Journal of Machine Learning Research 1 (year) pages Submitted 4/00; Published 10/00

Linear Thompson Sampling Revisited

Marc Abeille MARC.ABEILLE@INRIA.FR

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR

SequeL Team, Inria Lille - Nord Europe

Editor: Gergely Neu, Vicenç Gómez, Csaba Szepesvari

1. Introduction

The Thompson Sampling (TS) is a general scheme designed to address the exploration-exploitation trade-off in a wide range of problems and in particular in the multi-armed bandit (MAB) framework (Bubeck and Cesa-Bianchi, 2012). The basic idea is to assume a *prior* distribution over an unknown parameter and to update it with the Bayes rule using the observations collected over time. Furthermore, at each step, a parameter is sampled randomly from the posterior distribution and the corresponding optimal action is chosen. The regret of TS (i.e., the difference between rewards of the algorithm and of the optimal action) has been analyzed both in the Bayesian and in the frequentist setting (i.e., when the true parameter is not a random variable but a fixed parameter). 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 linear bandit setting, most of the literature focused on the analysis of the Bayesian regret (see e.g., Russo and Van Roy (2014b), Osband and Van Roy (2015), Russo and Van Roy (2014a)). Notable exceptions are the analysis of TS in finite MDPs by Gopalan and Mannor (2014) and 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. As opposed to UCB-like algorithms, 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 (2012a)) classifying arms as saturated and unsaturated depending on whether their standard deviation is smaller or bigger than their gap to the optimal arm. 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 $O\big(d^{3/2}\sqrt{T}\big)$. In this paper, **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, **2)** we provide an alternative proof matching the result of Agrawal and Goyal (2012b) that provides additional insights on the structure of the problems and avoids using the unsaturated/saturated arm classification.

2. Preliminaries

We consider the stochastic linear bandit model. Let the arm set $\mathcal{X} \subset \mathbb{R}^d$ be a bounded closed subset of $\mathbb{R}^d$ with arms $\|x\| \leq 1$. When at any step t, the learner pulls an arm x_t, the random reward $r_{t+1} = x_t^\mathsf{T}\theta^\star + \xi_{t+1}$ is obtained, where $\theta^\star \in \mathbb{R}^d$ is a fixed but unknown vector

and ξ_{t+1} is a zero-mean noise. For any $\theta \in \mathbb{R}^d$, we denote the optimal arm and its value by $x^\star(\theta) = \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta$ and $J(\theta) = \sup_{x\in\mathcal{X}} x^\mathsf{T}\theta$. At each step t, the learner suffers a *regret* equal to the difference in expected reward between the optimal arm $x^\star$ and the arm x_t, and its objective is to minimize the *cumulative regret* up to step T, i.e., $R(T) = \sum_{t=1}^T \left(x^{\star,\mathsf{T}}\theta^\star - x_t^\mathsf{T}\theta^\star\right)$. In the following, we consider some assumptions on the structure of the problem.

Assumption 1 *There exists $S \in \mathbb{R}^+$ such that $\|\theta^\star\| \le S$. All the information observed up to time t is encoded in the filtration $\mathcal{F}_t^x = (\mathcal{F}_1, \sigma(x_1, r_2, \dots, r_t, x_t))$ and 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 $R \ge 0$.*

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 $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\mathsf{T}$ and $\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_{s+1}$. For any positive semidefinite matrix A, the weighted 2-norm $\|.\|_A$ is defined by $\|x\|_A^2 = x^\mathsf{T} A x$ where $x \in \mathbb{R}^d$. 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, for any $\mathcal{F}_t^x$-adapted sequence $(x_1, \dots, x_t)$, the RLS estimator $\widehat{\theta}_t$ is such that*

$$\|\widehat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta), \quad \text{and} \quad \forall x \in \mathbb{R}^d,\ |x^\mathsf{T}(\widehat{\theta}_t - \theta^\star)| \le \|x\|_{V_t^{-1}} \beta_t(\delta), \tag{1}$$

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

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

At step t, we define the ellipsoid $\mathcal{E}_t^{\mathrm{RLS}} = \left\{\theta \in \mathbb{R}^d \mid ||\theta - \widehat{\theta}_t||_{V_t} \le \beta_t(\delta')\right\}$ centered in $\widehat{\theta}_t$ with orientation defined by V_t and radius $\beta_t(\delta')$, where $\delta' = \delta/4T$ is designed to match the confidence level $1-\delta$ of the final regret bound. From Eq. 1 we have that $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$ w.h.p. Finally, we report a standard result of RLS that, together with Prop. 1, shows that the prediction error on the points x_t used to construct the estimator $\widehat{\theta}_t$ is cumulatively small.

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

$$\sum_{s=1}^{t} \|x_s\|_{V_s^{-1}}^2 \le 2\log\frac{\det(V_{t+1})}{\det(\lambda I)} \le 2d\log\left(1 + \frac{t}{\lambda}\right). \tag{3}$$

This result plays a central role in most of the proofs for LB, 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) proceed by dividing arms in saturated and unsaturated, we follow a different path that leverages on the structure of $J(\theta)$ and of TS.

3. Thompson Sampling for Linear Bandit

Agrawal and Goyal (2012b) define Thompson Sampling (TS) for LB 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 the RLS-estimate $\widehat{\theta}_t$ and the design matrix V_t, TS samples a *perturbed* parameter $\widetilde{\theta}_t$ as

$$\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t, \tag{4}$$

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. Notice that the resulting distribution on $\widetilde{\theta}_t$ 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 optimistic approaches, such as OFUL (Abbasi-Yadkori et al., 2011a), which require solving a bilinear optimization problem (i.e., $\arg\max_\theta \max_x x^\mathsf{T}\theta$).

