Gain estimation of linear dynamical systems using Thompson Sampling

Matias I. Müller
KTH Royal Institute of Technology

Cristian R. Rojas
KTH Royal Institute of Technology

Abstract

We present the gain estimation problem for linear dynamical systems as a multi-armed bandit. This is particularly a very important engineering problem in control design, where performance guarantees are casted in terms of the largest gain of the frequency response of the system. The dynamical system is unknown and only noisy input-output data is available. In a more general setup, the noise perturbing the data is non-white and the variance at each frequency band is unknown, resulting in a two-dimensional Gaussian bandit model with unknown mean and scaled-identity covariance matrix. This model corresponds to a two-parameter exponential family. Within a bandit framework, the set of means is given by the frequency response of the system and, unlike traditional bandit problems, the goal here is to maximize the probability of choosing the arm drawing samples with the highest norm of its mean. A problem-dependent lower bound for the expected cumulative regret is derived and a matching upper bound is obtained for a Thompson-Sampling algorithm under a uniform prior over the variances and the two-dimensional means.

1 INTRODUCTION

Control engineering has been one of the most relevant disciplines during the last century. The control problem can be seen as a Markov decision problem [1] with a continuous state space and where the transition probabilities are known and may depend on all the past values of the input sequence; the most common class of such systems are the so-called linear dynamical systems [2], which will be considered in this paper. The system aimed to be controlled is completely characterized by these probabilities which, in the control field, are obtained from input-output data using techniques from identification for control [3]. The different sources of uncertainty during modelling (such as noise present in the data) lead to a mismatch between the system to be controlled and its model. This mismatch is a complex-valued function in the frequency domain, and known as *modelling error*.

The success of control engineering lies on the robustness with which control laws can be designed to account for modelling errors. In fact, the performance of the control policy depends on the largest magnitude of the modelling error across the frequency axis [4], what we call the ℓ_2-gain or the $\mathcal{H}_\infty$-norm [5] of the modelling error. It is then a crucial step in the control design process to estimate this quantity. The difficulty of this problem lies on the nature of the modelling error itself, which indeed characterizes what the model cannot explain, introducing a challenging task.

In this work we focus on the problem of finding the maximum amplitude of the modelling error in a model-free manner by casting the problem as a multi-armed bandit. When a system is only accessible through input-output data, the problem of efficiently estimating the $\mathcal{H}_\infty$ norm involves finding the peak frequency as quickly as possible, so the $\mathcal{H}_\infty$ norm can be estimated by applying a sinusoidal input of such frequency as input. In our setup, data is collected sequentially by adaptively designing the input sequence. The model-free condition is motivated by the fact that the system under study (being the one for which we want to derive its $\mathcal{H}_\infty$-norm) is equal to the modelling error which is, by its nature, unknown. The alternative approach requires one to derive an explicit model for the modelling error (from input-output data), as a transfer function or a state-space model, and then apply standard $\mathcal{H}_\infty$-norm computation methods such as the ones in [6] and [7]. We avoid the process of deriving a (new) model for the modelling error, which would be naturally uncertain,

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

introducing an additional modelling error. However, as discussed in [8], it is not clear if there is a gap, from an information-theoretic standpoint, between the sample complexity of the model-free approach and the one incurred by deriving a model for the modelling error.

The presented problem corresponds to a nonlinear bandit with two-dimensional Gaussian feedback, since the measurements of the modelling error are complex-valued. The agent is allowed to perform an experiment at each round, from which it collects input-output data, disturbed by additive Gaussian noise, in the time domain followed by a Fourier transformation. The input is a sinusoidal signal of a frequency adaptively chosen by the agent, denoting the played arm. The twist of this problem is that, unlike traditional bandit problems, the goal here is to find the arm whose complex mean has the largest magnitude. The problem presented in this work is interesting not only because it models an important engineering problem, but also since it introduces a non-traditional bandit problem in which the best arm is defined by a nonlinear function of its Gaussian outcomes' parameters.

Our contributions: Summarizing, the main contributions of our work are:

1. a model to the problem of gain estimation as a multi-armed bandit;
2. a lower bound on the asymptotic regret any uniformly good algorithm will incur in;
3. a thorough theoretical derivation of the matching upper bound on the asymptotic regret that Thompson Sampling incurs in for the nonlinear bandit with two-parameter bivariate Gaussian feedback, when the optimal arm is the one whose two-dimensional mean has the highest norm;
4. a numerical simulation illustrating the optimality of the algorithm.

The remainder of this paper is organized as follows: Section 2 formalizes the $\mathcal{H}_\infty$-norm estimation problem and describes it as a stochastic multi-armed bandit, while Section 3 describes the preliminaries on multi-armed bandits and Thompson Sampling. Concentration inequalities for the Gaussian model are derived in Section 4, and the optimality of Thompson Sampling is shown in Section 5. Finally, an illustrative example is introduced in Section 6 and conclusions are presented in Section 7. For brevity, proofs are appended in the supplementary material.

1.1 Related work

Multi-armed bandits (MAB) are a class of reinforcement learning problems formally introduced in [9], which exhibit the so-called exploration-exploitation dilemma. In these problems, an agent bets on an arm at each round, and this action generates an outcome the agent perceives. The outcome is a realization drawn from a parametrized distribution where the parameters for each arm are unknown to the agent. In the traditional bandit setup [10], the goal of the agent is to minimize the expected cumulative difference between the outcome of its choice and the one an oracle would have drawn by always choosing the optimal arm.

Fundamental limitations on the regret depend on how many of the parameters the agent knows beforehand, where a detailed analysis is provided in [11] for a general class of problems. Explicitly derived as lower bounds, these limitations motivate the search for *optimal algorithms* whose asymptotic performance match these restrictions. Different classes of optimal algorithms can be found in [10].

