A Notation Summary and Organization

A.1 Organization

After establishing notation below, the rest of the appendix is organized as follows. Appendix B provides additional algorithm details, including extension to backtracking and discussion of oracle complexity. Appendix C sketches implementation via finite-sample, finite-horizon oracles; notably, Appendix C.2 describes implementation with an oracle which does not require direct access to system states, but which rather "subsamples" outputs at various time steps.

Appendix D provides further discussion on the somewhat-nonstandard controllability assumption, Assumption 2.4, and demonstrates it holds generically. Appendix E contains assorted results about our assumptions and various other control-theoretic considerations. Appendix E also contains the proofs of various other supporting results, mainly on the characterization of optimal policies and their informativity. It also shows that random (continuous) initializations are informative with probability one. Finally, Appendix F provides further details for the various counterexamples presented in Section 3.

Part II turns to the proof of our main result, Theorem 2, as well as its more qualitative statement, Theorem 1. The high level proofs are given in Appendix G, with the following appendices establishing the main constituent results. Specifically, Appendix H establishes the proofs for the `DCL` framework and gradient descent for general objective functions. Appendix I substantiates the framework, and exhibits a `DCL` for our regularized loss for the `OE` problem, using a convex reformulation due to Scherer [1995]. Appendix J then establishes that informativity translates into bounds on the norm of the solutions to Lyapunov equations involving the closed loop matrix $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$. This is one of our most technically innovative arguments. Finally, Appendix K upper bounds the norms of various first- and second-order derivatives, via somewhat standard arguments.

A.2 Notation

We let lower case variables in script font $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ denote abstract parameters for optimization; standard vectors $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ are reserved for random variables and/or dynamical quantities. Matrices are denoted in bold, e.g $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. For vectors, $\|\mathbf{x}\|$ denotes the Euclidean norm, $\|\mathbf{X}\|$ denotes the matrix operator norm and $\|\mathbf{X}\|_{\mathrm{F}}$, the Frobenius norm.

We let $\mathcal{S}^{n-1}$ denote the unit sphere in $\mathbb{R}^n$. We denote the set of symmetric $n \times n$ matrices as $\mathbb{S}^n$; the set of nonstrictly positive semidefinite (PSD) matrices as $\mathbb{S}^n_+$, strictly positive definite (PD) matrices as $\mathbb{S}^n_{++}$, and invertible matrices as $\mathbb{GL}(n)$. Given $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{S}^n$, we let $\mathbf{X}_1 \preceq \mathbf{X}_2$ denote nonstrict PSD inequality, with $\mathbf{X}_1 \prec \mathbf{X}_2$ denoting strict inequality. Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\exp(\mathbf{A})$ denotes the matrix exponential. For $\mathbf{A}$ with real eigenvalues, $\lambda_i(\mathbf{A}), i = 1, \dots, n$ denotes its eigenvalues in descending order, with $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A}) = \lambda_n(\mathbf{A})$; when $\mathbf{A}$ has complex eigenvalues, $\lambda_i(\mathbf{A})$ are arranged in an arbitrary order. For general rectangular matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\sigma_i(\mathbf{A}), i = 1, \dots, n$ denotes its singular values in descending order. We use $\mathbf{I}_n$ to denote the identity matrix with dimension $n \times n$, and omit n when the dimension is clear from context.

We use parentheses to denote parameter concatenation: e.g. $\bar{\mathbf{X}} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) \in \mathbb{R}^{n_1 \times m_1} \times \mathbb{R}^{n_2 \times m_2} \times \mathbb{R}^{n_3 \times m_3}$ for $\mathbf{X}_i \in \mathbb{R}^{n_i \times m_i}$, and we define Euclidean norms of concatenation in the natural way (e.g. $\|\bar{\mathbf{X}}\|_{\ell_2} = \sqrt{\sum_i \|\mathbf{X}_i\|_{\mathrm{F}}^2}$ for the previous example $\bar{\mathbf{X}} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$).

A.3 Further Notational Review

In this section, we review some of the key notation used throughout.

System Variables	Definition
$\mathbf{x}(t)$	system state, dimension n
$\mathbf{y}(t)$	system observation, dimension m
$\mathbf{z}(t)$	system output, dimension p
$\mathbf{w}(t)$	process noise, dimension n
$\mathbf{v}(t)$	observation noise, dimension m

System Parameters	**Definition**
$\mathbf{A}$	system state transition matrix, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{w}(t)$.
$\mathbf{C}$	system observation matrix, $\mathbf{z}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{e}(t)$
$\mathbf{G}$	system output matrix, $\mathbf{z}(t) = \mathbf{G}\mathbf{x}(t)$
$\mathbf{W}_1$	process noise covariance, $\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{W}_1)$
$\mathbf{W}_2$	output noise covariance, $\mathbf{v}(t) \sim \mathcal{N}(0, \mathbf{W}_2)$

Nominal System Quantitites	**Definition**
$\mathbf{\Sigma}_{11,\mathrm{sys}}$	steady-state system covariance ($(1,1)$-block of any $\mathbf{\Sigma}_{\mathsf{K}}$, see below)
$\mathbf{P}_\star$	Solution to Riccati equation (Eq. (2.5))
$\sigma_\star$	$\lambda_{\min}(\mathbf{P}_\star)$ (strictly positive due to Lemma 3.3)
$\mathbf{L}_\star$	Optimal Kalman Gain (Eq. (2.5))
$C_{\mathtt{sys}}$	Upper bound on relevant problem parameters, Eq. (3.3) $\max\left\{\lVert\mathbf{A}\rVert, \lVert\mathbf{C}\rVert, \lVert\mathbf{G}\rVert, \lVert\mathbf{W}_2\rVert, \lVert\mathbf{W}_2^{-1}\rVert, \lVert\mathbf{W}_1^{-1}\rVert, \lVert\mathbf{\Sigma}_{11,\mathrm{sys}}\rVert, \sigma_\star^{-1}\right\}$

Policy Parameters	**Definition**
$\mathsf{K} = (\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$	parametrization of policy $\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}(t) = \mathbf{A}_{\mathsf{K}}\hat{\mathbf{x}}(t) + \mathbf{B}_{\mathsf{K}}\mathbf{y}(t), \quad \hat{\mathbf{z}}(t) = \mathbf{C}_{\mathsf{K}}\hat{\mathbf{x}}(t)$
$\mathsf{K}_\star$	cannonical realization of optimal policy $\mathsf{K}_\star = (\mathbf{A} - \mathbf{L}_\star\mathbf{C},\ \mathbf{L}_\star,\ \mathbf{G})$
$\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$	output estimation loss (Eq. (1.3))
$\mathsf{Sim}_{\mathbf{S}}(\mathsf{K})$	Similarity transform $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}) \mapsto (\mathbf{S}\mathbf{A}_{\mathsf{K}}\mathbf{S}^{-1}, \mathbf{S}\mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}\mathbf{S}^{-1})$.
$\mathbf{A}_{\mathrm{cl},\mathsf{K}}$	closed-loop system matrix (Eq. (2.1))
$\mathbf{\Sigma}_{\mathsf{K}}$	steady state covariance (Eq. (2.2))
$\mathbf{W}_{\mathrm{cl},\mathsf{K}}$	closed-loop noise matrix (Eq. (2.2))
$\mathbf{\Sigma}_{11,\mathrm{sys}}, \mathbf{\Sigma}_{12,\mathsf{K}}, \mathbf{\Sigma}_{22,\mathsf{K}}$	block-parition of $\mathbf{\Sigma}_{\mathsf{K}}$ (Eq. (2.3)) (note $\mathbf{\Sigma}_{11,\mathrm{sys}}$ does not depend on K.)
$\mathbf{\Sigma}$	typical variable name for matrix $\mathbf{\Sigma} \in \mathbb{S}_+^{2n}$
$\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{12}, \mathbf{\Sigma}_{22}$	block-parition of arbitrary $\mathbf{\Sigma}$ (Eq. (2.3))

Policy Classes	**Definition**
$\mathcal{K}_{\mathtt{stab}}$	class of all stabilizing policies ($\mathcal{K}_{\mathtt{stab}} := \{\mathsf{K} \in \mathcal{K}_{\mathtt{stab}} : \mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is Hurwitz$\}$)
$\mathcal{K}_{\mathtt{ctrb}}$	class of all controllable policies ($\mathcal{K}_{\mathtt{ctrb}} := \{\mathsf{K} \in \mathcal{K}_{\mathtt{stab}} : \mathbf{\Sigma}_{22,\mathsf{K}} \succ 0\}$)
$\mathcal{K}_{\mathtt{info}}$	class of all informative policies ($\mathcal{K}_{\mathtt{info}} := \{\mathsf{K} \in \mathcal{K}_{\mathtt{stab}} : \mathrm{rank}(\mathbf{\Sigma}_{12,\mathsf{K}}) = n\}$) (alternately, $\mathcal{K}_{\mathtt{info}} := \{\mathsf{K} \in \mathcal{K}_{\mathtt{ctrb}} : \mathrm{rank}(\mathbf{Z}_{\mathsf{K}} = n\}$)
$\mathcal{K}_{\mathtt{opt}}$	class of all optimal policies (similarity transforms of $\mathsf{K}_\star$)

Informativity & Reconditioning	**Definition**
$\mathbf{Z}_{\mathsf{K}}$	Informativity Matrix ($\mathbf{Z}_{\mathsf{K}} := \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^{\top}$)
$\mathcal{R}_{\mathtt{info}}(\mathsf{K})$	Informativity Regularizer, ($= \mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$ if $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$, ∞ otherwise)
$\mathcal{L}_\lambda(\mathsf{K})$	Regularized Loss ($\mathcal{L}_\lambda(\mathsf{K}) = \mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \lambda \cdot \mathcal{R}_{\mathtt{info}}(\mathsf{K})$)
$\mathsf{recond}(\mathsf{K})$	Reconditioning matrix, Eq. (3.2) $\mathsf{recond}(\mathsf{K}) := \mathsf{Sim}_{\mathbf{S}}(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$, where $\mathbf{S} = \mathbf{\Sigma}_{22,\mathsf{K}}^{-1/2}$

$\texttt{DCL}$ **Notation**	**Definition**
$\mathbb{R}$	extended reals, $\mathbb{R} \cup \{\infty\}$
f	function in question for $\texttt{DCL}$, argument takes argument $\boldsymbol{x} \in \mathbb{R}^d$
$\mathsf{dom}(f)$	given $f : \mathbb{R}^d \to \bar{\mathbb{R}}$, $\{\boldsymbol{x} : f(\boldsymbol{z}) \neq \infty\}$
$\inf(f)$	$\inf_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$
$\mathcal{K}$	typical notation for $\mathcal{K} \subset \mathbb{R}^d$
$\mathscr{C}^k(\mathcal{K})$	functions k-times differentiable on open set containing $\mathcal{K}$
$f_{\texttt{cvx}}$	convex function in $\texttt{DCL}$ takes in argument $\boldsymbol{z} \in \mathbb{R}^{d_z}$
$f_{\texttt{lft}}$	lifting function in $\texttt{DCL}$ takes in argument $(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathbb{R}^d \times \mathbb{R}^{d_\xi}$
Φ	Reparametrization $\mathbb{R}^{d+d_\xi} \to \mathbb{R}^{d_z}$
$(\Phi, f_{\texttt{cvx}}, f_{\texttt{lft}})$	triple defining a $\texttt{DCL}$ (Definition 4.1)
$\alpha_{\texttt{DCL}}$	weak-PL constant for $\texttt{DCL}$ (Theorem 3)

B Additional Algorithmic Details

B.1 Backtracking

In general, the smoothness constants may be difficult to compute in a model free fashion. We show that simple modification of our algorithm based on backtracking line search also inherits provable convergence guarantees. To this end, let $\mathcal{S}_{\mathrm{bkt}}$ be finite set of step sizes (to ensure the algorithm is always well defined, we assume that $\mathcal{S}_{\mathrm{bkt}}$ contains 0.) It is common practice to choosen $\mathcal{S}_{\mathrm{bkt}}$ to contain geometrically decreasing sizes (see,e.g. Wright et al. [1999, Chapter 3]). To choose the step sizes η_t, we search over $\mathcal{S}_{\mathrm{bkt}}$ to find the step which minimizes the objective subject to the constraint that $\mathbf{\Sigma}_{22,\mathsf{K}}$ remains well-conditioned, i.e.

$$\mathsf{K}_{s+1} = \widetilde{\mathsf{K}}_s - \eta_s \nabla_s, \text{ where } \nabla_s = \nabla \mathcal{L}_\lambda(\widetilde{\mathsf{K}}_s) \text{ and} \tag{B.1}$$

$$\eta_s \in \arg\min_{\eta \in \mathcal{S}_{\mathrm{bkt}}} \left\{ \mathcal{L}_\lambda(\mathsf{K}) : \tfrac{1}{2}\mathbf{I}_n \preceq \mathbf{\Sigma}_{22,\mathsf{K}} \preceq \frac{3}{2}\mathbf{I}_n, \quad \text{where } \mathsf{K} := \widetilde{\mathsf{K}}_s - \eta \nabla_s \right\}. \tag{B.2}$$

Note that since $0 \in \mathcal{S}_{\mathrm{bkt}}$ and $\mathsf{K} = \widetilde{\mathsf{K}}_s$ has $\mathbf{\Sigma}_{22,\mathsf{K}} = \mathbf{I}_n$, the backtracking condition is at the very least met with $\eta_s = 0$. The following modifies Theorem 2, and is proven in Appendix G.4.

Theorem 2a. *Fix $\lambda > 0$, $\mathsf{K}_0 \in \mathcal{K}_{\texttt{info}}$. There are terms $\mathcal{C}_1, \mathcal{C}_2 \geq 1$, which are at most polynomial in $n, m, C_{\texttt{sys}}, \lambda, \lambda^{-1}$ and $\mathcal{L}_\lambda(\mathsf{K}_0)$ such, if $\mathcal{S}_{\mathrm{bkt}}$ contains a step size $\eta > 0$ satisfying stepsize $\eta \leq \frac{1}{\mathcal{C}_1}$, then the iterates produced by Algorithm 2 satisfy*

$$\mathcal{L}_{\texttt{OE}}(\mathsf{K}_s) - \min_{\mathsf{K}} \mathcal{L}_{\texttt{OE}}(\mathsf{K}) \leq \mathcal{L}_\lambda(\mathsf{K}_s) - \min_{\mathsf{K}} \mathcal{L}_\lambda(\mathsf{K}) \leq \frac{\mathcal{C}_2}{\eta} \cdot \frac{1}{s}, \quad \forall s \geq 1.$$

Algorithm 2 IR-PG with backtracking

```
Input: Initial $\mathsf{K}_0 \in \mathcal{K}_{\texttt{info}}$, step size $\eta > 0$, regularization parameter $\lambda > 0$
    % Define $\mathcal{L}_\lambda(\mathsf{K}) := \mathcal{L}_{\texttt{OE}}(\cdot) + \lambda \mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$
for each iteration $s = 0, 1, 2, \ldots$ do
    Recondition $\widetilde{\mathsf{K}}_s = \mathsf{recond}(\mathsf{K}_s)$, where $\mathsf{recond}(\cdot)$ is defined in Eq. (3.2).
    Compute $\nabla_s = \nabla \mathcal{L}_\lambda(\widetilde{\mathsf{K}}_s)$.
    Update $\mathsf{K}_{s+1} \leftarrow \widetilde{\mathsf{K}}_s - \eta_s \nabla_s$, where $\eta_s$ is the backtracking step from Eq. (B.2).
```

B.2 Oracle Complexity

At each iteration, one can compute the derivative of $\mathcal{L}_\lambda$ using one call to $\mathsf{orac}_{\mathrm{eval}}$ (which evaluates $\mathcal{L}_{\texttt{OE}}(\mathsf{K}_s)$ and $\mathbf{\Sigma}_{\mathsf{K}}$), and one call to $\mathsf{orac}_{\mathrm{grad}}$, which computes the gradients of these quantities. This is true because $\nabla \mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$ admits a closed form in terms of $\mathbf{\Sigma}_{\mathsf{K}}$ and its gradient. The balancing step also requires only evaluation $\mathbf{\Sigma}_{\mathsf{K}}$, and can use an evaluation query called for the gradient. Thus, gradient descent variant (Algorithm 1) calls one evaluation and one gradient oracle per iteration. With backtracking (Algorithm 2), the backtracking step requires an evaluation query for all $|\mathcal{S}_{\mathrm{bkt}}|$ filters of the form $\widetilde{\mathsf{K}}_s - \eta\nabla_s$. In total, therefore, each iteration uses 1 call to $\mathsf{orac}_{\mathrm{grad}}$, and $|\mathcal{S}_{\mathrm{bkt}}| + 2$ calls to $\mathsf{orac}_{\mathrm{eval}}$.

C Further Details on Evaluation Oracle

Given that our primary focus is on understanding the *landscape properties* of the OE problem, we leave the precise details of finite sample considerations to future work. In this section, we provide brief remarks on how one might approximate the cost and gradients from finitely-many, finite-horizon samples. Subsequently, we describe how to implement cost and gradient evaluations without direct access to the state covariance matrix $\mathbf{\Sigma}_{\mathsf{K}}$ assumed in the body of the work.

C.1 Finite-sample considerations

Gradient descent with inexact gradients. In the finite-sample regime, one uses statistical approximations to the gradients and, in the case where the stepsize is determined by line search, function evaluations. A straightforward modification of our generic analysis of gradient descent under weak-PL, Proposition G.2, can establish robustness to these inexact queries. Robustness of gradient descent to error is well-known in the literature, even in generic problem settings (see Scaman and Malherbe [2020]; this is also related to the stability properties established in Hardt et al. [2016]).

Time discretization. Using digital controllers, one must implement the filter in discrete time. Given a discretization incremement δ, the (Euler) discretized filter dynamics for filter $\mathsf{K} = (\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$ is

$$\hat{\mathbf{z}}_{k;\delta} = \mathbf{C}_{\mathsf{K}}\hat{\mathbf{x}}_{k;\delta}, \quad \hat{\mathbf{x}}_{k+1;\delta} = (\mathbf{I}_n + \delta\mathbf{A}_{\mathsf{K}})\hat{\mathbf{x}}_{k;\delta} + \mathbf{B}_{\mathsf{K}}\mathbf{y}(k\delta), \quad \hat{\mathbf{x}}_{0;\delta} = 0. \tag{C.1}$$

Finite-horizon, finite-sample losses. Given independent trials indexed by $i = 1, 2, \dots, N$, and T such that $H = T/\delta$ is integral, we set

$$\hat{\mathcal{L}}_{\mathrm{OE}}(\mathsf{K}) := \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{z}^{(i)}(H\delta) - \hat{\mathbf{z}}^{(i)}_{H;\delta}\|^2$$

$$\hat{\mathbf{\Sigma}}_{\mathsf{K}} = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \mathbf{x}^{(i)}(H\delta) \\ \hat{\mathbf{x}}^{(i)}_{H;\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(i)}(H\delta) \\ \hat{\mathbf{x}}^{(i)}_{H;\delta} \end{bmatrix}^{\top}.$$

Using stability of the filter and nominal system and well-known properties of the Euler discretization,

$$\begin{aligned} |\mathbb{E}[\hat{\mathcal{L}}_{\mathrm{OE}}(\mathsf{K})] - \mathcal{L}_{\mathtt{OE}}(\mathsf{K})| &= +\mathcal{O}\left(\delta + e^{-\Omega(\delta H)}\right) \\ \|\mathbb{E}[\hat{\mathbf{\Sigma}}_{\mathsf{K}}] - \mathbf{\Sigma}_{\mathsf{K}}\| &= \mathcal{O}\left(\delta + e^{-\Omega(\delta H)}\right), \end{aligned} \tag{C.2}$$

which can be made arbitrarily close to being unbiased as $\delta \to 0$ and $\delta H \to \infty$. Here, the term $\mathcal{O}(\delta)$ comes from a standard error analysis of the Euler discretization (c.f. e.g Iserles [2008, Theorem 1.1]), and the exponentially decaying term $e^{-\Omega(\delta H)}$ from standard mixing time arguments Yu [1994]. Above, we surpress various problem dependent constants, including terms polynominal in dimension. By standard concentration inequalities (e.g. Tropp [2015]), we can obtain finite-sample concentration with high probability:

$$\begin{aligned} |\hat{\mathcal{L}}_{\mathrm{OE}}(\mathsf{K}) - \mathcal{L}_{\mathtt{OE}}(\mathsf{K})| &= \mathcal{O}\left(\delta + e^{-\Omega(\delta H)}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{N}}\right) \\ \|\mathbb{E}[\hat{\mathbf{\Sigma}}_{\mathsf{K}}] - \mathbf{\Sigma}_{\mathsf{K}}\| &= \mathcal{O}\left(\delta + e^{-\Omega(\delta H)}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{N}}\right). \end{aligned} \tag{C.3}$$

In particular, for δ sufficiently small and δN sufficiently large, invertibility of $\boldsymbol{\Sigma}_{22,\mathsf{K}}$ (i.e. $\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}}$) implies that $\hat{\boldsymbol{\Sigma}}_{22,\mathsf{K}}$ is invertible with high probability. We may then define:

$$\hat{\mathcal{R}}_{\texttt{info}}(\mathsf{K}) := \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{12,\mathsf{K}}(\hat{\boldsymbol{\Sigma}}_{22,\mathsf{K}})^{-1}\hat{\boldsymbol{\Sigma}}_{12,\mathsf{K}}^{\top}),$$

which yields the estimated regularized loss

$$\hat{\mathcal{L}}_{\lambda}(\mathsf{K}) := \hat{\mathcal{L}}_{\texttt{OE}}(\mathsf{K}) + \lambda \cdot \hat{\mathcal{R}}_{\texttt{info}}(\mathsf{K}).$$

Note that, since K passes nonlinearly into $\hat{\mathcal{R}}_{\texttt{info}}(\mathsf{K})$ and $\hat{\mathcal{L}}_{\lambda}(\mathsf{K})$, these losses are in general biased estimates of $\mathcal{R}_{\texttt{info}}(\mathsf{K})$ and $\mathcal{L}_{\lambda}(\mathsf{K})$. Invoking Eq. (C.3), together with some standard matrix (and matrix-inverse) perturbation arguments,

$$\begin{aligned}
&|\hat{\mathcal{L}}_{\lambda}(\mathsf{K}) - \mathcal{L}_{\lambda}(\mathsf{K})| \\
&\quad = |\hat{\mathcal{L}}_{\texttt{OE}}(\mathsf{K}) - \mathbb{E}[\hat{\mathcal{L}}_{\texttt{OE}}(\mathsf{K})] + \lambda \cdot \big(\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{12,\mathsf{K}}(\hat{\boldsymbol{\Sigma}}_{22,\mathsf{K}})^{-1}\hat{\boldsymbol{\Sigma}}_{12,\mathsf{K}}^{\top}) - \mathrm{tr}(\boldsymbol{\Sigma}_{12,\mathsf{K}}\boldsymbol{\Sigma}_{22,\mathsf{K}}^{-1}\boldsymbol{\Sigma}_{12,\mathsf{K}}^{\top})\big)| \\
&\quad \le \mathcal{O}\left(\delta + e^{-\Omega(\delta H)}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{N}}\right) + \text{(lower order terms)}.
\end{aligned}$$

Cost evaluations. In light of the above discussion, $\hat{\mathcal{L}}_{\lambda}(\mathsf{K})$ can be used to evaluate $\mathcal{L}_{\lambda}(\mathsf{K})$ provided the step size δ is sufficiently small, horizon H sufficiently large, and sample size N sufficiently large. This minimics the findings of Fazel et al. [2018], Mohammadi et al. [2021], Malik et al. [2019] in various related settings.

