Towards Qualitative System Identification Through Optimization

SIOPT23 | MS330 Data-Driven Optimization: Algorithms and Theoretical Guarantees

Wouter Jongeneel[1], based on joint work (IEEE TAC 2023) with Tobias Sutter[2] and Daniel Kuhn[1].

June 03, 2023

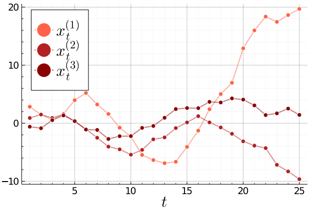
[1] École Polytechnique Fédérale de Lausanne (RAO), supported by the NCCR *Automation*, [2] University of Konstanz, Machine Learning and Optimization Group.

Consider a discrete-time system on $\mathbb{R}^n$

$$x_{t+1} = \theta x_t + w_t, \quad w_t \overset{i.i.d.}{\sim} \mathrm{distr}(0, S_w \succ 0) \tag{1}$$

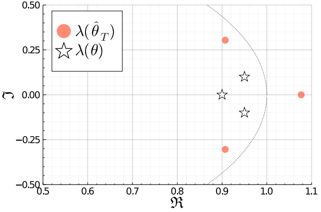
with **unknown** $\theta \in \mathbb{R}^{n\times n}$, being **asymptotically stable** $(\rho(\theta) < 1)$. Given measurements $(\widehat{x}_t)_{t\geq 0}$ of (1), consider the LS estimator of θ:

$$\widehat{\theta}_T = \left(\sum_{t=1}^T \widehat{x}_t \widehat{x}_{t-1}^\top\right)\left(\sum_{t=1}^T \widehat{x}_{t-1}\widehat{x}_{t-1}^\top\right)^{-1} \quad (T \geq n). \tag{2}$$



Collect single trajectory $\subset \mathbb{R}^3$

least squares estimation of θ, denoted $(\widehat{\theta}_T)$ $\rightsquigarrow$



Estimate $\widehat{\theta}_T$ is unstable?

But we understand the process $(\widehat{\theta}_t)_{t\geq 0}$ very well?

Outstanding work[1] on LS **statistical** identification of

$$x_{t+1} = \theta x_t + w_t \tag{3}$$

that is, bounds like $\mathbb{P}(\|\widehat{\theta}_t - \theta\|_p \leq \delta) \geq 1 - \beta$ for $t \geq T$.

At a lower level we first like to understand **qualitative behaviour**, that is, $\mathbb{P}(\widehat{\theta}_T$ qualitatively the same as $\theta)$?

Observation: ℓ_p-norms not appropriate for stability.

Let $\theta = \begin{pmatrix} \lambda & C \\ 0 & \lambda \end{pmatrix}$ for $\lambda \in (-1,1)$, $C \gg 1$ and $\widehat{\theta}_T = \begin{pmatrix} \lambda & C \\ \epsilon_T & \lambda \end{pmatrix}$ for $\epsilon_T > 0$.

Then, $\|\theta - \widehat{\theta}_T\|_2 = \epsilon_T$ yet $\lambda(\widehat{\theta}_T) = \{\lambda \pm \sqrt{C\epsilon_T}\}$ [$\rho(\cdot)$ not a norm].

[1] Substantial body of work on (sub-)optimal finite-sample concentration bounds for linear systems identified via least squares estimation [Sim+18; JP19; SR19; JP20; SRD21].

Perhaps we can truncate the Jordan normal form?

Naïve method to (asymptotically) stabilize $\theta' \notin \Theta := \{\theta \in \mathbb{R}^{n\times n} : \rho(\theta) < 1\}$: scale its unstable eigenvalues into $\mathbb{C}_{|z|<1}$. Consider the matrices

$$\theta' = \begin{bmatrix} 1.01 & 10 \\ 0.01 & 1 \end{bmatrix},\ \theta'_a = \begin{bmatrix} 0.84 & 4.77 \\ 0.005 & 0.84 \end{bmatrix},\ \theta'_b = \begin{bmatrix} 0.99 & 10 \\ 0 & 0.99 \end{bmatrix}.$$

Clipping off the unstable eigenvalues of θ' at $|\lambda| = 0.99$ yields θ'_a with $\rho(\theta'_a) = 0.99$ and $\|\theta' - \theta'_a\|_2 = 5.24$.

