Towards D-Optimal Input Design for Finite-Sample System Identification ⋆

Sándor Kolumbán * **Balázs Csanád Csáji** **

* *Department of Mathematics and Computer Science, Eindhoven University of Technology, Netherlands (e-mail: s.kolumban@tue.nl).*
** *Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences, Hungary (e-mail: balazs.csaji@sztaki.mta.hu)*

Abstract: Finite-sample system identification methods provide statistical inference, typically in the form of *confidence regions*, with rigorous non-asymptotic guarantees under minimal distributional assumptions. *Data Perturbation* (DP) methods constitute an important class of such algorithms, which includes, for example, Sign-Perturbed Sums (SPS) as a special case. Here we study a natural *input design* problem for DP methods in *linear regression* models, where we want to select the regressors in a way that the *expected volume* of the resulting confidence regions are minimized. We suggest a general approach to this problem and analyze it for the fundamental building blocks of all DP confidence regions, namely, for ellipsoids having confidence probability exactly 1/2. We also present experiments supporting that minimizing the expected volumes of such ellipsoids significantly reduces the average sizes of the constructed DP confidence regions.

Keywords: system identification, confidence regions, finite sample results, least squares, parameter estimation, distribution-free results, input design

1. INTRODUCTION

Finite-sample system identification (FSID) methods infer properties of stochastic dynamical systems, arch-typically build *confidence regions*, with rigorous *non-asymptotic guarantees* under *minimal statistical assumptions* (Carè et al., 2018). They are motivated, e.g., by results showing that applying asymptotic methods in finite sample settings can lead to misleading outcomes (Garatti et al., 2004).

Data Perturbation (DP) methods form a general class of FSID algorithms (Kolumbán et al., 2015; Kolumbán, 2016) which generalize the construction of the *Sign-Perturbed Sums* (SPS) method (Csáji et al., 2012, 2015; Kieffer and Walter, 2014; Weyer et al., 2017). While the core assumption of SPS is the distributional *symmetry* of the noise terms, DP methods can exploit other kinds of regularity and also work, for example, with *exchangeable* or *power defined* noise sequences. One of the key properties of DP methods is that, similarly to SPS, they are guaranteed to provide *exact* confidence regions (Kolumbán, 2016).

Input design is a subfield of *experiment design* (Goodwin and Payne, 1977; Pintelon and Schoukens, 2012; Rodrigues and Iemma, 2014), and it is an important area of system identification, as the choice of the input signal has a substantial influence on the observations. There could be many possible aims of input design, for example, reducing the bias of the estimator, making the experiments more informative about some parts and modes, or minimizing the variance of certain components (Ljung, 1999).

⋆ S. Kolumbán was supported by the NWO Gravitation Project NETWORKS; B. Cs. Csáji was supported by the Hung. Sci. Res. Fund (OTKA), KH_17 125698, the GINOP-2.3.2-15-2016-00002 grant, and the János Bolyai Research Fellowship, BO/00217/16/6.

In this paper we study a natural input design problem for DP methods for *linear regression* models. Our aim will be to choose the inputs in a way that the *expected volume* of the constructed confidence regions is minimized. By arguing that all DP confidence regions can be built by *unions and intersections of ellipsoids* having confidence probability *exactly* 1/2, we first analyze these ellipsoids, as they are the fundamental building blocks of DP regions.

Along the way we provide the first explicit formulation of these fundamental ellipsoids in terms of the regressor and perturbation matrices, the true parameter and the noise.

We show that minimizing their expected volume can be done by maximizing the *expected determinant* of a certain quadratic term. This result has strong connections to classical D-optimal input design, but our method builds only on finite sample results and, hence it also depends on the actual regularity of the noise, i.e., the transformation group with respect to the distribution is invariant.

The paper ends with numerical experiments demonstrating that minimizing the expected volume of such ellipsoids carries over to the general case and the resulting DP confidence regions have smaller volumes on average.

2. PRELIMINARIES

In this section we introduce some notations, formalize the model, and the input design problem for DP methods.

2.1 Notations

First, let us define a subset of positive integers up to and including a constant k as $[k] \triangleq \{1, \ldots, k\}$. The *cardinality* of a set S will be denoted by $|S|$. Thus, $|[k]| = k$.

For any matrix A, its transpose is denoted with A^{T}. I_n denotes the n dimensional identity matrix and $\|v\|$ denotes the Euclidien norm of a vector v, i.e., $\|v\|^2 = v^{\mathrm{T}}v$. We denote the set of all $n \times n$ *orthogonal* matrices by

$$O(n) \triangleq \{\, G \in \mathbb{R}^{n\times n} : G^{\mathrm{T}}G = I_n \,\}, \tag{1}$$

which forms a *group* with the usual matrix multiplication.