Gradient evaluations. To estimate gradients of the $\mathcal{L}_{\lambda}$, two strategies are possible. One can use the zeroth-order gradient estimator [Flaxman et al., 2005], where one estimates the gradient by evaluating

$$\hat{\nabla}\mathcal{L}_{\lambda}(\mathsf{K}) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{\mathsf{N}(r)}\hat{\mathcal{L}}_{\lambda}(\mathsf{K} + r\mathsf{U}^{(j)})\mathsf{U}^{(j)},$$

where $\mathsf{U}^{(j)} = (\mathsf{U}_A^{(j)}, \mathsf{U}_B^{(j)}, \mathsf{U}_C^{(j)})$ are i.i.d. parameter perturbations from a suitable, zero-mean distribution parameters (e.g. uniform on perturbation on the unit-Frobenius ball ($\|\mathsf{U}_A\|_{\mathrm{F}}^2 + \|\mathsf{U}_B\|_{\mathrm{F}}^2 + \|\mathsf{U}_C\|_{\mathrm{F}}^2 = 1$)), r a user-defined smoothing parameter that scales the perturbation, and $\frac{1}{\mathsf{N}(r)}$ a normalization constant. As in previous work, (Fazel et al. [2018], Malik et al. [2019], Mohammadi et al. [2021]), one can argue that this yields an estimator of the gradient with polynomial sample complexity. As in prior work, r must be chosen sufficiently small so that the perturbations do not render $\mathbf{A}_{\mathsf{K}}$ unstable.

Because we consider a filtering problem, rather than a control problem, it is possible to directly compute the gradients of $\hat{\mathcal{L}}_{\lambda}(\mathsf{K})$ by differentiating through the discretizated filter dynamics in Eq. (C.1) (provided $\hat{\boldsymbol{\Sigma}}_{22,\mathsf{K}} \succ 0$, so that the loss is defined and differentiable). Similar concentration techniques can be deployed to establish the accuracy of this estimator as well.

C.2 Implementation without access to system states

We now turn to the implementation of our algorithm without direct access to system states. For simplicity, this example considers continuous-time, infinite-horizon, and infinite-sample cost evaluations (and gradients). In essence, we provide a reduction to the oracle described in the main text.

Subsampled covariance matrix. In the subsampling oracle, we have access to evaluations and gradients of the following subsampled covariance matrix:

$$\bar{\boldsymbol{\Sigma}}_{\mathsf{K},\mathbf{t}} := \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T \begin{bmatrix}\mathbf{y}(t+t_1)\\ \mathbf{y}(t+t_2)\\ \dots\\ \mathbf{y}(t+t_k)\\ \hat{\mathbf{x}}_{\mathsf{K}}(t)\end{bmatrix}\begin{bmatrix}\mathbf{y}(t+t_1)\\ \mathbf{y}(t+t_2)\\ \dots\\ \mathbf{y}(t+t_k)\\ \hat{\mathbf{x}}_{\mathsf{K}}(t)\end{bmatrix}^{\top} \mathrm{d}t\right]. \in \mathbb{R}^{(k+1)n\times(k+1)n} \tag{C.4}$$

Here, $\mathbf{t} = (t_1, t_2, \dots, t_k)$ is a vector of increasing sampling times $0 = t_1 < t_2 < \cdots < t_k$. Introduce, for the sake of analysis, the observability matrix

$$\mathbf{V}_{\mathbf{t}} := \begin{bmatrix} \mathbf{C}\exp(t_1\mathbf{A}) \\ \mathbf{C}\exp(t_2\mathbf{A}) \\ \dots \\ \mathbf{C}\exp(t_k\mathbf{A}) \end{bmatrix},$$

where $\exp(\cdot)$ denotes the matrix exponential. We make the following assumption.

Assumption C.1. We assume that $\mathbf{t}$ is selected so that the observability matrix is full-rank: $\operatorname{rank}(\mathbf{V}_{\mathbf{t}}) = n$.

Importantly, Assumption C.1 holds *generically* when $(\mathbf{A}, \mathbf{C})$ is observable, as per Assumption 2.2. The following lemma makes this precise:

Lemma C.1. *Suppose* $(\mathbf{A}, \mathbf{C})$ *is observable, and that* $k \geq n$*. Then, the* $\{\mathbf{t} \in \mathbb{R}^k : \operatorname{rank}(\mathbf{V}_{\mathbf{t}}) < n\}$ *has Lebesgue measure zero. In particular, if* $\mathbf{t}$ *are drawn from a distribution with density with respect to the Lebesgue measure (e.g., drawn* k *points uniformly* $[0,1]$*, and order them in increasing order), then* $\mathbb{P}[\operatorname{rank}(\mathbf{V}_{\mathbf{t}}) = n] = 1$.

We establish the lemma at the end of the section.

Subsampled losses. One can compute that $\bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t}}$ can be partitioned in the following form

$$\bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t}} = \begin{bmatrix} \bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},11} & \bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},12} \\ \bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},12}^\top & \bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},22} \end{bmatrix} = \begin{bmatrix} * & \mathbf{V}_{\mathbf{t}}\mathbf{\Sigma}_{\mathsf{K},12} \\ \mathbf{\Sigma}_{\mathsf{K},12}^\top \mathbf{V}_{\mathbf{t}}^\top & \mathbf{\Sigma}_{\mathsf{K},22} \end{bmatrix}, \tag{C.5}$$

where $*$ is immaterial to the following discussion. We define

$$\bar{\mathbf{Z}}_{\mathsf{K}} := \bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},12}\bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},22}^{-1}\bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t},12}^\top \in \mathbb{S}_+^{nk}. \tag{C.6}$$

We define the subsample regularized loss as follows:

$$\mathcal{L}_{\lambda,\mathtt{sub}}(\mathsf{K}) = \mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \lambda\mathcal{R}_{\mathtt{sub}}(\mathsf{K}), \quad \mathcal{R}_{\mathtt{sub}}(\mathsf{K}) := \sum_{i=1}^{n} \lambda_i(\bar{\mathbf{Z}}_{\mathsf{K}})^{-1}. \tag{C.7}$$

Notice that $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$ is reminiscent of the regularizer $\mathcal{R}_{\mathtt{info}}(\mathsf{K})$ uses the state covariance oracle. It is clear that $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$, and thus $\mathcal{L}_{\lambda,\mathtt{sub}}(\mathsf{K})$ can be evaluated for any K. These quantities to do need knowledge of $\mathbf{V}_{\mathbf{t}}$ to be evaluated.

Differentiability of $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$. We now show that $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$ is $\mathscr{C}^2$ for $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$. Introduce the matrix $\mathcal{P}$ to be any orthogonal projection matrix from the space spanned by the image of $\mathbf{V}_{\mathbf{t}}$ (which is rank n) to $\mathbb{R}^n$. Define $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{Z}}_{\mathsf{K}}$ by

$$\widetilde{\mathbf{V}} = \mathcal{P}\mathbf{V}_{\mathbf{t}}, \quad \widetilde{\mathbf{Z}}_{\mathsf{K}} = \mathcal{P}\bar{\mathbf{Z}}_{\mathsf{K}}\mathcal{P}^\top.$$

Since the row (and hence column) space of the symmetric matrix $\bar{\mathbf{Z}}_{\mathsf{K}}$ is equal to the column space of $\mathbf{V}_{\mathbf{t}}$, which is precisely the row space of $\mathcal{P}$, we see that

$$\lambda_i(\widetilde{\mathbf{Z}}_{\mathsf{K}}) = \lambda_i(\bar{\mathbf{Z}}_{\mathsf{K}}), \quad i \in [n],$$

so that

$$\mathcal{R}_{\mathtt{sub}}(\mathsf{K}) = \operatorname{tr}[\widetilde{\mathbf{Z}}_{\mathsf{K}}^{-1}]. \tag{C.8}$$

From Eq. (C.5), we can compute that $\widetilde{\mathbf{Z}}_{\mathsf{K}}$ is related to $\mathbf{Z}_{\mathsf{K}}$ via conjugation by $\widetilde{\mathbf{V}}$:

$$\widetilde{\mathbf{Z}}_{\mathsf{K}} = \widetilde{\mathbf{V}}\mathbf{Z}_{\mathsf{K}}\widetilde{\mathbf{V}}^\top,$$

so that

$$\mathcal{R}_{\mathtt{sub}}(\mathsf{K}) = \operatorname{tr}[(\widetilde{\mathbf{V}}\mathbf{Z}_{\mathsf{K}}\widetilde{\mathbf{V}}^\top)^{-1}],$$

showing that $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$ is $\mathscr{C}^2$. Thus, the subsampled oracle model affords both evaluations and derivatives of $\mathcal{L}_{\lambda,\mathtt{sub}}(\mathsf{K})$. Note that $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$ can be evaluated without knowledge of $\mathcal{P}$ and $\mathbf{V}_{\mathbf{t}}$ by using the original definition in Eq. (C.7).

Remark C.1. A similar approach to the computation above can be used to derive a closed-form expression for the derivative of $\mathcal{R}_{\mathtt{sub}}(\mathsf{K})$ in terms of the derivatives of *only* $\bar{\mathbf{\Sigma}}_{\mathsf{K},\mathbf{t}}$, and *not* in terms of the observation matrix $\mathbf{V}_{\mathbf{t}}$ (which we do not have access to in this model).

In view of the identity $\mathcal{R}_{\mathtt{sub}}(\mathsf{K}) = \mathrm{tr}[\widetilde{\mathbf{Z}}_{\mathsf{K}}^{-1}]$ established above, we see that optimizing

$$\mathcal{L}_{\lambda,\mathtt{sub}}(\mathsf{K}) = \mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \mathrm{tr}[\widetilde{\mathbf{Z}}_{\mathsf{K}}^{-1}] \tag{C.9}$$

is equivalent to optimizing the state-covariance oracle loss $\mathcal{L}_{\lambda}(\mathsf{K})$ on the following similarity-transformed realization of the dynamics

$$\begin{aligned} &\tfrac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathbf{x}}(t) = \widetilde{\mathbf{A}}\widetilde{\mathbf{x}}(t) + \widetilde{\mathbf{w}}(t), \quad \mathbf{y}(t) = \widetilde{\mathbf{C}}\widetilde{\mathbf{x}}(t) + \mathbf{v}(t), \quad \mathbf{z}(t) = \widetilde{\mathbf{G}}\widetilde{\mathbf{x}}(t), \quad \widetilde{\mathbf{x}}(0) = 0, \\ &\quad \widetilde{\mathbf{w}}(t) \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \widetilde{\mathbf{W}}_1), \quad \mathbf{v}(t) \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{W}_2), \end{aligned} \tag{C.10}$$

where $\widetilde{\mathbf{A}} = \widetilde{\mathbf{V}}\mathbf{A}\widetilde{\mathbf{V}}^{-1}$, $\widetilde{\mathbf{C}} = \mathbf{C}\widetilde{\mathbf{V}}^{-1}$, $\widetilde{\mathbf{G}} = \mathbf{G}\widetilde{\mathbf{V}}^{-1}$, and $\widetilde{\mathbf{W}}_1 = \widetilde{\mathbf{V}}\mathbf{W}_1$ (it follows from Assumption C.1 and the definition of the projection $\mathcal{P}$ that $\widetilde{\mathbf{V}}$ is nonsingular). Indeed, $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ is invariant under similarity transformation of the true system, and if $\widetilde{\mathbf{\Sigma}}_{\mathsf{K}}$ is the associated covariance matrix (partitioned in the standard way), then we can verify

$$\widetilde{\mathbf{Z}}_{\mathsf{K}} = \widetilde{\mathbf{\Sigma}}_{\mathsf{K},12}\widetilde{\mathbf{\Sigma}}_{\mathsf{K},22}^{-1}\widetilde{\mathbf{\Sigma}}_{\mathsf{K},12}^{\top}.$$

Therefore, via this similarity-transformation, optimizing $\mathcal{L}_{\lambda,\mathtt{sub}}(\mathsf{K})$ on the dynamics Eq. (1.1) inherits all the guarantees of optimizing the loss $\mathcal{L}_{\lambda}(\mathsf{K})$ on the tilde-dynamics in Eq. (C.10).

We complete the section by providing the proof of Lemma C.1.

Proof of Lemma C.1. The proof is divided into two steps. First, we exhibit a $\mathbf{t}$ for which $\mathrm{rank}(\mathbf{V}_{\mathbf{t}}) = n$; then we use an analytic continuation argument to establish that, if such a $\mathbf{t}$ exists, then $\mathrm{rank}(\mathbf{V}_{\mathbf{t}}) = n$ Lebesgue almost everywhere.

Existence of a $\mathbf{t}$ for which $\mathrm{rank}(\mathbf{V}_{\mathbf{t}}) = n$. Without loss of generality, we may assume that $k = n$. Fix $\delta > 0$, and consider $t_i = (i-1)\delta$. Expanding the matrix exponential, we have

$$\begin{aligned} \mathbf{V}_{\mathbf{t}} &:= \mathbf{C} \cdot \begin{bmatrix} \mathbf{I}_m \\ \exp(\delta\mathbf{A}) \\ \exp(2\delta\mathbf{A}) \\ \dots \\ \exp((n-1)\delta\mathbf{A}) \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{I}_m & \delta\mathbf{I}_m & \frac{\delta^2}{2!}\mathbf{I}_m & \dots & \frac{\delta^{n-1}}{(n-1)!}\mathbf{I}_m \\ \mathbf{I}_m & 2\delta\mathbf{I}_m & \frac{(2\delta)^2}{2!}\mathbf{I}_m & \dots & \frac{(2\delta)^{n-1}}{(n-1)!}\mathbf{I}_m \\ \dots & & & & \\ \mathbf{I}_m & (n-1)\delta\mathbf{I}_m & \frac{((n-1)\delta)^2}{2!}\mathbf{I}_m & \dots & \frac{((n-1)\delta)^{n-1}}{(n-1)!}\mathbf{I}_m \end{bmatrix}}_{\mathcal{T}_{n,\delta}} \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \mathbf{C}\mathbf{A}^2 \\ \dots \\ \mathbf{C}\mathbf{A}^{n-1} \end{bmatrix}}_{\mathcal{O}_n} + \underbrace{\begin{bmatrix} 0 \\ \mathbf{C}\sum_{i\ge n}\frac{(\delta\mathbf{A})^i}{i!} \\ \mathbf{C}\sum_{i\ge n}\frac{(2\delta\mathbf{A})^i}{i!} \\ \dots \\ \mathbf{C}\sum_{i\ge n}\frac{((n-1)\delta\mathbf{A})^i}{i!} \end{bmatrix}}_{\mathcal{R}_{n,\delta}}. \end{aligned}$$

We show below that $\mathcal{T}_{n,\delta}$ is invertible, so it suffices to show that for some $\delta > 0$,

$$\mathcal{T}_{n,\delta}^{-1}\mathbf{V}_{\mathbf{t}} = \mathcal{O}_n + \mathcal{T}_{n,\delta}^{-1}\mathcal{R}_{n,\delta} \text{ has rank } n.$$

Since $(\mathbf{A}, \mathbf{C})$ is observable , $\mathrm{rank}(\mathcal{O}) = n$ (c.f. Zhou et al. [1996, Theorem 3.3]). Therefore, since the set of full-rank matrices is an open set, it suffices to show that $\lim_{\delta\to 0}\mathcal{T}_{n,\delta}^{-1}\mathcal{R}_{n,\delta} = \mathbf{0}$. Since $\|\mathcal{R}_{n,\delta}\| = \mathcal{O}\left(\delta^n\right)$ as $\delta \to 0$, it suffices to show that $\|\mathcal{T}_{n,\delta}^{-1}\| = \mathcal{O}\left(\frac{1}{\delta^{(n-1)}}\right)$. To this end, we factor

$$\mathcal{T}_{n,\delta} = \underbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{I}_m & \mathbf{I}_m & \frac{1}{2!}\mathbf{I}_m & \dots & \frac{1}{(n-1)!}\mathbf{I}_m \\ \mathbf{I}_m & 2\mathbf{I}_m & \frac{2^2}{2!}\mathbf{I}_m & \dots & \frac{2^{n-1}}{(n-1)!}\mathbf{I}_m \\ \dots & & & & \\ \mathbf{I}_m & (n-1)\mathbf{I}_m & \frac{(n-1)^2}{2!}\mathbf{I}_m & \dots & \frac{(n-1)^{n-1}}{(n-1)!}\mathbf{I}_m \end{bmatrix}}_{\mathcal{U}_n} \underbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \delta\mathbf{I}_m & 0 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \delta^2\mathbf{I}_m & \dots & \mathbf{0} \\ \dots & & & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \delta^{n-1}\mathbf{I}_m \end{bmatrix}}_{\mathcal{D}_{n,\delta}}.$$

Using row elimination, it is easy to observe that $\mathcal{U}_n$ is invertible for any $\delta > 0$. In addition, $\mathcal{D}_{n,\delta}$ is invertible, with $\|\mathcal{D}_{n,\delta}^{-1}\| = \frac{1}{\delta^{n-1}}$. Note that the invertibility of $\mathcal{U}_n$ and $\mathcal{D}_{n,\delta}$ establish the invertibility of $\mathcal{T}_{n,\delta}$, as promised. To conclude, we observe that since $\mathcal{U}_n$ does not depend on δ,

$$\|\mathcal{T}_{n,\delta}^{-1}\| \leq \|\mathcal{U}_n\|^{-1} \cdot \|\mathcal{D}_{n,\delta}^{-1}\| = \frac{1}{\delta^{n-1}} \|\mathcal{U}_n\|^{-1} = \mathcal{O}\left(\delta^{n-1}\right).$$

Proof for Lebesgue-almost-every $\mathbf{t}$. Having established the result for a fixed $\mathbf{t}$, define the function $f(\mathbf{t}) := \det(\mathbf{V}_{\mathbf{t}}^\top \mathbf{V}_{\mathbf{t}})$, with domain $\mathbf{t} \in \mathbb{R}^k$.[5] Then $f(\mathbf{t})$ is defined and analytic on all of $\mathbb{R}^k$. Moreover, $f(\mathbf{t}) = 0$ if and only if $\mathrm{rank}(\mathbf{V}_{\mathbf{t}}) \neq n$. Therefore, the previous part of the lemma establishes that there exists at least some $\mathbf{t} \in \mathbb{R}^k$ for which $f(\mathbf{t}) \neq 0$. The lemma is now a direct consequence of the identity theorem for analytic functions (Fact D.1). ☐

Part I

General Control-Theoretic Proofs

D Discussion of Controllability Assumption 2.4

D.1 Remarks of Assumption 2.4

Lemma D.1. *The following conditions are equivalent to Assumption 2.4:*

(a) There exists at least one $\mathsf{K}_\star \in \mathcal{K}_{\mathtt{opt}}$ *for which* $(\mathbf{A}_{\mathsf{K}_\star}, \mathbf{B}_{\mathsf{K}_\star})$ *is controllable.*

(b) $(\mathbf{A} - \mathbf{L}_\star \mathbf{C}, \mathbf{L}_\star)$ *is controllable.*

(c) $(\mathbf{A}, \mathbf{L}_\star)$ *is controllable.*

Proof. Point (a) follows since controllability is invariant under similarity transform; point (b) follows by taking $(\mathbf{A}_{\mathsf{K}_\star}, \mathbf{B}_{\mathsf{K}_\star})$ to be the the canonical realization of the optimal filter; point (c) follows since maps of the form $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \to (\widetilde{\mathbf{A}} + \widetilde{\mathbf{K}}\widetilde{\mathbf{B}}, \widetilde{\mathbf{B}})$ preserve controllability. ☐

Proposition D.2. *Fix any* $n, m \geq 1$, $\mathbf{W}_1 \in \mathbb{S}^n_{++}$, $\mathbf{W}_2 \in \mathbb{S}^m_{++}$, *and suppose that* $(\mathbf{A}, \mathbf{C})$ *are drawn from a distribution with density with respect to the Lebesgue measure, such that with probability* 1, $\mathbf{A}$ *is Hurwitz stable. Then* $\mathbb{P}[$*Assumption 2.4 holds for* $(\mathbf{A}, \mathbf{C}, \mathbf{W}_1, \mathbf{W}_2)] = 1$.

