<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *56th IEEE Conference on Decision and Control*.

Citation for the original published paper:

Müller, M I., Valenzuela, P E., Proutiere, A., Rojas, C R. (2017)
A stochastic multi-armed bandit approach to nonparametric H_∞ -norm estimation
In: *56th IEEE Conference on Decision and Control* (pp. 4632-4637). Institute of
Electrical and Electronics Engineers (IEEE)
<https://doi.org/10.1109/CDC.2017.8264343>

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-223861>

A Stochastic Multi-Armed Bandit Approach to Nonparametric $\mathcal{H}_\infty$ -Norm Estimation

Matias I. Müller, Patricio E. Valenzuela, Alexandre Proutiere and Cristian R. Rojas

Abstract—We study the problem of estimating the largest gain of an unknown linear and time-invariant filter, which is also known as the $\mathcal{H}_\infty$ norm of the system. By using ideas from the stochastic multi-armed bandit framework, we present a new algorithm that sequentially designs an input signal in order to estimate this quantity by means of input-output data. The algorithm is shown empirically to beat an asymptotically optimal method, known as Thompson Sampling, in the sense of its cumulative regret function. Finally, for a general class of algorithms, a lower bound on the performance of finding the $\mathcal{H}_\infty$ norm is derived.

I. INTRODUCTION

In automatic control, controller synthesis is based on the available knowledge about the process to be controlled, where mathematical models are built in order to represent the dynamics of the system, usually condensed as parametric or non-parametric models. The lack of knowledge to fully describe a phenomenon causes several problems when trying to take the best decision under limited information, and, for this reason, the task of designing a control law might be difficult due to different sources of uncertainty. More precisely, the mismatch between the model and the true process, or *modelling error*, must be considered when the closed-loop performance is aimed to be optimized. Previous works [1], [2], [3] have studied this topic under the name of robust control, and its solution involves spectral properties of the modelling error. Indeed, fundamental limitations of the closed-loop performance are posed in terms of the largest gain of the modelling error (or a weighted version of it), also known as its $\mathcal{H}_\infty$ norm.

Although most $\mathcal{H}_\infty$ -norm estimation algorithms rely on an explicit transfer function (or state space model) for the system (see, e.g., [4], [5]), this assumption does not hold when the largest gain of the modelling error is aimed to be estimated. When the system (or, equivalently, the modelling error) remains unknown, the $\mathcal{H}_\infty$ norm can be computed as the maximum quotient between the Euclidean norm of the noiseless output and the input, and it is well known that, for linear and time-invariant (LTI) systems, the input that maximizes this ratio is a sinusoid whose frequency is equal to the one where the largest gain of the system is located (called the *peak frequency* from now on). Then, the corresponding problem simplifies to finding this frequency, see, e.g., [6, Chapter 7]. However, if only noisy measurements can be collected, the accuracy of the estimation highly depends on the quality of the data. For instance, the applications of the iterative method provided in [7] are shown to be limited since it is inconsistent under the presence of noise.

The problem of estimating efficiently the $\mathcal{H}_\infty$ -norm of a system in a model-free manner involves finding as quickly as possible the peak frequency, so the $\mathcal{H}_\infty$ gain can then be estimated by applying to the system a sinusoid of such frequency as input. In this paper, we formulate the problem of quickly identifying the peak frequency as an on-line stochastic optimization problem that resembles a linear Multi-Armed Bandit (MAB) problem [8]. MAB problems [9] are sequential decision problems where in each round the decision maker selects an arm (for us a frequency) and observes a noisy realization of its corresponding reward whose distribution is initially unknown. The objective is to maximize the expected cumulative reward, and this requires to balance exploitation (arms with high rewards should be selected often enough) and exploration (all arms should be played enough times to learn their reward distributions). In the MAB problem arising from $\mathcal{H}_\infty$ -norm identification, the arms correspond to the possible input signals or to their spectral signatures, and the feedback consists of the spectral response to this input signal.

The main contribution of this work is to propose a new algorithm that estimates the $\mathcal{H}_\infty$ -norm of an unknown system by recursively designing the input signal. Motivated by the well known Thompson Sampling strategy [10], which assumes a prior distribution over the frequency response for each frequency, the spectral properties of the input signal are refined, at each round, according to the posterior probability of each frequency being the peak one. Additionally, we establish a lower bound in the performance for a wide family of algorithms which estimate the $\mathcal{H}_\infty$ -norm iteratively.

In the control design field, only a few successful multi-armed-bandit applications have been already reported, with special focus in *adaptive control* [11]. Indeed, the adaptive control problem can be addressed in terms of an exploration/exploitation trade-off, where the input signal is recursively designed with two proposes: identification and control. In fact, a standard MAB strategy, known as upper-confidence-bound (UCB) [12], has been proposed as a solution to an adaptive control problem under the name of "Bet on the Best" (BoB) [13]. Other incursions of MAB in automatic control include works as [14], [15], where the linear-quadratic Gaussian (LQG) control problem is analyzed from an MAB perspective.