Thompson Sampling [12] (TS) is one of the most interesting algorithms in MAB due to its excellent empirical finite-time performance for many models, compared to other optimal algorithms [13]. It corresponds to a Bayesian policy that keeps track of the posterior mean for each of the arms, where the agent decides the next action by sampling these posteriors and choosing the arm with the largest sample. Optimality of TS for the Bernoulli model [14] has been extended to the one-parameter one-dimensional exponential family bandit in [15]. Recently, [16] has developed a thoroughly analysis for this algorithm under a two-parameter (mean and variance) Gaussian bandit, concluding that its optimality crucially depends on the choice of the prior. In fact, the algorithm does not achieve optimality when a Jeffreys prior is employed, not even achieving logarithmic asymptotic regret. The authors also discuss that the main difficulty of proving optimality for the two-parameter Gaussian bandit lies in the posterior distribution for the mean being heavy-tailed, so no straightforward upper/lower bounds can be found. For the bandit model considered in our paper, the two-dimensional posterior mean distribution is a linear transformation of a multivariate t-distribution. Perhaps surprisingly, the latter can not be factorized as the multiplication of both one-dimensional marginal posteriors, even when their outcomes and prior distributions are statistically independent, as we show later. This counter-intuitive phenomenom forces us to derive multivariate concentration inequalities instead of recycling the ones in [16], making the extension of their work into the bivariate Gaussian bandit of highest norm not straightforward. A one-parameter Gaussian bandit model (known variance) for $\mathcal{H}_\infty$-norm estimation was firstly introduced in [17] when the policy allows to choose arms inside a simplex. The authors of [17] showed that the asymptotic regret lower bound for

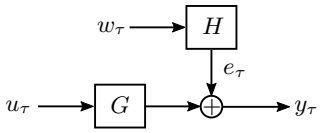

Figure 1: Mathematical relationship of a linear system with additive non-white noise.

these policies is equal to the one of policies playing only one arm per round, proving that playing several arms does not increase the performance of an optimal algorithm (*i.e.*, an algorithm whose asymptotic regret matches the lower bound).

Some iterative approaches have been already proposed in system identification for control [18, 19], where the input signal is allowed to be designed upon previous data at each round, based on the power-iterations method of numerical linear algebra. The particular problem of $\mathcal{H}_\infty$-norm estimation has gained some attention in computer science, where [8] has derived sharp asymptotic bounds on the error incurred by a method that firstly fits an L-FIR (finite impulse response) filter of L coefficients to N-length data, in terms of N.

2 MODEL SETUP AND PROBLEM STATEMENT

Following the set-up described in [3], let $g := (g_\tau)_{\tau=0}^\infty$ and $h := (h_\tau)_{\tau=0}^\infty$ denote the impulse responses of systems G and H, respectively, depicted in Fig. 1, where $g_\tau, h_\tau \in \mathbb{R}$, $\tau = 1, 2, \dots$. The systems are assumed to be linear and time-invariant (LTI) and causal ($g_\tau, h_\tau = 0\ \forall \tau < 0$). Then, in experiment t, the output signal $(y_\tau)_{\tau=0}^\infty$, as a function of the input $(u_\tau)_{\tau=0}^\infty$ and the zero-mean unit-variance and white[1] [2] Gaussian sequence $(w_\tau)_{\tau=0}^\infty$ is defined as

$$\begin{aligned} y_\tau &= (g * u)_\tau + \underbrace{(h * w)_\tau}_{=:e_\tau} \\ &= \sum_{\tau'=0}^{\infty} g_{\tau'} u_{\tau-\tau'} + \sum_{\tau'=0}^{\infty} h_{\tau'} w_{\tau-\tau'}, \end{aligned} \tag{1}$$

where each term corresponds to a convolution (denoted as $*$) between a signal and an impulse response. We also impose the assumption that both systems in Fig. 1 are stable, that is, $(u_\tau)_{\tau=0}^\infty$ satisfying $|u_\tau| < \infty$, $\forall \tau$, implies that $|y_\tau| < \infty$, $\forall \tau$. The latter implies that the Fourier transforms of g and h exist for each frequency $\omega \in [0, \pi]$, denoted by $G(\mathrm{e}^{j\omega})$ and $H(\mathrm{e}^{j\omega})$, also known as the *frequency responses* of G and H, respectively, with $j := \sqrt{-1}$.

The system in Fig. 1 is suitable to model the problem of collecting noisy data from the modelling error system $G = G_o - \hat{G}$, where G_o denotes the real system we are trying to derive a model for and where $\hat{G}$ is the actual model. Assuming that G_o and $\hat{G}$ are stable LTI systems, the difference is also stable and LTI, implying that measurements from $G = G_o - \hat{G}$ can be collected by exciting G_o and $\hat{G}$ independently with the same input sequence (u_τ) and then subtracting their outputs. When G denotes the modelling error, the goal is to estimate the $\mathcal{H}_\infty$-norm of G: $\|G\|_\infty := \max_{\omega\in[0,\pi]} \left|G(\mathrm{e}^{j\omega})\right|$, where both G and H are unknown to us. We assume that the maximum is attained in $(0, \pi)$.

In this work, we estimate $\|G\|_\infty$ recursively from input-output data collected in sequential experiments (rounds) $t = 1, \dots, T$. At each round $t \in \{1, \dots, T\}$, experiments are designed by defining an input sequence $u^t := (u_0^t, \dots, u_{N-1}^t)$ and collecting a noisy output $y^t = (y_0^t, \dots, y_{N-1}^t)$ disturbed by the additive *non-white* Gaussian sequence $e^t := (e_0^t, \dots, e_{N-1}^t)$ of zero mean. Experiments are performed independently of previous and future ones by waiting long enough between two consecutive experiments[2]. We allow experiments to be sequentially designed, that is, sequence u^t is mapped from previous input-output data $(u^1, y^1, \dots, u^{t-1}, y^{t-1})$. Furthermore, the input to G at round $t \in \{1, \dots, T\}$ is restricted to be a unit-norm sinusoidal sequence parametrized by frequency[3] $\omega \in [0, \pi]$. As we explain in the following paragraph, we make use of bandit technology to design an agent $\boldsymbol{\pi}$ able to design these experiments optimally by discretizing the frequency axis into K (with K large enough) equally spaced frequencies, denoting the possible arms the agent can choose at each round. Then, the discretized $\mathcal{H}_\infty$-estimation problem becomes $\|G\|_\infty := \max_{\omega\in[0,\pi]} \left|G(\mathrm{e}^{j\omega})\right| \approx \max_{k\in\{1,\dots,K\}} \left|G(\mathrm{e}^{j\omega_k})\right|$, where $\omega_k := 2\pi k/(2K+1)$. Hence, at every round t, u^t is completely characterized by its frequency $\omega_{k^{\boldsymbol{\pi}}(t)}$, with $k^{\boldsymbol{\pi}}(t)$ being selected by the agent $\boldsymbol{\pi}$ among the K different arms (frequencies). To avoid frequency leakage [3], N is set to $2K + 1$.