Proposition D.2 is proven in Appendix D.4.

D.2 A strictly smaller problem set

Assumption 2.4 states that any optimal $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ must be controllable, which implies that $\boldsymbol{\Sigma}_{22,\mathsf{K}} \succeq 0$, cf. Appendix E.1. This in turn ensures that $\mathbf{Z}_\star = \boldsymbol{\Sigma}_{12,\mathsf{K}} \boldsymbol{\Sigma}_{22,\mathsf{K}}^{-1} \boldsymbol{\Sigma}_{12,\mathsf{K}}^\top$ and the regularizer $\mathrm{tr}[\mathbf{Z}_\star^{-1}]$ are well-defined at optimality. Not all `OE` instances satisfy this property, as the following example demonstrates:

Example D.1. Consider the `OE` problem instance given by

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \mathbf{W}_1 = \begin{bmatrix} 48 & -36 \\ -36 & 48 \end{bmatrix}, \quad \mathbf{W}_2 = 1.$$

It is readily verified that this instance satisfies Assumptions 2.1 to 2.3. The optimal policy (up to similarity transformations) is given by

$$\mathbf{A}_{\mathsf{K}_\star} = \begin{bmatrix} -5 & -4 \\ 0 & -2 \end{bmatrix}, \quad \mathbf{B}_{\mathsf{K}_\star} = \mathbf{L}_\star = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \quad \mathbf{P}_\star = \begin{bmatrix} 16 & -12 \\ -12 & 12 \end{bmatrix}.$$

---

[5] Observe that, while we only select strictly increasing $\mathbf{t}$, this lemma does not need such a restriction.

Recall that the optimal policy is independent of $\mathbf{G}$, the value of which is irrelevant for this example. Straightforward calculations reveal that

$$[\mathbf{B}_{\mathsf{K}_\star} \quad \mathbf{A}_{\mathsf{K}_\star}\mathbf{B}_{\mathsf{K}_\star}] = \begin{bmatrix} 4 & -20 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{\Sigma}_{\mathsf{K}_\star} = \begin{bmatrix} 24 & -12 & 8 & 0 \\ -12 & 12 & 0 & 0 \\ 8 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

confirming that $(\mathbf{A}_{\mathsf{K}_\star}, \mathbf{B}_{\mathsf{K}_\star})$ is not controllable and $\mathbf{\Sigma}_{\mathsf{K}_\star}$ is rank-deficient.

D.3 Implications for the convex reformulation

In this subsection, we discuss the implications of Assumption 2.4 and Example D.1 for the convex reformulation of $\mathtt{OE}$ developed in Scherer et al. [1997]. In particular, it is natural to wonder whether the breakdown of the change of variables at the optimal policy for problems such Example D.1 pose a problem for the methods of Scherer et al. [1997]. Fortunately, they do not. The LMI formulations of Scherer et al. [1997] circumvent these degeneracies in the landscape by employing strict inequalities. As we detail below, one can always perturb the decision variables to satisfy these strict inequalities, even at points where $\mathbf{\Sigma}_{12,\mathsf{K}}$ is rank deficient, resulting in arbitrarily tight upper bounds on the true cost $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$.

Specifically, given the decision variables $\mathbf{S}, \mathbf{X}, \mathbf{Y}, \mathbf{K}, \mathbf{L}, \mathbf{M}$, and defining

$$\bar{\mathbf{A}} := \begin{bmatrix} \mathbf{AY} + \mathbf{BM} & \mathbf{A} \\ \mathbf{K} & \mathbf{AX} + \mathbf{LC} \end{bmatrix}, \ \bar{\mathbf{B}} := \begin{bmatrix} \mathbf{W}_1^{1/2} & \mathbf{0} \\ \mathbf{XW}_1^{1/2} & \mathbf{LW}_2^{1/2} \end{bmatrix}, \ \bar{\mathbf{C}} := [\mathbf{GY} - \mathbf{M} \quad \mathbf{G}], \ \bar{\mathbf{X}} = \begin{bmatrix} \mathbf{Y} & \mathbf{I} \\ \mathbf{I} & \mathbf{X} \end{bmatrix},$$

the approach of Scherer et al. [1997] proposes solving the following semidefinite program (SDP)

$$\begin{aligned} \min \quad & \mathrm{tr}(\mathbf{S}) \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{S} & \bar{\mathbf{C}} \\ \bar{\mathbf{C}}^\top & \bar{\mathbf{X}} \end{bmatrix} \succ \mathbf{0}, \quad \begin{bmatrix} \bar{\mathbf{A}} + \bar{\mathbf{A}}^\top & \bar{\mathbf{B}} \\ \bar{\mathbf{B}}^\top & -\mathbf{I} \end{bmatrix} \prec \mathbf{0}, \end{aligned} \tag{D.1}$$

which minimizes a convex upper bound on the $\mathtt{OE}$ cost. At optimality, to achieve $\mathrm{tr}(\mathbf{S}) = \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)$, the above linear matrix inequalities (LMIs) must be tight. Moreover, $\bar{\mathbf{X}}$ can then be interpreted as $\mathbf{\Sigma}_{\mathsf{K}_\star}^{-1}$, subject to a specific congruence transformation, cf. Scherer et al. [1997]. However, for problem instances such as Example D.1, $\mathbf{\Sigma}_{\mathsf{K}_\star}$ is rank deficient and thus $\mathbf{\Sigma}_{\mathsf{K}_\star}^{-1}$ does not exist. The convex reformulation circumvents this problem by through the use of strict LMIs: at optimality, the above inequalities remain strict, and $\mathrm{tr}(\mathbf{S}) > \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)$. In fact, for Example D.1, if one approximates the strict LMIs $\mathbf{F} \succ \mathbf{0}$, for generic $\mathbf{F}$, with non-strict $\mathbf{F} \succeq \varepsilon\mathbf{I}$ for $\varepsilon = 10^{-8}$, then Eq. (D.1) returns a solution satisfying $\mathrm{tr}(\mathbf{S}) - \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star) \approx 8 \times 10^{-6}$.

D.4 Proof of Proposition D.2

Our argument relies on the identity theorem for real-analytic functions. [6]

Fact D.1. Let $\mathcal{U}$ be an open, connected subset of $\mathbb{R}^k$, and $F : \mathcal{U} \to \mathbb{R}$ be an analytic function which is not identically zero. Then the set $\{\boldsymbol{x} \in \mathcal{U} : f(\boldsymbol{x}) = 0\}$ has Lebesgue measure zero.

Give $\mathbf{W}_1 \in \mathbb{S}^n_{++}, \mathbf{W}_2 \in \mathbb{S}^m_{++}$. Let $\mathsf{Hur}_n := \{\mathbf{A} \in \mathbb{R}^{n \times n} : \lambda_i(\mathbf{A}) < 0, \ \forall i \in [n]\}$ denote the set of Hurwitz matrices. We consider $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}} := \mathsf{Hur}_n \times \mathbb{R}^{m \times n}$. $\mathcal{U}_{\mathrm{asm}}$ is open and connected as a consequence of the following claim, due to Duan and Patton [1998]:

Claim D.3. *The set of Hurwitz matrices* $\mathsf{Hur}_n := \{\mathbf{A} : \lambda_i(\mathbf{A}) < 0\}$ *is a connected, open subset of* $\mathbb{R}^{n \times n}$.

We define our candidate function f_{asm} as follows. Given $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$, let

$$f_{\mathrm{asm}}(\mathbf{A}, \mathbf{C}) = \det(\sum_{i=0}^{n-1} \mathbf{A}^i \mathbf{L}_\star \mathbf{L}_\star \mathbf{A}^i)., \tag{D.2}$$

---

[6] For a proof, see e.g. `https://math.stackexchange.com/questions/1322858/zeros-of-analytic-function-of-several-real-variables`.

where $\mathbf{L}_\star$ solves is the associated optimal gain for $(\mathbf{A}, \mathbf{C}, \mathbf{W}_1, \mathbf{W}_2)$ (this exists for all Hurwitz $\mathbf{A}$). From Zhou et al. [1996, Theorem 3.3]), $(\mathbf{A}, \mathbf{L}_\star)$ is controllable if and only if $\mathrm{rank}[\mathbf{L}_\star \mid \mathbf{A}\mathbf{L}_\star \mid \mathbf{A}^2\mathbf{L}_\star \mid \dots \mathbf{A}^{n-1}\mathbf{L}_\star] = n$, which holds if and only if $f_{\mathrm{asm}}(\mathbf{A}, \mathbf{C}) \neq 0$. Hence, by Lemma D.1, we conclude

Claim D.4. *$(\mathbf{A}, \mathbf{C}) \in \mathsf{Hur}_n \times \mathbb{R}^{m\times n}$ satisfies Assumption 2.4 if and only if $f_{\mathrm{asm}}(\mathbf{A}, \mathbf{C}) \neq 0$, which holds if and only if $(\mathbf{A}, \mathbf{L}_\star)$ is controllable.*

To conclude, we must argue that (1) f_{asm} is analytic on $\mathcal{U}_{\mathrm{asm}}$, and (2) f_{asm} is not identically zero on $\mathcal{U}_{\mathrm{asm}}$; i.e. there exists *some* $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$ for which $(\mathbf{A}, \mathbf{L}_\star)$.

Analyticity of f_{asm}. For the first point, we have the following claim.

Claim D.5. *Fix matrices $\mathbf{W}_1 \succ 0, \mathbf{W}_2 \succ 0$. Then, the mapping $F_P : (\mathbf{A}, \mathbf{C}) \mapsto \mathbf{P}_\star$ to the solution $\mathbf{P}_\star$ to the Riccati equation below, as well as the map $F_L : (\mathbf{A}, \mathbf{C}) \mapsto \mathbf{L}_\star$ given below, are both real analytic on $\mathcal{U}_{\mathrm{asm}}$.*

$$\mathbf{A}\mathbf{P}_\star + \mathbf{P}_\star\mathbf{A}^\top - \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_\star + \mathbf{W}_1 = 0, \qquad \mathbf{L}_\star = \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}. \tag{D.3}$$

As a consequence, f_{asm} is real analytic on $\mathcal{U}_{\mathrm{asm}}$

Proof. Since $\mathbf{W}_1, \mathbf{W}_2$ are fixed, the map $F_0 : (\mathbf{P}_\star, \mathbf{C}) \mapsto \mathbf{L}_\star$ is polynomial, and thus analytic. Hence, $F_L = F_0 \circ F_P$ is analytic whenever F_P is. Similarly, f_{asm} is analytic whenever F_L is analytic, and hence whenever F_P is analytic.

To see that F_P is analytic, let us use the implicit function. $F_P(\mathbf{A}, \mathbf{C})$ is define by the zero of the equation

$$G(\mathbf{A}, \mathbf{C}, \mathbf{P}) = \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\top - \mathbf{P}\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P} + \mathbf{W}_1.$$

The total derivative of G is then

$$\begin{aligned}
&\mathrm{d}G(\mathbf{A}, \mathbf{C}, \mathbf{P})\\
&= \mathrm{d}\mathbf{A}\mathbf{P} + \mathbf{P}\mathrm{d}\mathbf{A}^\top - \mathbf{P}\mathrm{d}(\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}) + \mathbf{W}_1 + (\mathbf{A} - \mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P})\mathrm{d}\mathbf{P} + \mathrm{d}\mathbf{P}(\mathbf{A} - \mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P})^\top\\
&= \mathrm{d}\mathbf{A}\mathbf{P} + \mathbf{P}\mathrm{d}\mathbf{A}^\top - \mathbf{P}\mathrm{d}(\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}) + \mathbf{W}_1 + (\mathbf{A} - \mathbf{L}_\star\mathbf{C})\mathrm{d}\mathbf{P} + \mathrm{d}\mathbf{P}(\mathbf{A} - \mathbf{L}_\star\mathbf{C})^\top.
\end{aligned}$$

We see that a solution to $\mathrm{d}G(\mathbf{A}, \mathbf{C}, \mathbf{P}) = 0$ must have that $\mathrm{d}\mathbf{P}$ satisfies the following Lyapunov equation for $\mathbf{Y} := \mathrm{d}\mathbf{A}\mathbf{P} + \mathbf{P}\mathrm{d}\mathbf{A}^\top - \mathbf{P}\mathrm{d}(\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}) + \mathbf{W}_1$:

$$\widetilde{\mathbf{A}}\mathrm{d}\mathbf{P} + \widetilde{\mathbf{A}}\mathrm{d}\mathbf{P} + \mathbf{Y} = 0. \tag{D.4}$$

Since the Since $\widetilde{\mathbf{A}} := (\mathbf{A} - \mathbf{L}_\star\mathbf{C})$ is Hurwitz for a solution $\mathbf{L}_\star$ to Eq. (D.3), the solution $\mathrm{d}\mathbf{P}$ to Eq. (D.4) is unique. Hence, $\mathrm{d}G(\mathbf{A}, \mathbf{C}, \mathbf{P})$ satisfies the conditions of the implicit function theorem. In addition, G is analytic. This means that, in a neighborhood around any $(\mathbf{A}, \mathbf{C}) \in \mathsf{Hur}_n \times \mathbb{R}^{m\times n}$, there is an analytic function corresponding to $(\mathbf{A}, \mathbf{C}) \mapsto \mathbf{P}_\star$. By definition, this function coincides with F_P on that neighborhood, meaning F_P is also analytic. ☐

f_{asm} is not identically zero. To conclude, it suffices to show the existence of *some* $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$ for which f_{asm} doesn't vanish; i.e., some $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$ for which $(\mathbf{A}, \mathbf{L}_\star)$ is controllable. The following lemma is useful in our construction.

Lemma D.6. *Fix $\mathbf{W}_1 \in \mathbb{S}^n_{++}, \mathbf{W}_2 \in \mathbb{S}^m_{++}$, $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$, and let $\mathbf{P}_{\star,k}$ be the solution to the Lyapunov equation with $(\mathbf{A}, \frac{1}{k}\mathbf{C}, \mathbf{W}_1, \mathbf{W}_2)$. Then, $\lim_{k\to\infty} \mathbf{P}_{\star,k} = \mathbf{P}_{\star,\infty}$, where $\mathbf{P}_{\star,\infty}$ solves*

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\top + \mathbf{W}_1 = 0.$$

Proof. The sequence $\mathbf{P}_{\star,k}$ are the solution to the Ricatti equation $\mathcal{T}_k(\mathbf{P}) = 0$, where

$$\mathcal{T}_k(\mathbf{P}) := \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\top - \mathbf{P}\mathbf{C}_k^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P} + \mathbf{W}_1$$

Since $\mathbf{A}$ is stable, $\mathbf{P}_{\star,k}$ also the unique solution $\mathbf{P}$ to the Lyapunov equation $\widetilde{\mathcal{T}}_k(\mathbf{P}) = 0$ constructed by fixing $\mathbf{P} = \mathbf{P}_{\star,k}$ in the third term in $\mathcal{T}_k(\mathbf{P})$:

$$\widetilde{\mathcal{T}}_k(\mathbf{P}) := \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\top + \widetilde{\mathbf{W}}_{1,k}, \quad \widetilde{\mathbf{W}}_{1,k} := \left(\mathbf{W}_1 - \frac{1}{k}\mathbf{P}_{\star,k}\mathbf{C}_k^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_{\star,k}\right).$$

Since $\widetilde{\mathbf{W}}_{1,k} \preceq \mathbf{W}_1$, we have that $\mathbf{P}_{\star,k} \preceq \mathbf{P}_{\star,\infty}$. In addition, $\mathbf{P}_{\star,k} \succeq 0$ for all k. Thus, $\mathbf{P}_{\star,k}$ lie in the compact set $\mathcal{P} := \{\mathbf{P} \in \mathbb{S}^n : 0 \succeq \mathbf{P} \succeq \mathbf{P}_{\star,\infty}\}$, and hence it suffices to show that for any convergent subsequence $(\mathbf{P}_{\star,k_i})$ which converges to a limit $\widetilde{\mathbf{P}} \in \mathcal{P}$, $\widetilde{\mathbf{P}} = \mathbf{P}_{\star,\infty}$. To show show this, observe $\mathcal{T}_k(\cdot) \to \mathcal{T}_\infty(\cdot)$ uniformly on the compact set $\mathcal{P}$, and since $\mathcal{T}_\infty$ is continuous, it follows that

$$0 = \lim_{i \to \infty} \mathcal{T}_{k_i}(\mathbf{P}_{\star,k_i}) = \lim_{i \to \infty} \mathcal{T}_\infty(\mathbf{P}_{\star,k_i}) = \mathcal{T}_\infty(\widetilde{\mathbf{P}}).$$

Since $\mathcal{T}_\infty(\cdot)$ is a Lyapunov equation with A stable, the solution to $\mathcal{T}_\infty(\cdot) = 0$ is unique, and hence $\widetilde{P} = \mathbf{P}_{\star,\infty}$, as needed. □

Claim D.7. *Fix $\mathbf{W}_1 \in \mathbb{S}^n_{++}$, $\mathbf{W}_2 \in \mathbb{S}^m_{++}$. Then, there exists an $(\mathbf{A}, \mathbf{C}) \in \mathcal{U}_{\mathrm{asm}}$ for which $(\mathbf{A}, \mathbf{L}_\star)$ is controllable, where $\mathbf{L}_\star$ is as in Eq. (D.3). In particular, for this $(\mathbf{A}, \mathbf{C})$, $f_{\mathrm{asm}}(\mathbf{A}, \mathbf{C}) \neq 0$.*

Proof. By a change of basis of $\mathbb{R}^n$ and $\mathbb{R}^m$, we may assume without loss of generality that $\mathbf{W}_1 = \mathbf{I}_n$ and $\mathbf{W}_2 = \mathbf{I}_m$. Let $\mathbf{A} = \mathrm{Diag}(1, 2, \ldots, n)$, and let $\mathbf{C}_1 := \begin{bmatrix} \mathbf{1} & \mathbf{0}_n & \ldots & \mathbf{0}_n \end{bmatrix}^\top$, and set $\mathbf{C}_k = \frac{1}{k}\mathbf{C}_1$. It $\mathbf{P}_{\star,k}$ (resp. $\mathbf{L}_{\star,k}$) solve the Ricatti equation (resp. be the optimal gain) matrix for $(\mathbf{A}, \mathbf{C}_k)$. We show that for all k sufficiently large, $(\mathbf{A}, \mathbf{L}_{\star,k})$ is controllable (indeed, this establishes existence.)

It suffices to show that, for all k sufficiently large, $(\mathbf{A}, \widetilde{\mathbf{L}}_{\star,k})$ is controllable where $\widetilde{\mathbf{L}}_{\star,k} := k\mathbf{L}_{\star,k}$. From Eq. (D.3), the definition of $\mathbf{C}_k$, and assumption $\mathbf{W}_2 = \mathbf{I}_m$,

$$\widetilde{\mathbf{L}}_{\star,k} := k\mathbf{L}_{\star,k} = k\mathbf{P}_{\star,k}\mathbf{W}_2^{-1}\mathbf{C}_k^\top = \mathbf{P}_{\star,k}\mathbf{C}_1^\top$$

Since the set of controllable matrices is an open set, and since $\lim_{k \to \infty} \mathbf{P}_{k,\star} = \mathbf{P}_{\star,\infty}$ by Lemma D.6, we see that $(\mathbf{A}, \widetilde{\mathbf{L}}_{\star,k})$ is controllable for all k sufficiently large as long $(\mathbf{A}, \mathbf{P}_{\star,\infty}\mathbf{C}_1^\top)$ is controllable. Since $\mathbf{A}$ is diagonal, one can verify that $\mathbf{P}_{\star,\infty} = -\frac{1}{2}\mathbf{A}^{-1}$. In particular, $\mathbf{P}_{\star,\infty}\mathbf{C}_1^\top = \begin{bmatrix} -\frac{1}{2}\mathbf{A}^{-1}\mathbf{1} & \mathbf{0}_n & \ldots & \mathbf{0}_n \end{bmatrix}$; hence the first column of $\mathbf{P}_{\star,\infty}\mathbf{C}_1^\top$ does not lie in any $\mathbf{A}$-invariant subspace, so $(\mathbf{A}, \mathbf{P}_{\star,\infty}\mathbf{C}_1^\top) = (\mathbf{A}, \widetilde{\mathbf{L}}_{\star,k})$ is controllable for all k large. As noted above, this implies $(\mathbf{A}, \mathbf{L}_{\star,k})$ is controllable, so that by Claim D.4, $f_{\mathrm{asm}}(\mathbf{A}, \mathbf{C}_k) \neq 0$. □

Conclusion. Hence, we have established that f_{asm} is analytic, but not identically zero, on the open and commented domain $\mathcal{U}_{\mathrm{asm}}$. The proof follows. □

E Control Proofs

E.1 Controllability, stability, and nonsingularity of internal-state covariance

In Section 2, we restricted our attention to policies $\mathsf{K} \in \mathcal{K}_{\mathtt{stab}}$, that is, where the filter transition matrix $\mathbf{A}_{\mathsf{K}}$ was Hurwitz stable. This is equivalent to stability of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$, as shown by the following lemma.