However, θ'_b also has $\rho(\theta'_b) = 0.99$ but with $\|\theta' - \theta'_b\|_2 = 0.02$!

Problem is non-trivial, lots of related work

Towards a solution: early work by Maciejowski [Mac95], used in Sys. Id. [VD96] (*distorted*).

Lacy and Bernstein [LB02] ***approximate*** Θ by $\{\theta \in \mathbb{R}^{n\times n} : \|\theta\|_2 < 1\}$ (*convex*, but *conservative*), related: [LB03; BGS08; Tur+13] (conservative/expensive), regularization [Van+00; Van+01] (*tuning*), MLE approach [Ume+18] (expensive), more..

Related to the ***nearest stable matrix problem***

$$\Pi_\Theta(\theta') \in \arg\min_{\theta \in \operatorname{cl}\Theta} \|\theta' - \theta\|^2, \tag{4}$$

Solutions: *successive convex approximations* [ONV13], *low-rank matrix differential equations* [GL17], elegant *reparametrization* of $\widetilde{\Theta}$ [GKS19; CGS20], Nesterov and Protasov [NP20] solve (4) for *polyhedral norms* and *non-negative* θ'.

Projection problem (4) is mathematically beautiful but perhaps practically not ideal: Given that $\theta \in \Theta$ yet $\widehat{\theta}_T \notin \Theta$, do we want to project to ***the boundary*** $\partial\Theta$ [VD96, pp. 53–60, 125–129]?

Differently: one could try to design a ***LQR problem*** whose optimal feedback gain $K^\star \in \mathbb{R}^{n\times n}$ renders $\widehat{\theta}_T + K^\star$ stable.

Overlooked but early work: Tanaka and Katayama [TK05] propose a LQR objective that is *inversely proportional* to S_w (clear relation to $\partial\Theta$), yet, without all the analysis.

Additional benefit of LQR: well-understood [BLW91; LR95], fast and scalable ($n \approx 1000$), *structure preserving* [JK21], *e.g.*, $\ker(\widehat{\theta}_T) = \ker(\widehat{\theta}_T + K^\star)$.

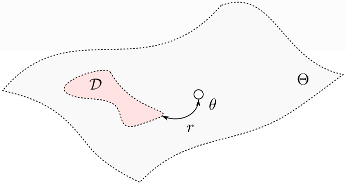
Moderate scale: $(a_T)_T$ such that $\lim_{T\to\infty} a_T = \infty$ yet $\lim_{T\to\infty} \frac{a_T}{T} = 0$.

Definition (Moderate Deviation Principle (MDP) [DZ09])

A sequence $(\widehat{\theta}_T)_T$ satisfies a MDP if there is a rate ("*distance*") function $I(\theta', \theta)$ such that for any Borel set $\mathcal{D} \subseteq \mathbb{R}^{n\times n}$:

$$-\underbrace{\inf_{\theta'\in \mathrm{int}\mathcal{D}} I(\theta', \theta)}_{\underline{r}} \leq \liminf_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\left(\widehat{\theta}_T \in \mathcal{D}\right)$$

$$\leq \limsup_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\left(\widehat{\theta}_T \in \mathcal{D}\right) \leq -\underbrace{\inf_{\theta'\in \mathrm{cl}\mathcal{D}} I(\theta', \theta)}_{\bar{r}}.$$



$\mathbb{P}_\theta(\widehat{\theta}_T \in \mathcal{D}) \leq e^{-\bar{r}\cdot a_T + o(a_T)}$

Finding $I(\cdot,\cdot)$ can be painful..

*Skipping technical assumptions, *e.g.*, $w \sim$ subG.