The paper is organized as follows: Section II describes the problem statement. Section III introduces the stochastic MAB framework addressing the peak-frequency identification as an online stochastic optimization problem. In Section IV a new algorithm for $\mathcal{H}_\infty$ -norm estimation, based on the stochastic MAB framework, is proposed. Section V provides an example on how well the new algorithm performs and Section VI presents conclusions. The proof of technical results are provided in the Appendix.

This work was partially supported by the Swedish Research Council under contracts 2015-04393 and 2016-06079.

Authors are with Department of Automatic Control, KTH Royal Institute of Technology, SE100-44 Stockholm, Sweden, {mimr2, pva, alepro, crroj}@kth.se

Fig. 1. Input-output mathematical relationship.

A. Notation

The sets of natural, real, positive real, and complex numbers are $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{C}$, respectively. The one-step time shift operator, in discrete time, is denoted by q . The real part of $z \in \mathbb{C}$ is $\text{Re } z$ and its imaginary part is $\text{Im } z$. The transpose and Hermitian (conjugate transpose) of a matrix are defined as $()^\top$ and $()^H$, respectively, and $\mathbf{I}$ symbolizes an identity matrix of appropriate dimensions. We denote by $\mathbb{E}\{\cdot\}$ the expected value of a random variable, and we make explicit its dependence on a distribution ν by writing $\mathbb{E}_\nu\{\cdot\}$. $X \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$ symbolizes that X has a Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{K}$. We write $Z \sim \mathcal{N}_\mathbb{C}(\boldsymbol{\mu}, \mathbf{K})$, $\boldsymbol{\mu} \in \mathbb{C}$, when $Z = X + jY$ has a one-dimensional complex Gaussian distribution, i.e., $(X, Y) \sim \mathcal{N}(\begin{bmatrix} \text{Re } \boldsymbol{\mu} \\ \text{Im } \boldsymbol{\mu} \end{bmatrix}, \mathbf{K})$. For a discrete-time second-order asymptotically wide-sense stationary random signal $(s_\tau)_{\tau=0}^N$, its norm is defined as $\|s\|_2^2 := \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\tau=1}^N \mathbb{E}\{s_\tau^2\}$. $\mathbf{1}_A$ denotes the indicator function, i.e., $\mathbf{1}_A = 1$ if A is true, and $\mathbf{1}_A = 0$ otherwise. $f(x) = o(g(x))$ denotes that $\lim_{x \rightarrow \infty} f(x)/g(x) = 0$. Finally, for a given set B , B^c denotes its complement.

II. PROBLEM STATEMENT

A. Problem set-up

Let G denote a stable LTI causal filter, defining an input-output relationship according to Fig. 1, whose frequency response $G(e^{j\omega})$, $\omega \in [-\pi, \pi]$, is unknown. During a single experiment indexed by $t = 1, \dots, T$, an arbitrary input $u^t := (u_\tau^t)_{\tau=0}^{N-1}$ is applied to the system, and data is collected from the output noisy measurements $y^t := (y_\tau^t)_{\tau=0}^{N-1}$. The measurement noise $e^t := (e_\tau^t)_{\tau=0}^{N-1}$ corresponds to a sequence of iid (independent and identically distributed) Gaussian random variables with zero mean and variance σ^2 . The input signal at experiment t can be arbitrarily designed from batches of N data samples collected in the previous experiments up to $t-1$, that is $(u^1, y^1, \dots, u^{t-1}, y^{t-1})$. Additionally, experiments $t = 1, \dots, T$ are performed independently, which can be achieved by waiting long enough between two consecutive experiments so that the natural response of the system decays to zero before performing the next experiment.

For our purposes, the frequency range $(0, \pi)$ is gridded into a number of K different equally spaced frequencies $(\omega_k)_{k=1, \dots, K}$, where $\omega_k := 2\pi k/(2K+1)$. We then approximate the $\mathcal{H}_\infty$ -norm of G as¹

$$\|G\|_\infty := \max_{\omega \in [0, \pi]} |G(e^{j\omega})| \approx \max_{k \in \{1, \dots, K\}} |G(e^{j\omega_k})|, \quad (1)$$

with K large enough. Let $U^t(\omega_k)$, $Y^t(\omega_k)$ and $E^t(\omega_k)$ be the discrete Fourier transforms (DFT) of the signals u^t , y^t

and e^t , respectively, at ω_k , defined as

$$U^t(\omega_k) := \frac{1}{\sqrt{N}} \sum_{\tau=0}^{N-1} u_\tau^t e^{-j\omega_k \tau}, \quad N=2K+1, \quad k = 1, \dots, K,$$

and analogously for $Y^t(\omega_k)$ and $E^t(\omega_k)$. For simplicity, we will focus on the frequency domain representation of the input rather than on its time-domain behaviour, introducing $p_k^t := |U^t(\omega_k)|^2$. Since we are interested in the output-to-input power ratio, there is no loss of generality in assuming a power constraint on the input; thus we assume $\sum_k p_k^t = 1$, for all t . The following remark captures an essential property for $\{E^t(\omega_k)\}_{k=1}^K$.