Given a (skinny, full rank) matrix $A \in \mathbb{R}^{n\times \cdot}$, the *orthogonal projection* matrix to the column space of A is

$$P_A \triangleq A[A^{\mathrm{T}}A]^{-1}A^{\mathrm{T}}, \tag{2}$$

and the projection to its *orthogonal complement* is

$$P_A^{\perp} \triangleq I_n - A[A^{\mathrm{T}}A]^{-1}A^{\mathrm{T}}. \tag{3}$$

Naturally, in both cases we have $P^2 = P$ and $P^{\mathrm{T}} = P$.

2.2 Linear Regression

Let us consider the following linear regression problem

$$Y \triangleq X\theta_0 + E, \tag{4}$$

where $X \in \mathbb{R}^{n\times n_\theta}$ is the regressor matrix (i.e., the input), $\theta_0 \in \mathbb{R}^{n_\theta}$ is an unknown (constant) true parameter vector, $E \in \mathbb{R}^n$ is a random noise vector, $Y \in \mathbb{R}^n$ is the (random) vector of observations, and n is the sample size.

The classical *least-squares* (LS) method estimates θ_0 given X and Y. The LS estimate, $\hat{\theta}_n$, can be written as

$$\hat{\theta}_n \triangleq \operatorname*{argmin}_{\theta\in\mathbb{R}^{n_\theta}} \| Y - X\theta \|^2 = [X^{\mathrm{T}}X]^{-1}X^{\mathrm{T}}Y, \tag{5}$$

if X is skinny and full rank, which we assume henceforth.

As the LS estimate, $\hat{\theta}_n$, depends on the random noise E, it is a random vector. Confidence regions can be used to quantify the reliability of the estimate, but information about the distribution of E is required to construct such confidence regions. The most commonly used results are in this respect asymptotic (as $n \to \infty$), and build on the fact that, under mild statistical assumptions, we have

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_e^2\Psi^{-1}\right), \tag{6}$$

as $n \to \infty$, assuming the following covariance matrix

$$\Psi \triangleq \lim_{n\to\infty} \frac{1}{n} X_n^{\mathrm{T}}X_n \tag{7}$$

exists and it is positive definite, where $\xrightarrow{d}$ denotes convergence in distribution, σ_e^2 is the variance of the marginal distributions of the (homoskedastic) noise vector E, and $\mathcal{N}(\mu, \Sigma)$ is the multidimensional Gaussian distribution with mean vector μ and covariance matrix Σ (Ljung, 1999). Here we used an index n for the regressor matrix, X_n, to emphasize its dependence on the sample size.

A drawback of such asymptotic approaches is that they are not guaranteed rigorously for finite samples. This motivates FSID methods (Carè et al., 2018), such as DP algorithms, which can deliver *exact* probabilistic statements for finite samples under mild statistical assumptions.

2.3 Data Perturbation Methods

Now, we briefly overview Data Perturbation (DP) methods (Kolumbán, 2016). The three main components of DP methods are: *i*) the model structure, which is in our case linear regression; *ii*) a compact group $(\mathcal{G}, \cdot)$ under which the noise distribution is invariant; and *iii*) a performance measure function $Z : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where Θ, $\mathcal{X}$ and $\mathcal{Y}$ are sets of parameters, inputs and outputs, respectively.

Throughout the paper we will consider noise distributions that are invariant under a subgroup of $(\mathcal{O}(n), \cdot)$, that is

Definition 1. (Group invariant noise). *Let $\mathcal{G}$ be a subgroup of orthogonal transforms $\mathcal{G} \subseteq \mathcal{O}(n)$. The noise vector E is said to be invariant under the group $(\mathcal{G}, \cdot)$ if and only if*

$$\mathbb{P}(\, E \in \mathcal{A} \,) = \mathbb{P}(\, GE \in \mathcal{A} \,) \tag{8}$$

for all $G \in \mathcal{G}$ and measurable $\mathcal{A} \subseteq \mathbb{R}^n$.

The arch-typical examples of such invariant noise classes are: *i*) *jointly symmetric* noise distributions where the group consists of diagonal matrices with ± 1 entries (cf. Sign-Perturbed Sums); and *ii*) *exchangeable* noise values where the group consists of permutation matrices.

We consider the *performance measure* $Z(\cdot,\cdot,\cdot)$ defined as

$$Z(\theta, X, Y) \triangleq (Y - X\theta)^{\mathrm{T}}X[X^{\mathrm{T}}X]^{-1}X^{\mathrm{T}}(Y - X\theta). \tag{9}$$

The DP method corresponding to a group $(\mathcal{G}, \cdot)$ and performance measure (9) can be used to construct confidence regions around the least-squares estimate with confidence level exactly $\alpha = 1 - q/m$, under the assumption that the distribution of E is invariant under $(\mathcal{G}, \cdot)$.

In order to construct such exact confidence sets around $\hat{\theta}_n$, let us define $G_0 \triangleq I_n$ and $\{G_i\}_{i\in[m-1]}$ as independent and uniformly chosen random matrices chosen from $\mathcal{G}$ (also independent from X and Y). We define $Y^{(i)}(\theta) \triangleq X\theta + G_i(Y - X\theta)$ and $Z_i(\theta) \triangleq Z(\theta, X, Y^{(i)}(\theta))$.