Lemma E.1. *$\mathbf{A}_{\mathsf{K}}$ is stable if and only if $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is stable, and $\mathbf{\Sigma}_{\mathsf{K}}$ is given by the solution of the Lyapunov equation Eq. (2.2).*

Proof. The equivalence of the stability of $\mathbf{A}_{\mathsf{K}}$ and $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ comes from the fact that, due to the block-triangular form of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ with blocks $\mathbf{A}$ and $\mathbf{A}_{\mathsf{K}}$, the eigenvalues of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ are just the union of those of $\mathbf{A}_{\mathsf{K}}$ and those of $\mathbf{A}$. All eigenvalues of $\mathbf{A}$ have negative real part by Assumption 2.1, so the non-negative real part of the eigenvalue of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ are equal to those of $\mathbf{A}_{\mathsf{K}}$. Thus, stability of $\mathbf{A}_{\mathsf{K}}$ and $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ are equivalent. That $\mathbf{\Sigma}_{\mathsf{K}}$ is given by the solution of the Lyapunov equation is standard, cf. [Zhou et al., 1996, Theorem 3.18]. □

Next, we show the equivalence between $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$ and controllability of $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$. We define controllability for (possibly unstable) $\mathbf{A}_{\mathsf{K}}$ as follows, cf., e.g., [Zhou et al., 1996, Theorem 3.1].

Definition E.1. The pair $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ is controllable if and only if there exists some $t > 0$ such that

$$\mathcal{G}^{[t]}_{\mathrm{cont},\mathsf{K}} := \int_0^t \exp(s\mathbf{A})\mathbf{B}_{\mathsf{K}}\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}^\top)\mathrm{d}s$$

is strictly positive definite.

Lemma E.2. *Suppose that Assumptions 2.1 and 2.3 hold. Then the following statements are equivalent.*

(a) The limiting covariance $\mathbf{\Sigma}_{22,\mathsf{K}}$, defined below, exists, and has $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$,

$$\mathbf{\Sigma}_{22,\mathsf{K}} = \lim_{t\to\infty} \mathbb{E}\left[\hat{\mathbf{x}}_{\mathsf{K}}(t)\hat{\mathbf{x}}_{\mathsf{K}}(t)^\top\right] \in \mathbb{S}^{2n}_{+}.$$

(b) $\mathbf{A}_{\mathsf{K}}$ is stable, and $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ is controllable.

(c) $\mathbf{A}_{\mathsf{K}}$ is stable and $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$.

Moreover, these equivalent conditions imply the limiting covariance $\mathbf{\Sigma}_{\mathsf{K}}$ is well-defined and given by the solution to Eq. (2.2).

Proof. The "moreover" statement is a consequence of Lemma E.1. We establish the equivalences of (a), (b), and (c).

(a) implies (b). We compute that

$$\mathbf{\Sigma}_{22,\mathsf{K}} = \lim_{t\to\infty} \mathbf{\Sigma}^{[t]}_{22,\mathsf{K}}, \quad \text{which is the bottom-diagonal block of } \mathbf{\Sigma}^{[t]}_{\mathsf{K}} = \int_0^t \exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_{\mathrm{cl},\mathsf{K}} \exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}s.$$

First, we show that $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ are controllable. Indeed, since $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$ and $\lim_{t\to\infty} \mathbf{\Sigma}^{[t]}_{22,\mathsf{K}}$ exists and is finite, we have that for this τ, $\mathbf{\Sigma}^{[\tau]}_{22,\mathsf{K}} \succ 0$. Thus by Lemma J.10, it follows that for some finite τ,

$$\int_0^\tau \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}_{\mathsf{K}})^\top \mathrm{d}s \succ 0. \tag{E.1}$$

Since $\mathbf{W}_2 \succ 0$ by Assumption 2.3, it therefore follows that

$$\mathcal{G}^{[\tau]}_{\mathrm{cont},\mathsf{K}} = \int_0^\tau \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}_{\mathsf{K}})^\top \mathrm{d}s \succ 0.$$

Next, we show stability. Since $\exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_{\mathrm{cl},\mathsf{K}} \exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \succeq 0$, existence of the limiting $\mathbf{\Sigma}_{22,\mathsf{K}}$ implies that for any vector of the form $\mathbf{v} = (0, \mathbf{v}_2) \in \mathbb{R}^{2n}$ for $\mathbf{v}_2 \in \mathbb{R}^n$,

$$\int_0^\infty \|\mathbf{v}^\top \exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}^{1/2}_{\mathrm{cl},\mathsf{K}}\|^2 \mathrm{d}s < \infty.$$

Note the $(2,2)$-bock of $\exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})$ is $\exp(s\mathbf{A}_{\mathsf{K}})$ (see Lemma J.3), and that, since $\mathbf{W}_{\mathrm{cl},\mathsf{K}}$ is block-diagonal,

$$\mathbf{W}_{\mathrm{cl},\mathsf{K}} := \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \end{bmatrix} \succeq \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \end{bmatrix}$$

Thus, considering a vector $\mathbf{v}$ of the form $(0, \mathbf{v}_2)$, for $\mathbf{v}_2 \in \mathbb{R}^n$,

$$\lim_{t\to\infty} \int_0^t \mathbf{v}_2^\top \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{v}_2 \mathrm{d}s < \infty,$$

which shows that the following limiting integral is well defined $\int_0^\infty \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}_{\mathsf{K}})^\top$. On the other hand, by Eq. (E.1), we must have that the following limiting integral is well-defined and strictly positive definite

$$\int_0^\infty \exp(s\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top \exp(s\mathbf{A}_{\mathsf{K}})\mathrm{d}s \succ 0.$$

Thus, Lemma J.6 implies that $\mathbf{A}_{\mathsf{K}}$ is Hurwitz stable. This (together with stability of $\mathbf{A}$) implies Hurwitz stability of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ (see below), and Lemma J.2 therefore guarantees that $\mathbf{\Sigma}_{\mathsf{K}}$ is the solution of the appropriate Lyapunov equation, given in Eq. (2.2).

(b) implies (c). From the computation in part (a), one can check that

$$\mathbf{\Sigma}_{22,\mathsf{K}} \succeq \lim_{t\to\infty} \int_0^\infty \exp(s\mathbf{A}_\mathsf{K})\mathbf{B}_\mathsf{K}\mathbf{W}_2\mathbf{B}_\mathsf{K}^\top \exp(s\mathbf{A}_\mathsf{K})\mathrm{d}s \succeq \lambda_{\min}(\mathbf{W}_2) \lim_{t\to\infty} \mathcal{G}^{[t]}_{\mathrm{cont},\mathsf{K}}.$$

Thus, controllability of $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$ implies $\mathcal{G}^{[t]}_{\mathrm{cont},\mathsf{K}} \succ 0$ for some finite t, which implies $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$.

(c) implies (a). From Lemma J.2, stability of $\mathbf{A}_\mathsf{K}$ implies stability of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$, which implies that the limiting covariance $\mathbf{\Sigma}_\mathsf{K}$ exists. In particular, the limiting (2,2)-block covariance exists. ☐

E.2 Characterization of optimal policies

We begin by reviewing some well-known properties of the optimal solution to the $\texttt{OE}$ problem.

Lemma E.3. *Under Assumptions 2.1 to 2.3, the* unique *(up to similarity transformations) optimal solution to the* $\texttt{OE}$ *problem is given by the policy*

$$\mathbf{A}_{\mathsf{K}_\star} = \mathbf{A} - \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}, \quad \mathbf{B}_{\mathsf{K}_\star} = \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}, \quad \mathbf{C}_{\mathsf{K}_\star} = \mathbf{G}, \tag{E.2}$$

where $\mathbf{P}_\star \succ 0$ *is the solution to the algebraic Riccati equation Eq. (2.5).*

Proof. A proof of this classical result can be found, e.g, in [Doyle et al., 1989, §IV.D]. We note that strict positive definiteness of $\mathbf{P}_\star$ is implied by the controllability of $(\mathbf{A}, \mathbf{W}_1)$, cf. [Doyle et al., 1989, §II.B]. Controllability of $(\mathbf{A}, \mathbf{W}_1)$ follows from $\mathbf{W}_1 \succ 0$, cf. Assumption 2.3. ☐

Fact E.1. The optimal solution to the $\texttt{OE}$ problem is independent of $\mathbf{G}$, and optimal for all values of $\mathbf{G}$.

Proof. The optimal policy given in Eq. (E.2) and the Riccati equation Eq. (2.5) are both independent of $\mathbf{G}$. Moreover, there are no restrictions placed on $\mathbf{G}$ (beyond the requirement that the number of columns matches the dimension of the state of the true system). ☐

E.3 Informativity of optimal policies

We begin with the following useful fact.

Fact E.2. Let $\mathsf{K}_\star \in \mathcal{K}_{\texttt{opt}}$ denote the realization of the optimal policy given in Eq. (E.2), i.e. with $\mathbf{C}_{\mathsf{K}_\star} = \mathbf{G}$. Then, under Assumptions 2.1 to 2.3, $\mathbf{\Sigma}_{12,\mathsf{K}_\star} = \mathbf{\Sigma}_{22,\mathsf{K}_\star}$.

Proof. All optimal policies $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$ must satisfy

$$\frac{\partial \mathcal{L}_{\texttt{OE}}(\mathsf{K})}{\partial \mathbf{C}_\mathsf{K}} = 2\mathbf{C}_\mathsf{K}\mathbf{\Sigma}_{22,\mathsf{K}} - 2\mathbf{G}\mathbf{\Sigma}_{12,\mathsf{K}} = 0. \tag{E.3}$$

In particular, for the realization of the optimal policy in Eq. (E.2) with $\mathbf{C}_{\mathsf{K}_\star} = \mathbf{G}$, this implies that $\mathbf{G}(\mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{12,\mathsf{K}_\star}) = 0$. By Fact E.1, this must hold for all $\mathbf{G}$, which implies that $\mathbf{\Sigma}_{22,\mathsf{K}_\star} = \mathbf{\Sigma}_{12,\mathsf{K}_\star}$. ☐

Lemma 3.1. *Under Assumptions 2.1 to 2.4,* $\mathcal{K}_{\texttt{opt}} \subset \mathcal{K}_{\texttt{info}} \subset \mathcal{K}_{\texttt{ctrb}}$*, and* $\mathcal{K}_{\texttt{info}}$ *is an open set.*

Proof. We prove each part in sequence.

Inclusion $\mathcal{K}_{\texttt{opt}} \subset \mathcal{K}_{\texttt{info}}$. Let $\mathsf{K}_\star \in \mathcal{K}_{\texttt{opt}}$ denote the realization of the optimal policy given in Eq. (E.2). By Assumption 2.4, all optimal policies are controllable, and so $\mathbf{\Sigma}_{22,\mathsf{K}_\star} \succ 0$. By Fact E.2, we have $\mathbf{\Sigma}_{12,\mathsf{K}_\star} = \mathbf{\Sigma}_{22,\mathsf{K}_\star} \succ 0$, which implies that $\mathbf{\Sigma}_{12,\mathsf{K}_\star}$ is full-rank. The rank of $\mathbf{\Sigma}_{12,\mathsf{K}}$ is invariant under similarity transformations of the policy; hence, $\mathbf{\Sigma}_{12,\mathsf{K}}$ is full-rank for all $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$.

Inclusion $\mathcal{K}_{\texttt{info}} \subset \mathcal{K}_{\texttt{ctrb}}$. Recall that $\mathcal{K}_{\texttt{ctrb}} := \{\mathsf{K} \in \mathcal{K}_{\texttt{stab}} : \mathbf{\Sigma}_{22,\mathsf{K}} \succ 0\}$. Hence, it suffices to show that if $\mathsf{K} \in \mathcal{K}_{\texttt{stab}}$ has $\mathrm{rank}(\mathbf{\Sigma}_{12,\mathsf{K}}) = n$, then $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$. This follows since $\mathbf{\Sigma}_\mathsf{K} \succeq 0$.

Openness. To see that $\mathcal{K}_{\texttt{info}}$ is open, we observe that $\mathcal{K}_{\texttt{stab}}$ is open (this follows from Claim D.3), and that $\mathsf{K} \mapsto \mathbf{\Sigma}_{\mathsf{K}}$ is continuous on $\mathcal{K}_{\texttt{stab}}$ (this is standard, and follows, for example, from arguments inAppendix E.7), Hence, the map $f : \mathsf{K} \mapsto \det(\mathbf{\Sigma}_{12,\mathsf{K}})$ is continuous on $\mathcal{K}_{\texttt{stab}}$, and thus $\mathcal{K}_{\texttt{info}} : \{\mathsf{K} \in \mathcal{K}_{\texttt{stab}} : \det(\mathbf{\Sigma}_{22,\mathsf{K}}) \neq 0\}$, being the inverse-image of the open set $\mathbb{R} \setminus \{0\}$ under f, is open. ☐

E.4 Maximality of $\mathbf{Z}_\star$

Lemma 3.2 (Existence of maximal $\mathbf{Z}_{\mathsf{K}}$). *Under Assumptions 2.1 to 2.3, there exists a unique* $\mathbf{Z}_\star \succ 0$ *such that* $\mathbf{Z}_\star = \mathbf{Z}_{\mathsf{K}}$ *if and only if* $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$*, and* $\mathbf{Z}_\star \succeq \mathbf{Z}_{\mathsf{K}}$ *for all* $\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}} \setminus \mathcal{K}_{\texttt{opt}}$*. Consequently,* $\mathcal{K}_{\texttt{opt}} \in \arg\min_{\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}}} \mathcal{R}_{\texttt{info}}(\mathsf{K})$.

Proof. We restrict our attention to $\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}}$, otherwise $\mathbf{\Sigma}_{22,\mathsf{K}}$ is not invertible and $\mathbf{Z}_{\mathsf{K}} = \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top$ is not well-defined. Recall that $\mathbf{Z}_{\mathsf{K}}$ is independent of the realization of K, i.e. $\mathbf{Z}_{\mathsf{K}}$ is invariant under similarity transformations of K.

First, observe that the $\texttt{OE}$ cost can be written as

$$\mathcal{L}_{\texttt{OE}}(\mathsf{K}) = \operatorname{tr}\left[\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix} \mathbf{\Sigma}_{\mathsf{K}} \begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix}^\top\right] = \operatorname{tr}[\mathbf{G}\mathbf{\Sigma}_{11,\mathrm{sys}}\mathbf{G}^\top] - 2\operatorname{tr}[\mathbf{G}\mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{C}_{\mathsf{K}}^\top] + \operatorname{tr}[\mathbf{C}_{\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}\mathbf{C}_{\mathsf{K}}^\top], \tag{E.4}$$

where $\mathbf{\Sigma}_{\mathsf{K}}$ satisfies the Lyapunov equation in Eq. (2.2). Minimizing Eq. (E.4) w.r.t. $\mathbf{C}_{\mathsf{K}}$ (keeping $\mathbf{A}_{\mathsf{K}},\ \mathbf{B}_{\mathsf{K}}$ fixed) gives

$$\operatorname{tr}[\mathbf{G}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \underbrace{\mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top}_{=\mathbf{Z}_{\mathsf{K}}})\mathbf{G}^\top] = \min_{\mathbf{C}_{\mathsf{K}}} \ \mathcal{L}_{\texttt{OE}}((\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})). \tag{E.5}$$

Let $\mathsf{K}_\star \in \mathcal{K}_{\texttt{opt}}$, and denote $\mathbf{Z}_\star = \mathbf{\Sigma}_{12,\mathsf{K}_\star}\mathbf{\Sigma}_{22,\mathsf{K}_\star}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}_\star}^\top$. Then by optimality of $\mathsf{K}_\star$ we have

$$\operatorname{tr}[\mathbf{G}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_{\mathsf{K}})\mathbf{G}^\top] \ \geq \ \operatorname{tr}[\mathbf{G}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)\mathbf{G}^\top] \implies \operatorname{tr}[\mathbf{G}(\mathbf{Z}_\star - \mathbf{Z}_{\mathsf{K}})\mathbf{G}^\top] \ \geq \ 0, \tag{E.6}$$

with equality if and only if $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$, due to uniqueness (of the transfer function) of the optimal policy, cf. Lemma E.3. By Fact E.1, this holds for all $\mathbf{G}$, which implies that $\mathbf{Z}_\star - \mathbf{Z}_{\mathsf{K}} \succeq 0$, again with equality if and only if $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$. This completes the first part of the proof.

To show that $\mathsf{K}_\star$ minimizes $\mathcal{R}_{\texttt{info}}(\mathsf{K}) = \operatorname{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$, we distinguish between two cases: those in which $\mathbf{Z}_{\mathsf{K}}$ is invertible, and those in which it is not. Consider the former, and assume K is such that $\mathbf{Z}_{\mathsf{K}}$ is invertible. Observe that $\mathbf{Z}_\star$ is always invertible: by Assumption 2.4 we have that $\mathbf{\Sigma}_{22,\mathsf{K}_\star} \succ 0$, and by Lemma 3.1 we have that $\mathbf{\Sigma}_{12,\mathsf{K}_\star}$ is full-rank. Therefore, $\mathbf{Z}_\star = \mathbf{\Sigma}_{12,\mathsf{K}_\star}\mathbf{\Sigma}_{22,\mathsf{K}_\star}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}_\star}^\top$ is also full-rank. We then have the following:

$$\mathbf{Z}_\star \succeq \mathbf{Z}_{\mathsf{K}} \implies \mathbf{Z}_{\mathsf{K}}^{-1} \succeq \mathbf{Z}_\star^{-1} \implies \operatorname{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}] \ \geq \ \operatorname{tr}[\mathbf{Z}_\star^{-1}] \implies \mathcal{R}_{\texttt{info}}(\mathsf{K}) \ \geq \ \mathcal{R}_{\texttt{info}}(\mathsf{K}_\star), \tag{E.7}$$

with equality if and only if $\mathsf{K} \in \mathcal{K}_{\texttt{opt}}$. This implies that $\mathsf{K}_\star \in \mathcal{K}_{\texttt{opt}}$ minimizes $\mathcal{R}_{\texttt{info}}(\cdot)$ over all $\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}}$ such that $\mathbf{Z}_{\mathsf{K}}$ is invertible.

Next, we consider the case in which K is such that $\mathbf{Z}_{\mathsf{K}}$ is not invertible. In this case, $\mathcal{R}_{\texttt{info}}(\mathsf{K}) := \infty$, and so $\mathcal{R}_{\texttt{info}}(\mathsf{K}) \ \geq \ \mathcal{R}_{\texttt{info}}(\mathsf{K}_\star)$ holds trivially. This completes the proof that $\mathcal{R}_{\texttt{info}}(\mathsf{K}) \ \geq \ \mathcal{R}_{\texttt{info}}(\mathsf{K}_\star)$ for all $\mathsf{K} \in \mathcal{K}_{\texttt{ctrb}}$. ☐

E.5 Positivity and characterization of $\sigma_\star$

Lemma 3.3. *Let* $\mathbf{P}_\star$ *be the solution to the Riccati equation in Eq. (2.5). Then under Assumptions 2.1 to 2.3,* $\sigma_\star := \lambda_{\min}(\mathbf{P}_\star)$ *is strictly positive. Moreover,* $\mathbf{P}_\star = \mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star$.

Proof. Strict positivity of $\sigma_\star := \lambda_{\min}(\mathbf{P}_\star)$ follows directly from Lemma E.3, which states that $\mathbf{P}_\star \succ 0$.