Lemma (Least squares MDP [JSK23])

If $(\widehat{\theta}_T)_{T\geq 0}$ is a sequence of ***least squares*** *estimators, then $(\frac{T}{a_T}(\widehat{\theta}_T - \theta) + \theta)_T$ satisfies a* ***MDP*** *with* ***rate function***

$$I(\theta', \theta) = \frac{1}{2}\mathrm{tr}\left(S_w^{-1}(\theta' - \theta)S_\theta(\theta' - \theta)^\top\right).$$

for[2] $S_\theta = \theta S_\theta \theta^\top + S_w = \sum_{k=0}^{\infty} \theta^k S_W (\theta^k)^\top$.

proof: Make the results from [YS09] explicit.

Complication: $I(\theta', \theta)$ non-convex in $\theta \in \Theta$.

[2] System: $x_{t+1} = \theta x_t + w_t$, State covariance: $S_\theta = \lim_{t\to\infty} \mathbb{E}_\theta[x_t x_t^\top]$, noise covariance: $S_w = \mathbb{E}[w_t w_t^\top]$.

Motivated by the MDP-related optimality results[3], we construct [JSK23] a ***reverse** I-**projection*** defined through

$$\mathcal{P}(\theta') \in \arg\min_{\theta\in\Theta} I(\theta',\theta) \text{ for } I(\theta',\theta) = \tfrac{1}{2}\mathrm{tr}\left(S_w^{-1}(\theta'-\theta)S_\theta(\theta'-\theta)^\top\right). \qquad (5)$$

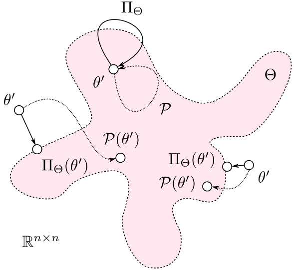


Figure 1: Schematic visualization of $\Pi_\Theta(\theta') \in \arg\min_{\theta\in\mathrm{cl}\Theta} \|\theta'-\theta\|^2$ and $\mathcal{P}(\theta')$ for different estimator realizations θ' inside and outside of Θ.

[3] VMK21; SVK20; BV21.

Observations regarding $\mathcal{P}(\theta') \in \arg\min_{\theta\in\Theta} I(\theta', \theta)$:

(i) $I(\theta', \theta)$ trades of *distance* (weighted $\|\cdot\|_F$) against *stability* (S_θ);

(ii) By exploiting that $I(\widehat{\theta}_T, \theta) = (a_T/T)I(\sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta, \theta)$, one can show that the PDF $\varrho_{\theta,T}$ of $\widehat{\theta}_T$ satisfies

$$\varrho_{\theta,T}(\widehat{\theta}_T) \approx \exp(-I(\widehat{\theta}_T, \theta) \cdot T). \tag{6}$$

Thus, $\mathcal{P}(\widehat{\theta}_T)$ maximizes the RHS of (6) across all $\theta \in \Theta$ (**MLE-like**).

In addition, by using ideas due to Jedra and Proutiere [JP20], one can show that for Gaussian noise, $I(\theta', \theta)$ can be interpreted as the long-run average expected log-likelihood ratio between observations generated under $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$.

On an intuitive level, we can motivate

$$\mathcal{P}(\theta') \in \arg\min_{\theta \in \Theta} l(\theta', \theta) \text{ for } l(\theta', \theta) = \frac{1}{2}\mathrm{tr}\left(S_w^{-1}(\theta' - \theta)S_\theta(\theta' - \theta)^\top\right).$$

but can we formalize this?

Better yet, can we work with this operator?

Theorem (Efficient identification with stability guarantees [JSK23])

For any $\theta \in \Theta$, the reverse I-projection has the following properties.

(i) ***Asymptotic consistency.*** $\lim_{T\to\infty} \mathcal{P}(\widehat{\theta}_T) = \theta \quad \mathbb{P}_\theta$*-a.s.*

(ii) ***Finite sample guarantee.*** *There are constants $\tau \geq 0$ and $\rho \in (0,1)$ that depend only on θ such that*

$$\mathbb{P}_\theta \left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \kappa(S_w) 2\varepsilon n^{\frac{1}{2}} \tau (1-\rho^2)^{-\frac{1}{2}} \right) \geq 1 - \beta$$

for all $\beta, \varepsilon \in (0,1)$ and $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$.