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 $\widehat{\theta}_{t+1}$
end for

Figure 1: Thompson sampling algorithm.

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

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

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

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

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

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. 4 we prove that any distribution satisfying the conditions in Def. 3 introduces the right amount of randomness to achieve the desired regret without actually satisfying any Bayesian assumption. Def. 3 includes the Gaussian prior as well as other distributions such as 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. B for exact values of c, c', and p for uniform and Gaussian distributions).

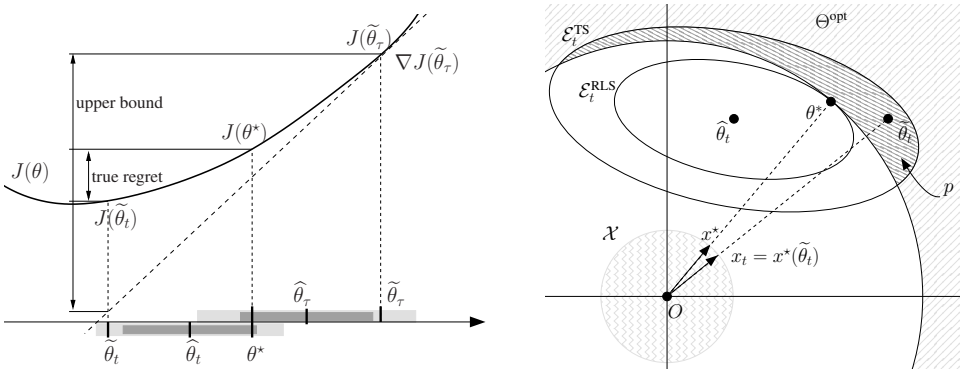

Figure 2: Illustration of the steps of the proof in $\mathbb{R}^1$ and $\mathbb{R}^2$ when $\mathcal{X}$ is the ball of radius 1 for which $J(\theta) = \|\theta\|$ and $x^\star(\theta) = \theta/\|\theta\|$. *Left:* The regret at step t could be bounded by the gradient of the function J at a previous 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 $\mathcal{D}^{\mathrm{TS}}$, which provides enough "mass" beyond $\theta^\star$, which corresponds to larger values of J.

4. Theoretical Analysis

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 4 *Under Asm. 1, the cumulative regret of* TS *over T steps is bounded as*

$$R(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')(1+2/p)\big)\sqrt{2Td\log\big(1+\frac{T}{\lambda}\big)} + \frac{2\gamma_T(\delta')}{p}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}, \tag{5}$$

with probability $1-\delta$ where $\delta' = \frac{\delta}{4T}$.

This bound is of order $\widetilde{O}(d^{3/2}\sqrt{T}/p)$ and it matches the result of Agrawal and Goyal (2012b). 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. Thus we introduce the filtration $\mathcal{F}_t$ as the accumulated information up to time t before the sampling procedure, i.e., $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, r_2, x_2, \ldots, x_{t-1}, r_{t-1}))$. 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 $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_2^x \subset \mathcal{F}_3 \subset \mathcal{F}_3^x, \ldots$. We are now ready to introduce the high-probability events we use throughout the rest of the proof.

Definition 5 *Let $\delta \in (0,1)$ and $\delta' = \delta/(4T)$ and $t \in [1,T]$. We define the event (RLS estimate concentration) $\widehat{E}_t = \big\{\forall s \leq t,\ \|\widehat{\theta}_s - \theta^\star\|_{V_s} \leq \beta_s(\delta')\big\}$ and the event (parameter $\widetilde{\theta}_s$ concentrates around $\widehat{\theta}_s$) $\widetilde{E}_t = \big\{\forall s \leq t,\ \|\widetilde{\theta}_s - \widehat{\theta}_s\|_{V_s} \leq \gamma_s(\delta')\big\}$, where $\gamma_s(\delta') = \beta_s(\delta')\sqrt{cd\log(c'd/\delta')}$.*

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 $E = \widehat{E} \cap \widetilde{E}$. From Prop. 1 and Def. 3 it follows that the joint event $\widehat{E} \cap \widetilde{E}$ holds with

probability $1-\delta/2$ (see Lemma 15). 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 decompose the regret and bound it as

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

with probability $1-\delta/2$. In the interest of space we only report the formal proof to bound R_t^{TS}, while the bound on R_t^{RLS} and the overall regret is postponed to App. E.

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

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

Let $\Theta^{\mathrm{opt}} = \{\theta : J(\theta) \ge J(\theta^\star)\}$ be the set of optimistic parameters (i.e., parameters whose corresponding optimal value is larger than the true one). We can bound the previous expression by the expectation over any random choice of θ in Θ^{opt}, that is

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

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. D).

Proposition 6 *For any set of arm $\mathcal{X}$, $J(\theta) = \sup_x x^\mathsf{T}\theta$ 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 $\mathcal{X}$. As a result, we can use the convexity of J and directly relate R_t^{TS} to its gradient as

$$\begin{aligned} R_t^{\mathrm{TS}} &\le \mathbb{E}\Big[\sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}} \nabla J(\widetilde{\theta})^\mathsf{T}(\widetilde{\theta}-\theta)\mathbb{1}\{E_t\}\Big|\mathcal{F}_t, \widetilde{\theta}\in\Theta^{\mathrm{opt}}\Big] \\ &\le \mathbb{E}\Big[\|\nabla J(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta}-\theta\|_{V_t}\mathbb{1}\{\widetilde{E}_t\}\Big|\mathcal{F}_t, \widetilde{\theta}\in\Theta^{\mathrm{opt}}, \widehat{E}_t\Big]\mathbb{P}(\widehat{E}_t), \end{aligned}$$

where we use Cauchy-Schwarz and we "push" the event $\widehat{E}_t$ into the conditioning.

