A Stochastic-Optimization-Based Adaptive-Sampling Scheme for Data-Driven Stability Analysis of Switched Linear Systems

Alexis Vuille, Guillaume O. Berger and Raphaël M. Jungers

Abstract— **We introduce a novel approach based on stochastic optimization to find the optimal sampling distribution for the data-driven stability analysis of switched linear systems. Our goal is to address limitations of existing approaches, in particular, the fact that these methods suffer from ill-conditioning of the optimal Lyapunov function, which was shown in recent work to be a direct consequence of the way the data is collected by sampling uniformly the state space. In this work, we formalize the notion of optimal sampling distribution, using the perspective of stochastic optimization. This allows us to leverage tools from stochastic optimization to estimate the optimal sampling distribution, and then use it to collect samples for data-driven stability analysis of the system. We show in numerical experiments (on challenging systems of dimension up to five) that the overall procedure is highly favorable in terms of data usage compared to existing methods using fixed sampling distributions. Finally, we introduce a heuristic that combines data points from previous samples, and show empirically that this allows an additional substantial reduction in the number of samples required to achieve the same stability guarantees.**

Index Terms— **data-driven methods, stochastic optimization, statistical learning, stability analysis, switched linear systems**

I. Introduction

In recent years, data-driven methods have gained a lot of attention for the study of cyber-physical systems because of the increasing number of applications in which no model of the system is available. At the same time, data has become more and more accessible due to the outbreak of cheap, accurate sensors, user feedback and open-source databases. Finally, statistical learning, the mathematical field of learning from data, has known many great advances in recent years both in theory and practice [1]. All this together opened the door to a new era in control theory where control and system analysis is made from data harvested from observation of the system and comes with formal guarantees of correctness; we refer the reader to [2, Chapter 11] for an introduction and further references on data-driven verification and control of cyber-physical systems.

In this paper, we consider a prototypical class of cyber-physical systems, known as *switched linear systems* [3]. These systems consist of several linear modes among which the system can switch over time. They appear naturally in a wide range of applications [4], or as approximations of more complex systems. A crucial question in the study of switched linear systems is their stability analysis [3], which turns out to be a very challenging problem in general even when the model of the system is available [5]. For instance, approximating the rate of convergence of the system (known as the Joint Spectral Radius, or JSR) is known to be NP-hard [5]. Nevertheless, several approximation techniques have been proposed in the last decades leading to good results in the model-based setting [5], [6].

However, a model of the system is not always available. Therefore, several approaches were proposed in recent years for the data-driven analysis of the JSR of switched linear systems [7]–[11]. These approaches use advanced tools from statistical learning, such as scenario optimization [12], to learn a Lyapunov function for the system to derive bounds on the JSR and provide probabilistic guarantees on the correctness of the bound.

However, the classical approaches [7]–[9] strongly suffer from a bad choice of the distribution used to sample the data. This was shown in [11]—where additionally an intuitive approach for finding a better sampling distribution was proposed, demonstrating huge gain in data usage. In this paper, we leverage these promising preliminary results in two complementary directions. First, we formalize the notion of *best* sampling distribution, using the perspective of stochastic optimization. Namely, we search for the sampling distribution that gives *in average* the best conditioning of the learned Lyapunov function (shown in [11] to be one of the main sources of conservatism). Second, we use stochastic optimization techniques (namely stochastic gradient descent) to learn this best sampling distribution form data. We show with numerical experiments that the overall procedure allows us to certify stability using less samples than other fixed-distribution data-driven methods [8]. Finally, we introduce a heuristic that combines data points from previous samples, empirically demonstrating a substantial reduction in the number of samples required to achieve the same stability guarantees, and improving upon the heuristic method of [11] in all experiments.

Related Works

Data-driven stability analysis of switched linear systems is studied in [7]–[10]. These approaches suffer from the curse of dimensionality, strongly amplified by the fact that the data points are sampled uniformly, as demonstrated in our recent work [11]. In [11], we introduced the method of adaptive sampling to alleviate the dependency on the sampling, showing a reduction of sample complexity of several orders of magnitude. Yet, the adaptive sampling methodology in [11] is strongly heuristic. It consists in two

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 864017 – L2C.

GB is a FNRS Postdoctoral Researcher, RJ is a FNRS honorary Research Associate. We are from UCLouvain, Louvain-la-Neuve, Belgium (e-mail: firstname.lastname@uclouvain.be).

phases. 1) Sample points uniformly, and learn a Lyapunov function from these samples. 2) Use this Lyapunov function to obtain a new sampling distribution; sample from this distribution and derive the stability guarantees. Although simple, this strategy showed great potential. However, key questions such as "what is actually a good sampling distribution?", or "can we do better by doing more than two steps (each time using the sampling distribution obtained at the previous step)?". This work aims to address these questions by (i) formalizing the notion of optimal distribution for data-driven stability analysis of switched linear systems, from the lens of stochastic optimization, and (ii) providing numerical techniques to compute the optimal distribution. We demonstrate in numerical examples the strong benefits of the proposed solution over the previous approaches.