The confidence region $\mathcal{C}_\alpha$ is defined as

$$\mathcal{C}_\alpha(X, Y) \triangleq \{\theta \in \mathbb{R}^{n_\theta} : |\{i : Z_0(\theta) >_\pi Z_i(\theta)\}| < q\}, \tag{10}$$

where the relation $>_\pi$ is the usual ordering $>$ with random tie-breaking according to the random (uniformly chosen) permutation π of $\{0, \ldots, m-1\}$. That is if $Z_i(\theta) = Z_j(\theta)$, we have $Z_i(\theta) >_\pi Z_j(\theta)$ if and only if $\pi(i) > \pi(j)$.

Let us highlight two important properties of such regions. First, the confidence level of these regions is *exact* for any finite sample size n, $\mathbb{P}(\, \theta_0 \in \mathcal{C}_\alpha \,) = 1 - q/m$. Furthermore, due to the random choices of $\{G_i\}_{i\in[m-1]}$ and π, $\mathcal{C}_\alpha$ is a *random* set, even if we condition on X and Y.

2.4 Input Design

The reliability of the obtained estimate, $\hat{\theta}_n$, depends on the input signal, X, as it can also seen from (6). Thus, it is a natural expectation that better estimates can be obtained by designing X, in case we can choose the inputs.

In order to design the input signal, X, first a proper design criterion needs to be selected. The most prominent choices are, among others, minimizing $\det\left([X^TX]^{-1}\right)$, which defines the *D-optimal design*, or $\operatorname{trace}\left([X^TX]^{-1}\right)$ defining the *A-optimal design*. The weighted trace minimization $\operatorname{trace}\left(W[X^TX]^{-1}\right)$ is a typical approximation for optimizing for a specific use of the estimate that can be measured as a scalar value (Goodwin and Payne, 1977).

Here, we focus on D-optimal input design, which can be interpreted as aiming for minimal volume confidence regions. If the noise, E, was normally distributed, then the minimal volume confidence region of the best linear unbiased estimator would be an ellipsoid whose volume was proportional to $\det\left([X^TX]^{-1}\right)$. Hence, minimizing this determinant would achieve the goal of minimizing the volume of the confidence region (Box and Draper, 1987).

More formally, the D-optimal input design problem for confidence level α can be formulated as

$$\begin{aligned} \mathcal{P}_d : \underset{X\in\mathcal{K}}{\text{minimize}} \quad & \text{vol}(\mathcal{C}_\alpha(X,Y)) \\ \text{subject to} \quad & \mathbb{P}(\,\theta_0 \in \mathcal{C}_\alpha(X,Y)\,) \geq \alpha \end{aligned} \tag{11}$$

where $\mathcal{K}$ denotes the set of admissible input signals.

We note that for (homoskedastic) Gaussian noises, the volume of $\mathcal{C}_\alpha(X,Y)$ can be written as $c(\alpha)\sigma_e^2 \det\left([X^TX]^{-1}\right)$, for any confidence level α. This results in the optimization criterion $\det\left([X^TX]^{-1}\right)$ independently of the confidence level. However, it is not evident if the optimization objective should be independent of the confidence level for other families of distributions (even for multimodal ones).

3. D-OPTIMAL INPUT DESIGN FOR DP METHODS

The classical D-optimal input design for linear regression, given in (11), simplified to a deterministic optimization problem, because the volume is determined by X. However this is no longer true for DP methods, as DP confidence regions are random even for fixed X and Y.

Here, taking the randomness of DP regions into account, we suggest formulating the input design problem as

$$\begin{aligned} \mathcal{P}_r : \underset{X\in\mathcal{K}}{\text{minimize}} \quad & \mathbb{E}\left[\,\text{vol}^2(\mathcal{C}_\alpha(X,Y)) \mid \text{vol}(\mathcal{C}_\alpha(X,Y)) < \infty\,\right] \\ \text{subject to} \quad & \mathbb{P}(\,\theta_0 \in \mathcal{C}_\alpha(X,Y)\,) = \alpha \end{aligned} \tag{12}$$

where $\mathcal{C}_\alpha(X,Y)$ is the DP confidence region given by (10), and assuming that each $X \in \mathcal{K}$ is skinny and full rank, to ensure that the performance measure (9) is well-defined.

Note that since DP methods are capable of providing *exact* confidence sets, we constrain the construction to such sets.

If we sample $\{G_i\}_{i\in[m-1]}$ with replacement, there is always a nonzero (though exponentially decaying, practically negligible) probability that a DP confidence region is equal to the whole parameter space. This means that (unconditionally) the expected size of the confidence set is infinite. Hence, to make the problem well-defined, we condition on the event that the region is not the whole space.

Finally, the specific choice of $\mathbb{E}\left[\,\text{vol}^2(\mathcal{C}_\alpha) \mid \text{vol}(\mathcal{C}_\alpha) < \infty\,\right]$ could be replaced without too much difficulty by any function that maps the distribution of $\text{vol}(\mathcal{C}_\alpha)$ to a scalar.