For every k, U_k^t, Y_k^t and E_k^t denote the discrete Fourier transforms (DFT) of u^t, y^t, e^t, respectively, at frequency $\omega_{k(t)}$ where $U_k^t = U^t(\omega_k) := \frac{1}{\sqrt{N}} \sum_{\tau=0}^{N-1} u_\tau^t \mathrm{e}^{-j\omega_{k(t)}\tau}$, and analogously for y^t and e^t. The agent has access to both U^t and Y^t, but not to E^t, where $U_{k(t)}^t = 1$, and $U_i^t = 0, i \neq k(t)$.

[1] A zero-mean sequence (w_t) is said to be white if $\mathbb{E}\{w_i w_k\} = \delta_{ij}$.

[2] This is assumed so the natural response of the system, due to initial conditions introduced by the previous experiments, exponentially decays to zero, making the plant static in terms of inputs $(u^1, u^2, \dots)$ and outputs $(y^1, y^2, \dots)$. For practicality, one can directly reset the system (if possible) or use a controller that brings the state of the system to zero in a finite amount of time.

[3] This assumption is general since $\|G\|_\infty = \sup_{u:\|u\|\neq 0} \|y\|/\|u\|$ is attained by a sinusoidal sequence of frequency $\omega^\star = \arg\max_{\omega\in[0,\pi]} \left|G(\mathrm{e}^{j\omega})\right|$ [20, Chapter 7].

Remark 1 *$\{E_k^t\}_{k=1}^K$ is a circularly symmetric [21, Section 3.7] complex zero-mean and white sequence, whose real and imaginary parts are statistically independent for every $k \in \{1,\dots,K\}$. Moreover, for every k, the real and imaginary parts of E_k^t are zero-mean Gaussian with variance $\sigma_k^2/2 := \mathbb{E}\{|E_k^t|^2\}/2 = |H(\mathrm{e}^{j\omega_k})|^2/2$. Notice that the sequence $(\sigma_k^2)_k$ is unknown since H is unknown.*

The *outcome* perceived by the agent when it plays arm k at experiment t is the $\mathbb{R}^2$-vector

$$\begin{aligned} X_k^t = X^t(\omega_k) &:= \begin{bmatrix} \mathrm{Re}\,\frac{Y_k^t}{U_k^t} & \mathrm{Im}\,\frac{Y_k^t}{U_k^t} \end{bmatrix}^\top \\ &= \begin{bmatrix} \mathrm{Re}\left\{G(\mathrm{e}^{j\omega_k}) + \frac{E_k^t}{U_k^t}\right\} & \mathrm{Im}\left\{G(\mathrm{e}^{j\omega_k}) + \frac{E_k^t}{U_k^t}\right\} \end{bmatrix}^\top, \end{aligned} \tag{2}$$

for every $k \in \{1,\dots,K\}$ and $t \in \{1,\dots,T\}$. Since $G(\mathrm{e}^{j\omega_k})$ is deterministic, it follows that

$$\mathbb{E}\left\{X_k^t \,|\, U_k^t\right\} = \begin{bmatrix} \mathrm{Re}\, G(\mathrm{e}^{j\omega_k}) & \mathrm{Im}\, G(\mathrm{e}^{j\omega_k}) \end{bmatrix}^\top, \tag{3}$$

$$\mathrm{var}\left\{X_k^t\right\} = \mathbb{E}\left\{\frac{|E_k^t|^2}{|U_k^t|^2}\right\} = \sigma_k^2. \tag{4}$$

Therefore, the outcome X_k^t corresponds to a bivariate Gaussian vector

$$X_k^t \,|\, U_k^t \;\sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I}_2/2), \tag{5}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{\mu}_k := [\mathrm{Re}\, G(\mathrm{e}^{j\omega_k}) \quad \mathrm{Im}\, G(\mathrm{e}^{j\omega_k})]^\top$, and where the best arm is denoted by $k^\star := \arg\max_{k\in\{1,\dots,K\}} \|\boldsymbol{\mu}_k\| = \arg\max_{k\in\{1,\dots,K\}} \left|G(\mathrm{e}^{j\omega_k})\right|$.

The selection of $k(t)$ is adaptive, depending on the collected outcomes and played arms on previous rounds, *i.e.*, for class of policies considered in this paper, $k(t)$ is $\mathcal{F}_t$-measurable, where $\mathcal{F}_t$ is the sigma algebra generated by $(k(1), X^1_{k(1)}, \dots, k(t-1), X^{t-1}_{k(t-1)})$. Denote $\boldsymbol{\Pi}$ as the set composed by these policies. For a given policy $\boldsymbol{\pi}$, its performance is indexed by the cumulative expected regret $\mathbb{E}\{R^{\boldsymbol{\pi}}(\boldsymbol{\mu}, T)\} = \sum_{t=1}^T \mathbb{E}\left\{\Delta_{k^{\boldsymbol{\pi}}(t)}\right\}$ it incurs , where $\Delta_k := \|\boldsymbol{\mu}_{k^\star}\| - \|\boldsymbol{\mu}_k\|$, with $\|\cdot\|$ being the Euclidean norm, and where $k^{\boldsymbol{\pi}}(t)$ is the arm at round t under policy $\boldsymbol{\pi}$. The bandit problem is then summarized as [10]

$$\min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}\{R^{\boldsymbol{\pi}}(\boldsymbol{\mu}, T)\} = \min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{k\neq k^\star} \mathbb{E}\{N_k^{\boldsymbol{\pi}}(T)\}\,\Delta_k, \tag{6}$$

where $N_k^{\boldsymbol{\pi}}(t) = \sum_{t=1}^t \mathbb{1}\{k^{\boldsymbol{\pi}}(t) = k\}$ is the number of times arm k has been played up to round t under policy $\boldsymbol{\pi}$.