Remark 1 The DFT $\{E^t(\omega_k)\}_{k=1}^K$ is circularly symmetric [16, Section 3.7] as pointed out in [17, Section 3]. In other words, sequence $\{E^t(\omega_k)\}_k$ is white and $E^t(\omega_k) \sim \mathcal{N}_\mathbb{C}(0, \sigma^2 \mathbf{I})$, $k = 1, \dots, K$, i.e., its real and imaginary components are statistically independent.

At the t -th experiment, we observe the outcome

$$X_k^t = X^t(\omega_k) := \frac{Y^t(\omega_k)}{U^t(\omega_k)} = G(e^{j\omega_k}) + \frac{E^t(\omega_k)}{U^t(\omega_k)},$$

for all $k = 1, \dots, K$. From Remark 1, and given its structure, we conclude that $(X_k^t)_k$ is a sequence of complex Gaussian random variables whose real and imaginary components are statistically independent. Additionally, the fact that $G(e^{j\omega_k})$ is fixed allows us to conclude that

$$\begin{aligned} \mathbb{E}\{X_k^t | U^t\} &= \mathbb{E}\left\{G(e^{j\omega_k}) + \frac{E^t(\omega_k)}{U^t(\omega_k)}\right\} = G(e^{j\omega_k}), \\ \text{cov}\{X_k^t | U^t\} &= \mathbb{E}\left\{\frac{|E^t(\omega_k)|^2}{|U^t(\omega_k)|^2}\right\} = \frac{\sigma^2}{p_k^t}, \end{aligned}$$

which is summarized as

$$X_k^t | U^t \sim \mathcal{N}_\mathbb{C}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}/2p_k^t), \quad k = 1, \dots, K, \quad (2)$$

where $\boldsymbol{\mu}_k := G(e^{j\omega_k})$. Moreover, the X_k^t 's are independent for different k 's, conditioned on U^t .

B. Objective

At experiment t , we select the power levels $p^t := (p_1^t, \dots, p_K^t) \in \Lambda$ of the input signal at the different frequencies $(\omega_1, \dots, \omega_K)$, where $\Lambda := \{p \in [0, 1]^K : \sum_k p_k = 1\}$ is the simplex. We then observe a realization of the outcome sequence $X^t := (X_1^t, \dots, X_K^t)$ where, for any k , $X_k^t | p^t \sim \mathcal{N}_\mathbb{C}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}/2p_k^t)$. At each experiment t , the estimation of the right-hand side of (1) is denoted as $\hat{\beta}^t$. The selection of p^t may be adaptive, and depend on the power levels and observed output signals in previous experiments. We denote by Π the set of such possible adaptive power level selection strategies, i.e., if $\pi \in \Pi$, the power levels $p^{t, \pi}$ selected under π at the t -th experiment is $\mathcal{F}_t^\pi$ -measurable, where $\mathcal{F}_t^\pi$ is the sigma algebra generated by $(p^{1, \pi}, X^1, \dots, p^{t-1, \pi}, X^{t-1})$.

Our objective is to design a strategy $\pi \in \Pi$ that identifies as fast as possible the system gain $|\boldsymbol{\mu}_{k^*}|$ where $k^* := \arg \max_{k \in \{1, \dots, K\}} |\boldsymbol{\mu}_k|$ is also unknown. More precisely, we may consider the two following (and related) problems:

¹We have omitted 0 and π from the frequency grid $(\omega_k)_{k=1, \dots, K}$ for simplicity, as for these frequencies the noise $E(e^{j\omega})$, defined later, is necessarily real, so its stochastic properties differ from those at other frequencies where it is complex. Including these frequencies can be done at the expense of having a slightly more complicated algorithm.

- **Gain Estimation (GE):** After the T experiments, $\pi \in \Pi$ returns an estimation of the peak frequency index $\hat{k}^\pi$ ($\hat{k}^\pi$ is $\mathcal{F}_{T+1}^\pi$ -measurable), and we wish to solve:

$$(GE) \quad \min_{\pi \in \Pi} \mathbb{E} \left\{ |\mu_{k^*}|^2 - |\mu_{\hat{k}^\pi}|^2 \right\}^2.$$

- **Regret Minimization (RM):** We compare the power level selection strategy π to an Oracle strategy always concentrating its power on the peak frequency, and we quantify its performance using the following notion of regret:

$$\mathcal{R}^\pi(\mu, T) = T |\mu_{k^*}|^2 - \sum_{t=1}^T \sum_{k=1}^K \mathbb{E} \{ p_k^{t,\pi} \} |\mu_k|^2, \quad (3)$$

where $\mu := (\mu_k)_{k=1}^K$. Hence, here our goal is to solve:

$$(RM) \quad \min_{\pi \in \Pi} \mathcal{R}^\pi(\mu, T).$$

Note that we choose $|\mu_k|^2$ rather than $|\mu_k|$ in the above definitions to have a proper and simple notion of regret: $\mathcal{R}^\pi(\mu, T)$ should be the expected difference between the outcome of an optimal algorithm always selecting $p_k^t = \mathbf{1}_{k=k^*}$ and that of the algorithm π . Now when $p_k^t = \mathbf{1}_{k=k^*}$, $\sum_k p_k^t \mathbb{E} \{ |X_k^t|^2 \} = |\mu_{k^*}|^2 + K\sigma^2$, and under any arbitrary p_k^t , $\sum_k p_k^t \mathbb{E} \{ |X_k^t|^2 \} = \sum_k p_k^{t,\pi} |\mu_k|^2 + K\sigma^2$, which is consistent with our definition of regret.