3.1 Structure of DP Confidence Regions

In order to effectively influence the expected sizes of DP confidence regions, first we need to understand the structure of the constructed regions. Recall that DP regions having confidence level $\alpha = 1 - q/m$ are given by (10).

Let us define $\mathcal{C}^i_{1/2}(X,Y)$ for $i \in [m-1]$ as

$$\mathcal{C}^i_{1/2}(X,Y) \triangleq \{\,\theta \in \mathbb{R}^{n_\theta} : Z_0(\theta) <_\pi Z_i(\theta)\,\}, \tag{13}$$

and let $[\mathcal{M}]_q$ denote the set of all subsets of $[m-1]$ with cardinality exactly q, that is

$$[\mathcal{M}]_q \triangleq \{\,S \subseteq [m-1] : |S| = q\,\}. \tag{14}$$

Using these notations, an equivalent characterization of the DP confidence region, $\mathcal{C}_\alpha(X,Y)$, can be given as

$$\mathcal{C}_\alpha(X,Y) = \bigcup_{S\in[\mathcal{M}]_q} \Big(\bigcap_{i\in S} \mathcal{C}^i_{1/2}(X,Y)\Big). \tag{15}$$

Hence, any confidence region with (rational) confidence probability $1-q/m$ can be constructed from $m-1$ instances of $1/2$ confidence regions by taking q of these in every possible way forming their intersections and taking the union of these intersections. In order to understand how to optimize the volume of general DP confidence regions, first we should study the structure of the sets $\mathcal{C}^i_{1/2}(X,Y)$.

Theorem 2. (Structure of $\mathcal{C}_{1/2}(X,Y)$). *Let $X = QR$ be the thin QR-decomposition of the regressor matrix X and assume that the noise is $\mathcal{G}$-invariant. Then the $1/2$ confidence region for the linear regression problem*

$$Y = X\theta_0 + E = QR\theta_0 + E \tag{16}$$

generated by the orthogonal matrix G is

$$\mathcal{C}_{1/2} = \left\{\,\theta : (\theta-\theta_c)^{\mathrm{T}} A_{Q,R}(\theta-\theta_c) \leq r_Q\,\right\}, \tag{17}$$

where $A_{Q,R}$, θ_c and r_Q are given by

$$A_{Q,R} \triangleq R^{\mathrm{T}}Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}QR\,, \tag{18}$$

$$r_Q \triangleq \|P_{[Q,G^{\mathrm{T}}Q]}E\|^2 - \|P_QE\|^2, \tag{19}$$

$$\theta_c \triangleq \theta_0 + A^{-1}R^{\mathrm{T}}Q^{\mathrm{T}}(P_Q - P_{G^{\mathrm{T}}Q})E. \tag{20}$$

Proof. See Appendinx A for a sketch of the proof.

This theorem shows that the $1/2$ confidence regions are *ellipsoids* with a center point that is θ_0 shifted with a linear function of E. The radius r depends on the norm of E projected to different subspaces of $\mathbb{R}^n$.

A plausible heuristic for optimizing the volumes of DP confidence regions could be to optimize the volumes of their building blocks, the $1/2$ confidence regions. Nonetheless, it is not obvious that optimizing the $1/2$ regions would result in optimal volumes for sets constructed by (15). In what follows we are going to explore this direction.

3.2 Optimization Objective

The goal of this section is to analyze the objective function, in order to design efficient algorithms to minimize the expected volumes of the $1/2$ confidence regions.

Since $\mathcal{C}_{1/2}$ is an ellipsoid, its (squared) volume is

$$\text{vol}^2(\mathcal{C}_{1/2}) \propto \det(A_{R,Q})^{-1} r_Q. \tag{21}$$

Thus, in order to decrease the volume, $\det(A_{R,Q})$ should be increased and r_Q decreased. What makes this a non-trivial task is that $A_{R,Q}$ and r_Q are intertwined through Q which is part of the input over which we try to optimize.

The following results establish conditions under which this coupling between $A_{R,Q}$ and r_Q can be neglected and there is a guaranteed distribution-free solution to problem (12).

Definition 3. (QR decoupled constraints). *The set of admissible inputs $\mathcal{K}$ is called QR decoupled if the admissibility of X can be verified based only on the R factor of the thin QR-decomposition of X. That is $\exists\, \mathcal{K}_R \subseteq \mathbb{R}^{n_\theta \times n_\theta}$ such that*

$$(X = QR \in \mathcal{K}) \;\Rightarrow\; (R \in \mathcal{K}_R), \tag{22}$$

$$(R \in \mathcal{K}_R) \;\Rightarrow\; (\forall Q \in \mathbb{R}^{n\times n_\theta}, Q^{\mathrm{T}}Q = I_{n_\theta} : QR \in \mathcal{K}). \tag{23}$$

Theorem 4. (Input design for QR decoupled constraints). *If the noise distribution is invariant under a subgroup $\mathcal{G}$ of $\mathcal{O}(n)$ and the admissible set of inputs $\mathcal{K}$ is QR decoupled then the optimizer of problem (12) can be obtained as*

$$R^* \;=\; \operatorname*{argmin}_{R\in\mathcal{K}_R} \det^{-1}(R)\,, \tag{24}$$

$$Q^* \;=\; \operatorname*{argmin}_{Q} \mathbb{E}\left[\det^{-1}(Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}Q) \mid G \neq \pm I_n\right]. \tag{25}$$

Proof. See Appendinx B for a sketch of the proof.