Step 2 (From gradient of $J(\theta)$ to optimal arm $x^\star(\theta)$). In the next lemma we show that there is a direct relationship between $\nabla J(\theta)$ and the optimal arm corresponding to θ by direct construction (proof in App. D).

Lemma 7 *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. E).

Lemma 8 *Let $\Theta_t^{opt} := \{\theta \in \mathbb{R}^d \mid J(\theta) \geq J(\theta^\star)\}$ be the set of optimistic parameters and $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$ with $\eta \sim \mathcal{D}^{\mathrm{TS}}$, then $\forall t \geq 1$, $\mathbb{P}\big(\widetilde{\theta}_t \in \Theta^{opt} \mid \mathcal{F}_t, \widehat{E}_t\big) \geq p$.*

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

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

If we define $f(\widetilde{\theta}) = \|x^\star(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta} - \theta\|_{V_t} \mathbb{1}\{\widetilde{E}_t\}$ and reintegrating event $\widehat{E}_t$, we have

$$R_t^{\mathrm{TS}} \leq \frac{1}{p}\mathbb{E}\Big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \|\widetilde{\theta} - \theta\|_{V_t} \mathbb{1}\{E_t\}\Big|\mathcal{F}_t\Big],$$

where $1/p$ can be interpreted as the expected time between any two optimistic samples. Since $\widetilde{\theta}$ is sampled according to the standard TS sampling distribution, then it belongs to $\mathcal{E}_t^{\mathrm{TS}}$ and

$$R_t^{\mathrm{TS}} \leq \frac{1}{p}\mathbb{E}\Big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \sup_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta' - \theta\|_{V_t} \mathbb{1}\{E_t\}\Big|\mathcal{F}_t\Big] \leq \frac{2\gamma_t(\delta')}{p}\mathbb{E}\Big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} \mathbb{1}\{E_t\}\Big|\mathcal{F}_t\Big].$$

Finally, we can use Azuma's inequality to obtain the final bound

$$R^{\mathrm{TS}}(T) \leq \frac{2\gamma_T(\delta')}{p}\sum_{t=1}^{T}\|x_t\|_{V_t^{-1}} + \sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}},$$

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

5. Conclusions

In this paper we developed an alternative proof for TS in LB with novel insights on the core elements of the algorithm (*randomization* and *optimism*) and of the structure of the problem (*support function* $J(\theta)$), significantly simplifying the proof by Agrawal and Goyal (2012b) and relying on basic convex geometry arguments. While the current analysis explicitly shows the role of optimism (and the over-sampling property of $\mathcal{D}^{\mathrm{TS}}$ by a factor of $\sqrt{d}$) in the regret, it is still an open question whether this is an unavoidable price to pay for improved computational complexity w.r.t. optimistic approaches. In the proof we rely on the convexity of J and this suggests that analysis could be extended to other settings which leads to convex J function. The critical step would be to have equivalent results as Prop. 1 for the estimator and Lemma 7 to relate the gradient of J to the arms played by TS. Finally, we notice that this analysis has interesting connections with the study of the Follow-the-Perturbed-Leader scheme by Abernethy et al. (2014, 2015) in adversarial full and bandit information settings.

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, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 807–823, 2014.

Jacob D. Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2197–2205, 2015.

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.

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.

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.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. *arXiv preprint arXiv:1406.7498*, 2014.

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. An information-theoretic analysis of thompson sampling. *CoRR*, abs/1403.5341, 2014a.

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

Appendix A. A Geometric Intuition 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 x_t) influence the regret. For the sake of illustration, we consider $\mathcal{X} = \{x \in \mathbb{R}^2, \|x\| \leq 1\}$, i.e., the unit ball in $\mathbb{R}^2$, such that the optimal arm is just the projection of θ on to the ball, i.e., $x^\star(\theta) = \theta/\|\theta\|$, 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[1]

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

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 and the TS parameters $\widetilde{\theta}_t$ 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 the RLS. As opposed to R^{TS}, the prediction error of the RLS 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 observing 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. **??**, 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.[2] 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 $R^{\mathrm{TS}}(T)$. While for optimistic algorithms this term is bounded by 0 at any step ($J(\widetilde{\theta}_t) \geq J(\theta^*)$ by construction w.h.p.), for TS we have to control the deviations caused by the random sampling of $\widetilde{\theta}_t$. This

1. This decomposition is often used for Bayesian regret (see e.g.,(Bubeck and Liu, 2013; Russo and Van Roy, 2014b)).
2. This passage can be done formally using Azuma's inequality exploiting the martingale property of the noise sequence, which is at the core of the derivation of the concentration inequality.

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.

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 zero) to obtain a no-regret algorithm. Nonetheless, the desired regret bound of $O(\sqrt{T})$ is obtained only if $\lambda_{\min,t}$ decreases as $O(1/t)$. 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. Fortunately, inspecting the definition of R_t^{TS} shows 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. 6 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 second factor in the previous expression is small 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. 7 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. 3 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)

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

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

$$R_t^{\mathrm{TS}} \le \|x_\tau\|_{V_\tau^{-1}} \Big(\|\widetilde{\theta}_\tau - \theta^\star\|_{V_\tau} + \sup_{\theta' \in \mathcal{E}_t^{\mathrm{TS}}} \|\theta^\star - \theta'\|_{V_t} \Big) \le \big(\beta_\tau(\delta') + 2\gamma_\tau(\delta')\big) \|x_\tau\|_{V_\tau^{-1}},$$

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

Step 3 (optimism). This bound shows the importance that TS is optimistic with high frequency. In fact, 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. Then, since $\mathbb{E}[\nu_k] = 1/p$ we have $R^{\mathrm{TS}}(T) \le \widetilde{O}(1/p\sqrt{T})$ w.h.p. by Cauchy-Schwarz and Prop. 2. Unfortunately, sampling $\widetilde{\theta}_t$ from e.g., the RLS ellipsoid $\mathcal{E}_t^{\mathrm{RLS}}$ may have a very small probability of being optimistic (see for instance Fig. 2, where sampling uniformly in $\mathcal{E}_t^{\mathrm{RLS}}$ would have 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. 8) and the final desired regret.