In (GE), we minimize the gain estimation error after T experiments, whereas in (RM), we minimize the cumulative error when estimating the gain after each of the T experiments. As discussed in [18], solving (RM) is more challenging than (GE), since the former accounts for the sample complexity, *i.e.*, the number of T iterations. This is exactly the case when a parameter has to be tuned up for a complex system seen as a black box, whose performance can only be assessed by a costly simulation/experiment. We then focus on (RM) in the remaining of the paper.

III. STOCHASTIC MULTI-ARMED BANDITS

The problem (RM) resembles a linear stochastic bandit problem [8]: the continuous set of arms (here power profiles) is the simplex Δ ; the *reward* at experiment t is the scalar product of the selected power profile p^t and $(|X_1^t|^2, \dots, |X_K^t|^2)$. However, in [8] the feedback provided to the decision maker after she selects p^t is actually a noisy version of the reward $\sum_k p_k^t |\mu_k|^2$, whereas in our problem, she receives a much more detailed feedback. More precisely, the feedback consists in a complex K -dimensional vector X^t , whose k -th component is centred on μ_k and has a variance inversely proportional to the power allocated to frequency ω_k . This difference in feedback is crucial: in [8], the regret is known to scale at least as $\Theta(\sqrt{T})$. We show in this section that for (RM), the regret of an optimal algorithm scales logarithmically with the time horizon T as if the set of arms were discrete.

Before we state this result, it is insightful to restrict our attention to power profile selection algorithms that concentrate their power on a single frequency in each experiment. Under such a policy π , we have for any t , $p^{t,\pi} \in \{0, 1\}^K$ and $\sum_k p_k^{t,\pi} = 1$. Denote by Π_{SF} (SF stands for 'Single Frequency') the subset of Π of such policies. Now under $\pi \in \Pi_{\text{SF}}$, at each experiment, a frequency k^t is chosen and the corresponding output $X_{k^t}^t \sim \mathcal{N}(\mu_{k^t}, \sigma^2 \mathbf{I}/2)$ is observed.

A. Regret Lower Bound and Optimal Algorithms in Π_{SF}

When restricting our attention to Π_{SF} , we recover the classical stochastic bandit problem with a finite number of arms (here frequencies): $\min_{\pi \in \Pi_{\text{SF}}} \mathcal{R}^\pi(\mu, T)$ is just the regret minimization problem introduced by Robbins [9]. For the latter, we have K arms, and when arm k^t is played at round (or experiment) t , we observe $X_{k^t}^t$ and collect a *reward* equal to $|X_{k^t}^t|^2$.

1) *Regret Lower Bound in Π_{SF} :* In their seminal paper [9], Lai and Robbins derive a tight asymptotic (when T grows large) regret lower bound satisfied by any *uniformly efficient* algorithm in Π_{SF} . An algorithm π is uniformly efficient if for any μ , the number of experiments $N_k^\pi(t)$ a suboptimal arm (here frequency) is selected up to experiment t satisfies $\mathbb{E} \{ N_k^\pi(T) \} = o(T^\alpha)$ for any $\alpha > 0$. Let ν_k denote the distribution $\mathcal{N}_C(\mu_k, \sigma^2 \mathbf{I}/2)$ of the observation when the frequency ω_k is selected. It is shown in [9] that under any uniformly efficient algorithm π in Π_{SF} :

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E} \{ N_k^\pi(T) \}}{\log T} \geq \frac{1}{\mathbb{D} \{ \nu_k \| \nu_{k^*} \}}, \quad \text{for all } k \neq k^*,$$

where $\mathbb{D} \{ \nu_k \| \nu_{k^*} \}$ denotes the KL-divergence number between the distributions ν_k and ν_{k^*} . In particular, the regret under the algorithm π satisfies:

$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}^\pi(\mu, T)}{\log T} \geq \sum_{k \neq k^*} \frac{|\mu_{k^*}|^2 - |\mu_k|^2}{\mathbb{D} \{ \nu_k \| \nu_{k^*} \}}.$$

Observe that here, since the ν_k 's are Gaussian distributions, we have: $\mathbb{D} \{ \nu_k \| \nu_{k^*} \} = |\mu_{k^*} - \mu_k|^2 / \sigma^2$.

An algorithm $\pi \in \Pi_{\text{SF}}$ with a regret as low as that predicted by the above lower bound is said to be asymptotically optimal. Some asymptotically optimal algorithms include KL-UCB [19] and Thompson Sampling [20].