Note that the conditioning on the event $G \neq \pm I_n$ in (25) is there to ensure that the confidence region does not coincide with the whole parameter space. Therefore, it allows the expected volume of the confidence region to be finite.

Lemma 5. (Indistinguishable choices of Q). *If $R \in \mathbb{R}^{n_\theta\times n_\theta}$, $Q_1, Q_2 \in \mathbb{R}^{n\times n_\theta}$ such $Q_1^{\mathrm{T}}Q_1 = Q_2^{\mathrm{T}}Q_2 = I_{n_\theta}$, the noise is invariant under a subgroup $\mathcal{G}$ of $\mathcal{O}(n)$ and*

$$\exists\, G' \in \mathcal{G} \;:\; Q_1 = G'Q_2, \tag{26}$$

then, we have that

$$(A_{R,Q_1}, r_{Q_1}) \;\overset{d}{=}\; (A_{R,Q_2}, r_{Q_2}), \tag{27}$$

where the distributional equality is understood with respect to the uniformly chosen G, from subgroup $\mathcal{G}$ of $\mathcal{O}(n)$, that appears in the definition of matrix $A_{R,Q}$ and vector r_Q.

Proof. This lemma is given without proof but it can be shown using the randomization property of groups, for example, using Lemma 2.9 of (Kolumbán, 2016).

Lemma 5 is a key ingredient of the proof of Theorem 4. We also highlight it here, as it has some important consequences. It shows that there is a whole set of Q matrices that are optimal. Given the orthogonality constraint for Q it follows that solving the optimization in (25) is difficult.

If the noise is invariant under $\mathcal{O}(n)$, e.g., the Gaussian distribution, then every Q_1 is indistinguishable from any other Q_2, because there is always a $G \in \mathcal{O}(n)$, such that $Q_1 = GQ_2$. Hence, improvements can be achieved only if the noise is invariant under a proper subgroup of $\mathcal{O}(n)$.

3.3 Comparison with Asymptotic Results

There are a few interesting observations that can be made about Theorem 4 when we compare the optimal DP solutions with the asymptotically optimal choices. The asymptotic input design criterion asked for minimizing

$$\det\left([X^{\mathrm{T}}X]^{-1}\right) = \det^{-2}(R)\,, \tag{28}$$

thus the solution obtained by (24) results in an R^* matrix that is also optimal in the asymptotic sense.

It is easy to see that $\det^{-1}(Q^{\mathrm{T}}P^{\perp}_{G^TQ}Q) \geq 1$. This means that the expected volumes of DP confidence regions are always greater than or equal to that of the confidence regions based on the asymptotic theory. This is a manifestation of the well-known fact that the asymptotic confidence regions can underestimate the uncertainty of the parameter estimates if the noises are not Gaussian.

The choice of Q is irrelevant in the asymptotic sense, but Theorem 4 shows that for finite sample sizes, its choice matters. A heuristic argument can be given to show that $\mathbb{E}[\det^{-1}(Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}Q)] \to 1$ as $n \to \infty$ under some assumptions on how $|\mathcal{G}| \to \infty$. This is again consistent with the asymptotic theory postulating that the choice of Q becomes less relevant as the sample size increases.

We note that this nice property, i.e., that the DP-optimal input is also optimal in the asymptotic sense, is only guaranteed in the case when the constraint set $\mathcal{K}$ is a QR decoupled. If this not the case, then a DP-optimal input X might have a QR decomposition in which the R factor does not coincide with the standard asymptotic solution.

3.4 Optimizing R and Q

Traditionally, the orientation of the confidence ellipsoid is not taken into account for D-optimal input design (Box and Draper, 1987). This principle also prevails in the FSID input design problem we analyze here, since the orientation of the ellipsoid is irrelevant in general, only the eigenvalues of kernel $A_{Q,R}$ and the radius r_Q determine the volume.

The main novelty w.r.t. classical input design is that if the noise is invariant under a proper subgroup of $\mathcal{O}(n)$, then some Q values should be preferred to others.

In general, it is unlikely that (25) has an analytical solution, thus, we are going to apply a suitable approximation.