Notation: $\mathbb{R}^{n\times n}_{\succ 0}$ denotes the set of positive definite $n\times n$ matrices. Given $P \in \mathbb{R}^{n\times n}_{\succ 0}$, we denote the quadratic norm of P as $\|x\|_P = \sqrt{x^T P x}$. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \ldots, N\}$. We denote the unit sphere in $\mathbb{R}^n$ by $\mathbb{S}^{n-1}$. Given two datasets of size N, $\mathcal{X} := \{x_i\}_{i=1}^N$ and $\mathcal{X}' := \{x_i'\}_{i=1}^N$, we denote their element-wise pairing as $\mathcal{X}\|\mathcal{X}' := \{(x_i, x_i')\}_{i=1}^N$.

II. Problem Statement

We consider a discrete-time *switched linear system* with m modes:

$$x(t+1) \in \{Ax(t) : A \in \mathcal{A}\}, \tag{1}$$

wherein $\mathcal{A} = \{A_1, \ldots, A_m\} \subseteq \mathbb{R}^{n\times n}$ is a set of m matrices in $\mathbb{R}^{n\times n}$. Since $\mathcal{A}$ characterizes the system in (1), in the following, we will often refer to the system simply by $\mathcal{A}$. A *trajectory* of $\mathcal{A}$ is a function $x : \mathbb{N} \to \mathbb{R}^n$ such that for all $t \in \mathbb{N}$, the condition in (1) holds.

We are interested in the stability of system (1). We remind that (1) is *asymptotically stable* if all trajectories of $\mathcal{A}$ converge to the origin. The rate of exponential convergence is called the *Joint Spectral Radius* (JSR) of $\mathcal{A}$.

Definition 1: The *joint spectral radius* of $\mathcal{A}$, denoted by $\rho(\mathcal{A})$, is the infimum of all $r \geq 0$ for which there exists $C \geq 1$ such that every trajectory x of $\mathcal{A}$ satisfies that for all $t \in \mathbb{N}$, $\|x(t)\| \leq Cr^t\|x(0)\|$.

A. Quadratic Approximation of the JSR

The JSR is notoriously difficult to approximate, even when the matrices in $\mathcal{A}$ are known [5]. One way to obtain an upper bound on the JSR is by finding a quadratic *Lyapunov* function for the system. The contraction rate associated with this function then provides an upper bound.

Definition 2: Given a positive definite matrix $P \in \mathbb{R}^{n\times n}_{\succ 0}$, we define the *contraction rate* of $\mathcal{A}$ with respect to P by

$$\rho(\mathcal{A}, P) = \max_{x\in\mathbb{R}^n\setminus\{0\},\, A\in\mathcal{A}} \frac{\|Ax\|_P}{\|x\|_P}.$$

The contraction rate is an upper bound on the JSR (see, e.g., [5, Proposition 2.8]):

Theorem 1: For any $P \in \mathbb{R}^{n\times n}_{\succ 0}$, $\rho(\mathcal{A}) \leq \rho(\mathcal{A}, P)$.

Hence, quadratic Lyapunov functions allow us to bound the JSR. However, this approach, as other *model-based* approaches for approximating the JSR, require the knowledge of $\mathcal{A}$. This is a limitation in several applications, thereby justifying the use of data-driven methods.

B. Data-Driven Analysis and Random Data Collection

In the setting considered in this paper (first introduced in [7]), we collect data by setting the system to an initial state x and observing the state y after one time step. Repeating this process N times yields a data set comprising N one-step trajectories $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}^n$, where $y_i = Ax_i$ for some $A \in \mathcal{A}$, for each $i \in [N]$. We assume that the mode selection is a stochastic process, wherein each mode in $\mathcal{A}$ has a nonzero probability of being applied independently at each sample:[1]

Assumption 1: There exists $\alpha \in (0, 1]$ such that for all $A \in \mathcal{A}$ and $i \in [N]$,

$$\mathbb{P}\left[\, y_i = Ax_i \mid x_i,\ \{(x_j, y_j)\}_{j\neq i} \,\right] \geq \alpha.$$

Regarding the choice of the initial state of each sample (x_i, y_i), we assume that the initial state x_i can be chosen for each $i \in [N]$. In particular, we will choose x_i randomly according to some distribution that we can design. In particular, we will consider the standard Gaussian distribution (reminded below), possibly after applying a change of basis:[2]

Definition 3: A random variable X with value in $\mathbb{R}^n$ has *standard Gaussian distribution* if its probability density function (pdf) f satisfies for all $x \in \mathbb{R}^n$, $f(x) \propto e^{-\frac{1}{2}\|x\|^2}$.

Remark 1: As we will see in Section III, the fact that the sampling distribution can be chosen is key in our framework since our approach to improve data-scalability is to learn an optimal sampling distribution. This assumption is realistic in a wide range of applications, namely when one has access to the system has a (stochastic) input–output black-box.