Remark 2 *Our definition of regret involves the measurement of a random variable of mean $\|\boldsymbol{\mu}_k\|$ for each arm $k \in \{1,\dots,K\}$. However, the feedback the agent receives is not the reward, but the outcome $X^t_{k(t)}$. This supposes an important difference to traditional MAB problems.*

Remark 3 *The gain estimation problem can be casted in different ways. In the traditional approach, we wish to solve $\min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}\{(\|\boldsymbol{\mu}_{k^\star}\| - \|\boldsymbol{\mu}_{\hat{k}^{\boldsymbol{\pi}}(T)}\|)^2\}$, where $\hat{k}^{\boldsymbol{\pi}}(T)$ is an $\mathcal{F}_{T+1}(T)$-measurable estimation of $k^\star$ under policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$. As discussed in [22], solving (6) is much more challenging because it accounts for the sample complexity. Additionally, the traditional problem only minimizes the estimation after T experiments, whereas in (6) we minimize the cumulative error when estimating the gain after each of the T experiments.*

3 UNIFORMLY GOOD POLICIES AND THOMPSON SAMPLING

In the following, and without loss of generality, we assume that $k^\star = 1$. This is possible because smoothness of the frequency responses G, H is not exploited by the considered class of algorithms. As discussed in [23], the regret can be pushed down if the agent were allowed to consider smoothness as prior knowledge.

Definition 1 *A policy $\boldsymbol{\pi}$ is said to be* uniformly efficient *[10] $N_k^{\boldsymbol{\pi}}(t) = \mathrm{o}(t^\alpha)$, for every $\alpha > 0$, and for every suboptimal arm k.*

When $k = 1$ is the unique optimal arm, among other mild regularity conditions, a lower bound derived in [11] for the asymptotic number of times each suboptimal arm is played holds. Using a similar reasoning, we derive the following lower bound for the regret incurred by any uniformly efficient algorithm aimed at solving (6).

Lemma 1 *Under every uniformly efficient algorithm $\boldsymbol{\pi}$, the expected cumulative regret satisfies*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\{R^{\boldsymbol{\pi}}(T)\}}{\log T} \geq \sum_{k=2}^K \frac{\|\boldsymbol{\mu}_1\| - \|\boldsymbol{\mu}_k\|}{\log\left(1 + \frac{(\|\boldsymbol{\mu}_1\| - \|\boldsymbol{\mu}_k\|)^2}{\sigma_k^2}\right)}. \tag{7}$$

Proof: For any $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_k$ satisfying $\|\boldsymbol{\mu}_1\| > \|\boldsymbol{\mu}_i\|$, $i = 2, \dots,$ it is known [10] that

$$\begin{aligned} &\liminf_{T\to\infty} \frac{\mathbb{E}\{N_k(T)\}}{\log T} \\ &\qquad \geq \frac{1}{\inf_{\boldsymbol{\mu}', \sigma'^2 \,:\, |\boldsymbol{\mu}'| > |\boldsymbol{\mu}_1|} \mathbb{D}\left\{\boldsymbol{\mu}_k, \sigma_k^2 || \boldsymbol{\mu}', \sigma'^2\right\}}, \end{aligned} \tag{8}$$

where $\mathbb{D}\left\{\boldsymbol{\mu}_k, \sigma_k^2 || \boldsymbol{\mu}', \sigma'^2\right\}$ denotes the Kullback-Leibler [24] (KL) divergence between two bivariate distributions parametrized by $(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I}_2)$ and $(\boldsymbol{\mu}', \sigma'^2 \boldsymbol{I}_2)$,

respectively, and is given by

$$\mathbb{D}\left\{\boldsymbol{\mu}_k, \sigma_k^2 || \boldsymbol{\mu}', \sigma'^2\right\} = \log \frac{\sigma'^2}{\sigma_k^2} + \frac{\sigma_k^2 + \|\boldsymbol{\mu}_k - \boldsymbol{\mu}'\|}{\sigma'^2} - 1.$$

It then follows that

$$\inf_{\boldsymbol{\mu}', \sigma'^2 \,:\, |\boldsymbol{\mu}'| > |\boldsymbol{\mu}_1|} \log \frac{\sigma'^2}{\sigma_k^2} + \frac{\sigma_k^2 + \|\boldsymbol{\mu}_k - \boldsymbol{\mu}'\|}{\sigma'^2} - 1 = \log\left(1 + \frac{(\|\boldsymbol{\mu}_1\| - \|\boldsymbol{\mu}_k\|)^2}{\sigma_k^2}\right). \quad (9)$$

The proof is completed by observing that $\mathbb{E}\{R^{\boldsymbol{\pi}}(T)\} = \sum_{k=1}^K \Delta_k \mathbb{E}\{N_k^{\boldsymbol{\pi}}(T)\}$[10]. ∎

Thompson Sampling (TS) is a Bayesian bandit policy. It starts with a prior distribution over the unknown parameters parametrizing the distribution of each arm. These parameters are deterministic but unknown, and the prior distribution represents our belief on each value being the actual parameter. At each experiment, the agent collects data and updates our belief, known as the *posterior distribution* of the parameters. TS keeps track of the posterior distribution for the mean of the rewards, drawing one sample of these posteriors at each round in order to play the arm with the highest sample during the next round. Algorithm 1 summarizes this procedure.

Algorithm 1 Thompson Sampling

```
Input: ρ^1 = (ρ_1^1, ..., ρ_K^1) (prior distribution for each reward mean)
for t = 1 to T do
    for k = 1 to K do
        Draw one sample μ̃_k ~ ρ_k^t
    end for
    Play arm k^TS(t) = arg max_k μ̃_k
    Collect the outcome from arm k^TS(t)
    Update the posterior ρ^{t+1} (given F_{t+1})
end for
```

In our case, the prior distribution is selected as $f_{\boldsymbol{\mu}_k, \sigma_k^2}(\boldsymbol{\mu}_k, \sigma_k^2) \propto 1$, $(\boldsymbol{\mu}_k, \sigma_k^2) \in \mathbb{R}^2 \times (0, \infty)$, $k \in \{1, \ldots, K\}$. This corresponds to an improper prior (since it is not integrable) that, in spirit, assigns the same confidence to each pair in $\mathbb{R}^2 \times (0, \infty)$.