2) *The Thompson Sampling (TS) Algorithm:* Thompson Sampling (TS) belongs to the family of Bayesian algorithms, and starts with a prior distribution on the μ_k 's, *e.g.* $\mu_k \sim \mathcal{N}_C(0, \lambda^2 \mathbf{I})$ for some $\lambda > 0$. Based on this prior, we can compute the initial probability distribution ρ^1 of the optimal arm (ρ_k^1 is the initial probability that k is optimal). For each experiment t , TS selects an arm k randomly sampled from the posterior distribution ρ^t of the optimal arm, *i.e.*, $\rho_k^t := \mathbb{P} \{ k^* = k | \mathcal{F}_t \}$ where $\mathcal{F}_t$ represents the information gathered in previous experiments. TS then observes the outcome of its arm selection, and computes ρ^{t+1} accordingly. The procedure is summarized in Algorithm 1, see *e.g.* [21].

Algorithm 1 Thompson Sampling

- 1: Input: $\rho^1 = (\rho_1^1, \dots, \rho_K^1)$ (prior distribution for k^*)
 - 2: **for** $t = 1$ to T **do**
 - 3: Sample $k^t \sim \rho^t$
 - 4: Observe the outcome $X_{k^t}^t$
 - 5: Update the posterior ρ^{t+1}
 - 6: **end for**
-

B. Regret Lower Bound for Algorithms in Π

As mentioned above, (RM) can be interpreted as a linear stochastic MAB problem with continuous set of arms and specific feedback. In linear bandits [8], the minimal regret usually scales as $\Theta(\sqrt{T})$ as T grows large, but for (RM), thanks to the nature of the feedback, this regret may scale

logarithmically with T . Indeed the TS algorithm belongs to $\Pi_{\text{SF}} \subset \Pi$ and exhibits logarithmic regret. The following theorem provides an asymptotic regret lower bound for (RM) for strategies in Π . Before stating this result, we re-define the notion of uniformly efficient algorithm: an algorithm $\pi \in \Pi$ is uniformly efficient if for any μ , $\sum_{t=1}^T \mathbb{E} \{p_k^{t,\pi}\} = o(T^\alpha)$, $\forall \alpha > 0$ and $k \neq k^*$, where we recall that $p_k^{t,\pi}$ denotes the power allocated to frequency ω_k under π .

Theorem 1 For any uniformly efficient algorithm $\pi \in \Pi$, the cumulative expected power allocated at some frequency ω_k satisfies:

$$\liminf_{T \rightarrow \infty} \frac{\sum_{t=1}^T \mathbb{E} \{p_k^t\}}{\log T} \geq \frac{\sigma^2}{|\mu_{k^*} - \mu_k|^2}, \quad \text{for all } k \neq k^*.$$

In particular, the regret under π satisfies:

$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}^\pi(\mu, T)}{\log T} \geq \sum_{k \neq k^*} \sigma^2 \frac{|\mu_{k^*}|^2 - |\mu_k|^2}{|\mu_{k^*} - \mu_k|^2}.$$

The proof of Theorem 1 can be found in the Appendix.

The above theorem implies that to get an asymptotically optimal algorithm for (RM), one may restrict to strategies in Π_{SF} , i.e., selecting only one frequency at each experiment. It is important to note that the above regret lower bound is valid when T is very large only. When T is small, it might be much wiser to spread the power on different frequencies, and exploit this information to minimize the regret. This hand-waving argument motivates the development of our algorithm, introduced in the following section.

IV. THE WEIGHTED THOMPSON SAMPLING (WTS) ALGORITHM

In this section, we devise WTS, an algorithm inspired by Thompson Sampling, that leverages the ability to gather information on *all* frequencies in a single experiment. WTS maintains a posterior distribution ρ^t of the optimal frequency, and allocates in each experiment its power to frequencies accordingly. Next, we explain in detail how WTS works.

A. Posterior Update

The posterior distribution ρ^t is computed by means of Monte Carlo methods, and by keeping track of the posterior distribution of the mean μ of $(X_k^t)_k$, given the gathered information in the $(t-1)$ first experiments. More precisely, we initially assume that for any k , μ_k has a Gaussian prior $\mu_k \sim \mathcal{N}_C(0, \lambda^2 \mathbf{I})$. Then after gathering the information $X_k^1, \dots, X_k^{t-1}$, it is well known [22] that the posterior distribution of μ_k is:

$$\mu_k | (X_k^\ell)_{\ell=1}^{t-1} \sim \mathcal{N}_C(\mathbf{m}_k^t, v_k^t \mathbf{I}), \quad (4a)$$

$$\mathbf{m}_k^t = \frac{\lambda^2 \sum_{\ell=1}^{t-1} p_k^\ell X_k^\ell}{\sigma^2 + \lambda^2 \sum_{\ell=1}^{t-1} p_k^\ell}, \quad (4b)$$

$$v_k^t = \frac{\lambda^2}{1 + \lambda^2 / \sigma^2 \sum_{\ell=1}^{t-1} p_k^\ell}. \quad (4c)$$

Then to compute ρ^t , we generate M samples of the μ_k 's according to their respective posterior distributions, and set ρ_k^t to be proportional to the number of samples where $|\mu_k|$ is the greatest.

Fig. 2. Frequency response of the unknown system G in Section V, together with the power distribution for the input signal at $T = 1000$.

B. Input Design

WTS chooses the input signal such that $|U^t(\omega_k)|^2 = p_k^t = \rho_k^t$, $k = 1, \dots, K$, instead of allocating all the power at one frequency, drawn from ρ^t , as TS would.