C. Fixed Sampling Distribution

The data-driven approach in [7] (refined in [8]) provides probabilistic upper bounds on the JSR from data collected from a fixed sampling distribution. See also [9], [10] for similar data-driven approaches using a fixed sampling distribution. We remind here the main result of [8] because this will be useful for the rest of this paper.

The approach of [7], [8] works as follows. Given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ consisting of N one-step trajectories, we formulate the problem of finding a positive definite matrix P with the smallest *data-based contraction rate* defined by

$$\hat{\rho}(\mathcal{D}, P) = \max_{i\in[N]} \frac{\|y_i\|_P}{\|x_i\|_P}.$$

[1] Note that even though we consider the mode selection to be a stochastic process for data collection, the probabilistic guarantees we will obtain still hold for all possible trajectories of the system (even ones for which the mode selection is not a stochastic process). However, without assumption 1, that would be not the case as a mode could be never sampled in our dataset for all datasets of finite size. Similarly, while we consider in the rest of the paper different kinds of distributions for the initial states to collect the data and analyse the stability, the stability analysis and the confidence bounds remain valid for all possible trajectories and all possible initial states.

[2] Because of the scaling invariance, a data point (x_i, y_i) carries the same information as the data point $(\lambda x_i, \lambda y_i)$ for every $\lambda \neq 0$. This is why sampling with respect to the standard Gaussian distribution is equivalent in this problem to sampling with respect to the uniform distribution on the unit sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$.

Hence, we aim to solve the optimization problem

$$\min_{P \in \mathcal{P}} \hat{\rho}(\mathcal{D}, P). \tag{2}$$

where $\mathcal{P}$ is a closed subset of $\mathbb{R}^{n\times n}_{\succ 0}$ (we assume that $I \in \mathcal{P}$). Note that (2) can be solved efficiently, as it is a quasi-convex optimization problem [8]. The optimal cost of (2) is denoted by $\gamma_\star(\mathcal{D})$, and the optimal solution, if it exists, by $P_\star(\mathcal{D})$. We assume without loss of generality that if $P_\star(\mathcal{D})$ exists, then it is unique (this can be done by using a tie-breaking rule [8]). When $\mathcal{D}$ is clear from the context, we write $\gamma_\star$ and $P_\star$ instead of $\gamma_\star(\mathcal{D})$ and $P_\star(\mathcal{D})$.

Under some mild assumption (Assumption 2) on $\mathcal{A}$, one can guarantee an upper bound on the JSR of $\mathcal{A}$ with high confidence (Theorem 2).[3]

Definition 4: A matrix is said to be *Barabanov* if it is diagonalizable and all its eigenvalues have the same modulus.

Assumption 2: The matrices in $\mathcal{A}$ are not Barabanov.

Theorem 2 ([8]): Let d be the dimension of $\operatorname{span}(\mathcal{P})$, and $N \geq d$. Let $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^n$ be sampled i.i.d. following the standard Gaussian distribution. Let Assumptions 1 and 2 hold. Let $\beta \in (0, 1]$. Then, with probability $1 - \beta$ on the sampling of $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^N$, it holds that[4]

$$\rho(\mathcal{A}) \leq \rho(\mathcal{A}, P_\star) \leq \gamma_\star \cdot f(\beta, \kappa(P_\star), N, d, \alpha, n), \tag{3}$$

wherein

- $\kappa(P) = \sqrt{\frac{\det(P)}{\lambda_{\min}(P)^n}}$;
- $f(\beta, k, N, d, \alpha, n) = \frac{1}{\sqrt{1 - I^{-1}\left(\frac{k}{\alpha}\Phi^{-1}(\beta; d-1; N); \frac{n-1}{2}; \frac{1}{2}\right)}}$;
- $I^{-1}(y; a; b)$ is the inverse incomplete regularized beta function, i.e., it is the unique $x \in [0, 1]$ such that
$$I(x; a; b) := \frac{\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t}{\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t} = y;$$
- $\Phi^{-1}(\beta; \zeta; N)$ is the unique $\epsilon \in [0, 1]$ such that
$$\Phi(\epsilon; \zeta; N) := \sum_{i=0}^{\zeta} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} = \beta.$$
- α is defined in Assumption 1.

Remark 2: The *inflation factor* $f(\beta, \kappa(P_\star), N, d, \alpha, n)$ is the source of the additional conservatism of the data-driven approach, compared to the model-based approach. Unfortunately, this factor increases exponentially with the value of $\kappa(P_\star)$. Hence, it is crucial to make the value of $\kappa(P_\star)$ as small as possible. This is the purpose of the adaptive sampling approach, described next.

D. Adaptive Sampling Distribution

The two-step approach was proposed in [11] as an effective method to reduce the value of $\kappa(P_\star)$ through adaptive sampling. The key idea behind this approach is that a change of basis of the state variable x can change the value of $P_\star$ and thereby the value of $\kappa(P_\star)$. Hence, our approach aims to find the change of coordinates for which $\kappa(P_\star)$ is the smallest.