The drawback of using TS to solve our problem is that the posterior mean of the rewards, say $f_{\|\boldsymbol{\mu}_k\| \,|\, \mathcal{F}_t}$, might be hard to obtain. However, we can overcome this issue in two steps. Firstly, for every $\ell \in \{1, \ldots, T\}$, we condense $\mathcal{F}_{\ell+1}$ into the sufficient statistics

$$\bar{\boldsymbol{x}}_k(\ell) = \bar{\boldsymbol{x}}_{k, N_k(\ell)} := \frac{1}{N_k(\ell)} \sum_{t=1}^{\ell} X_k^t \mathbb{1}\{k(t) = k\}$$

$$S_k(\ell) := S_{k, N_k(\ell)} = \sum_{t=1}^{\ell} \mathbb{1}\{k(t) = k\} \|X_k^t - \bar{\boldsymbol{x}}_{k, N_k(\ell)}\|^2,$$

and then $\bar{\boldsymbol{x}}_{k,n} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I}/(2n))$ and $S_{k,n}/(\sigma_k^2/2) \sim \chi^2_{2(n-1)}$. Secondly, samples from $f_{\|\boldsymbol{\mu}_k\| \,|\, \mathcal{F}_t}$ can be obtained by sampling $f_{\boldsymbol{\mu}_k \,|\, \bar{\boldsymbol{x}}_k(t) = \boldsymbol{x}, S_k(t) = s}$ and then taking the norm of the sample. Let $\tilde{\boldsymbol{\mu}}_k(t) \sim f_{\boldsymbol{\mu}_k \,|\, \bar{\boldsymbol{x}}_k(t) = \boldsymbol{x}, S_k(t) = s}$ denote a random vector whose pdf (probability density function) is the posterior distribution for $\boldsymbol{\mu}_k$ given $\mathcal{F}_t$. The following result shows that, for every $k \in \{1, \ldots, K\}$, given $\bar{\boldsymbol{x}}_k(t)$ and $S_k(t)$, $\sqrt{2((N_k(t) - 2)n/S_k(t)}(\tilde{\boldsymbol{\mu}}_k(t) - \bar{\boldsymbol{x}}_{N_k}(t))$ has a bivariate t-distribution with parameter $\nu = 2(N_k(t) - 2)$.

Lemma 2 *Let the outcomes $X_{k(t)}^t$ be generated as in* (5)*, and consider the improper prior $f_{\boldsymbol{\mu}, \sigma^2}(\boldsymbol{\mu}, \sigma^2) \propto 1$. Then, for every $k \in \{1, \ldots, K\}$, the posterior mean density function, given $n := N_k(t) \geq 2$ samples and sufficient statistics $\hat{\theta}_{k,n} = (\bar{\boldsymbol{x}}_{k,n}, S_{k,n})$, is*

$$\rho_k^t(\boldsymbol{\mu}) := f_{\boldsymbol{\mu}_k \,|\, \bar{\boldsymbol{x}}_{k,n} = \boldsymbol{x}, S_{k,n} = s}(\boldsymbol{\mu}) = \frac{n(n-2)}{\pi s}\left(1 + \frac{n\|\boldsymbol{x} - \boldsymbol{\mu}\|^2}{s}\right)^{-n+1}. \quad (10)$$

Remark 4 *After t experiments, $\|G\|_\infty$ can be estimated as $\|\bar{\boldsymbol{x}}_{\beta(t)}(t)\|$ where $\beta(t) = \arg\max_{k \in \{1, \ldots, K\}} N_k(t)$. However, different methods to perform the estimation at round t may be considered, as we explain in Section 6.*

4 CONCENTRATION INEQUALITIES

In this section we provide concentration inequalities for the tail upper bounds of the sufficient statistics and for the posterior mean conditioned on them.

Lemma 3 *For every arm $k \in \{1, \ldots, K\}$, with $N_k(t) = n \geq 2$, and $\epsilon > 0$, it holds that*

$$\mathrm{P}\{\|\bar{\boldsymbol{x}}_{k,n}\| \geq \|\boldsymbol{\mu}_k\| + \epsilon\} \leq \mathrm{e}^{-n\epsilon^2/\sigma^2},$$

$$\mathrm{P}\left\{S_{k,n} \geq n(\sigma^2 + \epsilon)\right\} \leq \left(1 + \frac{\epsilon}{\sigma_k^2}\right)^{-1} \mathrm{e}^{-nh(\epsilon/\sigma_k^2)},$$

where $h(x) = x - \log(1 + x) > 0, \quad \forall x > 0$.

Proof: Consider any arbitrary arm $k \in \{1, \ldots, K\}$, and let $n \geq 2$ arbitrary. For the first inequality, observe that

$$\begin{aligned}\mathrm{P}\{\|\bar{\boldsymbol{x}}_{k,n}\| \geq \|\boldsymbol{\mu}_k\| + \epsilon\} &\leq \mathrm{P}\{\|\bar{\boldsymbol{x}}_{k,n} - \boldsymbol{\mu}_k\| \geq \epsilon\} \\ &= \int_{\boldsymbol{z} : \|\boldsymbol{z}\| \geq \epsilon} \frac{n}{\pi\sigma_k^2} \mathrm{e}^{-n\|\boldsymbol{z}\|/\sigma_k^2} \\ &= 2\int_{n\epsilon^2/\sigma_k^2}^{\infty} \mathrm{e}^{-y} dy = \mathrm{e}^{-n\epsilon^2/\sigma_k^2}.\end{aligned}$$

For the second inequality, we relax Chernoff's bound:

$$\mathrm{P}\left\{S_{k,n} \geq n(\sigma_k^2 + \epsilon)\right\}$$

$$\leq \mathrm{e}^{\inf_{\lambda<1/\sigma_k^2} \log \mathbb{E}\left\{\mathrm{e}^{\lambda S_{k,n}}\right\} - \lambda n(\sigma_k^2+\epsilon)}$$
$$= \left(\frac{n}{n-1}\right)^{n-1} \left(1+\frac{\epsilon}{\sigma_k^2}\right)^{n-1} \mathrm{e}^{-(1+n\epsilon/\sigma_k^2)}, \quad (11)$$

and the bound follows from $\left(\frac{n}{n-1}\right)^{n-1} < \mathrm{e}^1$ being strictly increasing in n. ∎

Lemma 4 *For every $k = 1, \ldots, K$, satisfying $N_k(t) = n \geq 2$, it holds that*

$$\mathrm{P}\left\{\|\tilde{\boldsymbol{\mu}}_k - \bar{\boldsymbol{x}}_{k,n}\| \geq \delta \,|\, S_{k,n} = s\right\} \leq \left(1+\frac{n\delta^2}{s}\right)^{-n+2}. \quad (12)$$

Proof: From Lemma 2, and by a polar change of coordinates,

$$\mathrm{P}\left\{\|\tilde{\boldsymbol{\mu}}_i - \bar{\boldsymbol{x}}_{k,n}\| \geq \delta \,|\, S_{k,n} = s\right\}$$
$$= \int_{\boldsymbol{z}:\ \|\boldsymbol{z}\|>\delta} \frac{n(n-2)}{\pi s}\left(1+\frac{n\|\boldsymbol{z}\|_2^2}{s}\right)^{-n+1} d\boldsymbol{z}$$
$$= \frac{n(n-2)}{\pi s}\int_\delta^\infty 2\pi \left(1+\frac{nr^2}{s}\right)^{-n+1} r\, dr$$
$$= \left(1+\frac{n\delta^2}{s}\right)^{-n+2}. \quad (13)$$

∎

5 OPTIMALITY OF THOMPSON SAMPLING

In this section we present the main result of our work. We derive an upper bound for (6) under TS by splitting the non-expected regret into three different terms parametrized by some arbitrary $\epsilon(T) > 0$ depending on the time horizon T. It is shown then that as $T \to \infty$, one of the terms achieves the lower bound, while the other two increase slower than $\log T$.