The pseudo-code of WTS is presented in Algorithm 2.

Algorithm 2 Weighted Thompson Sampling

- 1: Input: M, λ, σ
- 2: $\rho_k^1 = 1/K$, for all $k = 1, \dots, K$ (prior distribution)
- 3: **for** $t = 1$ to T **do**
- 4: Input design: select the power profile $p^t = \rho^t$
- 5: Perform the experiment and obtain $(X_k^t)_k$
- 6: Update the posterior of μ_k for all k :
- 7: $\mathbf{m}_k^{t+1} = \frac{\lambda^2 \sum_{\ell=1}^t p_k^\ell X_k^\ell}{\sigma^2 + \lambda^2 \sum_{\ell=1}^t p_k^\ell}$
- 8: $v_k^{t+1} = \lambda^2 / (1 + \lambda^2 / \sigma^2 \sum_{\ell=1}^t p_k^\ell)$
- 9: Update the posterior ρ^{t+1} :
- 10: Draw $\mathbf{s}^l \sim \mathcal{N}_C(\mathbf{m}_k^{t+1}, v_k^{t+1} \mathbf{I})$, $l = 1, \dots, M$
- 11: $\rho_k^{t+1} = \frac{1}{M} \sum_{l=1}^M \mathbf{1}_{\arg \max_i \{|\mathbf{s}_i^l|\} = k}$
- 12: $\hat{\beta}^t := \max_{k \in \{1, \dots, K\}} (\sum_{\ell=1}^t p_k^\ell X_k^\ell) / \sum_{\ell=1}^t p_k^\ell$
- 13: **end for**

Note that in Algorithm 2, the variance of X_k^t becomes undefined as soon as $p_k^t = 0$. Infinite variance measurements are considered as missing data.

V. SIMULATION EXAMPLE

Consider a 10-th order LTI system G whose unknown frequency response, depicted in Figure 2, has several local maxima. The new algorithm performance is evaluated by choosing $\sigma = 0.1$ and by searching among $K = 100$ different frequencies, with the unknown optimal arm (frequency) corresponding to $k^* = 39$ ($\omega_{39} = 1.225$ [rad/s]). Additionally, $M = 1000$ Monte Carlo simulations are used to *accurately* estimate ρ^t .

At round $T = 1000$, the power allocated by the algorithm at each frequency is presented in Figure 2, showing that WTS is indeed able to allocate most of the power of the input signal around the peak frequency.

Figure 3 compares the performance between WTS and the optimal TS algorithm in terms of the regret (3), where Figure 3(a) shows how well WTS performs for large T . Figure 3(b) depicts the same comparison for smaller values of T , where the median as well as 3σ -confidence intervals for

Fig. 3. Regret performance comparison between the new algorithm and Thompson Sampling, when using $\sigma = 0.1$, and the frequency response for G is the one depicted in Figure 2. Figure (a) shows the performance comparison when T growing large, while (b) does it when smaller values of T are considered. In (b), light yellow and light blue areas symbolize the confidence bound defined by 3 times the standard deviation of the regret for TS and WTS, respectively.

the cumulative regrets have been plotted based on 500 Monte Carlo realizations. Although both algorithms have similar regrets at the beginning, WTS rapidly balances exploration and exploitation showing a better performance than TS, as T grows. TS is indeed worst than WTS in terms of exploration, since it only explores one frequency at each experiment. Also, the regret of WTS exhibits less variability than that of TS, implying that the performance of WTS is more robust to sample variability.

Finally, WTS is analyzed in terms of $\mathcal{H}_\infty$ -norm estimation mean-squared error (MSE) by comparing it with the iterative method studied in [7], known as Power Iterations (PI). As depicted in Figure 4, the convergence of PI is faster than WTS. However, due to the inconsistency of PI (in the presence of noise), as shown in [7], its regret grows linearly with T . Our new algorithm WTS achieves a better performance in terms of accuracy for large T , outperforming PI in both MSE and regret.

VI. CONCLUSIONS

In this paper it has been explored how the $\mathcal{H}_\infty$ -norm estimation problem can be addressed from a stochastic multi-armed bandit perspective, where a lower bound has been derived on achievable performance of iterative $\mathcal{H}_\infty$ -norm estimators, and a new algorithm, called Weighted Thompson Sampling (WTS) to find the peak frequency has been intro-

Fig. 4. Mean-squared error comparison.

duced. This procedure provides a collection of data sets to accurately estimate the largest gain. Simulation results show the excellent performance of the new method when compared to state-of-the-art algorithms.

When the goal is to find the peak frequency, two different methods that adaptively design the spectrum of the input signal have been addressed: one of them allocating all the available power at one single frequency, while the other strategy spreads the power among different frequencies. Even though both strategies achieve the same asymptotic performance rate, WTS (as well as other power spreading strategies) may achieve a much more better performance for a finite number of experiments.