(iii) ***Efficient computation.*** *For any $\theta' \notin \Theta$ and $S_w, Q \succ 0$ there is a $p \geq 1$, such that for all $\delta > 0$ we have that*

$$\theta^\star_\delta = \theta' + \mathrm{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1}) \in \Theta \text{ with } \|\mathcal{P}(\theta') - \theta^\star_\delta\|_2 \leq O(\delta^p).$$

Not particularly surprising since $\mathcal{P}(\theta') = \theta'$ for all $\theta' \in \Theta$.

Formally, this follows readily from continuity, convexity and $\lim_{T\to\infty} \widehat{\theta}_T = \theta$ $\mathbb{P}_\theta$-almost surely [CK98].

In fact, we can prove:

Lemma (Properties of $I(\theta', \theta)$ [JSK23])

The rate function $I(\theta', \theta)$ has the following properties.

(i) *$I(\theta', \theta)$ is analytic in $(\theta', \theta) \in \Theta' \times \Theta$.*

(ii) *If $\theta' \in \Theta$, then the sublevel set $\{\theta \in \Theta : I(\theta', \theta) \leq r\}$ is compact for every $r \geq 0$.*

(iii) *If $\theta' \in \Theta$, then $I(\theta', \theta)$ tends to infinity as θ approaches the boundary of Θ.*

To quantify "*where*" in Θ the matrix θ lives: we say that the system matrix $\theta \in \Theta$ is (τ, ρ)-***stable*** [KTR19, Def. 1] for some $\tau \geq 1$ and $\rho \in (0, 1)$ if $\|\theta^k\|_2 \leq \tau \rho^k$ for all $k \in \mathbb{N}$.

As θ is (τ, ρ)-stable:

$$\begin{aligned} I(\widehat{\theta}_T, \mathcal{P}(\widehat{\theta}_T)) \leq & I(\widehat{\theta}_T, \theta) = \tfrac{1}{2}\mathrm{tr}\left(S_w^{-1}(\widehat{\theta}_T - \theta)S_\theta(\widehat{\theta}_T - \theta)^\top\right) \\ \leq & \tfrac{1}{2}\mathrm{tr}(S_w^{-1})\|\widehat{\theta}_T - \theta\|_2^2\|S_\theta\|_2 \\ \leq & \tfrac{1}{2}n\kappa(S_w)\|\widehat{\theta}_T - \theta\|_2^2\frac{\tau^2}{1-\rho^2}, \end{aligned}$$

Combine with a Pinsker-type inequality [JSK23]: for any $\theta' \in \Theta'$ and $\theta \in \Theta$ we have $\|\theta' - \theta\|_2^2 \leq 2\kappa(S_w)I(\theta', \theta)$.

Interestingly, we can get a (similar) probabilistic grip on $\|\widehat{\theta}_T - \theta\|_2$ using the MDP *or* contemporary identification results [SR19].

$\mathcal{P}(\theta') \in \arg\min_{\theta\in\Theta} I(\theta',\theta)$ is *non*-convex, how does that work?

Consider $\min_{\theta\in\Theta}\{\mathrm{tr}(QS_\theta) : I(\theta',\theta) \leq r\}$ for the *smallest* radius $r = \underline{r}$ that preserves feasibility.

Taking a Lagrangian (penalty) viewpoint, equivalent to

$$\begin{aligned}
&\min_{\theta\in\Theta} \mathrm{tr}(QS_\theta) + \delta^{-1} I(\theta',\theta)\\
&= \min_{\theta\in\Theta} \lim_{T\to\infty} T^{-1}\mathbb{E}_\theta\Big[\textstyle\sum_{k=0}^{T-1} x_k^\top Q x_k + \frac{1}{2\delta} x_k^\top (\theta'-\theta)^\top S_w^{-1}(\theta'-\theta)x_k\Big]\\
&= \min_{L\in\mathbb{R}^{n\times n}} \lim_{T\to\infty} T^{-1}\mathbb{E}_{\theta'+L}\Big[\textstyle\sum_{k=0}^{T-1} x_k^\top \left(Q + \frac{1}{2\delta} L^\top S_w^{-1} L\right) x_k\Big],
\end{aligned}$$

Here, $\theta = \theta' + L$, with L the *feedback* term.