The objective of (25) is to optimize a (conditional) expectation with respect to G that is a uniformly distributed random variable over $\mathcal{G}$. Here, we propose the following Monte Carlo approximation to handle (25)

$$\widehat{Q}^* \;\triangleq\; \operatorname*{argmin}_{Q} \frac{1}{K_G}\sum_{i=1}^{K_G} \operatorname{tr}(Q^{\mathrm{T}}P_{G_i^{\mathrm{T}}Q}Q), \tag{29}$$

where K_G is a user-chosen parameter and $\{G_i\}_{i\in[K_G]}$ are K_G elements from $\mathcal{G}$ chosen uniformly at random.

This formulation can be interpreted as the empirical mean of $\operatorname{tr}(Q^{\mathrm{T}}P_{G^{\mathrm{T}}Q}Q)$. Constructing the empirical mean using the $\det^{-1}$ appearing in (25) proved to be numerically unstable, nevertheless, we can obtain (29) as an approximation of (25) considering that

$$\mathbb{E}\left[\det^{-1}(Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}Q)\right] \;\approx\; \mathbb{E}\left[\det(Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}Q)\right]^{-1} \tag{30}$$

$$\approx\; \left(1 - \mathbb{E}\big[\operatorname{tr}(Q^{\mathrm{T}}P_{G_i^{\mathrm{T}}Q}Q)\big]\right)^{-1}. \tag{31}$$

Let $(\lambda_i)_{i=1}^{n_\theta}$ be the eigenvalues of the matrix $Q^{\mathrm{T}}P_{G^{\mathrm{T}}Q}Q$. It is easy to see that $\lambda_i \geq 0$ and they are small for the minimizer of (25). Since, after rearrangement

$$Q^{\mathrm{T}}P^{\perp}_{G^{\mathrm{T}}Q}Q = I_{n_\theta} - Q^{\mathrm{T}}P_{G^{\mathrm{T}}Q}Q, \tag{32}$$

the determinant can be written as

$$\det(Q^{\mathrm{T}} P^{\perp}_{G^{\mathrm{T}}Q} Q) = \prod_{i=1}^{n_\theta} (1 - \lambda_i) \tag{33}$$

$$= 1 - \sum_{i=1}^{n_\theta} \lambda_i + O(\max_i(\lambda_i^2)), \tag{34}$$

where neglecting the quadratic terms is reasonable. This rationalizes the approximation in (31), while the step in (30) is meaningful because the function $1/x$ is approximately linear in the neighborhood of $x = 1$.

As (29) may have multiple global optima, we try to find one of its minima by starting local optimizations from randomly chosen initial Q matrices and using the best local optimizer so obtained. Since it is possible to choose the initial Q matrix uniformly over the feasible set, this algorithm can provide a good approximation to (25).

4. NUMERICAL EXPERIMENTS

This section contains an example to illustrate the effectiveness of the proposed input design approach. The sample problem is specified by $n_\theta = 2$, $\theta_0 = [1, 1]^{\mathrm{T}}$ and $n = 10, 20, 50$. The noise is i.i.d. Laplacian with density

$$f(E) \triangleq \prod_{i=1}^{n} \frac{\lambda}{2} e^{-\lambda |e_i|}, \quad E = [e_1, \ldots, e_n]^{\mathrm{T}}, \quad \lambda = 10. \tag{35}$$

This distribution is jointly symmetric, and thus invariant under the group of matrices with ± 1 diagonal entries. Note that the decay of tails of the density function is only $e^{-|x|}$, which is much slower than e^{-x^2}, i.e., the case of Gaussian noise. As a result, the asymptotic Gaussian confidence region is expected to underestimate the uncertainty of the estimates. We are going to examine the volume of $1 - 1/10$ confidence regions for θ_0 constructed by the Sign-Perturbed Sums (SPS) method (Csáji et al., 2015).

Table 1 contains the aggregated results of 3000 independent experiments for various confidence region constructions and sample sizes. As it was discussed in Section 3, we excluded the cases when the region was the whole space, thus, the expectations and variances are understood conditionally that the regions are non-degenerate.

We constrained X to matrices with $\mathrm{tr}(X^{\mathrm{T}}X) \leq 2$. This constraint is QR decoupled and the D-optimal choice for R is $R = I_2$, which we used for *all* constructions. For SPS with input design we obtained Q using the approximation algorithm outlined in Section 3.4 with $K_G = 250$ and using 1000 randomly initialized local optimizations.

As the maximum allowed energy of the input signal was the same in each experiment, namely $\mathrm{tr}(X^{\mathrm{T}}X) \leq 2$, independently of the sample size, the regions do not shrink as the sample size increases, and are directly comparable.

Note that SPS, both with and without input design, provides *exact* confidence regions, hence, the first two lines of the $\mathbb{P}(\theta_0 \in C_\alpha)$ column for each sample size are around the desired 0.9. On the other hand, the classical confidence regions based on the asymptotic Gaussianity of the (scaled) estimation error are not guaranteed rigorously and underestimate the real uncertainty of the parameters, resulting in lower than required empirical coverage values.