To show that $\mathbf{P}_\star = \mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star$, we will first show that $\mathbf{P}_\star = \mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{\Sigma}_{22,\mathsf{K}_\star}$, where $\mathsf{K}_\star$ denotes the realization given in Eq. (E.2). Let $\mathbf{\Sigma}_{\mathsf{K}_\star}$ be given by the solution to the Lyapunov equation $\mathbf{A}_{\mathrm{cl},\mathsf{K}_\star}\mathbf{\Sigma}_{\mathsf{K}_\star} + \mathbf{\Sigma}_{\mathsf{K}_\star}\mathbf{A}_{\mathrm{cl},\mathsf{K}_\star}^\top + \mathbf{W}_{\mathrm{cl},\mathsf{K}_\star} = 0$, as in Eq. (2.2). The (2,2) block of this Lyapunov equation is given by

$$\mathbf{A}_{\mathsf{K}_\star}\mathbf{\Sigma}_{22,\mathsf{K}_\star} + \mathbf{\Sigma}_{22,\mathsf{K}_\star}\mathbf{A}_{\mathsf{K}_\star}^\top + \mathbf{B}_{\mathsf{K}_\star}\mathbf{C}\mathbf{\Sigma}_{12,\mathsf{K}_\star} + \mathbf{\Sigma}_{12,\mathsf{K}_\star}^\top\mathbf{C}^\top\mathbf{B}_{\mathsf{K}_\star}^\top + \mathbf{B}_{\mathsf{K}_\star}\mathbf{W}_2\mathbf{B}_{\mathsf{K}_\star}^\top = 0. \tag{E.8}$$

Substituting $\mathbf{A}_{\mathsf{K}_\star} = \mathbf{A} - \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}$ and $\mathbf{B}_{\mathsf{K}_\star} = \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}$ into Eq. (E.8) gives

$$(\mathbf{A} - \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C})\mathbf{\Sigma}_{22,\mathsf{K}_\star} + \mathbf{\Sigma}_{22,\mathsf{K}_\star}(\mathbf{A} - \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C})^\top + \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{\Sigma}_{12,\mathsf{K}_\star} + \mathbf{\Sigma}_{12,\mathsf{K}_\star}^\top\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_\star \\ + \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_\star = 0. \tag{E.9}$$

Subtracting the (1,1) block of the Lyapunov equation Eq. (2.2), given by $\mathbf{A}\mathbf{\Sigma}_{11,\text{sys}} + \mathbf{\Sigma}_{11,\text{sys}}\mathbf{A}^\top + \mathbf{W}_1 = 0$, from Eq. (E.9) and collecting terms leads to

$$\mathbf{A}(\mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{11,\text{sys}}) + (\mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{11,\text{sys}})\mathbf{A}^\top + \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}(\mathbf{\Sigma}_{12,\mathsf{K}_\star} - \mathbf{\Sigma}_{22,\mathsf{K}_\star}) \\ + (\mathbf{\Sigma}_{12,\mathsf{K}_\star} - \mathbf{\Sigma}_{22,\mathsf{K}_\star})^\top\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_\star + \mathbf{P}_\star\mathbf{C}^\top\mathbf{W}_2^{-1}\mathbf{C}\mathbf{P}_\star - \mathbf{W}_1 = 0. \tag{E.10}$$

Next, from Fact E.2 we have $\mathbf{\Sigma}_{12,\mathsf{K}_\star} = \mathbf{\Sigma}_{22,\mathsf{K}_\star}$ for this particular realization of $\mathsf{K}_\star$, given in Eq. (E.2). Making this substitution, and adding the Riccati equation Eq. (2.5) to Eq. (E.10) gives

$$\mathbf{A}(\mathbf{P}_\star + \mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{11,\text{sys}}) + (\mathbf{P}_\star + \mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{11,\text{sys}})\mathbf{A}^\top = 0. \tag{E.11}$$

Clearly, $\mathbf{P}_\star + \mathbf{\Sigma}_{22,\mathsf{K}_\star} - \mathbf{\Sigma}_{11,\text{sys}} = 0$ is a valid solution to Eq. (E.11). As $\mathbf{A}$ is stable, the solution to the Lyapunov equation Eq. (E.11) is unique, and hence $\mathbf{P}_\star = \mathbf{\Sigma}_{11,\text{sys}} - \mathbf{\Sigma}_{22,\mathsf{K}_\star}$. Recall once more that due to Fact E.2 we have $\mathbf{\Sigma}_{12,\mathsf{K}_\star} = \mathbf{\Sigma}_{22,\mathsf{K}_\star}$, with $\mathbf{\Sigma}_{22,\mathsf{K}_\star}$ being symmetric. Therefore,

$$\mathbf{\Sigma}_{22,\mathsf{K}_\star} = \mathbf{\Sigma}_{12,\mathsf{K}_\star} = \mathbf{\Sigma}_{12,\mathsf{K}_\star}\mathbf{\Sigma}_{22,\mathsf{K}_\star}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}_\star}^\top =: \mathbf{Z}_\star,$$

and so $\mathbf{P}_\star = \mathbf{\Sigma}_{11,\text{sys}} - \mathbf{\Sigma}_{22,\mathsf{K}_\star} = \mathbf{\Sigma}_{11,\text{sys}} - \mathbf{Z}_\star$. Though we arrived at this conclusion via a specific realization Eq. (E.2) of the optimal policy $\mathsf{K}_\star$, both $\mathbf{\Sigma}_{11,\text{sys}}$ and $\mathbf{Z}_\star$ are independent of the realization of the optimal policy. ☐

E.6 Information-theoretic interpretation of $\mathbf{Z}_\mathsf{K}$

Recall that

$$\mathbf{\Sigma}_\mathsf{K} = \lim_{t\to\infty}\mathbb{E}\left[\begin{bmatrix}\mathbf{x}(t)\\ \hat{\mathbf{x}}(t)\end{bmatrix}\begin{bmatrix}\mathbf{x}(t)\\ \hat{\mathbf{x}}(t)\end{bmatrix}\right].$$

Since $(\mathbf{x}(t), \hat{\mathbf{x}}(t))$ are jointly Gaussian with zero mean, $(\mathbf{x}(t), \hat{\mathbf{x}}(t))$ converge in distribution to a limiting Gaussian distribution

$$\begin{bmatrix}\mathbf{x}_\infty\\ \hat{\mathbf{x}}_\infty\end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathsf{K}), \quad \mathbf{\Sigma}_\mathsf{K} = \begin{bmatrix}\mathbf{\Sigma}_{11,\text{sys}} & \mathbf{\Sigma}_{12,\mathsf{K}}\\ \mathbf{\Sigma}_{12,\mathsf{K}}^\top & \mathbf{\Sigma}_{22,\mathsf{K}}\end{bmatrix}.$$

The conditional covariance of $\mathbf{x}_\infty$ given $\hat{\mathbf{x}}_\infty$ is then given by the formula

$$\operatorname{Cov}[\mathbf{x}_\infty \mid \hat{\mathbf{x}}_\infty] = \mathbf{\Sigma}_{11,\text{sys}} - \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}} = \mathbf{\Sigma}_{11,\text{sys}} - \mathbf{Z}_\mathsf{K}.$$

In other words, $\mathbf{Z}_\mathsf{K}$ describes the reduction in covariance of $\mathbf{x}_\infty$ provided by the information in $\hat{\mathbf{x}}_\infty$.

E.7 Random Stable Initializations Are Informative

Lemma E.4. *Fix $\mathbf{C}_\mathsf{K}$, and suppose that the $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$ is chosen from some probability distribution $\mathbb{P}$ with density with respect to the Lebesgue measure on $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ satisfying $\mathbb{P}[\mathbf{A}_\mathsf{K}$ is Hurwitz$] = 1$. Then, $\mathbb{P}[\mathsf{K} \in \mathcal{K}_{\mathtt{info}}] = 1$.*

Proof. Let Hur_n denote the set of Hurwitz matrices in $\mathbb{R}^n \times n$. Note that if $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}) \in \mathsf{Hur}_n \times \mathbb{R}^{n\times m}$, then $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ if and only if $\operatorname{rank}(\mathbf{\Sigma}_{22,\mathsf{K}}) = n$ and $\operatorname{rank}(\mathbf{\Sigma}_{12,\mathsf{K}}) = n$. In fact, since $\mathbf{\Sigma}_\mathsf{K} \succeq 0$, The Schur complement test implies that $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ if and only if $\operatorname{rank}(\mathbf{\Sigma}_{12,\mathsf{K}}) = n$ (as this also implies $\operatorname{rank}(\mathbf{\Sigma}_{22,\mathsf{K}}) = n$). Thus, if $f(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$ is the mapping from $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$ to $\det(\mathbf{\Sigma}_{12,\mathsf{K}})$, then, given $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}) \in \mathsf{Hur}_n \times \mathbb{R}^{n\times m}$, $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ if and only if $f(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}) \neq 0$.

As shown in Claim D.3, the set Hur_n is open and connected, so the $\mathcal{U} := \mathsf{Hur}_n \times \mathbb{R}^{n\times m}$. Moreover, f does not identically vanish on $\mathcal{U}$: indeed, for any $(\mathbf{A}_{\mathsf{K}_\star}, \mathbf{B}_{\mathsf{K}_\star})$ corresponding to some $\mathsf{K}_\star \in \mathcal{K}_{\mathtt{opt}}$, we have $\operatorname{rank}(\mathbf{\Sigma}_{12,\mathsf{K}_\star}) = n$ by Lemma 3.1, so $f(\mathbf{A}_{\mathsf{K}_\star}, \mathbf{B}_{\mathsf{K}_\star}) \neq 0$.

Therefore, to prove our result, it suffices to show that f is an analytic function of $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$, and apply the identity theorem (Fact D.1). In fact, we show $f(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K})$ is an *rational function*. The following claim is useful.

Claim E.5. *Let* $\bar{F} : \mathsf{Hur}_{2n} \times \mathbb{S}^{2n} \to \mathbb{S}^{2n}$ *be the map for which* $\bar{F}(\bar{\mathbf{A}}, \bar{\mathbf{W}})$ *is the solution to the Lyapunov equation* $\bar{\mathbf{A}}\mathbf{\Gamma}+\mathbf{\Gamma}\bar{\mathbf{A}}+\bar{\mathbf{W}} = 0$*. Then* $\bar{F}$ *is a rational function with no poles on* $\mathsf{Hur}_{2n} \times \mathbb{S}^{2n}$.

Proof. Since this solution to the Lyapunov equation is unique for $\bar{\mathbf{A}} \in \mathsf{Hur}_{2n}$, we see that the map $\mathcal{T}_{\bar{\mathbf{A}}} : \mathbf{\Gamma} \mapsto \bar{\mathbf{A}}\mathbf{\Gamma} + \mathbf{\Gamma}\bar{\mathbf{A}}$ is invertible, and hence $\bar{F}(\bar{\mathbf{A}}, \bar{\mathbf{W}}) = \mathcal{T}_{\bar{\mathbf{A}}}^{-1}(\bar{\mathbf{W}})$. It follows that $\bar{F}(\bar{\mathbf{A}}, \bar{\mathbf{W}})$ is a rational function (notice the entries of $\mathcal{T}_{\bar{\mathbf{A}}}$ are linear in $\bar{\mathbf{A}}$, and thus the inverse is a rational function of $\mathcal{T}_{\bar{\mathbf{A}}}$ using the adjugate formula for matrix inverses). It has no polls because $\mathcal{T}_{\bar{\mathbf{A}}}$ is invertible for $\bar{\mathbf{A}} \in \mathsf{Hur}_{2n}$ ☐

By composing the rational $\bar{F}(\cdot, \cdot)$ in the above claim with the polynomial-function $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}) \mapsto (\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{W}_{\mathrm{cl},\mathsf{K}})$, we see that $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}) \mapsto \mathbf{\Sigma}_{\mathsf{K}}$ is a rational function function on $\mathsf{Hur}_n \times \mathbb{R}^{n\times m}$. In particular, $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}) \mapsto \mathbf{\Sigma}_{\mathsf{K}}$ is an analytic function. Thus, $f(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$, being a polynomial in $\mathbf{\Sigma}_{\mathsf{K}}$, is also rational. This concludes the proof. ☐

F Details for examples in Section 3

F.1 Details for Example 3.1

That $\mathsf{K}_{\mathrm{bad}}$ is a suboptimal stationary point follows from [Tang et al., 2021, Theorem 4.2], as `OE`is a special case of `LQG`. Nonetheless, it is straightforward to verify that $\mathsf{K}_{\mathrm{bad}}$ is indeed a stationary point. Specifically, one can readily verify that the controllability Gramian

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11,\mathrm{sys}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

satisfies the Lyapunov equation

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathrm{bad}} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{11,\mathrm{sys}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{\Sigma}_{11,\mathrm{sys}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathrm{bad}} \end{bmatrix}^\top + \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},$$

and that the observability Gramian

$$\mathcal{O} = \begin{bmatrix} \mathcal{O}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

satisfies the Lyapunov equation

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathrm{bad}} \end{bmatrix}^\top \begin{bmatrix} \mathcal{O}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathcal{O}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathrm{bad}} \end{bmatrix} + \begin{bmatrix} \mathbf{G}\mathbf{G}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0}.$$

It is then straightforward to confirm that

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}})}{\partial \mathbf{A}_{\mathrm{bad}}} &= 2\mathcal{O}_{12}^\top \mathbf{\Sigma}_{12} + 2\mathcal{O}_{22}\mathbf{\Sigma}_{22} \\ &= 2 \times \mathbf{0} \times \mathbf{0} + 2 \times \mathbf{0} \times \mathbf{0} = \mathbf{0}, \\ \frac{\partial \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}})}{\partial \mathbf{B}_{\mathrm{bad}}} &= 2(\mathcal{O}_{12}^\top \mathbf{\Sigma}_{11,\mathrm{sys}}\mathbf{C}^\top + \mathcal{O}_{22}\mathbf{\Sigma}_{12}^\top \mathbf{C}^\top + \mathcal{O}_{22}\mathbf{B}_{\mathrm{bad}}\mathbf{W}_2) \\ &= 2(\mathbf{0}^\top \times \mathbf{\Sigma}_{11,\mathrm{sys}}\mathbf{C}^\top + \mathbf{0} \times \mathbf{0}^\top \times \mathbf{C}^\top + \mathbf{0} \times \mathbf{0} \times \mathbf{W}_2) = \mathbf{0}, \\ \frac{\partial \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}})}{\partial \mathbf{C}_{\mathrm{bad}}} &= 2(\mathbf{C}_{\mathrm{bad}}\mathbf{\Sigma}_{22} - \mathbf{G}\mathbf{\Sigma}_{12}) \\ &= 2(\mathbf{0} \times \mathbf{0} - \mathbf{G} \times \mathbf{0}) = \mathbf{0}. \end{aligned}$$

Moreover, $\mathbf{T}\mathbf{B}_{\mathrm{bad}} = \mathbf{0}$ and $\mathbf{C}_{\mathrm{bad}}\mathbf{T}^{-1} = \mathbf{0}$ for all similarity transformations $\mathbf{T}$. Given that $\mathbf{B}_\star$ and $\mathbf{C}_\star$ are nonzero, it is clear that $\mathsf{K}_{\mathrm{bad}}$ is not equivalent to $\mathsf{K}_\star$ under any similarity transformation. Hence, $\mathsf{K}_{\mathrm{bad}}$ is suboptimal.

F.2 The perils of enforcing minimality

A classical result due to Brockett [1976] states that the set of minimal n-th order single input-single output transfer functions is the disjoint union of $n + 1$ open sets. Moreover, it is impossible for a continuous path through parameter space to pass from one of these open sets to another without entering a region corresponding to a non-minimal transfer function. This implies that if one were to regularize so as to ensure minimality of the filter at every iteration, the search will remain confined in the open set in which it is initialized, unable to reach the set containing the optimal filter, unless there is some mechanism (e.g. sufficiently large step size) by which to "hop" over the boundary of non-minimality, from one region to another. We now illustrate the possibility of this phenomenon (of remaining trapped in such a region) on a simple second-order ($n = 2$) example. We begin by characterizing the three open sets that partition the space of minimal second-order transfer functions; cf. [Brockett, 1976, §II] for derivation.

Fact F.1. Every strictly proper second-order transfer function with no pole-zero cancellations belongs to exactly one of the following three open sets, characterized as follows:

1. Both poles are real, and both residues are positive. This set is simply connected.
2. Poles are complex, or if both poles are real, then the residues have opposite signs. This set is not simply connected.
3. Both poles are real, and both residues are negative. This set is simply connected.

For the purpose of the following example, we shall refer to these sets as regions 1 to 3.

Example F.1. Consider `OE` instance given by:

$$\mathbf{A} = \begin{bmatrix} -1.2901 & -0.2626 \\ -0.2626 & -0.2814 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0.5710 & -0.5093 \end{bmatrix}, \quad \mathbf{G} = \mathbf{C},$$
$$\mathbf{W}_1 = \begin{bmatrix} 3.0940 & -1.5716 \\ -1.5716 & 1.2422 \end{bmatrix}, \quad \mathbf{W}_2 = 1.$$

It may be verified by straightforward calculations that the optimal filter $\mathsf{K}_\star$ for this instance belong to region 1. Let K_0 denote the filter from which policy search is initialized. K_0 is given by:

$$\mathbf{A}_{\mathsf{K}_0} = \begin{bmatrix} -9.863 & -20.19 \\ 17.4 & -4.143 \end{bmatrix}, \quad \mathbf{B}_{\mathsf{K}_0} = \begin{bmatrix} -1.499 \\ -16.44 \end{bmatrix}, \quad \mathbf{C}_{\mathsf{K}_0} = \begin{bmatrix} 11.56 & -2.97 \end{bmatrix}.$$

Similarly, it may be readily verified that K_0 belong to region 2.

We apply policy search to Example F.1, using four different regularization strategies:

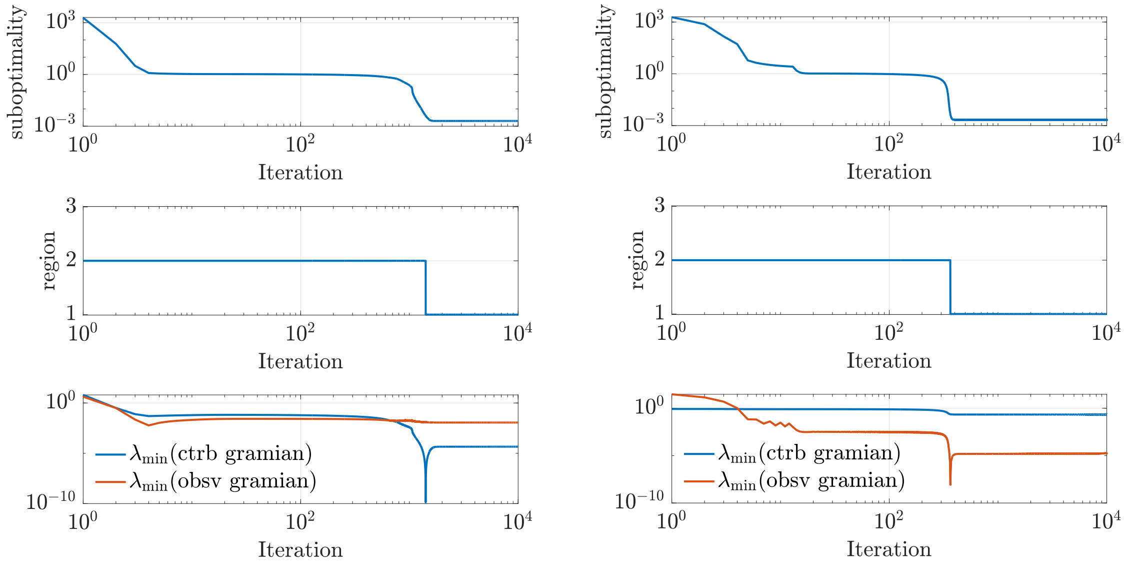
a. No regularization, i.e. gradient descent on $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$.

b. Regularization for controllability, i.e. gradient descent on $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \lambda \mathcal{R}_{\mathrm{ctr}}(\mathsf{K})$, where $\mathcal{R}_{\mathrm{ctr}}(\mathsf{K}) := \|\mathbf{Y}_{\mathrm{ctr},\mathsf{K}} - \mathbf{Y}_{\mathrm{ctr},\mathsf{K}}^{-1}\|_{\mathrm{F}}^2$ and $\mathbf{Y}_{\mathrm{ctr},\mathsf{K}}$ is the controllability Gramian for $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ satisfying the Lyapunov equation $\mathbf{A}_{\mathsf{K}}\mathbf{Y}_{\mathrm{ctr},\mathsf{K}} + \mathbf{Y}_{\mathrm{ctr},\mathsf{K}}\mathbf{A}_{\mathsf{K}}^\top + \mathbf{B}_{\mathsf{K}}\mathbf{B}_{\mathsf{K}}^\top = 0$.