For $\delta \downarrow 0$, we approach the original problem solution[4].

[4] If θ' has unimodular eigenvalues, then, approximation is the best we can do.

Let $P_\delta \succ 0$ be a fixed point of $P_\delta = Q + \theta'^{\mathsf{T}} P_\delta \theta' - \theta'^{\mathsf{T}} P_\delta (P_\delta + (2\delta S_w)^{-1})^{-1} P_\delta \theta'$, then

$$
\begin{aligned}
\theta^\star_\delta &= \theta' + \mathrm{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1}) \\
&= \Lambda_\delta^{-1}\theta' \text{ for } \Lambda_\delta = (I_n + 2\delta S_w P_\delta) \text{ (topological ramifications).}
\end{aligned}
$$

Observe that $\theta^\star_\delta \in \Theta$ for any δ!

By using [Pol86, Lem. 3.2], one can show that P_δ and consequently also $\theta^\star_\delta$ are real-analytic (C^ω) in $\delta > 0$.

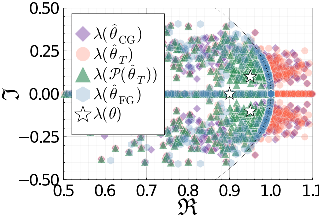


Figure 2: Comparison to CG [BGS08] and FG [GKS19] methods.

We compute

$$\mathcal{P}(\theta') \in \arg\min_{\theta\in\Theta} l(\theta',\theta) \text{ for } l(\theta',\theta) = \tfrac{1}{2}\mathrm{tr}\left(S_w^{-1}(\theta'-\theta)S_\theta(\theta'-\theta)^\top\right).$$

(approximately) through

$$\theta^\star_\delta = \theta' + \mathrm{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1}) \text{ for “small” } \delta,$$

$\mathrm{dlqr}(\cdot)$ standard routine in MATLAB, Julia, Python and so forth,

but how to pick $Q \succ 0$ and does $\delta \downarrow 0$ not look problematic?

The ***algebraic Riccati equation*** (ARE)

$$P_\delta = Q + \theta'^{\top} P_\delta \left(I_n + 2\delta S_w P_\delta\right)^{-1} \theta' \tag{7}$$

is well-understood [BLW91; LR95]. Fixed-point (DP) schemes can be unstable, elegant solution proposed in the 80s.

To start, define the pair of matrices $S_1, S_2 \in \mathbb{R}^{2n \times 2n}$ by

$$S = \{S_1, S_2\} = \left\{ \begin{pmatrix} \theta' & 0_{n\times n} \\ -Q & I_n \end{pmatrix}, \begin{pmatrix} I_n & 2\delta S_w \\ 0_{n\times n} & \theta'^{\top} \end{pmatrix} \right\}. \tag{8}$$

Now, consider the *generalized eigenvalue* problem

$$S_1 x = \lambda S_2 x, \quad x \in \mathbb{C}^{2n}\ \lambda \in \mathbb{C}. \tag{9}$$

Unroll $S_1 X = S_2 X J$ as

$$\begin{pmatrix} \theta' & 0_{n\times n} \\ -Q & I_n \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = \begin{pmatrix} I_n & 2\delta S_w \\ 0_{n\times n} & \theta'^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} J^s & 0_{n\times n} \\ 0_{n\times n} & J^u \end{pmatrix}$$

Eigenvalues come in reciprocal pairs, so without unimodular λ we have a *stable* and *unstable* subspace.

Lemma (Structure of ARE solutions [PLS80, Lem. 1])

Let P_δ be a solution to (7) and let $X^s = [X_{11}^{\mathsf{H}}\ X_{21}^{\mathsf{H}}]^{\mathsf{H}} \in \mathbb{C}^{2n\times n}$ denote a basis for the stable eigenspace. Then, $P_\delta = X_{21} X_{11}^{-1}$ and $\theta_\delta^\star = X_{11} J^s X_{11}^{-1} \in \Theta$.

Follows by direct computation (inspired by the maximum principle).