Theorem 1 *Under the improper prior $f_{\boldsymbol{\mu}_k,\sigma_k^2}(\boldsymbol{\mu}_k, \sigma_k^2) \propto 1$, the regret incurred by TS satisfies*

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left\{R^{\mathrm{TS}}(T)\right\}}{\log T} \leq \sum_{k=2}^K \frac{\|\boldsymbol{\mu}_1\| - \|\boldsymbol{\mu}_k\|}{\log\left(1+\frac{(\|\boldsymbol{\mu}_1\|-\|\boldsymbol{\mu}_k\|)^2}{\sigma_k^2}\right)}. \quad (14)$$

Proof: Define the following events for $0 < \epsilon < \min_k(\|\boldsymbol{\mu}_1\| - \|\boldsymbol{\mu}_k\|)/2$:

$$\mathcal{A}(t) := \{\|\tilde{\boldsymbol{\mu}}^\star(t)\| \geq \|\boldsymbol{\mu}_1\| - \epsilon\},$$
$$\mathcal{B}_k(t) := \{\|\bar{\boldsymbol{x}}_k(t)\| \leq \|\boldsymbol{\mu}_1\| + \epsilon, S_k(t) \leq n(\sigma_k^2+\epsilon)\},$$

where $\tilde{\boldsymbol{\mu}}_k(t)$ follows the posterior distribution in Lemma 2, and where $\|\tilde{\boldsymbol{\mu}}^\star(t)\| := \max_k \|\tilde{\boldsymbol{\mu}}_k(t)\|$. Let $\Delta_{\max} := \max_k \Delta_k$, and let $k^{\mathrm{TS}}(t)$ denote the arm played at round t under TS. Define also $\bar{T} := 3K$. The non-expected cumulative regret can be then written as

$$R^{\mathrm{TS}}(\boldsymbol{\mu}, T) = \sum_{t=1}^T \Delta_{k^{\mathrm{TS}}(t)}$$
$$\leq \sum_{t=1}^{\bar{T}}\sum_{k=2}^K \Delta_k \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k\right\}$$
$$+ \sum_{t=1}^{\bar{T}}\sum_{k=2}^K \Delta_k \Big(\mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}(t)\right\} + \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}^c(t)\right\}\Big)$$
$$\leq \bar{T}\sum_{k=2}^K \Delta_k + \sum_{k=2}^K\sum_{t=1}^{\bar{T}} \Delta_k \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}(t)\right\}$$
$$+ \Delta_{\max}\sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) \neq 1, \mathcal{A}^c(t)\right\}$$
$$= \Delta_{\max}\sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) \neq 1, \mathcal{A}^c(t)\right\}$$
$$+ \sum_{k=2}^K \Delta_k \Big(\sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}(t), \mathcal{B}_k(t)\right\} + \sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}(t), \mathcal{B}_k(t)^c\right\} + \bar{T}\Big)$$
$$\leq \Delta_{\max}\sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) \neq 1, \mathcal{A}^c(t)\right\}$$
$$+ \sum_{k=2}^K \Delta_k \Big(\sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{A}(t), \mathcal{B}_k(t)\right\} + \sum_{t=\bar{T}+1}^T \mathbb{1}\left\{k^{\mathrm{TS}}(t) = k, \mathcal{B}_k(t)^c\right\} + \bar{T}\Big). \quad (15)$$

Now, by Lemmas 5, 6, and 7 (all of them appended in the suplementary material), expectation in (15) yields

$$\frac{\mathbb{E}\left\{R^{\mathrm{TS}}(\boldsymbol{\mu}, T)\right\}}{\log T} \leq \frac{1}{\log\left(1+\frac{(\|\boldsymbol{\mu}_1\|-\|\boldsymbol{\mu}_k\|-2\epsilon)^2}{\sigma_k^2+\epsilon}\right)} + \frac{-1+\mathrm{O}(\epsilon^{-2})+\mathrm{O}(\epsilon^{-6})}{\log T}, \quad (16)$$

so the result follows by choosing $\epsilon \leq \log^{-a} T$, $1/6 > a > \mathrm{e}^{(\min_k \|\boldsymbol{\mu}_1\|-\|\boldsymbol{\mu}_k\|)/2}$. ∎

6 APPLICATION TO $\mathcal{H}_\infty$-NORM ESTIMATION

The previous sections establish a framework for selecting the optimal frequency as often as possible, however, they do not provide an estimate of $\|\boldsymbol{\mu}_{k^\star}\| := \max_{k\in\{1,\dots,K\}} \|\boldsymbol{\mu}_k\| \approx \|G\|_\infty$. In this section we discuss the challenge of how TS can be employed to estimate the $\mathcal{H}_\infty$-norm of a system, given collected data on previous experiments, in an asymptotically efficient[4] manner. The theoretical relationship between regret-optimal algorithms and asymptotically efficient estimation is still ongoing research. We remark that there exist several ways to proceed that can also lead to asymptotically efficient estimates but we only address two ways here.

Let $\beta := \|G\|_\infty$ denote the target parameter, and let $\hat{k}^{(1)}(t) := \arg\max_{k\in\{1,\dots,K\}} N_k^{\mathrm{TS}}(t)$ and $\hat{k}^{(2)} := \arg\max_{k\in\{1,\dots,K\}} \|\bar{\boldsymbol{x}}_k(t)\|$ denote two different estimators of the *best arm* at round t. In this case, $\hat{k}^{(1)}$ takes the best arm as that one played the most up to round t, while $\hat{k}^{(2)}$ is that arm whose empirical mean is the largest at round t. The two considered estimators are then

$$\hat{\beta}^{(i)}(t) := \|\bar{\boldsymbol{x}}_{\hat{k}^{(i)}}\|, \quad i \in \{1,2\}. \tag{17}$$

Algorithm 2 summarizes the procedure within a bandit set-up.