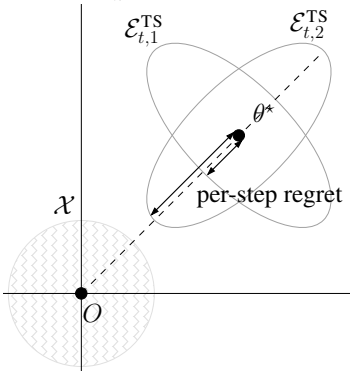

Figure 3: While $\mathcal{E}_{t,1}^t$ and $\mathcal{E}_{2,t}^t$ have an equivalent accurate estimation of $\theta^\star$, $\mathcal{E}_{1,t}^t$ has smaller regret than $\mathcal{E}_{2,t}^t$.

Appendix B. Examples of TS distributions

Example 1: Uniform distribution $\eta \sim \mathcal{U}_{B_d(0,\sqrt{d})}$. The uniform distribution satisfies the concentration property with constants $c = 1$ and $c' = \frac{e}{d}$ by definition. Since the set $\{\eta | u^\mathsf{T}\eta \ge 1\} \cap B_d(0,\sqrt{d})$ is an hyper-spherical cap for any direction u of $\mathbb{R}^d$, the the anti-concentration property is satisfied provided that the ratio between the volume of an

hyper-spherical cap of height $\sqrt{d}-1$ and the volume of the ball of radius $\sqrt{d}$ is constant (i.e., independent from d). Using standard geometric results (see Prop. 16), one has that for any vector $\|u\|=1$

$$\mathbb{P}(u^\mathsf{T}\eta \geq 1) = \frac{1}{2} I_{1-\frac{1}{d}}\Big(\frac{d+1}{2}, \frac{1}{2}\Big), \tag{6}$$

where $I_x(a,b)$ is the incomplete regularized beta function. In Prop. 17 we prove that

$$I_{1-\frac{1}{d}}\Big(\frac{d+1}{2}, \frac{1}{2}\Big) \geq \frac{1}{8\sqrt{6\pi}},$$

and hence we obtain $p = \frac{1}{16\sqrt{6\pi}}$. ■.

Example 2: Gaussian case $\eta \sim \mathcal{N}(0, I_d)$. The concentration property comes directly from the Chernoff bound for standard Gaussian random variable together with union bound argument. For any $\alpha > 0$, we have

$$\mathbb{P}(\|\eta\| \leq \alpha\sqrt{d}) \geq \mathbb{P}(\forall 1 \leq i \leq d,\ |\eta_i| \leq \alpha) \geq 1 - d\mathbb{P}(|\eta_i| \geq \alpha).$$

Standard concentration inequality for Gaussian random variable gives, $\forall \alpha > 0$,

$$\mathbb{P}(|\eta_i| \geq \alpha) \leq 2e^{-\alpha^2/2}.$$

Plugging everything together with $\alpha = \sqrt{2\log\frac{2d}{\delta}}$ gives the desired result with $c = c' = 2$. Let η_i be the i-th component of η for any $1 \leq i \leq d$. Then $\eta_i \sim \mathcal{N}(0,1)$. Since η is rotationally invariant, for any direction u of $\mathbb{R}^d$ and an appropriate choice of basis, we have $\mathbb{P}(u^\mathsf{T}\eta \geq 1) \geq \mathbb{P}(\eta_1 \geq 1)$. From standard Gaussian properties (see Thm 2 of Chang et al. (2011)) we have

$$\mathbb{P}(\eta_1 \geq 1) = \frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\right) \geq \frac{1}{4\sqrt{e\pi}}$$

which ensures the anti-concentration property with $p = \frac{1}{4\sqrt{e\pi}}$. ■

Appendix C. Properties of convex function

Proposition 9 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and C be a closed convex set of $\mathbb{R}^d$. Then, on C, f reaches its maximum on the boundary of C.*

Proof Let's denote as $int(C)$ and $bound(C)$ the interior and the boundary of the closed convex set C respectively. Assume that $\exists x^\star \in int(C)$ such that $f(x^\star) > f(x)$ for any $x \in bound(C)$ and $f(x^\star) \geq f(y)$ for any $y \in int(C)$.

Then define $y = x^\star + \epsilon(x^\star - x)$ for some $x \in bound(C)$. By definition of the open set $int(C)$, $\exists \epsilon > 0$ such that $y \in int(C)$. Moreover, $x^\star \in [y, x]$ e.g.

$$x^\star = (1-t)x + ty, \quad t = \frac{1}{1+\epsilon} \in]0,1[$$

Using the convexity of f on has

$$f(x^\star) \leq (1-t)f(x) + tf(y) < (1-t)f(x^\star) + tf(y)$$
$$f(x^\star) < f(y)$$

which is impossible by assumption. ■

Proposition 10 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Let $B_d(0,1)$ be the unit $d-$dimensional ball and $S_d(0,1)$ the associated unit sphere.*
Given a point $x \in S_d(0,1)$, define as $\mathcal{H}(x)$ the hyperplan tangent to $B_d(0,1)$ at the point x. $\mathcal{H}(x)$ split $\mathbb{R}^d$ into two complementary subspace $\mathcal{G}(x)$ and $\mathcal{G}^{\perp}(x)$ where $\mathcal{G}(x)$ does not contain the unit ball by convention.
Then for any $x^\star \in S_d(0,1)$ such that $f(x^\star) \geq f(x)$ for all $x \in B_d(0,1)$, one has

$$\forall y \in \mathcal{G}(x^\star), \quad f(y) \geq f(x^\star)$$

Proof We first notice that from Proposition 9 $x^\star$ is well defined since the maximum is reached on the boundary. The associated subspace $\mathcal{G}(x^\star)$ is then

$$\mathcal{G}(x^\star) := \{y = x^\star + u, u \in \mathbb{R}^d \mid u^\mathsf{T} x^\star \geq 0\}.$$