Table 1. Empirical coverage and volume statistics (conditioned on non-degenerate regions)

	$\mathbb{E}[\,\mathrm{vol}(C_\alpha)\,]$	$\mathrm{Var}[\,\mathrm{vol}(C_\alpha)\,]$	$\mathbb{P}(\,\theta_0 \in C_\alpha\,)$
Sample size, $n = 10$			
SPS, uniform Q	0.5304	0.9047	0.8968
SPS, designed $\widehat{Q}^*$	0.3041	0.2571	0.9069
Asymptotic	0.0703	0.0025	0.8137
Sample size, $n = 20$			
SPS, uniform Q	0.1479	0.0164	0.9012
SPS, designed $\widehat{Q}^*$	0.1166	0.0061	0.9069
Asymptotic	0.0730	0.0014	0.8612
Sample size, $n = 50$			
SPS, uniform Q	0.1039	0.0031	0.8971
SPS, designed $\widehat{Q}^*$	0.0979	0.0025	0.8992
Asymptotic	0.0730	0.0005	0.8847

The expected volumes with designed inputs, $\widehat{Q}^*$, are significantly smaller than the ones based on a uniformly chosen Q. The improvements are between 6% and 42%, even though in both cases the optimal choice for R is used. Moreover, the variance of the regions are also decreased with designing Q. Though, the expected volumes of the asymptotically designed regions are the smallest, their confidence probabilities are not correct, as it was noted.

In order to give a more detailed view on the influence of input design on the shape of the confidence regions, Figure 1 shows the inclusion probability $\mathbb{P}(\theta \in \mathcal{C}_\alpha)$ for different parameter values and confidence region constructions.

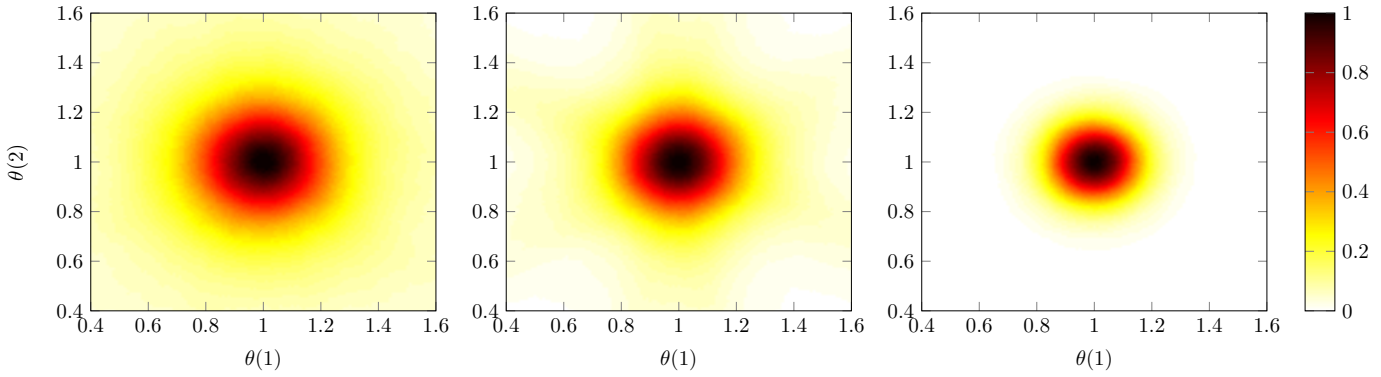
Figures 1(a) and 1(b) contain the heatmaps corresponding to SPS with uniformly random chosen Q and designed $\widehat{Q}^*$, respectively. The heatmap of the asymptotic confidence region is also given on Figure 1(c) for comparison.

5. CONCLUSIONS

Finite-sample system identification (FSID) aims at providing methods with rigorous non-asymptotic guarantees under minimal statistical assumptions. Data-Perturbation (DP) methods generalize the Sign-Perturbed Sums (SPS) algorithm and can construct exact confidence regions using some mild distributional regularities of the noise.

In this paper we studied a natural input design problem for DP methods in which we aim at minimizing the expected volume of the confidence region. We explored the possibility of achieving this by individually minimizing the volumes of the fundamental building blocks of DP regions, namely, ellipsoids with confidence probability exactly $1/2$.

It was shown that even handling such ellipsoids is hard in a finite sample setting, but can be achieved under certain decoupling assumptions, which also leads to nice connections with the classical asymptotic theory. A Monte Carlo approximation was suggested to numerically solve the optimization and simulation experiments were presented indicating that minimizing the expected volumes of these fundamental ellipsoids carries over to the general case and reduces the expected sizes of DP confidence regions.



(a) SPS, uniformly chosen Q (b) SPS, designed $\widehat{Q}^*$ (c) Asymptotic theory

Fig. 1. The probability of various parameter values, $\theta \in \mathbb{R}^2$, being included in the (random) $\mathcal{C}_{0.9}$ confidence region, for $n = 10$, depending on the method used to construct the region, and whether matrix Q is designed. Though the asymptotic theory provides the smallest regions, it underestimates the uncertainty of the parameters, cf. Table 1.

REFERENCES

Box, G.E.P. and Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons.