APPENDIX PROOF OF THEOREM 1

A. Preliminaries

We first state and prove two lemmas that capture important properties of the KL divergence, introducing the following notations: $\mathbb{P} := \mathbb{P}_\mu$ (resp. $\mathbb{P}' := \mathbb{P}_{\mu'}$) is a probability measure for $(X_k^t)_k$ parametrized by the set of means μ (resp. μ'), $\mathbb{E}\{\cdot\}$ (resp. $\mathbb{E}'\{\cdot\}$) is the expectation under μ (resp. μ'), and $\nu := \nu_\mu$ (resp. $\nu' := \nu_{\mu'}$) is a pdf for $(X_k^t)_{t=1}^T$ under μ (resp. μ') whose marginal Gaussian pdf's are $\nu_{\mu_k}^t$ (resp. $\nu_{\mu'_k}^t$).

Lemma 1 Let $\mu := (\mu_1, \dots, \mu_K) \in \mathbb{C}^K$ and $\mu' := (\mu'_1, \dots, \mu'_K) \in \mathbb{C}^K$ be two sets of means for the outcome $(X_k^t)_k$. Then, the KL divergence between the parametrized densities ν and ν' satisfies

$$\mathbb{D}\{\nu || \nu'\} = \frac{1}{\sigma^2} \sum_{t=1}^T \sum_{k=1}^K \mathbb{E}\{p_k^t\} |\mu_k - \mu'_k|^2. \quad (5)$$

Proof: The log-likelihood function under μ is given by

$$\begin{aligned} \ell(\mu) &:= \log \nu(X^1, \dots, X^T) = \sum_{t=1}^T \sum_{k=1}^K \log \nu_{\mu_k}(X_k^t) \\ &= c + \sum_{t=1}^T \sum_{k=1}^K \log p_k^t - \frac{1}{\sigma^2} \sum_{t=1}^T \sum_{k=1}^K p_k^t |X_k^t - \mu_k|^2. \end{aligned}$$

where c is a constant independent of μ . The expression for $\ell(\mu')$ can be obtained analogously. Then,

$$\mathbb{D}\{\nu || \nu'\} = \mathbb{E}\{\ell(\mu) - \ell(\mu')\} = \sum_{t=1}^T \sum_{k=1}^K \frac{\mathbb{E}\{p_k^t\} |\mu'_k - \mu_k|^2}{\sigma^2}.$$

$$= \frac{T}{T^{\alpha-1}} \quad \text{for all } \alpha > 0. \quad (9)$$

Lemma 2 Under the set-up of Lemma 1, and for any event B , the KL divergence between ν and ν' satisfies

$$\mathbb{D}\{\nu \parallel \nu'\} \geq d(\mathbb{P}\{B\}, \mathbb{P}'\{B\}), \quad (6)$$

where $d(q, s) := q \log(q/s) + (1-q) \log((1-q)/(1-s))$ is the binary relative entropy.

Proof: By Jensen's inequality [23],

$$\mathbb{P}'\{B\} = \int_B \left(\frac{d\mathbb{P}'}{d\mathbb{P}} \right) d\mathbb{P} \geq e^{\mathbb{E}\{\ell(\mu') - \ell(\mu)|B\}} \mathbb{P}\{B\},$$

and then,

$$\mathbb{E}\{\ell(\mu) - \ell(\mu')|B\} \geq \log \frac{\mathbb{P}\{B\}}{\mathbb{P}'\{B\}}.$$

It now follows that

$$\begin{aligned} \mathbb{D}\{\nu \parallel \nu'\} &= \mathbb{E}\{\ell(\mu) - \ell(\mu')\} \\ &= \mathbb{E}\{\ell(\mu) - \ell(\mu')|B\} \mathbb{P}\{B\} + \mathbb{E}\{\ell(\mu) - \ell(\mu')|B^c\} \mathbb{P}\{B^c\} \\ &\geq \mathbb{P}\{B\} \log \frac{\mathbb{P}\{B\}}{\mathbb{P}'\{B\}} + \mathbb{P}\{B^c\} \log \frac{\mathbb{P}\{B^c\}}{\mathbb{P}'\{B^c\}} \\ &= \mathbb{P}\{B\} \log \frac{\mathbb{P}\{B\}}{\mathbb{P}'\{B\}} + (1 - \mathbb{P}\{B\}) \log \frac{1 - \mathbb{P}\{B\}}{1 - \mathbb{P}'\{B\}}. \end{aligned}$$

B. Proof of Theorem 1

We rely on the proof in [21], originally formulated for the classical stochastic MAB problem. Without loss of generality, let $\mu = (\mu_1, \mu_2, \dots, \mu_K)$ denote the *true* mean rewards, such that $|\mu_1|^2 > |\mu_2|^2 \geq \dots \geq |\mu_K|^2$. Let a denote a sub-optimal arm (i.e., $a \neq 1$). Also, consider a *confusion parameter* $\mu' = (\mu'_1, \mu'_2, \dots, \mu'_K)$ that satisfies $\mu'_k = \mu_k$, for all $k \neq a$, and $|\mu'_a|^2 > |\mu_1|^2$, for some $a \neq 1$ (i.e., according to μ , $a \neq 1$ is originally a suboptimal arm). By means of Lemma 1, it holds that

$$\mathbb{D}\{\nu \parallel \nu'\} = \frac{|\mu'_a - \mu_a|^2}{\sigma^2} \sum_{t=1}^T \mathbb{E}\{p_a^t\}. \quad (7)$$