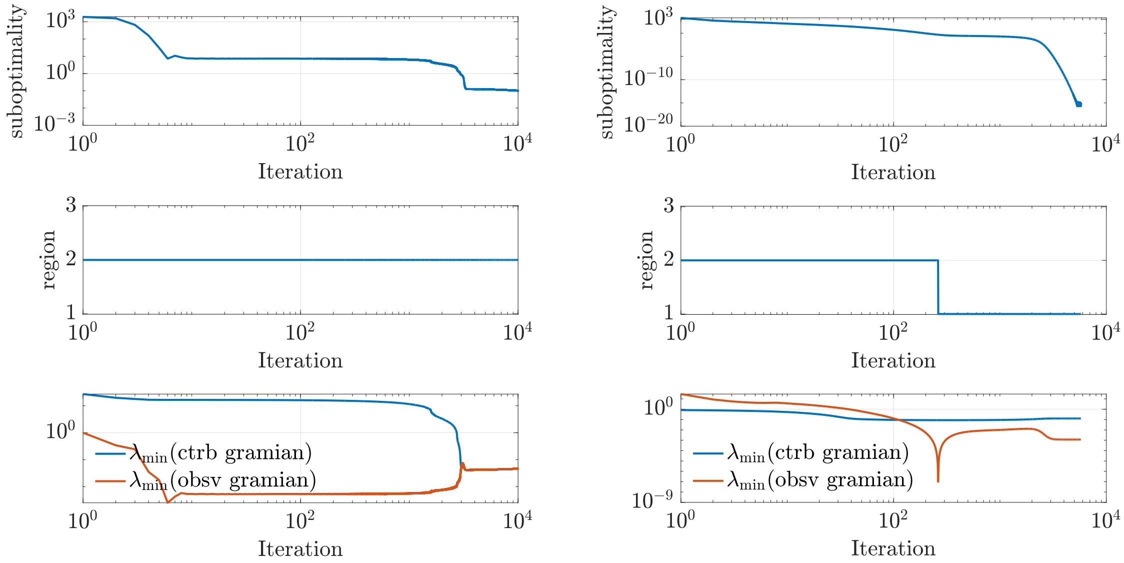
c. Regularization for minimality, i.e. gradient descent on $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \lambda(\mathcal{R}_{\mathrm{ctr}}(\mathsf{K}) + \mathcal{R}_{\mathrm{obs}}(\mathsf{K}))$, where $\mathcal{R}_{\mathrm{obs}}(\mathsf{K}) := \|\mathbf{Y}_{\mathrm{obs},\mathsf{K}} - \mathbf{Y}_{\mathrm{obs},\mathsf{K}}^{-1}\|_{\mathrm{F}}^2$ and $\mathbf{Y}_{\mathrm{obs},\mathsf{K}}$ is the observability Gramian for $(\mathbf{A}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$ satisfying the Lyapunov equation $\mathbf{A}_{\mathsf{K}}^\top\mathbf{Y}_{\mathrm{obs},\mathsf{K}} + \mathbf{Y}_{\mathrm{obs},\mathsf{K}}\mathbf{A}_{\mathsf{K}} + \mathbf{C}_{\mathsf{K}}^\top\mathbf{C}_{\mathsf{K}} = 0$.

d. The proposed algorithm IR-PG.

The results are presented in Fig. 1 below. Observe that while all other methods eventually cross from region 2 (containing the initial K_0) to region 1 (containing $\mathsf{K}_\star$), the method regularized to preserve minimality at each iteration remains "trapped" in region 2.



(a) No regularization. (b) Regularizing for controllability.



(c) Regularizing for minimality. (d) IR-PG.

Figure 1: Suboptimality, region of parameter space, and controllability/observability as a function of iteration for Example F.1 and four different regularization strategies. All searches are initialized at the same filter in region 2 of parameter space; the optimal filter is located in region 1. A backtracking line search is used in all instances. (a) with no regularization, the iterate crosses from region 2 to region 1 with a loss of controllability. (b) regularizing for controllability, the iterate now crosses from region 2 to region 1 with a loss of observability instead. (c) regularizing for minimality, the iterate never crosses from region 2 to region 1. (d) under the proposed method, IR-PG, the iterate crosses from region 2 to region 1 with a loss of observability, and quickly converges to the global optimum.

F.3 Insufficiency of controllability

Consider the $\mathtt{OE}$ instance given by

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{C} = \mathbf{I}_2, \quad \mathbf{W}_1 = 3 \times \mathbf{I}_2, \quad \mathbf{W}_2 = \mathbf{I}_2, \tag{F.1}$$

and the filter $\mathsf{K}_{\mathrm{bad}}$ given by

$$\mathbf{A}_{\mathrm{bad}} = \begin{bmatrix} -2 & 0 \\ \gamma & -\gamma \end{bmatrix}, \quad \mathbf{B}_{\mathrm{bad}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{C}_{\mathrm{bad}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{F.2}$$

The following shows that the true system Eq. (F.1) satisfies all our assumptions, and that the filter Eq. (F.2) is a critical point, but, because $\mathbf{\Sigma}_{12,\mathsf{K}_{\mathrm{bad}}}$ is not full rank, it is a strictly suboptimal first-order critical point of $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$. The following proposition is proven in Appendix F.4.

Proposition F.1. *For the* $\mathtt{OE}$ *instance Eq. (F.1) and any* $\gamma > 0$, *and any filter* $\mathsf{K}_{\mathrm{bad}}$ *of the form Eq. (F.2), the following are true:*

- *i.* *Eq. (F.1) satisfies Assumptions 2.1 to 2.4.*
- *ii.* $\mathsf{K}_{\mathrm{bad}} \in \mathcal{K}_{\mathtt{stab}}$.
- *iii.* $\mathsf{K}_{\mathrm{bad}}$ *is a first-order critical point:* $\nabla \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}}) = 0$.
- *iv.* *The filter is strictly suboptimal:* $\mathsf{K}_{\mathrm{bad}} \notin \mathcal{K}_{\mathtt{opt}}$.
- *v.* $\mathsf{K}_{\mathrm{bad}}$ *is controllable:* $\mathsf{K}_{\mathrm{bad}} \in \mathcal{K}_{\mathtt{ctrb}}$, $\Sigma_{\mathsf{K}_{\mathrm{bad}},22} \succ 0$.
- *vi.* $\mathbf{\Sigma}_{12,\mathsf{K}_{\mathrm{bad}}}$ *is not full rank.*

Moreover, $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ *does not depend on* γ, *showing that* $\mathcal{L}_{\mathtt{OE}}$ *does not have compact level sets.*

A proof of Proposition F.1 is given in Appendix F.4. Here, let us briefly describe the intuition behind this construction. To see that $\mathsf{K}_{\mathrm{bad}}$ is suboptimal, first notice that the true system in Eq. (F.1) comprises two independent, first-order subsystems. As the second row of $\mathbf{B}_{\mathrm{bad}}$ is zero, the output of the second subsystem will never enter the policy $\mathsf{K}_{\mathrm{bad}}$. In particular, the state of the policy will contain no information about the state of the second subsystem, resulting in suboptimal predictions concerning the second subsystem. To see that $\mathsf{K}_{\mathrm{bad}}$ is controllable, notice that the non-zero $(2, 1)$ entry of $\mathbf{A}_{\mathrm{bad}}$ allows the first component of the state of $\mathsf{K}_{\mathrm{bad}}$ to excite the second component. This ensures controllability of $\mathsf{K}_{\mathrm{bad}}$, even though the second state of $\mathsf{K}_{\mathrm{bad}}$ is not excited directly by the input to the policy (as the second row of $\mathbf{B}_{\mathrm{bad}}$ is zero). To see that $\mathsf{K}_{\mathrm{bad}}$ is a stationary point, first observe that the first row of the matrices comprising the policy $\mathsf{K}_{\mathrm{bad}}$ in Eq. (F.2) corresponds to the optimal policy (filter) for the first subsystem in Eq. (F.1), i.e. these are the optimal parameters that will provide the best possible prediction of the output of the first subsystem. Any single perturbation to one of these parameters will result in worse predictions and higher cost. Next, notice that the second row of $\mathbf{C}_{\mathrm{bad}}$ is zero; as such, any single perturbation to any parameter in the second row of $\mathbf{A}_{\mathrm{bad}}$ or $\mathbf{B}_{\mathrm{bad}}$ will not change the output of the policy, and therefore not change the cost. Finally, because the internal state of the policy contains no information about the state of the second subsystem in Eq. (F.1), any single perturbation to $\mathbf{C}_{\mathrm{bad}}$ will simply inject uncorrelated noise into the prediction for the second subsystem, thereby increasing the cost.

Moreover, as shown in Fig. 2 below, the minimum eigenvalue of the Hessian $\nabla^2 \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}})$ can be made arbitrarily close to zero by taking γ in Eq. (F.2) to be arbitrarily large. Existing results suggest that first order methods may take take $\Omega(\mathrm{poly}(\varepsilon))$-iterations to escape an approximate saddle point with minimum-Hessian eigenvalue ε [Jin et al., 2017, 2018, Carmon et al., 2018, Agarwal et al., 2017]; hence, these large-γ examples may prove challenging for first-order methods designed to escape approximate saddles. In addition, the non-compactness of the level sets for the $\mathtt{OE}$ objective may also lead to a number of pathologies.

Before closing, we note that one can similarly construct examples of policies that are observable, but not controllable, that correspond to suboptimal first-order critical points. For example, the policy

$$\mathbf{A}_{\mathrm{bad}} = \begin{bmatrix} -2 & 0 \\ 0 & -\gamma \end{bmatrix}, \quad \mathbf{B}_{\mathrm{bad}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{C}_{\mathrm{bad}} = \begin{bmatrix} 1 & \gamma \\ 0 & 0 \end{bmatrix}, \tag{F.3}$$

is observable for $\gamma > 0$, yet corresponds to a suboptimal first-order critical point of the `OE` loss for the true system Eq. (F.1). The intuition behind this construction is similar to that of Eq. (F.2) above. In particular, lack of controllability (notice that the $(2,1)$ entry of $\mathbf{A}_{\mathrm{bad}}$ is now zero) implies that the second component of the policy state decays to zero in steady state. As such, the non-zero $(1,2)$ entry of $\mathbf{C}_{\mathrm{bad}}$ does not disturb the optimal prediction for the first subsystem of Eq. (F.1). It does, however, ensure that $(\mathbf{C}_{\mathrm{bad}}, \mathbf{A}_{\mathrm{bad}})$ is observable.

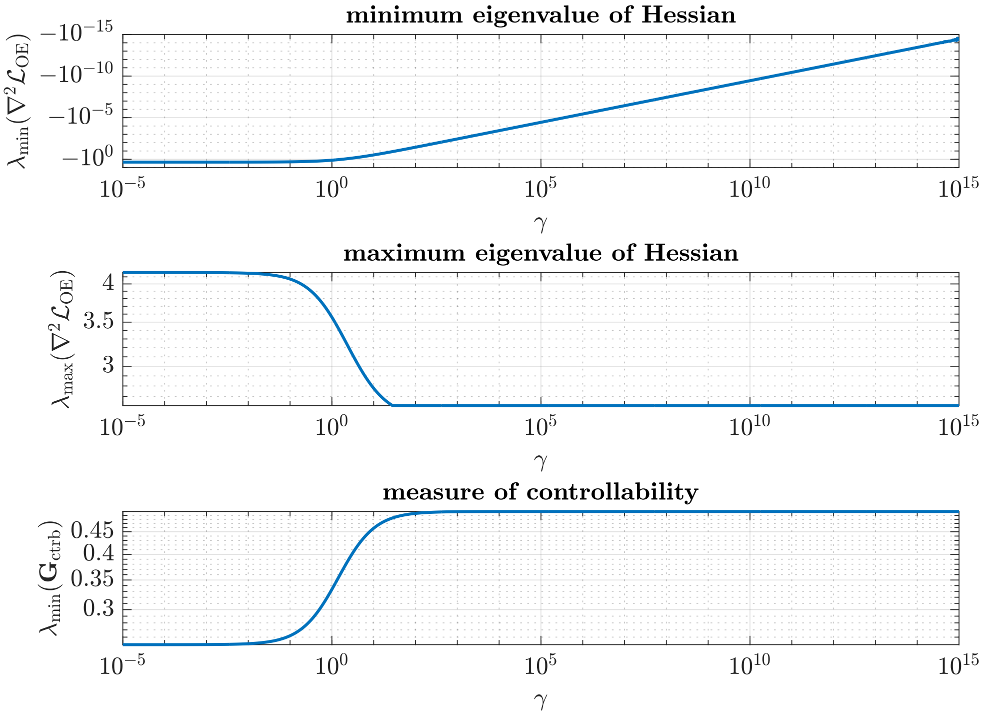


Figure 2: Spectral properties of the Hessian $\nabla^2\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_{\mathrm{bad}})$ in Proposition F.1 for various values γ, cf. $\mathbf{A}_{\mathrm{bad}}$ in Eq. (F.2). Here $\mathbf{G}_{\mathrm{ctrb}}$ denotes the controllability Gramian associated with $(\mathbf{A}_{\mathrm{bad}}, \mathbf{B}_{\mathrm{bad}})$.

F.4 Proof of Proposition F.1

Part i. Assumptions. The matrix $\mathbf{A}$ is Hurwitz stable, with eigenvalues -1 (repeated), meeting Assumption 2.1. The pair $(\mathbf{A}, \mathbf{C})$ is observable, as $\mathbf{C} = \mathbf{I}_2$, meeting Assumption 2.2. $\mathbf{W}_1$ and $\mathbf{W}_2$ are also clearly positive definite, meeting Assumption 2.3. Lastly, one can show that

$$(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}) = (-2\mathbf{I}_2, \mathbf{I}_2, \mathbf{I}_n)$$

is an optimal filter. Clearly $(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}})$ is controllable, so Assumption 2.4 is met.

Part ii. Stability. As $\mathbf{A}_{\mathrm{bad}}$ is lower diagonal, the eigenvalues are easily seen to be $(-2, -\gamma)$. Hence $\mathbf{A}_{\mathrm{bad}}$ is Hurwitz stable.

Part iii. First-Order Critical Point. Decompose

$$\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) = \mathbb{E}[\|\mathbf{x} - \hat{\mathbf{z}}_{\mathsf{K}}\|^2] = \underbrace{\mathbb{E}[|\mathbf{x}[1] - \hat{\mathbf{z}}_{\mathsf{K}}[1]|^2]}_{\mathcal{L}_1(\mathsf{K})} + \underbrace{\mathbb{E}[|\mathbf{x}[2] - \hat{\mathbf{z}}_{\mathsf{K}}[2]|^2]}_{\mathcal{L}_2(\mathsf{K})}, \tag{F.4}$$

where $(\mathbf{x}, \hat{\mathbf{z}}_{\mathsf{K}})$ are jointly distribution as $\mathcal{N}(0, \mathbf{\Sigma}_{\mathsf{K}})$. We show $\mathsf{K} = \mathsf{K}_{\mathrm{bad}}$ is a critical point of both $\mathcal{L}_1(\mathsf{K})$ and $\mathcal{L}_2(\mathsf{K})$. We start with $\mathcal{L}_1(\mathsf{K})$.

Claim F.2. *We have* $\nabla_{\mathsf{K}}\mathcal{L}_1(\mathsf{K})\big|_{\mathsf{K}=\mathsf{K}_{\mathrm{bad}}} = 0$.

Proof. It suffices to show that $\mathsf{K} = \mathsf{K}_{\mathrm{bad}}$ is global minimizer of $\mathcal{L}_1(\cdot)$. This can be checked by showing that $(a_{\mathsf{K}}, b_{\mathsf{K}}, c_{\mathsf{K}}) = (-2, 1, 1)$ is the optimal solution to the one-dimensional scalar `OE` problem with $(a, c, w_1, w_2) = (-1, 1, 3, 1)$ and $z = 1$. Solving the scalar Continuous Algebraic Riccati Equation, we see that an optimal filter is of the form $(a_{\mathsf{K}}, b_{\mathsf{K}}, c_{\mathsf{K}}) = (a - \ell, 1, 1)$, where $\ell = w_2^{-1} cp = p$, and $p > 0$ solves the continuous Algebriac Ricatti Equation

$$0 = ap + pa + p^2 b^2 w_2^{-1} + w_1 = -2p - p^2 + 3$$

Taking the positive solution to the quadratic $0 = p^2 + 2p - 3 = (p+3)(p-1)$, we have $p = 1$. Hence, the optimal filter has $l = w_2^{-1} cp = 1$. Hence, $(a - \ell, 1, 1) = (-1 - 1, 1, 1) = (-2, 1, 1)$ is an optimal solution to the scalar `OE` problem, as needed. □

Next, we address $\mathcal{L}_2(\mathsf{K})$. We begin with a lemma establishing the structure of $\mathbf{\Sigma}_{12,\mathsf{K}}$ for $\mathsf{K} = \mathsf{K}_{\mathrm{bad}}$, proven in Appendix F.5.

Lemma F.3. *For* $\mathsf{K} = \mathsf{K}_{\mathrm{bad}}$, *we have*

$$\mathbf{\Sigma}_{12,\mathsf{K}} = \begin{bmatrix} \frac{1}{2} & \frac{\gamma}{2(1+\gamma)} \\ 0 & 0 \end{bmatrix}.$$

We can now conclude by checking that $\mathsf{K}_{\mathrm{bad}}$ is a criticial point of $\mathcal{L}_2(\cdot)$.

Claim F.4. *We have* $\nabla_{\mathsf{K}} \mathcal{L}_2(\mathsf{K})\big|_{\mathsf{K} = \mathsf{K}_{\mathrm{bad}}} = 0$.

Proof. For simplicity, we drop the subscripts involving K.

$$\begin{aligned}
\mathcal{L}_2(\mathsf{K}) &= \mathbb{E}[|\mathbf{x}[2] - \hat{\mathbf{z}}[2]|^2] = \mathbb{E}[|\mathbf{x}[2] - \boldsymbol{e}_2^\top \mathbf{C}_{\mathsf{K}} \hat{\mathbf{x}}|^2] \\
&= \mathbb{E}[\mathbf{x}[2]^2] - 2\boldsymbol{e}_2^\top \mathbb{E}[\mathbf{x}\hat{\mathbf{x}}^\top] \mathbf{C}_{\mathsf{K}}^\top \boldsymbol{e}_2 + \boldsymbol{e}_2^\top \mathbf{C}_{\mathsf{K},2} \mathbb{E}[\hat{\mathbf{x}}\hat{\mathbf{x}}^\top] \boldsymbol{e}_2^\top \mathbf{C}_{\mathsf{K},2} \\
&= \mathbb{E}[\mathbf{x}[2]^2] - 2\boldsymbol{e}_2^\top \mathbf{\Sigma}_{12,\mathsf{K}} \mathbf{C}_{\mathsf{K}} \boldsymbol{e}_2 + \boldsymbol{e}_2^\top \mathbf{C}_{\mathsf{K}} \mathbf{\Sigma}_{22,\mathsf{K}} \mathbf{C}_{\mathsf{K}}^\top \boldsymbol{e}_2 \\
&= \underbrace{\mathbb{E}[\mathbf{x}[2]^2]}_{=\mathcal{L}_2(\mathsf{K}_{\mathrm{bad}})} - 2\boldsymbol{e}_2^\top (\mathbf{\Sigma}_{12,\mathsf{K}} - \mathbf{\Sigma}_{12,\mathrm{bad}})(\mathbf{C}_{\mathsf{K}} - \mathbf{C}_{\mathrm{bad}})^\top \boldsymbol{e}_2 + \boldsymbol{e}_2^\top (\mathbf{C}_{\mathsf{K}} - \mathbf{C}_{\mathrm{bad}})^\top \mathbf{\Sigma}_{22,\mathsf{K}} (\mathbf{C}_{\mathsf{K}} - \mathbf{C}_{\mathrm{bad}})^\top \boldsymbol{e}_2,
\end{aligned}$$

where above we use $\mathbf{C}_{\mathrm{bad}}^\top \boldsymbol{e}_2 = 0$ and, as shown in in Lemma F.3, $\boldsymbol{e}_2^\top \mathbf{\Sigma}_{12,\mathrm{bad}} = 0$. In particular, for a perturbation $\mathbf{\Delta}_{\mathsf{K}} = (\mathbf{\Delta}_A, \mathbf{\Delta}_B, \mathbf{\Delta}_C)$,

$$\begin{aligned}
&\mathcal{L}_2(\mathsf{K}_{\mathrm{bad}} + t\mathbf{\Delta}_{\mathsf{K}}) - \mathcal{L}_2(\mathsf{K}_{\mathrm{bad}}) \\
&= -t\boldsymbol{e}_2^\top (\mathbf{\Sigma}_{12,\mathsf{K}_{\mathrm{bad}} + t\mathbf{\Delta}_{\mathsf{K}}} - \mathbf{\Sigma}_{12,\mathrm{bad}}) \mathbf{\Delta}_{\mathbf{C}}^\top \boldsymbol{e}_2 + t^2 \boldsymbol{e}_2^\top \mathbf{\Delta}_C \mathbf{\Sigma}_{22,\mathsf{K}_{\mathrm{bad}} + t\mathbf{\Delta}_{\mathsf{K}}} \mathbf{\Delta}_C^\top \boldsymbol{e}_2 \\
&= -t\boldsymbol{e}_2^\top \mathbf{\Sigma}_{12,\mathrm{bad}} \mathbf{\Delta}_{\mathbf{C}}^\top \boldsymbol{e}_2 - t^2 \boldsymbol{e}_2^\top \mathbf{\Delta}_{12} \mathbf{\Delta}_{\mathbf{C}}^\top \boldsymbol{e}_2 + t^2 \boldsymbol{e}_2^\top \mathbf{\Delta}_C \mathbf{\Sigma}_{22,\mathrm{bad}} \mathbf{\Delta}_C^\top \boldsymbol{e}_2 + O(t^3) \\
&= -t^2 \boldsymbol{e}_2^\top \left(\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{\Sigma}_{\mathsf{K}_{\mathrm{bad}} + t\mathbf{\Delta}_{\mathsf{K}},12} \Big|_{t=0} \right) \mathbf{\Delta}_{\mathbf{C}}^\top \boldsymbol{e}_2 + t^2 \boldsymbol{e}_2^\top \mathbf{\Delta}_C \mathbf{\Sigma}_{22,\mathrm{bad}} \mathbf{\Delta}_C^\top \boldsymbol{e}_2 + O(t^3),
\end{aligned}$$

where again we use $\boldsymbol{e}_2^\top \mathbf{\Sigma}_{12,\mathrm{bad}} = 0$ by Lemma F.3. Thus, $\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{L}_2(\mathsf{K}_{\mathrm{bad}} + t\mathbf{\Delta}_{\mathsf{K}}) = 0$, showing $\nabla \mathcal{L}_2(\mathsf{K})\big|_{\mathsf{K} = \mathsf{K}_{\mathrm{bad}}} = 0$. □

Part vi. Suboptimality. By solving the continuous algebraic Ricatti equation (in the spirit of Claim F.2), one can show that

$$(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}) = (-2\mathbf{I}_2, \mathbf{I}_2, \mathbf{I}_n)$$

is *an* optimal filter. It is clear that there is no similarity transformation which relates this filter to $\mathsf{K}_{\mathrm{bad}} = (\mathbf{A}_{\mathrm{bad}}, \mathbf{B}_{\mathrm{bad}}, \mathbf{C}_{\mathrm{bad}})$ (for one, the the rank of $\mathbf{B}_{\mathsf{K}}$, $\mathbf{C}_{\mathsf{K}}$ would be preserved under such a similarity transform). Since optimal filters are unique up to similarity transform (Lemma E.3), $\mathsf{K}_{\mathrm{bad}}$ cannot be optimal.