Algorithm 2 $\mathcal{H}_\infty$-norm estimation using Thompson Sampling

1: Input: $\rho^1 = (\rho_1^1, ..., \rho_K^1)$ (prior distribution for each reward mean)
2: **for** $t = 1$ to T **do**
3: Play arm $k^{\mathrm{TS}}(t)$ (see Algorithm 1)
4: Obtain $\hat{k}^{(i)}(t)$
5: Set $\hat{\beta}^{(i)}(t) := \|\bar{\boldsymbol{x}}_{\hat{k}^{(i)}}\|, \quad i \in \{1,2\}.$
6: **end for**

In the following subsections, an illustrative example is provided where the analysis of Algorithm 2 is done separately in terms of regret and estimation analysis, and the estimation analysis includes a comparison to other well known algorithms for $\mathcal{H}_\infty$-norm estimation. The first method corresponds to a well established algorithm [18], [19] based on the power-iterations method to estimate the largest eigenvalue (in absolute value) of a matrix. This algorithm is well known for having a fast rate of convergence and by being asymptotically efficient in the absence of output noise, although it is also well known [19] that it leads to bias estimation under noisy measurements. The second algorithm is a method recently analyzed in [8] in terms of sample complexity. This rather old algorithm estimates a transfer function from input-output data and uses this information to compute its largest gain.

[4] Asymptotically efficiency means that the proposed estimator has an asymptotic estimation variance that matches the lower bound stated in [25], [26]

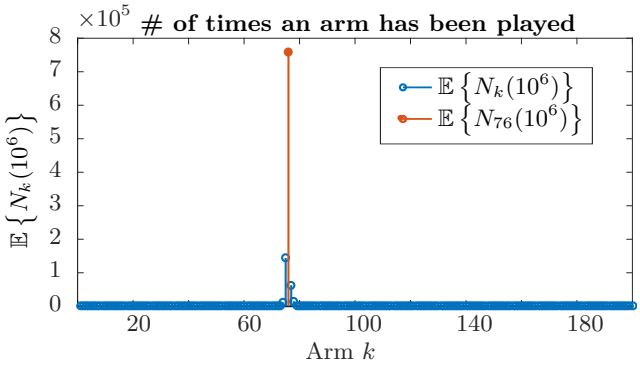

Figure 2: Expected number of times each of the 200 arms has been played at round $T = 10^6$ for each arm $k \in \{1, \dots, K\}$ (in blue) and for the optimal arm ($k^\star = 76$) in red.

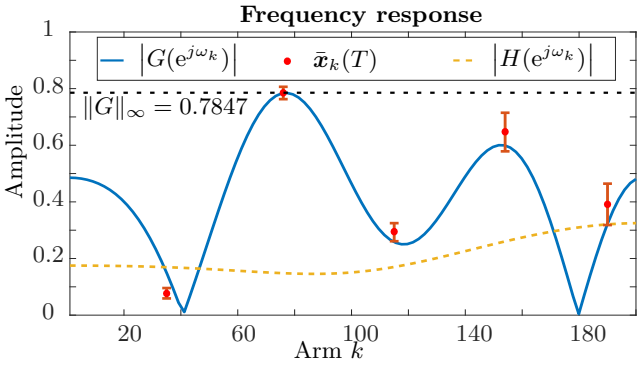

Figure 3: Frequency response of G together with the frequency response of G at arm $k^\star = 76$. Additionally, we present $\bar{\boldsymbol{x}}_i(T)$ as red dots and $S_i(T)/N_i^{\mathrm{TS}}(T)$ as red bars around $\bar{\boldsymbol{x}}_i(T)$, for $i \in \{35, 76, 115, 154, 190\}$.

6.1 An illustrative example

We consider systems G, H having frequency responses depicted in Fig. 3. The frequency axis is discretized into $K = 200$ equispaced frequencies. Under this setup, the optimal arm is $k^\star = 76$ ($\omega_{k^\star} = 2\pi\frac{k^\star}{2K+1} = 1.1908$ [rad/s]), satisfying $\|\boldsymbol{\mu}_{k^\star}\| = \max_k \|\boldsymbol{\mu}_k\| = \|\boldsymbol{\mu}_{76}\|$.

6.1.1 Regret analysis

We observe in Fig. 2 that the algorithm is able to find the optimal arm. Figure 3 shows the sufficient statistic for only five arms (including $k^\star$): the empirical mean $\bar{\boldsymbol{x}}_{k,N_k^{\mathrm{TS}}(T)}$ and the empirical variance $S_{k,N_k^{\mathrm{TS}}(T)}$. It can be seen that the estimation of the variance for each arm grows as the arm is played less often, as predicted. In this line, the length of the estimated variance is smaller on the optimal arm and higher for the suboptimal ones, since the latter are less often played.

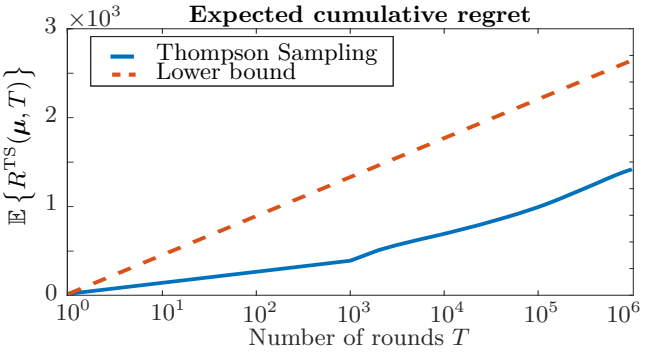

Figure 4: Comparison between the regret incurred by Thompson Sampling and the theoretical lower bound derived in Lemma 1.

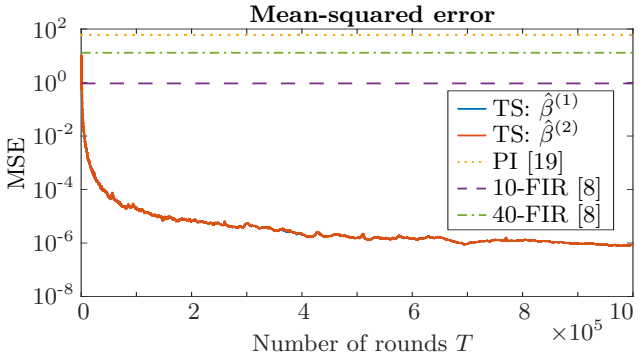

Figure 5: Performance of different $\mathcal{H}_\infty$-norm estimators in the mean-squared error sense.

