Another interesting setting to study is stochastic combinatorial optimization with semi-bandit feedback, where the arm set is the hypercube and each component of the linear combination $x^\mathsf{T}\theta^\star$ is observed. While Wen et al. [2015] derived a frequentist regret bound for a UCB-like strategy, only a Bayesian regret analysis for TS is available. Exploiting the fact that combinatorial optimization is a special case of linear optimization, our analysis could be adapted to derive frequentist regret bounds. In Sect. F we show that we can deal with more complex scenarios and we derive the first frequentist regret bound for TS in generalized linear models (GLM). Moreover, we can generalize our proof to the other convex optimization problems $\max_{x\in\mathcal{X}} f(x,\theta)$, with linear observations (i.e., $y = x^\mathsf{T}\theta + \xi$). If $f(x,\theta)$ is convex in θ, then $J(\theta)$ is convex as well, thus enabling the possibility to apply our line of proof. More precisely, the gradient of J to the arms played by TS should be related (step 2, Lem. 2) and the on-policy prediction error R^{RLS} measured w.r.t. f should be bounded (Prop. 1). Whenever these properties are satisfied, the regret result follows. Notice that while the original proof by Agrawal and Goyal [2012b] may be extended to cover some of these problems, its requirements are slightly stronger. In fact, the definition of saturated and unsaturated arms relies on the fact that $f(x,\widehat{\theta}_n)$ concentrates to $f(x,\theta)$ for any x, while in our case, we only need to bound R^{RLS}, which corresponds to an *on-policy* error, where prediction errors are measured *on* the specific arms selected by the algorithm. While this advantage may appear abstract, let consider the reinforcement learning case, where $f(x,\theta)$ is the value function of a policy x in an environment θ. In this case, $f(x,\theta^\star)$ may actually be unbounded for some x (i.e., the policy x does not control the system) and the definition of saturated/unsaturated arms could not be easily adjusted. This suggests that our proof could enable covering special RL cases as well. Finally, we remark that defining TS as a randomized algorithm and using convex geometry arguments in its analysis bears a strong resemblance with follow-the-pertubed-leader algorithm and its regret analysis in adversarial linear bandit [Abernethy et al., 2015], suggesting that the two approaches may be strongly related.

About optimism and oversampling. As illustrated in Sect. 4, in the current proof optimistic steps allows to bound the regret of non-optimistic steps. Nonetheless, it can be shown that some non-optimistic steps (even very *pessimistic*!) may indeed be as "informative" as optimistic steps and allow reducing the regret as well. Let consider a minor change in the line of proof, anticipating the use of the convexity of J, i.e.,

$$\begin{aligned} R_t^{\mathrm{TS}} &\le \sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}} \nabla J(\theta^\star)^\mathsf{T}(\theta^\star - \theta)\mathbb{1}\{E_t\} \\ &\le \|\nabla J(\theta^\star)\|_{V_t^{-1}} 2\gamma_t(\delta')\mathbb{1}\{E_t\}. \end{aligned}$$

If we sample a $\widetilde{\theta}$ such that the gradient at it $\nabla J(\widetilde{\theta})$ (i.e., which coincides with the corresponding optimal action $x^\star(\widetilde{\theta})$) has the same V_t^{-1}-norm as $\nabla J(\theta^\star)$, then we could apply the same reasoning as in the original sketch of the proof and bound the regret of any subsequent step. More formally, we can define the set $\Theta_t^{\mathrm{grad}} = \{\theta : \|\nabla J(\theta)\|_{V_t^{-1}} \ge \|\nabla J(\theta^\star)\|_{V_t^{-1}}\}$ of parameters that have larger gradient than $\theta^\star$'s. Similar to Θ^{opt}, if the probability of sampling $\widetilde{\theta}$ in Θ_t^{grad} is lower-bounded by a constant p', then the proof can be reproduced with exactly the same arguments and result. Even further, we could relax the requirement and define $\Theta_t^{\mathrm{grad}}(\alpha) = \{\theta : \|\nabla J(\theta)\|_{V_t^{-1}} \ge \alpha\|\nabla J(\theta^\star)\|_{V_t^{-1}}\}$, with $\alpha < 1$, which would allow even a bigger probability at the cost of an extra constant factor α in the final regret. As illustrated in Fig. 4, in the case $\mathcal{X} = \mathbb{R}^d$, $\Theta_t^{\mathrm{grad}}(\alpha)$ corresponds to a cone whose overlap with $\mathcal{E}^{\mathrm{TS}}$ may actually be even larger than for Θ^{opt}. This illustration shows that the set of *useful* explorative actions does not necessarily coincide with the set of optimistic parameters and that many more parameters in $\mathcal{E}^{\mathrm{TS}}$ may contribute to reduce the regret. This may explain the empirical success of TS and it may suggest that the oversampling by a factor $\sqrt{d}$ to ensure optimism may be a too strong requirement. Finally, we remark that a similar optimistic argument is employed by Agrawal and Goyal [2013] in MAB. Nonetheless, in Lemma 2 they prove that the probability of being optimistic increases over time. This may suggest that $\mathcal{E}^{\mathrm{TS}}$ needs to be only a *constant* fraction bigger than $\mathcal{E}^{\mathrm{RLS}}$, since the initial small probability of being optimistic would tend to a constant (or even to 1) later on during the learning process. Whether this argument holds and how to prove it remains an open question.

Acknowledgement This research is supported in part by a grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, CRIStAL (Centre de Recherche en Informatique et Automatique de Lille), and the French National Research Agency (ANR) under project ExTra-Learn n.ANR-14-CE24-0010-01.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b.

Jacob D. Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, pages 2197–2205, 2015.

Rajeev Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012b.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of AI&Stats*, 2013.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Sebastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems 26*, pages 638–646. 2013.

Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. Chernoff-type bounds for the gaussian error function. *Communications, IEEE Transactions on*, 59(11): 2939–2944, 2011.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257. 2011.

Chao-Ping Chen and Feng Qi. Completely monotonic function associated with the gamma functions and proof of wallis' inequality. *Tamkang Journal of Mathematics*, 36 (4):303–307, 2005.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of The 28th Conference on Learning Theory*, 2015.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT 2012)*, pages 199–213, 2012.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 26*, pages 1448–1456, 2013.

Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

Benedict C May, Nathan Korda, Anthony Lee, and David S Leslie. Optimistic bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13(1):2069–2106, 2012.

Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications: a contemporary approach*. Springer Science & Business Media, 2006.

Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *CoRR*, 2015.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17, 2016.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 1113–1122, 2015.