We want to show that $f(y) \geq f(x^\star)$ for any $y \in \mathcal{G}(x^\star)$. We introduce the increasing sequence of subspace

$$\mathcal{G}_n = \left\{ y = x^\star + u, u \in \mathbb{R}^d \mid u^\mathsf{T} x^\star \geq \frac{\|u\|}{2(n-1)} \right\}, \quad n \geq 2.$$

For any $y = x^\star + u$ in $\mathcal{G}_n$, we associate

$$x = x^\star - \frac{1}{2(n-1)} \frac{u}{\|u\|}.$$

By definition of y (and hence u), we have

$$\begin{aligned}
\|x\|^2 &= 1 + \frac{1}{2(n-1)}^2 - \frac{1}{2(n-1)\|u\|} u^\mathsf{T} x^\star \\
&= 1 + \frac{1}{2(n-1)} \left[\frac{1}{2(n-1)} - \frac{u^\mathsf{T}}{\|u\|} x^\star \right] \\
&\leq 1,
\end{aligned}$$

which means that $x \in \mathcal{B}_d(0,1)$. Moreover let $t = [2(n-1)\|u\| + 1]^{-1}$, $t \in]0,1[$ one has $x^\star = (1-t)x + ty$. Since $x \in \mathcal{B}_d(0,1)$ then

$$\begin{aligned}
f(x^\star) &\leq (1-t) f(x) + t f(y) \\
&\leq (1-t) f(x^\star) + t f(y) \\
\Rightarrow & f(x^\star) \leq f(y).
\end{aligned}$$

Since the statement of the proposition holds for any $\mathcal{G}_n$, then we obtain the desired result for $\mathcal{G}$ by continuity of f. Let $y \in \mathcal{G}(x^\star)$, $y = x^\star + u$. If $u^\mathsf{T} x^\star > 0$, then $\exists n \geq 2$ such that $y \in \mathcal{G}_n$ and the proposition is satisfied. Otherwise, if $u^\mathsf{T} x^\star = 0$, we introduce the sequences $\{u_n\}$

and $\{y_n\}$ defined as:

$$\begin{aligned} u_n &= u + \frac{\|u\|}{\sqrt{1 - \frac{1}{2(n-1)}^2}} \frac{x^\star}{2(n-1)} \\ &= u + \frac{\|u_n\|}{2(n-1)} x^\star, \\ y_n &= x^\star + u_n. \end{aligned}$$

By construction, $y_n \in \mathcal{G}_n$ and $y_n \to y$ as $n \to \infty$. Since the $f(y_n) \geq f(x^\star)$ for any $n \geq 2$ we obtain the desired result taking the limit since f is continuous as a convex function on $\mathbb{R}^d$. ■

Theorem 11 (A.D. Alexandrov) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, then it is twice differentiable almost everywhere with respect to the Lebesgue's measure.*

Proof This result is an extension of the Rademacher's theorem for convex functions. A proof can be found in Niculescu and Persson (2006), theorem 3.11.2. ■

Appendix D. Properties of support function (proof of Proposition 6 and Lemma 7)

We study the *support function* of a set C, which is a function $f_C : \mathbb{R}^d \to \mathbb{R}$ such that

$$f_C(\theta) = \sup_{x \in C} x^\mathsf{T} \theta \tag{7}$$

Those functions are at the core of convex geometry analysis.

Proposition 12 *Let $C \subset \mathbb{R}^d$ be a non-empty compact set and f_C the associated support function. Then,*

1. *f_C is real-valued and $\sup_{x \in C} x^\mathsf{T} \theta$ is attained in C,*
2. *f_C is convex,*
3. *f_C is continuous on $\mathbb{R}^d$ and twice differentiable almost everywhere with respect to the Lebesgue's measure.*

Proof

1. This comes directly from the compactness of C: since C is bounded, the support function is real-valued and since C is closed, the supremum is attained in C,
2. Let θ_1, θ_2 two vectors of $\mathbb{R}^d$, and $t \in (0, 1)$. By definition of the supremum, since f_C is real-valued:

$$f_C(t\theta_1 + (1-t)\theta_2) = \sup_{x \in C} \left(t x^\mathsf{T} \theta_1 + (1-t) x^\mathsf{T} \theta_2 \right) \leq t \sup_{x \in C} x^\mathsf{T} \theta_1 + (1-t) \sup_{x \in C} x^\mathsf{T} \theta_2$$

3. The continuity is consequence of the convexity of f_C on the open convex set $\mathbb{R}^d$ and the second order differentiability comes from Alexandrov's theorem 11.