Figure 4 shows the performance of TS comparing it with the predicted lower bound in Lemma 1. As predicted by Theorem 1, the asymptotic slope in the rate of growth of the expected cumulative regret under Thompson Sampling matches the slope described by the lower bound in Lemma 1.

6.1.2 Estimation analysis

Figure 5 depicts how the proposed algorithms perform, in a MSE (mean-squared error) sense, against two other methods for data-driven $\mathcal{H}_\infty$-norm estimation, named PI (power-iterations based algorithm [18]) and L-FIR (computation of the $\mathcal{H}_\infty$ norm of an L-length fitted FIR filter from input-output data). To make the test more challenging, the noise level is increased so that $\left|H(\mathrm{e}^{j\omega})\right|$ is always above $\left|G(\mathrm{e}^{j\omega})\right|$, $\forall\omega$ (the noise level is now 40 times larger than in Fig. 3). Covering $|G|$ with the noise level reveals that some algorithms are just unable to create accurate estimates.

We start by noticing the poor performance of the estimates provided by PI where, in line with what is predicted by [19], the stationary estimates are biased and the error covariance is significantly higher than the rest of the algorithms. On the other hand, we see that the quality of the estimates passing through an L-length FIR model do not provide efficient estimates and that, at least empirically, the estimation quality does not increase with the length of the filter, reinforcing what we have discussed in the introduction. On the other hand, we see that both proposed estimators $\hat{\beta}^{(i)}$, $i \in \{1, 2\}$ perform similarly (they actually overlap in Fig. 5) well compared to the other algorithms. We remark that the MSE in the estimation of $\hat{\beta}^{(i)}$, $i \in \{1, 2\}$ can not be lower that the discretization error $\|G\|_\infty - \|\boldsymbol{\mu}_{k^\star}\|$.

Finally, we notice that the proposed algorithm behaves similar in both cases, attaining a covariance that decreases to zero, suggesting that these methods could provide efficient estimates of β.

7 CONCLUSIONS

We have presented a novel application of multi-armed bandits to the problem of estimating the $\mathcal{H}_\infty$-norm of a linear dynamical system. The novelty of this approach lies in that the optimal arm is given by the one whose bivariate outcomes have the largest mean norm, arising an interesting variation of the two-parameter Gaussian bandit. We provide results with high theoretical quality by deriving a lower bound for this class of problems together with an algorithm which attains such regret. Additionally, we have proposed two different ways of obtaining an estimator from Thompson Sampling for the $\mathcal{H}_\infty$-norm of an LTI system under colored noise, which compare to other known algorithms for $\mathcal{H}_\infty$-norm estimation, outperforming the latter in a simulation study.

References

[1] M. L. Puterman, *Markov Decision Processes*. Willey, 2005.

[2] T. Söderström, *Discrete-time Stochastic Systems*. Springer, 2002.

[3] L. Ljung, *System Identification: Theory for the User*. Prentice Hall, 1999.

[4] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall, 1996.

[5] F. W. Fairman, *Linear Control Theory: The State Space Approach*. John Willey & Sons, 1998.

[6] N. Bruinsma and M. Steinbuch, "A fast algorithm to compute the $\mathcal{H}_\infty$-norm of a transfer function matrix," *Systems & Control Letters*, vol. 14, no. 4, pp. 287–293, 1990.

[7] M. N. Belur and C. Praagman, "An efficient algorithm for computing the $\mathcal{H}_\infty$ norm," *IEEE Transactions on Automatic Control*, vol. 56, pp. 1656–1660, July 2011.

[8] S. Tu, R. Boczar, and B. Recht, "On the approximation of Toeplitz operators for nonparametric $\mathcal{H}_\infty$-norm estimation," in *Proceedings of the American Control Conference (ACC)*, 2018.

[9] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[10] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[11] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17, no. 7, pp. 122–142, 1996.

[12] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3, pp. 285–294, 1933.

[13] O. Chapelle and L. Li, "An empirical evaluation of Thompson Sampling," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

[14] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT)*, pp. 199–213, Springer, 2012.

[15] N. Korda, E. Kaufmann, and R. Munos, "Thompson sampling for 1-dimensional exponential family bandits," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[16] J. Honda and A. Takemura, "Optimality of Thompson sampling for Gaussian bandits depends on priors," in *Proceedings of the 17th international conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[17] M. I. Müller, P. E. Valenzuela, A. Proutiere, and C. R. Rojas, "A stochastic multi-armed bandit approach to nonparametric $\mathcal{H}_\infty$-norm estimation," in *Proceedings of the 56th IEEE Conference on Decision and Control*, 2017.

[18] B. Walhberg, M. Barenthin, and H. Hjalmarsson, "Nonparametric methods for $\mathcal{L}_2$-gain estimation using iterative experiments," *Automatica*, vol. 46, no. 8, pp. 1376–1381, 2010.

[19] C. R. Rojas, T. Oomen, H. Hjalmarsson, and B. Wahlberg, "Analyzing iterations in identification with application to nonparametric -norm estimation," *Automatica*, vol. 48, no. 11, pp. 2776–2790, 2012.

[20] M. Barenthin, *Complexity Issues, Validation and Input Design for Control in System Identification*. PhD thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2008.

[21] J. C. Agüero, J. I. Yuz, G. C. Goodwin, and R. Delgado, "On the equivalence of time and frequency domain maximum likelihood estimation," *Automatica*, vol. 46, pp. 260–270, February 2010.

[22] J.-B. Grill, M. Valko, and R. Munos, "Black-box optimization of noisy functions with unknown smoothness," in *Neural Information Processing Systems*, 2015.

[23] S. Magureanu, "Lipschits bandits: Regret lower bounds and optimal algorithms," in *Proceedings of the 27th annual Conference on Learning Theory (COLT)*, 2014.

[24] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 758–764, 1951.

[25] R. C. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, no. 3, pp. 81–91, 1945.

[26] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[27] F. Dawson, "On the numerical value of $\int_0^h \mathrm{e}^{x^2} dx$," *Proceedings of the London Mathematical Society*, vol. 29, no. 1, pp. 519–522, 1898.