Eigenvector decomposition is not continuous [Lax07; Kat95].. so? One can shown that $\mathrm{fl}(T^{-1}AT) = T^{-1}AT + E$ for $\|E\|_2 \lesssim \mu\kappa_2(T)\|A\|_2$ [GL13].

So, ideally we find $T \in \arg\min_{T\in \mathrm{GL}(n,\mathbb{R})} \kappa_2(T)$, *e.g.*, $T \in \mathrm{O}(n,\mathbb{R})$. We cannot simply assume symmetry, instead we use:

Lemma (Gen. real Schur decomposition [GL13, Thm. 7.7.2])

For any $A, B \in \mathbb{R}^{n\times n}$ there exist $Q, Z \in \mathrm{O}(n,\mathbb{R})$ such that $Q^{\mathsf{T}}AZ$ is upper quasi-triangular and $Q^{\mathsf{T}}BZ$ is upper triangular.

Lemma (QZ alg. [PLS80, Thm. 8a], know since the 70s [Fat69; Wil71]!)

Consider for the pair (S_1, S_2) its gen. real Schur decomposition as proposed above, then, all ARE solutions are of the form $P = U_{21}U_{11}^{-1}$, for U:

$$U = \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} = Z \begin{pmatrix} I_n \\ 0_{n\times n} \end{pmatrix} = \begin{pmatrix} Z_{11} \\ Z_{21} \end{pmatrix}. \tag{10}$$

Symplectic perspective on the computation (4/5)

If $\theta' \in \mathrm{GL}(n,\mathbb{R})$ (a.s.), then $S_1 x = \lambda S_2 x \to S_2^{-1} S_1 x = \lambda x$, but $S_2^{-1} S_1 \in \mathrm{Sp}(2n,\mathbb{R}) = \{M \in \mathbb{R}^{2n\times 2n} : M^\top \Omega M = \Omega\}$ (**unlock structure**). Specifically, we can define the curve $M : \mathbb{R} \to \mathrm{Sp}(2n,\mathbb{R})$ by

$$\delta \mapsto M(\delta) = S_2^{-1} S_1 = \begin{pmatrix} \theta' + 2\delta S_w \theta'^{-\top} Q & -2\delta S_w \theta'^{-\top} \\ -\theta'^{-\top} Q & \theta'^{-\top} \end{pmatrix}. \tag{11}$$

Use (11) to compute "*optimal*" pair $(\delta, Q(\delta))$.

Lemma (Approximately geodesic [Jon22])

For $Q(\delta) = 2\delta \theta'^\top S_w \theta'$ *and* $\delta(t) = \delta_0 - t$, $t \in [0, \delta_0)$ *then* $t \mapsto M(\delta(t)) \approx \exp(tX) M(\delta_0)$ *is* δ*-approximately geodesic.*

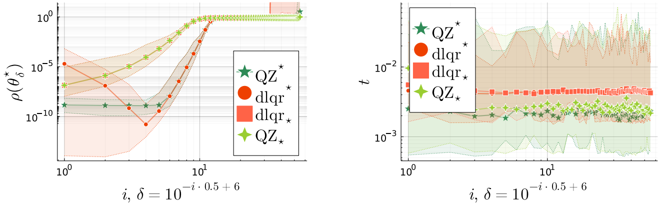
Exploit Lie group structure and define a *left-invariant* metric g via

$$\mathrm{g}_g(X,Y) = \langle X, Y\rangle_g = \langle d(L_{g^{-1}})_g(X), d(L_{g^{-1}})_g(Y)\rangle_e, \quad X, Y \in T_g\mathrm{Sp}(n,\mathbb{R}).$$

We have $\mathscr{L}_X \mathrm{g} = 0$ (Killing), leads to bounds.

Original QZ algorithm due to Moler and Stewart [MS73] (so Mathworks dlqr(·) not bad).

Shorthand notation: $Q^\star = 2^{-1}\delta^2\theta'^\top(2S_w)^{-1}\theta'$ (damped) and $Q_\star = 2\delta\theta'^\top S_w\theta'$ (approx. geodesic).



(a) Convergence for $n = 10$

(b) Time to compute $\theta^\star_\delta$ for $n = 10$

Figure 3: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ by means of the QZ algoritm or Julia's dlqr(·) routine, for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and $S_w = (1/n^2)I_n$. Each figure displays all available data.