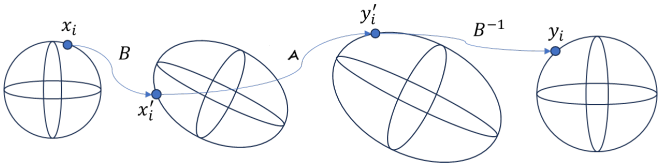

Fig. 1. Sampling with change of basis B.

More precisely, the change of basis and sampling distribution works as follows. Given an invertible matrix $B \in \mathbb{R}^{n\times n}$ and a data set $\mathcal{X} = \{x_i\}_{i=1}^N$ sampled i.i.d. from the standard Gaussian distribution, we define the data set in the basis B by $\mathcal{X}' := \{x_i' := Bx_i\}_{i=1}^N$. This transformed data set can then be fed to the system oracle, providing the data set $\mathcal{Y}' := \{y_i'\}_{i=1}^N$ Finally, we can apply the reverse change of basis to obtain the data set $\mathcal{Y} := \{y_i := B^{-1}y_i'\}_{i=1}^N$. This sampling process is illustrated in Figure 1. We denote the resulting datasets by $\mathcal{D} = \mathcal{X}\|\mathcal{Y}$ and $\mathcal{D}' = \mathcal{X}'\|\mathcal{Y}'$. The interest of the change of basis is that if B is chosen appropriately, then $\kappa(P_\star(\mathcal{D}))$ can be expected to be close to one. In fact, it is shown in [8, Proposition 8] that in the model-based setting (i.e., when $\mathcal{A}$ is known and $N \to \infty$), one can choose B so that $\kappa(P_\star(\mathcal{D})) = 1$ with probability one. However, in the data-driven context, $\mathcal{A}$ is unknown and we want to keep N small (thus finite). Therefore, [11] proposed a two-step approach, consisting in first guessing a change of basis B by using a initial dataset $\mathcal{D}_\circ$ (of size $N_\circ$), and then using this B to build the dataset $\mathcal{D}$ (of size N), and find $\gamma_\star(\mathcal{D})$ and $P_\star(\mathcal{D})$. This results in an approach that requires (empirically) much less data ($N_\circ + N$), compared to fixed-sampling distribution approaches, to provide stability guarantees.

Yet, despite promising empirical results, the construction of the change of basis B in [11] is essentially intuitive. Also, the notion of "good" change of basis is not clearly defined. We address these challenges in Section III below. Then, in Section IV, we build upon the results in Section III to provide a heuristic aimed at improving the data usage even further, as demonstrated empirically on numerical examples.

III. Adaptive Sampling Through the Lens of Stochastic Optimization

Building on the observations made in the previous section, we seek a change of basis B for which $\kappa(P_\star(\mathcal{D}))$ is expected to be small when $\mathcal{D}$ is built as explained in the previous section. This will lead to the definition of the optimal B as the solution of a stochastic optimization problem. Building on this, we will then propose an numerical method to find the optimal change of basis.

Given $B \in \mathbb{R}^{n\times n}$ invertible, we denote the distribution of the datasets $\mathcal{D} = \mathcal{X}\|\mathcal{Y}$ and $\mathcal{D}' = \mathcal{X}'\|\mathcal{Y}'$, built as explained in Section II, by $D^N(B)$ and $D'^N(B)$ respectively.

A. Stochastic Optimization Problem

We formalize the property of being the optimal change of basis B, which captures the fact that the *expected* value

[3] This guarantees the non-degeneracy property to apply the PAC bounds from Scenario Optimization. See for example [12].

[4] Note that the first inequality in (3) is always satisfied, while the second one is guaranteed to hold with probability at least $1 - \beta$.

of $\log(\kappa(P_\star(\mathcal{D})))$ will be minimized among all B in some subset $\mathcal{B}$ of invertible $n \times n$ matrices[5]:

$$\min_{B\in\mathcal{B}} \mathbb{E}_{\mathcal{D}\sim D^N(B)}[\log(\kappa(P_\star(\mathcal{D})))]. \tag{4}$$

We consider the logarithm as it leads in the end to a more numerically stable method to solve the problem. We now introduce a closely related problem, which will be used for optimization:

$$\min_{B\in\mathcal{B}} \mathbb{E}_{\mathcal{D}'\sim D'^N(B)}[\log(\kappa(B^\top P_\star(\mathcal{D})B))] \tag{5}$$

It turns out that solving (4) or (5) is equivalent:

Theorem 3: Problems (4) and (5) have the same set of optimal solutions.