■

Proposition 13 *Let $x(\theta) \in \arg\sup_{x\in C} x^\mathsf{T}\theta$, denote as $\nabla f_C(\theta)$ and $\partial f_C(\theta)$ the gradient (when it is uniquely defined) and the sub-gradient of f_C in $\theta \in \mathbb{R}^d$. Then,*

1. *for all $\theta \in \mathbb{R}^d$, $x(\theta) \in \partial f_C(\theta)$,*
2. *their exists a null set $\mathcal{N}$ with respect to the Lebesgue's measure such that $x(\theta) = \nabla f_C(\theta)$ for all $\theta \in \mathbb{R}^d \setminus \mathcal{N}$,*
3. *equivalentely, $x(\theta) = \nabla f_C(\theta)$ where the equality holds in the sense of the distribution.*

Proof Thanks to proposition 12, we know that the supremum is attained in $x(\theta) \in C$. Moreover, Alexandrov's theorem guarantee that $\mathcal{N}$ is a null-set. Since the sub-gradient is reduced to a singleton where the function is differentiable e.g. $\partial f_C(\theta) = \{\nabla f_C(\theta)\}$ for all $\theta \in \mathbb{R}^d \setminus \mathcal{N}$, one just need to show to $x(\theta) \in \partial f_C(\theta)$ for all $\theta \in \mathbb{R}^d$.
Since $f_C(\theta) = \max_{x\in C} x^\mathsf{T}\theta$, their exist at least one $x(\theta) \in C$ for which the maximum is attained i.e. $x(\theta)^\mathsf{T}\theta = f_C(\theta)$. Moreover, for any $\bar{\theta} \in \mathbb{R}^d$, $f_C(\bar{\theta}) \geq x(\theta)^\mathsf{T}\bar{\theta}$ by definition. Therefore,

$$f_C(\bar{\theta}) - x(\theta)\bar{\theta} \geq 0 := f_C(\theta) - x(\theta)^\mathsf{T}\theta$$
$$f_C(\bar{\theta}) \geq f_C(\theta) + x(\theta)^\mathsf{T}\left(\bar{\theta} - \theta\right), \quad \forall \bar{\theta} \in \mathbb{R}^d$$

which is the definition of the sub-gradient.

■

Appendix E. Regret Proofs

We collect here the main tools that we need to derive the proof. We first recall the Azuma's concentration inequality for super-martingale.

Proposition 14 *If a super-martingale $(Y_t)_{t\geq 0}$ corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| < c_t$ for some constant c_t for all $t = 1, \ldots, T$ then for any $\alpha > 0$,*

$$\mathbb{P}(Y_T - Y_0 \geq \alpha) \leq 2e^{-\frac{\alpha^2}{2\sum_{t=1}^T c_t^2}}$$

Lemma 15 *Under Asm. 1 we have $\mathbb{P}(\widehat{E} \cap \widetilde{E}) \geq 1 - \frac{\delta}{2}$.*

Proof [Proof of Lemma 15] We first bound the two events separately.

Bounding $\widehat{E}$. This bound is a straightforward application of Proposition 1 together with a union bound argument. Let $\delta' = \delta/(4T)$, then

$$\begin{aligned}
\forall 1 \le t \le T, \quad & \mathbb{P}\left(\|\widehat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta')\right) \ge 1 - \delta' \\
\text{from union bound}, \quad & \mathbb{P}\left(\bigcap_{t=1}^{T}\left\{\|\widehat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta')\right\}\right) \ge 1 - \sum_{t=1}^{T}\mathbb{P}\left(\|\widehat{\theta}_t - \theta^\star\|_{V_t} \ge \beta_t(\delta')\right) \\
\Rightarrow \quad & \mathbb{P}\left(\bigcap_{t=1}^{T}\left\{\|\widehat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta')\right\}\right) \ge 1 - \sum_{t=1}^{T}\delta' \\
\Rightarrow \quad & \mathbb{P}\big(\widehat{E}\big) \ge 1 - T\delta' = 1 - \frac{\delta}{4}.
\end{aligned}$$

Bounding $\widetilde{E}$. This bound comes directly from the concentration property of the TS sampling distribution. From the expression of $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$ where η_t is drawn i.i.d. from $\mathcal{D}^{\mathrm{TS}}$, we have

$$\forall 1 \le t \le T, \quad \mathbb{P}\left(\|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} \le \beta_t(\delta')\sqrt{cd\log\frac{c'd}{\delta'}}\right) = \mathbb{P}\left(\|\eta_t\| \le \sqrt{cd\log\frac{c'd}{\delta'}}\right).$$

Then from Definition 3, we have

$$\mathbb{P}\left(\|\eta_t\| \le \sqrt{cd\log\frac{c'd}{\delta'}}\right) \ge 1 - \delta'.$$

As before, a union bound over the two bounds ensures that

$$\mathbb{P}(\widetilde{E}) \ge 1 - T\delta' = 1 - \frac{\delta}{4}.$$

Finally, a union bound argument between the two terms leads to

$$\mathbb{P}(\widehat{E} \cap \widetilde{E}) \ge 1 - \frac{\delta}{2}.$$

■

Proof [Proof of Lemma 8] We need to study the probability that a $\widetilde{\theta}$ drawn at time t from the TS sampling distribution is optimistic, i.e., $J(\widetilde{\theta}) \ge J(\theta^\star)$, under event $\widehat{E}_t$. More formally let

$$p_t = \mathbb{P}\big(J(\widetilde{\theta}) \ge J(\theta^\star)|\mathcal{F}_t, \widehat{E}_t\big).$$

Using the definition of $\widehat{E}_t$ we have that $\theta^\star \in \mathcal{E}_t^{\mathrm{RLS}}$ (i.e., the true parameter vector belongs to the RLS ellipsoid) and then we can replace $J(\theta^\star)$ by the supremum over the ellipsoid as

$$p_t \ge \mathbb{P}\Big(J(\widetilde{\theta}) \ge \sup_{\theta \in \mathcal{E}_t^{\mathrm{RLS}}} J(\theta)\Big|\mathcal{F}_t, \widehat{E}_t\Big).$$

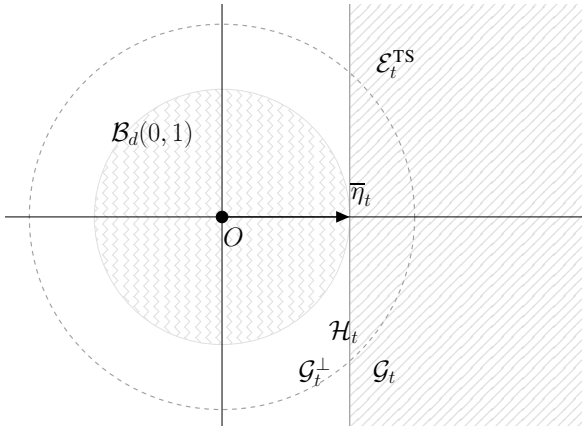

Figure 4: Illustration of the probability of selecting an optimistic $\widetilde{\theta}_t$.