Part v. Controllability and rank of $\mathbf{\Sigma}_{22,\mathsf{K}}$ ($\mathsf{K} \in \mathcal{K}_{\mathtt{ctrb}}$) As shown in Appendix E.1, $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$ provided that $(\mathbf{A}_{\mathrm{bad}}, \mathbf{B}_{\mathrm{bad}})$ is controllable. The latter can be verified since $\mathbf{B}_{\mathrm{bad}} = [e_1 \mid \mathbf{0}_2]$, and e_1 is not an eigenvector of $\mathbf{A}_{\mathrm{bad}}$.

Part vi. Rank of $\mathbf{\Sigma}_{12,\mathsf{K}}$ The computation in Lemma F.3 shows $\mathbf{\Sigma}_{12,\mathsf{K}}$ has rank 1.

This concludes the demonstration of points i-vi. To see uniform boundedness, we again decomposition $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) = \mathcal{L}_1(\mathsf{K}) + \mathcal{L}_2(\mathsf{K})$ as in Eq. (F.4). Since $\mathcal{L}_1(\mathsf{K})$ is globally minimized at $\mathsf{K} = \mathsf{K}_{\mathrm{bad}}$, and since $\hat{\mathbf{z}}[2] \equiv 0$ regardless of γ, we see $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ does not depend on γ. □

F.5 Proof of Lemma F.3

Proof. Writing out the Lyapunov equation (and using $*$ to ignore irrelevant blocks),

$$
\begin{aligned}
&-\begin{bmatrix}3\mathbf{I}_n & 0\\ * & *\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_{\mathsf{K}} & *\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{11,\mathrm{sys}} & \boldsymbol{\Sigma}_{12,\mathsf{K}}\\ \boldsymbol{\Sigma}_{12,\mathsf{K}}^\top & *\end{bmatrix} + \left(\begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_{\mathrm{bad}} & \mathbf{A}_{\mathrm{bad}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{11,\mathrm{sys}} & \boldsymbol{\Sigma}_{12,\mathsf{K}}\\ \boldsymbol{\Sigma}_{12,\mathsf{K}}^\top & *\end{bmatrix}\right)^\top\\
&= \begin{bmatrix}\mathbf{A}\boldsymbol{\Sigma}_{11,\mathrm{sys}} & \mathbf{A}\boldsymbol{\Sigma}_{12,\mathsf{K}}\\ \mathbf{B}_{\mathrm{bad}}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \mathbf{A}_{\mathrm{bad}}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top & *\end{bmatrix} + \left(\begin{bmatrix}\mathbf{A}\boldsymbol{\Sigma}_{11,\mathrm{sys}} & \mathbf{A}\boldsymbol{\Sigma}_{12,\mathsf{K}}\\ \mathbf{B}_{\mathrm{bad}}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \mathbf{A}_{\mathrm{bad}}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top & *\end{bmatrix}\right)^\top\\
&= \begin{bmatrix}\mathbf{A}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \boldsymbol{\Sigma}_{11,\mathrm{sys}}\mathbf{A}^\top & \mathbf{A}\boldsymbol{\Sigma}_{12,\mathsf{K}} + (\mathbf{B}_{\mathrm{bad}}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \mathbf{A}_{\mathrm{bad}}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top)^\top\\ * & *\end{bmatrix}
\end{aligned}
$$

Using $\mathbf{A} = -\mathbf{I}_2$, we have $-3\mathbf{I}_2 = -2\boldsymbol{\Sigma}_{11,\mathrm{sys}}$, so $\boldsymbol{\Sigma}_{11,\mathrm{sys}} = \frac{3}{2}\mathbf{I}_2$. Then,

$$
\begin{aligned}
0 &= \mathbf{A}\boldsymbol{\Sigma}_{12,\mathsf{K}} + (\mathbf{B}_{\mathrm{bad}}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \mathbf{A}_{\mathrm{bad}}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top)^\top\\
&= -\boldsymbol{\Sigma}_{12,\mathsf{K}} + \frac{1}{2}(3\mathbf{I}_2)\mathbf{B}_{\mathrm{bad}}^\top + \boldsymbol{\Sigma}_{12}\mathbf{A}_{\mathrm{bad}}^\top\\
&= \frac{3}{2}\mathbf{B}_{\mathrm{bad}}^\top + \boldsymbol{\Sigma}_{12,\mathsf{K}}(\mathbf{A}_{\mathrm{bad}} - \mathbf{I}_n)^\top,
\end{aligned}
$$

so that

$$
\boldsymbol{\Sigma}_{12,\mathsf{K}} = -\frac{3}{2}\mathbf{B}_{\mathrm{bad}}^\top(\mathbf{A}_{\mathrm{bad}} - \mathbf{I}_n)^{-\top}
$$

Next,

$$
\begin{aligned}
(\mathbf{A}_{\mathrm{bad}} - \mathbf{I}_n)^{-1} &= -\left(\begin{bmatrix}1 + a_\star & 0\\ -\gamma & 1+\gamma\end{bmatrix}\right)^{-1}\\
&= -\begin{bmatrix}(1+a_\star)^{-1} & 0\\ \frac{\gamma}{(1+a_\star)(1+\gamma)} & (1+\gamma)^{-1}\end{bmatrix}\\
&= \begin{bmatrix}-\frac{1}{3} & 0\\ \frac{-\gamma}{3(1+\gamma)} & \frac{-1}{1+\gamma}\end{bmatrix}
\end{aligned}
$$

So, substituing in the definition of $\mathbf{B}_{\mathrm{bad}}$

$$
\boldsymbol{\Sigma}_{12,\mathsf{K}} = -\frac{3}{2}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}^\top\begin{bmatrix}-\frac{1}{3} & 0\\ \frac{-\gamma}{3(1+\gamma)} & \frac{-1}{1+\gamma}\end{bmatrix}^\top = \begin{bmatrix}\frac{1}{2} & \frac{\gamma}{2(1+\gamma)}\\ 0 & 0\end{bmatrix}
$$

. □

F.6 Additional numerical examples

In this subsection we present the results of a number of additional numerical experiments illustrating the performance of IR-PG. All numerical experiments are carried out with a 2.4 GHz 8-Core Intel Core i9 processor with 64 GB of RAM.

Random generation of true systems. Each experimental trial begins with the random generation of a *true system* of the form Eq. (1.1). System parameters $\mathbf{A}, \mathbf{C}$ are randomly generated using `Matlab`'s `rss` function, with state dimension $n = 2$ and output dimension $m = 1$. The matrix $\mathbf{G}$ defining the mapping from state to performance output $\mathbf{z}$ is set to $\mathbf{G} = I$. The intensity of the system disturbances is randomly generated as $\mathbf{W}_1 = \mathbf{M}^\top\mathbf{M}$ with each entry of $\mathbf{M} \in \mathbb{R}^{n\times n}$ sampled from $\mathcal{N}(0,1)$. The intensity of the measurement noise is normalized to $\mathbf{W}_2 = 1$. To select suitable systems, we then rejection sample according to the following criteria: i) $\mathbf{A}$ must be strictly stable, and the observability Gramian $\mathcal{O}$ corresponding to $(\mathbf{A}, \mathbf{C})$ must satisfy $10^{-4} \le \lambda_{\min}(\mathcal{O}) \le 10^{-2}$; ii) $\mathbf{W}_1$ must satisfy $\lambda_{\max}(\mathbf{W}_1) \le 5$; iii) the optimal cost must satisfy $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star) \le 10^3$. The first criterion regulates the observability of the true system, which sets the difficulty of the filtering problem; the second ensures that the ratio between the disturbances and measurement noise remains “reasoanble”; and the third ensures that the problem instance is not “pathological”, as determined by excessively high cost of the optimal filter.

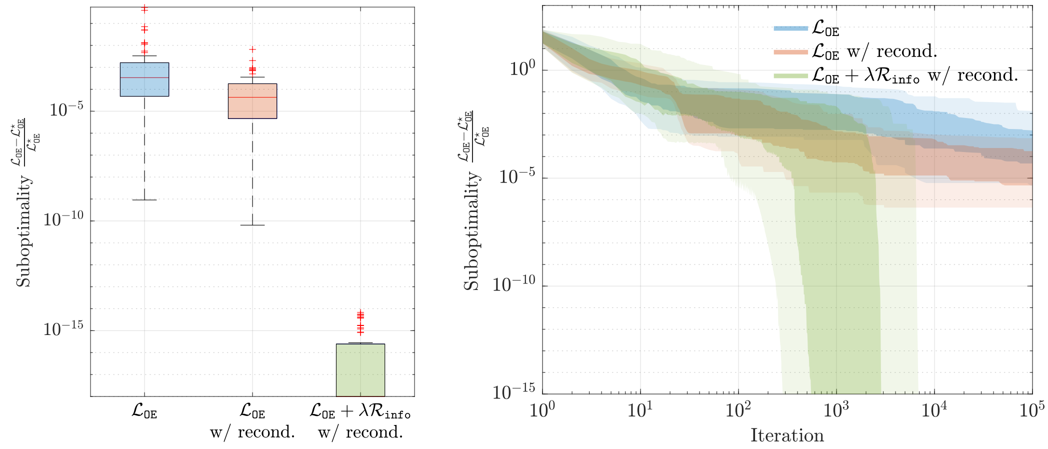
Remark F.1 (Choice of $\mathbf{G} = I$.)**.** As detailed in Section 3.2, IR-PG makes use of the regularizer $\mathcal{R}_{\mathtt{info}}$, defined in Eq. (3.1), the computation of which requires access to the true system states $\mathbf{x}$, as described in Section 2. To facilitate a more fair comparison with direct minimization of $\mathcal{L}_{\mathtt{OE}}$, we selected $\mathbf{G} = \mathbf{I}$ to effectively give the optimizer of $\mathcal{L}_{\mathtt{OE}}$ access to the true system states $\mathbf{x}$ as well. As a result, all algorithms compared in this section have access to the same information concerning the true system.

Random generation of initial filters. Next we randomly generate a filter K_0 from which to initialize gradient descent. To do so, we take the optimal (Kalman) filter $\mathsf{K}_\star$, and randomly perturb each of the parameters; specifically, we set $(\mathsf{K}_0)_i = (\mathsf{K}_\star)_i + \delta_i$ with $\delta_i \sim \mathcal{N}(0, 100)$ for the ith parameter. Before accepting this K_0, we rejection sample based on the following criteria: i) $\mathbf{\Sigma}_{\mathsf{K}_0}$ must satisfy $10^{-5} \leq \sigma_{\min}(\mathbf{\Sigma}_{12,\mathsf{K}_0}) \leq 10^{-3}$; ii) $\mathbf{\Sigma}_{\mathsf{K}_0}$ must satisfy $10^{-3} \leq \sigma_{\min}(\mathbf{\Sigma}_{22,\mathsf{K}_0}) \leq 1$; iii) the initial suboptimality must satisfy $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_0) \leq 100 \times \mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)$. The first criterion ensures that we do not begin from an initial guess for which the informativity is too low, nor a guess for which it is too high (which makes the search easier). The second criterion ensures that the initial filter is sufficiently controllable, to avoid initializations that are too close to suboptimal stationary points. The final criterion ensures that the initial guess is, in all other ways, "reasonable", as measured by suboptimality.

Optimization methods compared. Given a randomly generated true system, and random initial filter K_0, we then apply the following three optimization algorithms: i) gradient descent on $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$; ii) gradient descent on $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ with filter state normalization performed before each gradient step, cf. Eq. (3.2); iii) IR-PG, as detailed in Algorithm 1, with regularization parameter $\lambda = 10^{-4}$. See below for further discussion on the selection of λ. All methods are initialized from the same K_0, and make use of the same backtracking line search to select step sizes. Moreover, all algorithms have the same termination criteria. Each algorithm terminates when either: i) the Frobenius norm of the gradient of the cost function being minimized (either $\mathcal{L}_{\mathtt{OE}}$ or $\mathcal{L}_\lambda$) falls below a tolerance of 10^{-8}; ii) the step size selected by the line search falls below a tolerance of 10^{-16} for more than three consecutive iterations; or iii) the number of iterations (gradient descent steps) exceeds $100,000$.

Results. The results of 60 such experimental trials are depicted in Fig. 3. It is evident that simple "unregularized" gradient descent on $\mathcal{L}_{\mathtt{OE}}$ routinely fails to converge to the global optimum, in the allotted number of iterations. In fact, the median (normalized) suboptimality gap $\frac{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})-\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)}{\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)}$ exceeds 10^{-4}, and only a single trial achieves suboptimality less than 10^{-7}. Loss of informativity in these trials can be seen clearly in Fig. 4. The addition of the filter state reconditioning procedure of Eq. (3.2) offers only minimal improvement. In contrast, IR-PG converges reliably to high-quality solutions that are extremely close to the global optimum; the median normalized suboptimality gap was zero, to numerical precision. In fact, for one third of trials, the suboptimality gap was actually *negative* (by very small margins, e.g. 10^{-17}) indicating that IR-PG has reached the limits of numerical precision with which `Matlab`'s `icare` solves Riccati equations (which we use to compute $\mathsf{K}_\star$).

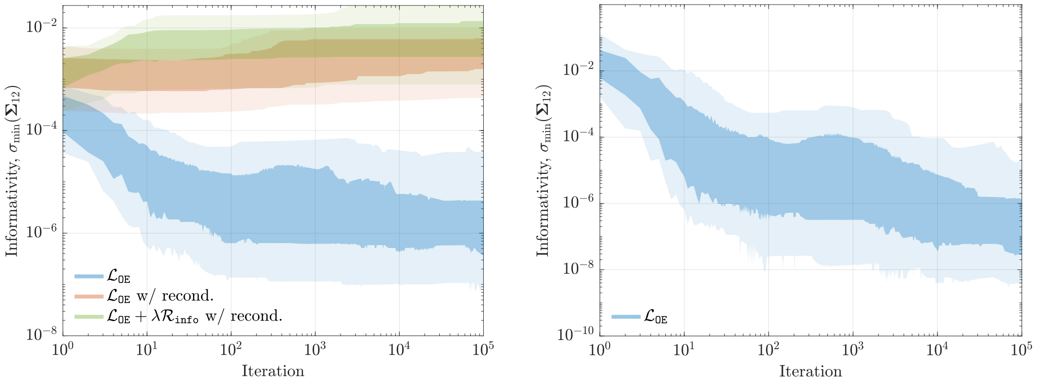
Selection of regularization parameter λ. Performance of IR-PG is in many instances insensitive to the value of λ selected. Experiments were conducted with $\lambda = 10^{-4}$. However, we observed that handful of experimental trails required λ to be chosen more judiciously, in particular, when the spectral properties of $\nabla^2\mathcal{R}_{\mathtt{info}}$ differ significantly from those of $\nabla^2\mathcal{L}_{\mathtt{OE}}$. Very small stepsizes may be required when $\lambda_{\max}(\nabla^2\mathcal{R}_{\mathtt{info}})$ is very large, which means the search may make slow progress in updating $\mathbf{C}_{\mathsf{K}}$, as $\mathcal{R}_{\mathtt{info}}$ is independent of $\mathbf{C}_{\mathsf{K}}$. We have observed good performance in practice by simply "turning off" the regularizer (i.e. setting $\lambda = 0$) when the stepsize becomes excessively small (e.g. drops below 10^{-16}).



(a) Normalized suboptimality at the termination of each algorithm.

(b) Normalized suboptimality as a function of iteration for each algorithm.

Figure 3: Performance of each algorithm as measured by the normalized suboptimality of the output estimation cost, $\frac{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})-\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)}{\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_\star)}$. 60 trials of the experimental procedure described in Appendix F.6 are plotted. In (b), the lightly shaded region covers the 10th to 90th percentiles, and the darker region covers the 25th to 75th percentiles.



(a) Informativity, as measured by $\sigma_{\min}(\mathbf{\Sigma}_{12})$.

(b) Conditioning, as measured by $\sigma_{\min}(\mathbf{\Sigma}_{22})$.

Figure 4: Properties of $\mathbf{\Sigma}$ for the same 60 trials plotted in Fig. 3. The lightly shaded region covers the 10th to 90th percentiles, and the darker region covers the 25th to 75th percentiles.

Part II

Proofs for Convergence Guarantee

G Proof of Theorems 1 and 2

G.1 Gradient descent with reconditioning

Before outlining the formal steps of our main results, we provide analyze gradient descent under the weak-PL condition. This generalizes Proposition 4.1 to accomodate the reconditioning step in IR-PG (Algorithm 1). All proofs are deferred to Appendix H.3.

Definition G.1 (Reconditioning matrix)**.** Given $f : \mathbb{R}^d \to \bar{\mathbb{R}}$, we say that $\mathbf{\Lambda} : \mathsf{dom}(f) \to \mathbb{S}^n_+$ is a reconditioning matrix for f if it is continuous on $\mathsf{dom}(f)$, and for every $\boldsymbol{x} \in \mathsf{dom}(f)$ such that $\mathbf{\Lambda}(\boldsymbol{x}) \succ 0$, there exists an $\boldsymbol{x}' \in \mathsf{dom}(f)$ such that $\mathbf{\Lambda}(\boldsymbol{x}') = \mathbf{I}_n$ and $f(\boldsymbol{x}') = f(\boldsymbol{x})$. We define the set $\mathsf{recond}_{\mathbf{\Lambda}}(\boldsymbol{x}) := \{\boldsymbol{x}' : f(\boldsymbol{x}') = f(\boldsymbol{x}), \quad \mathbf{\Lambda}(\boldsymbol{x}) = \mathbf{I}_n\}$ as the set of such points. We say $\boldsymbol{x}$ is reconditioned if $\mathbf{\Lambda}(\boldsymbol{x}) = \mathbf{I}_n$.

Observation G.1. $\mathbf{\Lambda}(\mathsf{K}) = \mathbf{\Sigma}_{\mathsf{K},22}$ is a reconditioning matrix for the loss $\mathcal{L}_{\lambda(\cdot)}$.

Proof. Since $\mathsf{dom}(\mathcal{L}_\lambda) = \mathcal{K}_{\mathtt{info}} \subset \mathcal{K}_{\mathtt{ctrb}}$, $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$ on $\mathsf{dom}(\mathcal{L}_\lambda)$. As observed in Eq. (3.2), there is a similarity transformation mapping $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ some K' with $\mathbf{\Sigma}_{\mathsf{K}',22} = \mathbf{I}_n$. Since $\mathcal{L}_\lambda$ is invariant under similarity transformation, it follows $\mathcal{L}_\lambda(\mathsf{K}') = \mathcal{L}_\lambda(\mathsf{K})$. ☐

Reconditioning serves to ensure that f need only be well-behaved (i.e. satisfy upper-smoothness and weak-PL for suitable constants) on a restricted set of approximately reconditioned parameters $\boldsymbol{x} : \mathbf{\Lambda}(\boldsymbol{x}) \approx \mathbf{I}_n$.

The following proposition is the guiding template for the overall convergence analysis. Its proof is given in Appendix H.3.1.