Proof: Observe that

$$\begin{aligned} P_\star(\mathcal{X}\|\mathcal{Y}) &= \arg\min_P \max_{x_i\in\mathcal{X},y_i\in\mathcal{Y}} \frac{\|y_i\|_P}{\|x_i\|_P} \\ &= \arg\min_P \max_{x_i'\in\mathcal{X}',y_i'\in\mathcal{Y}'} \frac{\|B^{-1}y_i'\|_P}{\|B^{-1}x_i'\|_P} \\ &= B^\top \arg\min_{P'} \max_{x_i'\in\mathcal{X}',y_i'\in\mathcal{Y}'} \frac{\|y_i'\|_{P'}}{\|x_i'\|_{P'}} B \\ &= B^\top P_\star(\mathcal{X}'\|\mathcal{Y}')B, \end{aligned}$$

where the second equality come from the definition of the sets $\mathcal{X}, \mathcal{X}', \mathcal{Y}, \mathcal{Y}'$ and the third equality is obtained by doing a change of variable $P' := B^{-\top}PB^{-1}$. ■

B. Stochastic Gradient Descent

We solve (5) by employing a stochastic gradient algorithm. Since the distribution involved in the objective function of (5) depends on the decision variable B, we use a stochastic gradient algorithm accounting for *decision-dependent distribution*; see, e.g., [13].

Concretely, the method works as follows: given an estimate of the optimal sampling distribution B_k, we sample a dataset $\mathcal{D}_k'$ according to $D'^N(B_k)$ and use this dataset to compute the gradient direction as

$$g_k := \nabla_B \log(\kappa(B^\top P_\star(\mathcal{D}_k')B)).$$

We then update the sampling distribution as $B_{k+1}' = B_k - \eta_k g_k$, for some predefined step size $\eta_k > 0$. If needed, we project on B_{k+1}' on $\mathcal{B}$, giving B_{k+1}. As a first estimate of the optimal sampling distribution, we use $B_0 := I$, which corresponds to no change of basis.

Using the formula of $\kappa(P)$, we can derive an explicit formula for the gradient[6]:

$$\nabla_B \log(\kappa(B^\top PB)) = B^{-T} - \frac{nPBvv^\top}{\lambda_{\min}(B^\top PB)}, \tag{6}$$

where $v := v_{\min}(B^\top P_k B)^\top$ is the eigenvector associated to $\lambda_{\min}(B^\top P_k B)$, the minimum eigenvalue of $B^\top P_k B$. The proof can be found in the Appendix in the extended version on arXiv [14].

After T steps of the stochastic gradient descent, we have our final estimate B_T of the optimal sampling distribution. We use it to sample a dataset $\mathcal{D}_T$ according to $D^N(B_T)$, and we compute the probabilistic upper bound as $\gamma_\star(\mathcal{D}_T) \cdot f(\beta, \kappa(P_\star(\mathcal{D}_T)), N, d, \alpha, n)$. This gives Algorithm 1.

Algorithm 1 Stochastic Optimization Upper Bound

```
Input:
        - A black-box switched linear system $\mathcal{A}$.
        - $T$: number of iterations.
        - $(\eta_k)_{k=0}^{T-1}$: step sizes.
        - $N_{batch}$: batch size.
Set $B_0 := I$
for $k = 0$ to $T - 1$ do
    Sample $\mathcal{D}_k' := \{(x_i', y_i')\}_{i=1}^{N_{batch}} \sim D'^{N_{batch}}(B_k)$.
    Using $\mathcal{D}_k'$, solve (2) to compute $P_k := P_\star(\mathcal{D}_k')$
    Compute the gradient: $g_k := \nabla_B \log(\kappa(B^\top P_k B))$
    Update the estimate: $B_{k+1}' := B_k - \eta_k g_k$
    Project $B_{k+1}'$ on $\mathcal{B}$ to obtain $B_{k+1}$
end for
Sample $\mathcal{D}_T := \{(x_i, y_i)\}_{i=1}^{N_{batch}} \sim D^{N_{batch}}(B_T)$.
Solve (2) to get $\gamma_\star(\mathcal{D}_T)$ and $P_\star(\mathcal{D}_T)$.
Output: $\gamma_\star(\mathcal{D}_T) \cdot f(\beta, \kappa(P_\star(\mathcal{D}_T)), N_{batch}, d, \alpha, n)$
```

It turns out that if $\kappa(P_\star(\mathcal{D}_T))$ is expected to be close to one (which can be estimated from the objective value of Algorithm 1), then it can be beneficial to use $\mathcal{P} = \{I\}$ (i.e., fix $P = I$), because the small increase of $\gamma_\star$ (since $\mathcal{P}$ is more restricted) will be compensated by the fact that $f(\beta, k, N, d, \alpha, n)$ is smaller when $d = 1$:

Proposition 1: Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. It holds that $\gamma_\star \le \hat{\rho}(\mathcal{D}, I) \le \gamma_\star \kappa(P_\star)$.

Proof: The first inequality is direct from (2). To prove the second inequality, first assume without loss of generality that $\lambda_{\min}(P_\star) = 1$. Denote $k = \kappa(P_\star)$. Then, observe that $\det(P_\star) \le k^2$, which implies that $I \preceq P_\star \preceq k^2 I$. Hence, we get that for each $i \in [N]$,

$$\|y_i\|^2 \le y_i^\top P_\star y_i \le \gamma_\star^2 x_i^\top P_\star x_i \le \gamma_\star^2 k^2 \|x_i\|.$$

This shows that $\hat{\rho}(\mathcal{D}, I) \le k\gamma_\star$. ■

Corollary 1: Let $B \in \mathbb{R}^{n\times n}$ be invertible. For any $\epsilon > 1$, it holds with probability $1 - \delta/\log(\epsilon)$ on $\mathcal{D} \sim D^N(B)$ that $\hat{\rho}(\mathcal{D}, I) \le \epsilon\gamma_\star(\mathcal{D})$, where $\mathbb{E}_{\mathcal{D}\sim D^N(B)}[\log(\kappa(P_\star(\mathcal{D})))] = \delta$.