Carè, A., Csáji, B.Cs., Campi, M., and Weyer, E. (2018). Finite-sample system identification: An overview and a new correlation method. *IEEE Control Systems Letters*, 2(1), 61 – 66.

Csáji, B.Cs., Campi, M., and Weyer, E. (2012). Non-asymptotic confidence regions for the least-squares estimate. In *Proceedings of the 16th IFAC Symposium on System Identification*, 227 – 232.

Csáji, B.Cs., Campi, M., and Weyer, E. (2015). Sign-perturbed sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Transactions on Signal Processing*, 63(1), 169 – 181.

Garatti, S., Campi, M., and Bittanti, S. (2004). Assessing the quality of identified models through the asymptotic theory – when is the result reliable? *Automatica*, 40(8), 1319–1332.

Goodwin, G.C. and Payne, R.L. (1977). *Dynamic System Identification: Experiment Design and Data Analysis*. Mathematics in Science and Engineering. Elsevier.

Kieffer, M. and Walter, E. (2014). Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS. *Automatica*, 50, 507 – 512.

Kolumbán, S. (2016). *System Identification in Highly Non-Informative Environment*. University Press.

Kolumbán, S., Vajk, I., and Schoukens, J. (2015). Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets. *Automatica*, 51, 326 – 331.

Ljung, L. (1999). *System Identification - Theory for the User*. Prentice Hall, 2nd edition.

Pintelon, R. and Schoukens, J. (2012). *System Identification: A Frequency Domain Approach*. Wiley-IEEE Press, 2nd edition.

Rodrigues, M.I. and Iemma, A.F. (2014). *Experimental Design and Process Optimization*. CRC Press.

Weyer, E., Campi, M.C., and Csáji, B.Cs. (2017). Asymptotic properties of SPS confidence regions. *Automatica*, 82, 287 – 294.

Appendix A. PROOF SKETCH FOR THEOREM 2

Here, the main steps of proving Theorem 2 are given without the detailed calculations. The first key ingredient is eq. (2.37) from Kolumbán (2016) expressing $Z_i(\theta)$ as

$$\begin{aligned} Z_i(\theta) = &(Y - X\theta_0)^{\mathrm{T}} G^{\mathrm{T}} X [X^{\mathrm{T}} X]^{-1} X^{\mathrm{T}} G (Y - X\theta_0) + \\ &+ 2 (Y - X\theta_0)^{\mathrm{T}} G^{\mathrm{T}} X [X^{\mathrm{T}} X]^{-1} X^{\mathrm{T}} G X (\theta_0 - \theta) + \\ &+ (\theta_0 - \theta)^{\mathrm{T}} X^{\mathrm{T}} G^{\mathrm{T}} X [X^{\mathrm{T}} X]^{-1} X^{\mathrm{T}} G X (\theta_0 - \theta). \end{aligned}$$

This can be used to write $Z_0(\theta) - Z_1(\theta)$ as a quadratic function of $\Delta = \theta_0 - \theta$. This quadratic function can be re-written into another form as a function of $\Delta + \Delta_0$ for some Δ_0, such that the linear term is eliminated (completing the square). The confidence region $\mathcal{C}_{1/2}$ is the 0 level set of $Z_0(\theta) - Z_1(\theta)$, so it corresponds to some non-zero level-set of the completed square. Performing some linear algebraic manipulations will lead to the statement of the theorem.

Appendix B. PROOF SKETCH FOR THEOREM 4

One of the main ingredients in proving Theorem 4 is the following lemma, that can be shown by simple algebraic manipulations and using the properties of projections.

Lemma 6. For any Q and G, and $\widetilde{Q} = TQ$, where T is an orthogonal matrix, we have that

$$\widetilde{Q}^{\mathrm{T}} P^{\perp}_{G^{\mathrm{T}}\widetilde{Q}} \widetilde{Q} = Q^{\mathrm{T}} P^{\perp}_{T^{\mathrm{T}} G^{\mathrm{T}} T Q} Q. \tag{B.1}$$

We have already provided Lemma 5 which asserts that Q_1 and Q_2 cannot be distinguished from each other if they can be transformed into one another by an element of the considered invariance group $\mathcal{G}$.

As we already mentioned in Section 3.2, every such pair of Q's can be transformed into each other by an orthogonal matrix, so the last step is to show how Q_1 and Q_2 can be compared if $Q_1 = TQ_2$ holds only for $T \notin \mathcal{G}$. We can define $\mathcal{G}_T \triangleq \{\tilde{G} \in \mathbb{R}^{n\times n} : \exists G \in \mathcal{G} : \tilde{G} = TG\}$. The main obstacle in the proof is that the distribution of the noise E appears in the comparison of Q_1 and Q_2. By making a one to one correspondence between distributions invariant under $\mathcal{G}$ and those invariant under $\mathcal{G}_T$ this can be bypassed and it follows that the comparison can be made solely based on the value of Q and the group $\mathcal{G}$, as given in Theorem 4.