Take away:

(I) LDPs allow for probabilistic results adapted to the process;
(II) Steady-state covariances can be linked to optimal control;
(III) Geometric thinking pays off towards fast and reliable algorithms.

Further topics:

(i) (Topological identification [JSK22]): for $\theta \in \Theta \cap \mathrm{GL}(n, \mathbb{R})$ we have

$$\mathbb{P}(\mathcal{P}(\widehat{\theta}_T) \overset{t}{\not\sim} \theta) \lesssim e^{-\mathcal{O}(\sigma_{\min}(\theta)^2 a_T)}.$$

(ii) Hyperbolic nonlinear systems and Lyapunov exponents.

Thank you! Questions?

wjongeneel.nl (slides will appear here)
wouter.jongeneel@epfl.ch

[Sim+18] Max Simchowitz et al. “Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification”. In: *Conference On Learning Theory*. 2018, pp. 439–473.

[JP19] Yassir Jedra and Alexandre Proutiere. “Sample complexity lower bounds for linear system identification”. In: *IEEE Conference on Decision and Control*. 2019, pp. 2676–2681.

[SR19] Tuhin Sarkar and Alexander Rakhlin. “Near optimal finite time identification of arbitrary linear dynamical systems”. In: *International Conference on Machine Learning*. 2019, pp. 5610–5618.

[JP20] Yassir Jedra and Alexandre Proutiere. “Finite-time identification of stable linear systems optimality of the least-squares estimator”. In: *IEEE Conference on Decision and Control*. 2020, pp. 996–1001.

[SRD21] Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. “Finite Time LTI System Identification”. In: *The Journal of Machine Learning Research* 22.1 (2021).

[Mac95] J.M. Maciejowski. “Guaranteed stability with subspace methods”. In: *Systems & Control Letters* 26.2 (1995), pp. 153–156.

[VD96] Peter Van Overschee and Bart De Moor. *Subspace Identification for Linear Systems: Theory-Implementation-Applications*. Kluwer Academic Publishers, 1996.

[LB02] S. L. Lacy and D. S. Bernstein. “Subspace identification with guaranteed stability using constrained optimization”. In: *American Control Conference*. 2002, pp. 3307–3312.

[LB03] S. L. Lacy and D. S. Bernstein. “Subspace identification with guaranteed stability using constrained optimization”. In: *IEEE Transactions on Automatic Control* 48.7 (2003), pp. 1259–1263.

[BGS08] Byron Boots, Geoffrey J Gordon, and Sajid M. Siddiqi. “A Constraint Generation Approach to Learning Stable Linear Dynamical Systems”. In: *Neural Information Processing Systems*. 2008, pp. 1329–1336.

[Tur+13] K. Turksoy et al. "Guaranteed stability of recursive multi-input-single-output time series models". In: *American Control Conference*. 2013, pp. 77–82.

[Van+00] T. Van Gestel et al. "Imposing stability in subspace identification by regularization". In: *IEEE Conference on Decision and Control*. Vol. 2. 2000, pp. 1555–1560.

[Van+01] T. Van Gestel et al. "Identification of stable models in subspace identification by using regularization". In: *IEEE Transactions on Automatic Control* 46.9 (2001), pp. 1416–1420.

[Ume+18] Jack Umenberger et al. "Maximum likelihood identification of stable linear dynamical systems". In: *Automatica* 96 (2018), pp. 280–292.

[ONV13] Francois-Xavier Orbandexivry, Yurii Nesterov, and Paul Van Dooren. "Nearest stable system using successive convex approximations". In: *Automatica* 49.5 (2013), pp. 1195–1203.

[GL17] Nicola Guglielmi and Christian Lubich. "Matrix stabilization using differential equations". In: *SIAM Journal on Numerical Analysis* 55.6 (2017), pp. 3097–3119.

[GKS19] Nicolas Gillis, Michael Karow, and Punit Sharma. "Approximating the nearest stable discrete-time system". In: *Linear Algebra and its Applications* 573 (2019), pp. 37–53.