Proof: Apply Markov's inequality on $\log(\frac{\hat{\rho}(\mathcal{D},I)}{\gamma_\star(\mathcal{D})})$ which is always nonnegative by Proposition 1, and its expectation is smaller than or equal to δ by Proposition 1. ■

C. Open Questions on Convergence and Optimality

1) Uniqueness of the solution: Experimentally, (4) and (5) seem to admit only one local optimum (up to a nonzero scaling) which is the global optimum. Indeed, one can see on the Figure 2 that the data-driven JSR converges the toward model-based one. However, a formal proof of this claim would be valuable.

[5] In our experiments, we consider $\mathcal{B}$ to be a compact subset of positive symmetric definite matrice of the from $\mathcal{B} := \{B \in \mathbb{R}^{n\times n}_{\succ 0} : I \preceq B = B^T \preceq \alpha I\}$, for some $\alpha > 1$, which guarantees the existence of a solution.

[6] For some values of $P^\top BP$ (with Lebesgue zero measure) it's is only a subdiferential; see Appendix in the extended version on arXiv [14].

2) Convergence: Furthermore, experimental observations indicate that stochastic gradient descent seem to converge, provided the step sizes satisfy the conditions $\sum_{k=0}^{\infty} \eta_k = +\infty$ and $\lim_{k \to +\infty} \eta_k = 0$.

Establishing a theorem that guarantees convergence under specific conditions would be highly desirable. The main challenge for this lies in the fact that the probability distribution involved in the expectation depends on the decision variable. There exist some works that obtain such guarantees but they require the objective to be strongly convex [13], which is not our case. We will address these questions in the future.

IV. Reusing Previous Samples—a Heuristic Approach

The stochastic gradient descent algorithm has a key practical limitation in that each iteration uses only the new samples without exploiting the previous ones. This can result in a prohibitively high overall number of samples. This was already observed in previous work, leading to extended notions of the SGD, such as *multi-pass* SGD [15].

In this section, we propose a modification of Algorithm 1 that reuses previous samples, and is ad-hoc to our problem. Although not theoretically grounded, we show experimentally that this modified algorithm outperforms Algorithm 1 (and multi-pass SGD[7]) in terms of number of data needed to provide stability guarantees.[8]

The two key differences of the modified algorithm are that (i) all previous and current samples are used in $\mathcal{D}_k$ to compute $\gamma_\star(\mathcal{D}_k)$ and $P_\star(\mathcal{D}_k)$, and (ii) instead of moving in the direction of the gradient of the objective function with respect to B, we move in the direction of $P_\star(\mathcal{D}_k)^{-1/2}$, which corresponds to the minimizer of $\log(\kappa(B^\top P_\star(\mathcal{D}_k)B)$. This approach was obtained as a heuristic, testing different options for choosing the update direction.

More precisely, our approach is as follows. Starting with $B_0 = I$, we sample an initial dataset $\mathcal{D}_0 := \{(x'_i, y'_i)\}_{i=1}^{N_0}$ according to $D'^{N_0}(B_0)$. Then, at each step k, given $\mathcal{D}_k$ and B_k, we update B in the direction of $P_\star(\mathcal{D}_k)^{-1/2}$:

$$B_{k+1} := (1 - \eta_k)B_k + \eta_k P_\star(\mathcal{D}_k)^{-1/2}.$$

Next, we draw a new sample $\{(x'_k, y'_k)\}$ from $D'(B_{k+1})$ and augment the dataset: $\mathcal{D}_{k+1} := \mathcal{D}_k \cup \{(x'_k, y'_k)\}$. This process continues until a predefined convergence criterion is satisfied. In practice, we consider the algorithm converged when the average difference between successive iterations falls below a predefined threshold $\epsilon > 0$, or when a maximum number of iterations is reached. This is implemented in Algorithm 2.

V. Experimental Results

We demonstrate the effectiveness of our methods (Algorithms 1 and 2) on a synthetic example and on a consensus problem. We also compare them with 1) the approach without adaptive sampling from [8], and 2) the state-of-the-art resampling technique from [11].