By recalling the definition of the TS sampling process, we can write $\widetilde{\theta} = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta$, where $\eta \sim \mathcal{D}^{\text{TS}}$ and for notational convenience, we define the function $f_t(\eta) = J(\widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta)$. Let $\overline{\theta}_t = \arg\sup_{\theta \in \mathcal{E}_t^{\text{RLS}}} J(\theta)$ and $\overline{\eta}_t$ be the corresponding η (i.e., $\overline{\eta}_t$ is such that $\overline{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\overline{\eta}_t$). Since the supremum is taken within $\mathcal{E}_t^{\text{RLS}}$, $\overline{\eta}_t$ belongs to the unit ball (i.e., $\overline{\eta}_t \in \mathcal{B}_d(0,1)$). As a result, we can rewrite the previous expression as

$$p_t \geq \mathbb{P}\Big(f_t(\eta) \geq f_t(\overline{\eta}_t)\Big|\mathcal{F}_t, \widehat{E}_t\Big).$$

Since the function f_t inherits all the properties of J, notably its convexity in η, we know that the supremum on a convex closed set is reached at least at one point $\bar{\eta}_t$ and that it belongs to the boundary (see Prop. 9), which in our case corresponds to $\|\overline{\eta}_t\| = 1$. Moreover, let $\mathcal{H}_t(\overline{\eta}_t)$ be the hyperplane tangent to $\overline{\eta}_t$. $\mathcal{H}_t(\bar{\eta}_t)$ splits $\mathbb{R}^d$ in two complementary subspaces $\mathcal{G}_t$ and $\mathcal{G}_t^{\perp}$ where $\mathcal{G}_t$ does not contain the unit ball by convention. Again, the convexity of f_t ensures that $f_t(\eta) \geq f_t(\bar{\eta}_t)$ for all $\eta \in \mathcal{G}_t$ as proved in Prop. 10. As illustrated in Fig. 4 the probability of being optimistic is now reduced to the probability that η drawn from $\mathcal{D}^{\text{TS}}$ falls into $\mathcal{G}_t$, which corresponds to

$$p_t \geq \mathbb{P}\Big(\eta \in \mathcal{G}_t\Big|\mathcal{F}_t, \widehat{E}_t\Big).$$

Let u_t be the vector defining the hyperspace $\mathcal{H}_t(\overline{\eta}_t)$, notice that the subspace u_t is entirely defined by the filtration $\mathcal{F}_t$ and the event $\widehat{E}_t$ and it is thus independent from $\overline{\eta}_t$. As a result, we finally obtain

$$p_t \geq \mathbb{P}\Big(u_t^{\mathsf{T}}\eta \geq 1\Big|\mathcal{F}_t, \widehat{E}_t\Big) \geq p,$$

where the last step immediately follows from property 1 of Def. 3 of the TS sampling distribution. ■

Proof [Proof of Theorem 4]

We first bound the two regret terms $R^{\mathrm{TS}}(T)$ and $R^{\mathrm{RLS}}(T)$.

Bound on $R^{\mathbf{TS}}(T)$. We collect the bounds on each term R_t^{TS} and obtain

$$R^{\mathrm{TS}}(T) \leq \sum_{t=1}^{T} R_t^{\mathrm{TS}} \mathbb{1}\{E_t\} \leq \frac{2\gamma_T(\delta')}{p} \sum_{t=1}^{T} \mathbb{E}\big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} | \mathcal{F}_t\big]. \tag{8}$$

Since this term contains an expectation, we cannot directly apply Proposition 2 and we first need to rewrite to the total regret $R^{\mathrm{TS}}(T)$ as

$$R^{\mathrm{TS}}(T) \leq \frac{2\gamma_T(\delta')}{p} \Bigg(\sum_{t=1}^{T} \|x_t\|_{V_t^{-1}} + \underbrace{\sum_{t=1}^{T} \Big(\mathbb{E}\big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} | \mathcal{F}_t\big] - \|x_t\|_{V_t^{-1}} \Big)}_{R_2^{\mathrm{TS}}} \Bigg). \tag{9}$$

From Prop. 2, the first term is bounded as,

$$\sum_{t=1}^{T} \|x_t\|_{V_t^{-1}} \leq \sqrt{T} \Bigg(\sum_{t=1}^{T} \|x_t\|^2_{V_t^{-1}} \Bigg)^{1/2} \leq \sqrt{2Td \log\big(1 + \frac{T}{\lambda}\big)}.$$

We now proceed applying Azuma inequality 14 to the second term which is a martingale by construction. Under assumption **??**, $\|x_t\| \leq 1$ for all $t \geq 1$, so since $V_t^{-1} \leq \frac{1}{\lambda} I$ one gets,

$$\mathbb{E}\big[\|x^\star(\widetilde{\theta})\|_{V_t^{-1}} | \mathcal{F}_t\big] - \|x_t\|_{V_t^{-1}} \leq \frac{2}{\sqrt{\lambda}}, \quad a.s.$$