[CGS20] Neelam Choudhary, Nicolas Gillis, and Punit Sharma. "On approximating the nearest Ω-stable matrix". In: *Numerical Linear Algebra with Applications* 27.3 (2020), e2282.

[NP20] Yu. Nesterov and V. Yu. Protasov. "Computing Closest Stable Nonnegative Matrix". In: *SIAM Journal on Matrix Analysis and Applications* 41.1 (2020), pp. 1–28.

[TK05] Hideyuki Tanaka and Tohru Katayama. "Stochastic subspace identification guaranteeing stability and minimum phase". In: *IFAC Proceedings Volumes* 38.1 (2005), pp. 910–915.

[BLW91] Sergio Bittanti, Alan J. Laub, and Jan C. Willems, eds. *The Riccati Equation*. Berlin, Heidelberg: Springer-Verlag, 1991.

[LR95] Peter Lancaster and Leiba Rodman. *Algebraic Riccati Equations*. Oxford University Press, 1995.

[JK21] Wouter Jongeneel and Daniel Kuhn. "On Topological Equivalence in Linear Quadratic Optimal Control". In: *European Control Conference*. 2021, pp. 2002–2007.

[DZ09] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2009.

[JSK23] Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. "Efficient Learning of a Linear Dynamical System With Stability Guarantees". In: *IEEE Transactions on Automatic Control* 68.5 (2023), pp. 2790–2804.

[YS09] Miao Yu and Shen Si. "Moderate deviation principle for autoregressive processes". In: *Journal of Multivariate Analysis* 100.9 (2009), pp. 1952–1961.

[VMK21] Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. "From Data to Decisions: Distributionally Robust Optimization Is Optimal". In: *Management Science* 67.6 (2021), pp. 3387–3402.

[SVK20] Tobias Sutter, Bart P. G. Van Parys, and Daniel Kuhn. "A General Framework for Optimal Data-Driven Optimization". In: *arXiv preprint, arXiv:2010.06606* (2020).

[BV21] M Bennouna and Bart PG Van Parys. "Learning and decision-making with data: Optimal formulations and phase transitions". In: *arXiv preprint arXiv:2109.06911* (2021).

[CK98] Marco C Campi and P. R. Kumar. "Adaptive linear quadratic Gaussian control: The cost-biased approach revisited". In: *SIAM Journal on Control and Optimization* 36.6 (1998), pp. 1890–1907.

[KTR19] Karl Krauth, Stephen Tu, and Benjamin Recht. "Finite-time Analysis of Approximate Policy Iteration for the Linear Quadratic Regulator". In: *Neural Information Processing Systems*. 2019, pp. 8514–8524.

[Pol86] Jan Willem Polderman. “A note on the structure of two subsets of the parameter space in adaptive control problems”. In: *Systems & Control Letters* 7.1 (1986), pp. 25–34.

[PLS80] T. Pappas, A. Laub, and N. Sandell. “On the numerical solution of the discrete-time algebraic Riccati equation”. In: *IEEE Transactions on Automatic Control* 25.4 (1980), pp. 631–641.

[Lax07] Peter D. Lax. *Linear Algebra and its Applications*. Wiley, 2007.

[Kat95] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 1995.

[GL13] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 2013.

[Fat69] A. Fath. “Computational aspects of the linear optimal regulator problem”. In: *IEEE Transactions on Automatic Control* 14.5 (1969), pp. 547–550.

[Wil71] J. Willems. “Least squares stationary optimal control and the algebraic Riccati equation”. In: *IEEE Transactions on Automatic Control* 16.6 (1971), pp. 621–634.

[Jon22] Wouter Jongeneel. “Stability via reverse I-projections, a symplectic perspective on the computation”. In: (2022). URL: http://wjongeneel.nl/pub/num_revI_2022.pdf.

[MS73] Cleve B Moler and Gilbert W Stewart. “An algorithm for generalized matrix eigenvalue problems”. In: *SIAM Journal on Numerical Analysis* 10.2 (1973), pp. 241–256.

[JSK22] Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. “Topological Linear System Identification via Moderate Deviations Theory”. In: *IEEE Control Systems Letters* 6 (2022), pp. 307–312.