Algorithm 2 Heuristic Optimization Upper Bound

```
Input:
        - A black-box switched linear system $\mathcal{A}$.
        - $N_0$: initial sample size.
        - $N$: total sample budget.
        - $T$: number of iterations.
        - $(\eta_k)_{k=0}^{T-1} \subseteq [0,1]$: step sizes.
        - $\epsilon > 0$, $K \in \mathbb{N}$: convergence criteria parameters.
Initialize $B_0 := I$.
Sample $\mathcal{D}'_0 := \{(x'_i, y'_i)\}_{i=1}^{N_0} \sim D'^{N_0}(B_0)$.
for $k = 0$ to $T - 1$ do
    Solve (2) using $\mathcal{D}'_k$ to compute $P_k := P_\star(\mathcal{D}'_k)$.
    Compute the direction as: $g_k := B_k - P_k^{-1/2}$.
    Update: $B_{k+1} := B_k - \eta_k g_k$.
    if $\sum_{j=k-K}^{k} \|B_{j+1} - B_j\| \leq \epsilon$, terminate loop.
    Sample $\{(x'_k, y'_k)\} \sim D'^1(B_{k+1})$
    Augment the dataset: $\mathcal{D}_{k+1} = \mathcal{D}_k \cup \{(x'_k, y'_k)\}$.
end for
Sample $\mathcal{D}_T := \{(x_i, y_i)\}_{i=1}^{N-|\mathcal{D}_T|} \sim D^{N-|\mathcal{D}_T|}(B_T)$.
Solve (2) to get $\gamma_\star(\mathcal{D}_T)$ and $P_\star(\mathcal{D}_T)$.
Output: $\gamma_\star(\mathcal{D}_T) \cdot f(\beta, \kappa(P_\star(\mathcal{D}_T)), N - |\mathcal{D}_T|, d, \alpha, n)$
```

A. Synthetic Example

We applied the four data-driven approaches on a randomly generated system of dimension 3 with three modes. To obtain statistics, we averaged the results over 25 experiences. For the two-step approach from [11], we used the heuristic from this paper for dataset splitting and parameters $\delta_1 = 10^2$, $\delta_2 = 1$. For Algorithm 1, we used $N_{\text{batch}} = 200$, $T = \lfloor N/N_{\text{batch}} \rfloor$, and $\eta_k = 0.3/(k+1)$. For Algorithm 2, we used $T = \lfloor N/2 \rfloor$, $\eta_k = 0.3$, with convergence criterion parameters $\epsilon = 10^{-4}$ and $K = 10$. We considered $\mathcal{P} := \{P \in \mathbb{R}_{\succ 0}^{n \times n} : I \preceq P \preceq 10^2 I\}$ and for Algorithm 1, $\mathcal{B} := \{B \in \mathbb{R}_{\succ 0}^{n \times n} : I \preceq B = B^T \preceq 10^2 I\}$.

The results—depicted in Figure 2—clearly showcase the advantages of the adaptive sampling approach, leading to significantly better guarantees compared to the basic non-adaptive sampling method [8]. While Algorithm 1 outperforms the non-adaptive method, its performance remains sub-optimal compared to the other adaptive sampling techniques ([11] and Algorithm 2). However, Algorithm 2, which reuses all previous samples in a heuristic manner, achieves strong performance, surpassing the state-of-the-art method [11].

B. Consensus Network

We consider the problem of consensus in a hidden switching interaction network, illustrated in Figure 4. This problem can be equivalently formulated as a linear switched system in dimension $n = 5$, and determining whether consensus is achieved translates to analyzing the stability of this switched linear system [16]. Since the network is hidden and its model

[7] The reason for our algorithm to outperform multi-pass SGD can be that the distribution is decision-dependent, thereby making the gradient computed from data collected with a different distribution less relevant.

[8] By heuristic method we mean that the computed change of basis has no whatsoever guarantee to converge to an optimal change of basis but nevertheless confidence bound from Theorem 1 still apply as it holds for any change of basis used.

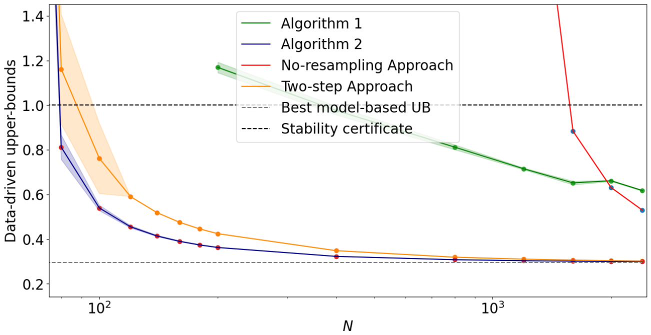

Fig. 2. Comparison of the upper bounds provided by the various algorithms, as a function of the total number of samples N, with $|\mathcal{A}| = 3$, $\alpha = \frac{1}{|\mathcal{A}|}$, $n = 3$, and $\beta = 5\%$. Results are averaged over 25 experiences on one randomly generated system $\mathcal{A}$. The shaded area show the mean ± 1 standard deviation.