On the other hand, using Lemma 2 with $B = B_T := \{\sum_{t=1}^T p_1^t < T - T^\gamma\}$, $0 < \gamma < 1$, we have that

$$\begin{aligned} \mathbb{P}\{B_T\} &= \mathbb{P}\left\{T - \sum_{t=1}^T p_1^t > T^\gamma\right\} \stackrel{(a)}{=} \mathbb{P}\left\{\sum_{k=2}^K \sum_{t=1}^T p_k^t > T^\gamma\right\} \\ &\stackrel{(b)}{\leq} \frac{1}{T^\gamma} \sum_{k=2}^K \sum_{t=1}^T \mathbb{E}\{p_k^t\} \\ &\stackrel{(c)}{=} o(T^{\alpha-\gamma}), \quad \forall \alpha > 0, \quad (8) \end{aligned}$$

where (a) follows from the fact that $\sum_{t=1}^T \sum_{k=1}^K p_k^t = T$, (b) follows from Markov's inequality, and (c) is a consequence of the fact that the algorithm is uniformly efficient. Similarly, under μ' (where arm 1 is suboptimal), it follows that

$$\begin{aligned} \mathbb{P}'\{B_T^c\} &= \mathbb{P}'\left\{\sum_{t=1}^T p_k^t > T - T^\gamma\right\} \\ &\leq \frac{1}{T - T^\gamma} \sum_{t=1}^T \mathbb{E}'\{p_k^t\} \end{aligned}$$

Combining Lemma 2 with (8) and (9), and taking $\alpha < \gamma$ sufficiently small, it follows that

$$\begin{aligned} d(\mathbb{P}\{B_T\}, \mathbb{P}'\{B_T\}) &\geq \mathbb{P}\{B\} \log \frac{\mathbb{P}\{B_T\}}{\mathbb{P}'\{B_T\}} + (1 - \mathbb{P}\{B_T\}) \log \frac{1 - \mathbb{P}\{B_T\}}{1 - \mathbb{P}'\{B_T\}} \\ &= o(1) + (1 - \alpha) \log(T). \end{aligned}$$

Finally, using the latter equation together with (7),

$$\liminf_{T \rightarrow \infty} \frac{\sum_{t=1}^T \mathbb{E}\{p_a^t\}}{\log T} \geq \frac{\sigma^2(1 - \alpha)}{|\mu'_a - \mu_a|^2}. \quad (10)$$

The proof finishes by letting $\alpha \rightarrow 0$ and $\mu'_a \rightarrow \mu_1$ in (10).

REFERENCES

- [1] B. A. Francis, *A course in $\mathcal{H}_\infty$ control theory*. Springer, 1987.
- [2] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall, 1996.
- [3] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control System Design*. Prentice Hall, 2000.
- [4] 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.
- [5] M. N. Belur and C. Praagman, "An efficient algorithm for computing the $\mathcal{H}_\infty$ norm," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1656–1660, July 2011.
- [6] M. Barenthin, "Complexity issues, validation and input design for control in system identification," Ph.D. dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2008.
- [7] C. R. Rojas, T. Oomen, H. Hjalmarsson, and B. Wahlberg, "Analyzing iterations in identification with application to nonparametric $\mathcal{H}_\infty$ -norm estimation," *Automatica*, vol. 48, no. 11, pp. 2776 – 2790, 2012.
- [8] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *21st Annual Conference on Learning Theory (COLT)*, 2008, pp. 355–366.
- [9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, pp. 4–22, 1985.
- [10] 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.
- [11] K. J. Astrom and B. Wittenmark, *Adaptive Control*, 2nd ed. Dover Publications, 2008.
- [12] 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.
- [13] M. C. Campi, A. Lecchini, and S. M. Savaresi, "Virtual reference feedback tuning: a direct method for the design of feedback controllers," *Automatica*, vol. 38, no. 8, pp. 1337–1346, 2002.
- [14] M. C. Campi and P. R. Kumar, "Adaptive linear quadratic Gaussian control: The cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.
- [15] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *24th Annual Conference on Learning Theory (COLT)*, 2011, pp. 1–26.
- [16] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.
- [17] J. Agüero, J. Yuz, G. Goodwin, and R. Delgado, "On the equivalence of time and frequency domain maximum likelihood estimation," *Automatica*, vol. 46, no. 2, pp. 260–270, February 2010.
- [18] J.-B. Grill, M. Valko, and R. Munos, "Black-box optimization of noisy functions with unknown smoothness," in *Neural Information Processing Systems*, 2015.
- [19] O.-A. Maillard, R. Munos, and G. Stoltz, "A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences," in *24th Annual Conference on Learning Theory (COLT)*, 2011, pp. 497–514.
- [20] 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)*. Springer, 2012, pp. 199–213.
- [21] E. Kaufmann, "Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources," Ph.D. dissertation, Paris Institute of Technology, Paris, France, 2014.
- [22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.
- [23] Z. Cvetkovski, *Inequalities*. Springer, 2012.