This provides an upper-bound on each element of R_2^{TS} which holds with probability at least $1 - \frac{\delta}{2}$ as

$$R_2^{\mathrm{TS}} \leq \sqrt{\frac{8T}{\lambda} \log \frac{4}{\delta}}.$$

Bound on $R^{\mathbf{RLS}}(T)$. The bound on R^{RLS} is derived as previous results in (Abbasi-Yadkori et al., 2011b; Agrawal and Goyal, 2012b). We decompose the term in a *sampling prediction error* and a RLS *prediction error* as follow

$$R^{\mathrm{RLS}}(T) \leq \sum_{t=1}^{T} |x_t^{\mathsf{T}} (\widetilde{\theta}_t - \widehat{\theta}_t)| \mathbb{1}\{E_t\} + \sum_{t=1}^{T} |x_t^{\mathsf{T}} (\widehat{\theta}_t - \theta^\star)| \mathbb{1}\{E_t\}$$

By definition of the concentration event E_t,

$$|x_t^{\mathsf{T}} (\widetilde{\theta}_t - \widehat{\theta}_t)| \mathbb{1}\{E_t\} \leq \|x_t\|_{V_t^{-1}} \gamma_t(\delta'), \quad |x_t^{\mathsf{T}} (\widehat{\theta}_t - \theta^\star)| \mathbb{1}\{E_t\} \leq \|x_t\|_{V_t^{-1}} \beta_t(\delta'),$$

so from proposition 2,

$$R^{\mathrm{RLS}}(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')\big) \sqrt{2Td \log\big(1 + \frac{T}{\lambda}\big)}. \tag{10}$$

Final bound. We finally plug everything together since from lemma 15 the concentration event holds with probability at least $1 - \frac{\delta}{2}$. Using the bound on $R^{\mathrm{TS}}(T)$ and a union bound argument one obtains the desired result which holds with probability at least $1 - \delta$. ■

Appendix F. Hyperspherical cap and beta function

Proposition 16 *Let $V_d(R)$ be the volume of the $d-$dimensional ball of radius R and let $V_d^{cap}(h)$ the volume of the hyperspherical cap of heigh $h = R - r > 0$. Then,*

$$V_d^{cap}(h) = \frac{1}{2} V_d(R) I_{1-(\frac{r}{R})^2}\left(\frac{d+1}{2}, \frac{1}{2}\right)$$

where $I_x(a,b)$ is the incomplete regularized beta function.

Proof The proof can be found in Li (2011). ■

Proposition 17 *Let $I_x(a,b)$ is the incomplete regularized beta function,*

$$\forall d \geq 2, \quad I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{8\sqrt{6\pi}}$$

Proof The incomplete regularized beta function can be expressed in terms of the beta function $B(a,b)$ and the incomplete beta function $B_x(a,b)$ where

$$\begin{aligned} B_x(a,b) &= \int_0^x t^{a-1}(1-t)^{b-1} dt \\ B(a,b) &= B_1(a,b) \\ I_x(a,b) &= \frac{B_x(a,b)}{B(a,b)} \end{aligned}$$

Hence we seek for a lower bound on $B_{1-\frac{1}{d}}\left(\frac{d+1}{2} + \frac{1}{2}\right)$ and an upper bound for $B\left(\frac{d+1}{2} + \frac{1}{2}\right)$.

1. Let first find an lower bound for the incomplete beta function. Since $t \to t^{\frac{d-1}{2}}(1-t)^{-1/2}$ is positive and increasing on $[0,1]$, for any $d \geq 2$,

$$\begin{aligned} B_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) &\geq \int_{1-\frac{3}{2d}}^{1-\frac{d}{2}} t^{\frac{d-1}{2}}(1-t)^{-1/2} dt \\ &\geq \frac{1}{2d}\left(\frac{3}{2d}\right)^{-1/2}(1-\frac{3}{2d})^{\frac{d-1}{2}} \\ &\geq \frac{1}{\sqrt{6d}}(1-\frac{3}{2d})^{\frac{d-1}{2}} \\ &\geq \frac{1}{\sqrt{6d}}(1-\frac{3}{2d})^{\frac{d}{2}} \end{aligned}$$

From the increasing property of $x \to (1-\frac{\alpha}{x})^x$ for any $\alpha < 1$ the sequence $\left\{(1-\frac{3}{2d})^{\frac{d}{2}}\right\}_{d\geq 2}$ is increasing and

$$B_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{\sqrt{6d}}(1-\frac{3}{2\times 2})^{\frac{2}{2}} = \frac{1}{4\sqrt{6d}}$$

2. Now we seek for an upper bound for $B\left(\frac{d+1}{2}+\frac{1}{2}\right)$. Since $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ one has:

$$B\left(\frac{d+1}{2}+\frac{1}{2}\right)=\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}=\sqrt{\pi}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}$$

From Chen and Qi (2005) we have the following inequalities for the gamma function $\forall n\geq 1$:

$$\frac{\Gamma(n+1/2)}{\Gamma(n+1)}\leq(n+1/4)^{-1/2}$$
$$\frac{\Gamma(n+1/2)}{\Gamma(n+1)}\geq(n+4/\pi-1)^{-1/2}$$

Together with $\Gamma(x+1)=x\Gamma(x)$ and treating separately cases where d is even or not, one gets $\forall d\geq 2$

$$\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}\leq\frac{2}{\sqrt{d}}$$

3. Using the obtained upper and lower bound we get:

$$I_{1-\frac{1}{d}}\left(\frac{d+1}{2},\frac{1}{2}\right)\geq\frac{\sqrt{d}}{2\sqrt{\pi}\times 4\sqrt{6d}}\geq\frac{1}{8\sqrt{6\pi}}$$

■