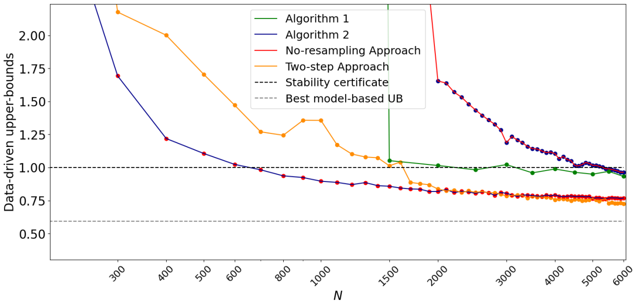

Fig. 3. Data-driven upper bounds with confidence $1 - \beta = 95\%$ on the JSR of the consensus network as a function of the total number of samples used N.

is unknown, we need to resort to data-driven methods to verify stability.

For the two-step approach from [11], we used the dataset-splitting heuristic and parameters $\delta_1 = 10^3$, $\delta_2 = 1$. For Algorithm 1, we used $N_{\text{batch}} = 500$, $T = \lfloor N/N_{\text{batch}} \rfloor$, and $\eta_k = 0.3/(k+1)$. For Algorithm 2, we used $T = \lfloor N/2 \rfloor$, $\eta_k = 0.3$, $\epsilon = 10^{-4}$ and $K = 10$. We consider $\mathcal{P} := \{P \in \mathbb{R}^{n \times n}_{\succ 0} : I \preceq P \preceq 10^3 I\}$ and for Algorithm 1, $\mathcal{B} := \{B \in \mathbb{R}^{n \times n}_{\succ 0} : I \preceq B = B^T \preceq 10^3 I\}$.

Figure 3 illustrates the total number of data points required to certify consensus with a confidence level of $1 - \beta = 95\%$ using the four data-driven methods. The results highlight the substantial benefits of adaptive sampling approach over the basic no-resampling approach, as well as the superiority of the heuristic method (Algorithm 2) over the two-step approach. Without resampling, 5400 data points were necessary to certify stability. By employing the stochastic optimization approach, this requirement is reduced to 2200 data points. The two-step approach from [11] further improves this to 1600 data points. However, Algorithm 2 achieves a breakthrough, requiring only 600 data points to ensure the system's stability with 95% confidence.

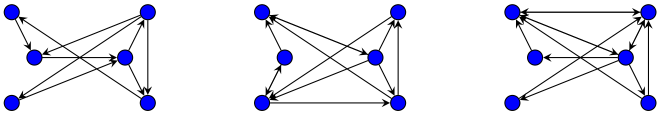

Fig. 4. Three (unknown) interaction networks for the consensus problem in Sec. V-B.

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms*. Cambridge, UK: Cambridge University Press, 2014.

[2] S. Mitra, *Verifying cyber-physical systems: a path to safe autonomy*. Cambridge, MA: MIT Press, 2021.

[3] Z. Sun and S. S. Ge, *Stability theory of switched dynamical systems*. London: Springer, 2011.

[4] D. Liberzon, *Switching in systems and control*. Boston, MA: Birkhäuser, 2003.

[5] R. M. Jungers, *The joint spectral radius: theory and applications*. Berlin: Springer, 2009.

[6] A. A. Ahmadi, R. M. Jungers, P. A. Parrilo, and M. Roozbehani, "Joint spectral radius and path-complete graph Lyapunov functions," *SIAM Journal on Control and Optimization*, vol. 52, no. 1, pp. 687–717, 2014.

[7] J. Kenanian, A. Balkan, R. M. Jungers, and P. Tabuada, "Data driven stability analysis of black-box switched linear systems," *Automatica*, vol. 109, p. 108533, 2019.

[8] G. O. Berger, R. M. Jungers, and Z. Wang, "Chance-constrained quasi-convex optimization with application to data-driven switched systems control," in *Learning for Dynamics and Control 2021*, ser. Proceedings of Machine Learning Research, vol. 144, 2021, pp. 571–583.

[9] A. Rubbens, Z. Wang, and R. M. Jungers, "Data-driven stability analysis of switched linear systems with sum of squares guarantees," *IFAC-PapersOnLine*, vol. 54, no. 5, pp. 67–72, 2021.

[10] Z. Wang, G. O. Berger, and R. M. Jungers, "Data-driven feedback stabilization of switched linear systems with probabilistic stability guarantees," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 4400–4405.

[11] A. Vuille, G. O. Berger, and R. M. Jungers, "Data-driven stability analysis of switched linear systems using adaptive sampling," *IFAC-PapersOnLine*, vol. 58, no. 11, pp. 31–36, 2024.

[12] G. C. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 742–753, 2006.

[13] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119. PMLR, July 2020, pp. 7599–7609. [Online]. Available: https://proceedings.mlr.press/v119/perdomo20a.html

[14] A. Vuille, G. O. Berger, and R. M. Jungers, "A stochastic-optimization-based adaptive-sampling scheme for data-driven stability analysis of switched linear systems," 2025. [Online]. Available: https://arxiv.org/abs/2508.21617

[15] Y. Lei, T. Hu, and K. Tang, "Generalization performance of multi-pass stochastic gradient descent with convex loss functions," *Journal of Machine Learning Research*, vol. 22, no. 25, pp. 1–41, 2021.

[16] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on automatic control*, vol. 48, no. 6, pp. 988–1001, 2003.