Proposition G.2. *Let* $f : \mathbb{R}^d \to \bar{\mathbb{R}}$, $\boldsymbol{x}_0 \in \mathsf{dom}(f)$*, and let* $\mathbf{\Lambda}$ *be a reconditioning matrix for* f *such that* $\mathbf{\Lambda}(\boldsymbol{x}_0) \succ 0$*. Define* $\mathcal{K}(\boldsymbol{x}_0)$ *as the following reconditioned level set, which we assume is closed:*

$$\mathcal{K}(\boldsymbol{x}_0) := \left\{ \boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \le f(\boldsymbol{x}_0) \text{ and } \|\mathbf{\Lambda}(\boldsymbol{x}) - \mathbf{I}_n\|_{\mathrm{op}} \le \frac{1}{2} \right\}. \tag{G.1}$$

Assume that the function $\boldsymbol{x} \mapsto \mathbf{\Lambda}(\boldsymbol{x})$ *is* $L_{\mathrm{cond},\boldsymbol{x}_0}$*-Lipschitz as a mapping from* $(\mathbb{R}^d, \|\cdot\|) \to (\mathbb{S}^n_+, \|\cdot\|_{\mathrm{op}})$ *and that* f *is* $\beta_{\boldsymbol{x}_0}$*-upper-smooth,* $L_{f,\boldsymbol{x}_0}$*-Lipschitz, and satisfies the* $\alpha_{\boldsymbol{x}_0}$*-weak PL condition for points in* $\mathcal{K}(\boldsymbol{x}_0)$*. Lastly, let* $\{\eta_k\}_{k=0}^\infty$ *be a series of step sizes such that* $0 < \inf_k \eta_k \le \sup_k \eta_k \le \min\{\frac{1}{\beta_{\boldsymbol{x}_0}}, \frac{1}{2L_{f,\boldsymbol{x}_0} L_{\mathrm{cond},\boldsymbol{x}_0}}\}$*. If iterates are chosen according to,*

$$\widetilde{\boldsymbol{x}}_k \in \mathsf{recond}_{\mathbf{\Lambda}}(\boldsymbol{x}_k), \quad \boldsymbol{x}_{k+1} = \widetilde{\boldsymbol{x}}_k - \eta_k \nabla f(\widetilde{\boldsymbol{x}}_k), \tag{G.2}$$

or the more general condition,

$$\widetilde{\boldsymbol{x}}_k \in \mathsf{recond}_{\mathbf{\Lambda}}(\boldsymbol{x}_k), \quad \boldsymbol{x}_{k+1} \text{ satisfies } f(\boldsymbol{x}_{k+1}) \le f(\widetilde{\boldsymbol{x}}_k - \eta_k \nabla f(\widetilde{\boldsymbol{x}}_k)), \tag{G.3}$$

then for all $k \ge 1$ *it holds that*

$$f(\boldsymbol{x}_k) \le \frac{2}{\alpha_{\boldsymbol{x}_0}^2 \eta} \cdot \frac{1}{k}, \quad \text{where } \eta := \inf_{k \ge 1} \eta_k. \tag{G.4}$$

Proposition G.2 can also be used to establish that every $\boldsymbol{x}_0 \in \mathsf{dom}(f)$ is in the path-connected component of some $\boldsymbol{x}^\star \in \arg\min(f)$. To do so, we need the matrix operator to be connected in the following sense:

Definition G.2. We say that a reconditioning matrix $\mathbf{\Lambda} : \mathsf{dom}(f) \to \mathbb{S}^n_+$ is *connected* if there exists a parametrized operator $\overline{\mathsf{recond}}_{\mathbf{\Lambda}}(\cdot,\cdot) : \mathsf{dom}(f) \times [0,1] \to \mathbb{S}^n_+$ such that (a) $\overline{\mathsf{recond}}_{\mathbf{\Lambda}}(\cdot,\cdot)(\boldsymbol{x}, 0) = \boldsymbol{x}$ (b) $\overline{\mathsf{recond}}(\boldsymbol{x}, 1) = \mathsf{recond}(\boldsymbol{x})$, and (c) for all $\boldsymbol{x} \in \mathsf{dom}(f)$, $t \mapsto \overline{\mathsf{recond}}(\boldsymbol{x}, t)$ is connected, and its image lies in $\mathsf{dom}(f)$.

Observation G.3. The reconditioning matrix $\mathbf{\Lambda}(\mathsf{K}) = \mathbf{\Sigma}_{\mathsf{K},22}$ for the loss $\mathcal{L}_{\lambda(\cdot)}$ is connected.

Proof. Define $\overline{\mathsf{recond}}(\mathsf{K}, t) := \mathsf{Sim}_{\mathbf{S}_t}(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K})$, where $\mathbf{S}_t = \mathbf{\Sigma}_{22,\mathsf{K}}^{-t/2}$. Since similarity transforms preserve membership in $\mathcal{K}_{\mathtt{info}}$, and since $t \mapsto \overline{\mathsf{recond}}(\mathsf{K}, t)$ is continuous and coincides with K at $t = 0$ (resp. $\mathsf{recond}(\mathsf{K})$ at $t = 1$), the observation follows. ☐

The following proposition, proved in Appendix H.3.2, establishes path-connectedness for connected reconditioning matrices.

Proposition G.4. *Consider the set up of Proposition G.2 with* $\boldsymbol{x}_0 \in \mathsf{dom}(f)$*, and in addition, suppose (a) that* $\mathbf{\Lambda}(\cdot)$ *is continuous reconditioning matrix and (b) the set* $\mathcal{K}(\boldsymbol{x}_0)$ *is compact. Then, there exists an* $\boldsymbol{x}^\star \in \arg\min(f)$ *and a path* $\gamma : [0,1] \to \mathsf{dom}(f)$ *such that* $\gamma(0) = \boldsymbol{x}_0$ *and* $\gamma(1) = \boldsymbol{x}^\star$*.*

Proposition 4.1 can be recovered as the special case when the reconditioning matrix $\mathsf{recond}_{\mathbf{\Lambda}}(\boldsymbol{x}) \equiv \mathbf{I}_d$ is always the identity. In this case, the reconditioning step is vacuous. Moreover $L_{\mathrm{cond},\boldsymbol{x}_0} = 0$ ($\mathsf{recond}_{\mathbf{\Lambda}}$ is constant), and it is straightforward to modify the proof of Proposition G.2 to dispense with the dependence on $L_{f,\boldsymbol{x}_0}$.

G.2 Proof of Theorem 2

With key ingredients of the analysis in mind, we now finish the proof of Theorem 2 by illustrating the existence of a DCL for the regularized OE problem, and establishing smoothness and Lipschitzness of the objective when restricted to the reconditioned set so as to apply Proposition G.2. More specifically, we first establish the relevant properties "locally", in that they depend on the choice of the filter K, and then prove a uniform bound over all K in the reconditioned set at the very end. A recurring theme is that both the weak-PL and the smoothness properties are controlled by the informativity, as measured by $\|\mathbf{Z}_{\mathsf{K}}^{-1}\|$. These are terms are also controlled by $\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|$, which we show below are bounded in terms of $\|\mathbf{\Sigma}_{22,\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|$, which are both bounded due to the reconditioning step.

As shorthand, we let $\mathrm{poly}_{\mathrm{op}}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \kappa)$ denote a term which is at most a polynomial function of the operator norm of the matrix arguments $\|\mathbf{X}_1\|, \|\mathbf{X}_2\|, \ldots$, and a polynomial in the scalar argument κ; $\|\cdot\|_{\ell_2}$ denotes the Euclidean norm (e.g. on parameters $\mathsf{K} = (\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$). All results below assume $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$, and that $\mathbf{\Sigma}_{\mathsf{K}}$ is invertible (we verify this condition in Lemma G.7 below.)

A DCL for the regularized OE objective. While it is by now well-known within the controls community that the OE problem admits a convex reformulation [Scherer et al., 1997], we prove a stronger result showing that this reformulation is in fact a DCL. We prove the following result in Appendix I.

Proposition 4.2. *For any $\lambda \geq 0$ (non-strict), the objective $\mathcal{L}_\lambda(\mathsf{K})$ admits a DCL $(f_{\mathtt{cvx}}, f_{\mathtt{lft}}, \Phi)$ where the lifted parameter takes the form $(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}}) \in \mathcal{K}_{\mathtt{info}} \times \mathbb{S}^{2n}_{++}$, $\mathcal{L}_\lambda(\mathsf{K}) = f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}}) = \min_{\mathbf{\Sigma} \in \mathbb{S}^{2n}_{+}} f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma})$, and where*

$$\begin{aligned}\sigma_{d_z}(\nabla\Phi(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}})) &\geq 1/\mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{\Sigma}_{\mathsf{K}}, \mathbf{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}, \mathcal{L}_{\mathtt{OE}}(\mathsf{K})\right)\\ \|\Phi(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}})\|_{\ell_2} &\leq \left(\max\{n, \sqrt{mn}\} + \sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})}\right) \cdot \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{\Sigma}_{\mathsf{K}}, \mathbf{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}\right).\end{aligned}$$

Furthermore, the norms of the parameters $\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}$ satisfy the following bounds:

$$\max\{\|\mathbf{A}_{\mathsf{K}}\|_{\mathrm{op}}, \|\mathbf{B}_{\mathsf{K}}\|_{\mathrm{op}}\} \leq \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}, \mathbf{\Sigma}_{\mathsf{K}}, \mathbf{\Sigma}_{\mathsf{K}}^{-1}\right), \quad \|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}} \leq \sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})/\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|}. \tag{4.1}$$

Recall that the domain of $\mathcal{L}_\lambda(\mathsf{K})$ is the set $\mathcal{K}_{\mathtt{info}}$, on which $\mathbf{Z}_{\mathsf{K}}$ and (as noted above) $\mathbf{\Sigma}_{\mathsf{K}}$ are invertible. Hence, all quantities in the above lemma are well-defined. Having established the existence of a DCL, a direct application of Theorem 3 shows that this objective satisfies the weak-PL property.

Corollary G.1 (Weak-PL Property of $\mathcal{L}_\lambda$)**.** *For any $\lambda \geq 0$ and $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$,*

$$\begin{aligned}\|\nabla\mathcal{L}_\lambda(\mathsf{K})\| &\geq \frac{1}{C_{\mathtt{PL}}(\mathsf{K}) \cdot \max\{n, \sqrt{mn}\}} \cdot (\mathcal{L}_\lambda(\mathsf{K}) - \inf(\mathcal{L}_\lambda)), \quad \text{where}\\ C_{\mathtt{PL}}(\mathsf{K}) &= \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}, \mathbf{\Sigma}_{\mathsf{K}}, \mathbf{\Sigma}_{\mathsf{K}}^{-1}, \mathcal{L}_{\mathtt{OE}}(\mathsf{K})\right).\end{aligned} \tag{G.5}$$

Smoothness and Lipschitzness of $\mathcal{L}_\lambda(\mathsf{K})$. To verify these regularity conditions, we need to bound the norms of various quantities, which are themselves the solutions to Lyapunov equations involving the closed-loop system matrix $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ (defined in Eq. (2.1)). The main step is therefore to show that the solutions to these Lyapunov equations are uniformly bounded, as per the following lemma (proof in Appendix J).

Proposition 4.3 (Stability of $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$)**.** *Suppose that $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$. Then, for any matrix $\mathbf{Y} \in \mathbb{S}^{2n}$, the solution $\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}$ to the Lyapunov equation $\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}} + \mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top + \mathbf{Y} = 0$ satisfies*

$$\|\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}\|_\circ \leq C_{\mathtt{lyap}}(\mathsf{K}) \cdot \|\mathbf{Y}\|_\circ, \quad \text{where } C_{\mathtt{lyap}}(\mathsf{K}) = \mathrm{poly}_{\mathrm{op}}\left(\mathbf{\Sigma}_{\mathsf{K}}, \mathbf{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}, \mathbf{W}_1^{-1}, \mathbf{W}_2^{-1}, \mathbf{C}\right),$$

and where $\|\cdot\|_\circ$ denotes either the operator, Frobenius, or nuclear norm.

Using this intermediate result, we can bound the norms of the various derivatives which govern the smoothness and Lipschitz constants for the regularized $\mathtt{OE}$ problem. We present the proof of the following result in Appendix K, as well as formal explanations of the notation of the norms below.

Proposition G.5 (Smoothness and Lipschitzness). *For any* $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$*,* $\mathcal{L}_\lambda(\cdot)$ *is* $\mathscr{C}^2$ *in an open neighorhood containing* $\mathcal{K}$*, and*

$$\|\nabla^2 \mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2 \to \ell_2} \le C_{\mathtt{grad},2}(\mathsf{K}) \cdot C_{\mathtt{lyap}}(\mathsf{K})^2 \cdot (1+\lambda) \qquad \text{(Local smoothness)}$$

$$\|\nabla \mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2} \le C_{\mathtt{grad},1}(\mathsf{K}) \cdot C_{\mathtt{lyap}}(\mathsf{K}) \cdot (1+\lambda)\sqrt{n} \qquad \text{(Lipschitz loss)}$$

$$\|\nabla \mathbf{\Sigma}_{22,\mathsf{K}}\|_{\ell_2 \to \mathrm{op}} \le C_{\Sigma,1}(\mathsf{K}) \cdot C_{\mathtt{lyap}}(\mathsf{K}), \qquad \text{(Lipschitz reconditioning)}$$

where $C_{\Sigma,1}(\mathsf{K}) = \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}, \mathbf{W}_2)$*, where*

$$C_{\mathtt{grad},1}(\mathsf{K}), C_{\mathtt{grad},2}(\mathsf{K}) = \mathrm{poly}_{\mathrm{op}}(\mathbf{Z}_\mathsf{K}^{-1}, \mathbf{\Sigma}_{22,\mathsf{K}}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2),$$

where $C_{\mathtt{lyap}}(\mathsf{K})$ *is as in Proposition 4.3, and where the gradient norms are in the Euclidean geometry.*

Concluding the proof: uniform parameter bounds. Note again that bounds above are local, in that they depend on the choice of filter K. To finish the proof of Theorem 2, we prove a uniform bound over all filters K which lie in the set considered by Proposition G.2, namely.

$$\mathcal{K}_0 := \left\{\mathsf{K} \in \mathcal{K}_{\mathtt{info}} : \mathcal{L}_\lambda(\mathsf{K}) \le \mathcal{L}_\lambda(\mathsf{K}_0) \text{ and } \frac{1}{2}\mathbf{I}_n \preceq \mathbf{\Sigma}_{22,\mathsf{K}} \preceq 2\mathbf{I}_n\right\}. \tag{G.6}$$

Immediately, we see that on this set $\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\| \le 2$, and that

$$\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) \le \mathcal{L}_\lambda(\mathsf{K}) \le \mathcal{L}_\lambda(\mathsf{K}_0), \quad \|\mathbf{Z}_\mathsf{K}^{-1}\| \le \mathrm{tr}[\mathbf{Z}_\mathsf{K}^{-1}] = \mathcal{R}_{\mathtt{info}}(\mathsf{K}) \le \frac{1}{\lambda}\mathcal{L}_\lambda(\mathsf{K}) \le \frac{1}{\lambda}\mathcal{L}_\lambda(\mathsf{K}_0).$$

As a consequence, we can bound the terms appear in the bounds above as follows (see Appendix G.5):

Lemma G.6. *The terms* $C_{\mathtt{PL}}(\mathsf{K}), C_{\mathtt{lyap}}(\mathsf{K}), C_{\Sigma,1}(\mathsf{K}), C_{\mathtt{grad},1}(\mathsf{K}), C_{\mathtt{grad},2}(\mathsf{K})$ *appearing above are all bounded by at most* $\mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_\mathsf{K}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2, \mathbf{W}_2^{-1}, \mathbf{W}_1^{-1}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda})$.

Lastly, we control the dependence on $\mathbf{\Sigma}_\mathsf{K}$ and $\mathbf{\Sigma}_\mathsf{K}^{-1}$. The follow lemma is proven in Appendix G.6.

Lemma G.7. *Let* $\sigma_\star > 0$ *be as in we mean Lemma 3.3. Then, for any* $\mathsf{K} \in \mathcal{K}_{\mathtt{ctrb}}$*, it holds that: (a)* $\mathbf{\Sigma}_\mathsf{K} \succ 0$ *is invertible, (b)* $\|\mathbf{\Sigma}_\mathsf{K}^{-1}\| \le 2\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\| + 2\sigma_\star^{-1}\max\{1, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|\}$*, and (c)* $\|\mathbf{\Sigma}_\mathsf{K}\| \le 2\max\{\|\mathbf{\Sigma}_{22,\mathsf{K}}\|, \|\mathbf{\Sigma}_{11,\mathrm{sys}}\|\}$.

In particular, on $\mathcal{K}_0$, where $\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}\| \le 2$, we have $\|\mathbf{\Sigma}_\mathsf{K}\|, \|\mathbf{\Sigma}_\mathsf{K}^{-1}\| \le \mathrm{poly}_{\mathrm{op}}(\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|, \sigma_\star^{-1})$, so that the terms $C_{\mathtt{PL}}(\mathsf{K}), C_{\mathtt{lyap}}(\mathsf{K}), C_{\Sigma,1}(\mathsf{K}), C_{\mathtt{grad},1}(\mathsf{K}), C_{\mathtt{grad},2}(\mathsf{K})$ are all at most polynomial in

$$C_{\mathtt{sys}} := \max\{\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|, \|\mathbf{A}\|, \|\mathbf{C}\|, \|\mathbf{G}\|, \|\mathbf{W}_2\|, \|\mathbf{W}_2^{-1}\|, \|\mathbf{W}_1^{-1}\|, \sigma_\star^{-1}\}, \tag{G.7}$$

as well as in $\mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}$. Thus, from Corollary G.1 and Proposition G.5, we verify the conditions of Proposition G.2 uniformly on the set $\mathcal{K}_0$.

Corollary G.2. *The loss function* $\mathcal{L}_\lambda$ *satisfies* α*-weak PL and* β*-upper smoothness on* $\mathcal{K}_0$ *with*

$$\alpha^{-1} \le \max\{n, \sqrt{mn}\} \cdot \mathrm{poly}(C_{\mathtt{sys}}, \mathcal{L}_\lambda(\mathsf{K}_0), \tfrac{1}{\lambda}), \quad \beta \le \mathrm{poly}(C_{\mathtt{sys}}, \mathcal{L}_\lambda(\mathsf{K}_0), \tfrac{1}{\lambda}, \lambda),$$

where $C_{\mathtt{sys}}$ *is defined in Eq. (G.7). In addition, on* $\mathcal{K}_0$*,* $\mathcal{L}_\lambda$ *is* $L \le \sqrt{n}\mathrm{poly}(C_{\mathtt{sys}}, \mathcal{L}_\lambda(\mathsf{K}_0), \lambda, \frac{1}{\lambda})$ *Lipschitz , and* $\mathsf{K} \mapsto \mathbf{\Sigma}_{22,\mathsf{K}}$ *is at most* $L_\Sigma \le \mathrm{poly}(C_{\mathtt{sys}}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}, \lambda)$ *Lipschitz as a mapping from* $(\mathcal{K}_{\mathtt{info}}, \|\cdot\|_{\ell_2}) \to (\mathbb{S}^n, \|\cdot\|_{\mathrm{op}})$*.*

Lastly, we establish compact level sets. The subtlety here is not only showing that $\mathcal{K}_0$ is bounded (this is rather direct from Proposition 4.2), but also closed.

Lemma G.8. *Let set* $\mathcal{K}_0$ *in Eq. (G.6) is compact.*

The upper bound on $\mathcal{L}_\lambda(\mathsf{K}_s) - \min_\mathsf{K} \mathcal{L}_\lambda(\mathsf{K})$ in Theorem 2 is now a direct consequence of instantiating Proposition G.2 $\eta = \eta_s$ with the bounds in the above Corollary G.2, and noting that $\mathcal{K}_0$ is closed by Lemma G.8.

The inequality $\mathcal{L}_{\mathtt{OE}}(\mathsf{K}_s) - \min_\mathsf{K} \mathcal{L}_{\mathtt{OE}}(\mathsf{K}) \le \mathcal{L}_\lambda(\mathsf{K}_s) - \min_\mathsf{K} \mathcal{L}_\lambda(\mathsf{K})$ is just a consequence of Corollary 3.1. □

G.3 Proof of Theorem 1

Due to the $\mathtt{DCL}$ exhbited by Proposition 4.2, and in particular Corollary G.1, we find that any $\lambda \geq 0$ and $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ for which $\nabla\mathcal{L}_\lambda(\mathsf{K}) = 0$ must be optimal (in applying the corollary, we again note that $\mathbf{Z}_\mathsf{K}$ is guaranteed to be invertible of $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$, and $\mathbf{\Sigma}_\mathsf{K}$ invertible by Lemma G.7). By taking $\lambda = 0$, we have $\nabla\mathcal{L}_\lambda(\mathsf{K}) = \nabla\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$, proving the theorem. Path connectedness follows from Proposition G.4, again noting that $\mathcal{K}_0$ is compact (Lemma G.8). □

G.4 Proof of Theorem 2a

The proof is nearly identical to that of Theorem 2. The only difference is that the step sizes are selected according to backtracking line search. We apply Proposition G.2 where η_s (in the statement of the proposition) is set to any $\eta \in \mathcal{S}_{\mathrm{bkt}}$ for all s satisfying the same upper bound $\eta \leq \frac{1}{\mathcal{C}_1}$ required in Theorem 2. Since since backtracking line search selects the step which attains the greatest direction of descent, at each iteration, we have

$$\mathcal{L}_\lambda(\mathsf{K}_{t+1}) \leq \mathcal{L}_\lambda(\widetilde{\mathsf{K}}_t - \eta\nabla\mathcal{L}_\lambda(\widetilde{\mathsf{K}}_t)).$$

Hence, backtracking satisfies the descent condition Eq. (G.3), and the theorem follows.

G.5 Proof of Lemma G.6

Recall that, for $\mathcal{K} \in \mathcal{K}_0$,

$$\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) \leq \mathcal{L}_\lambda(\mathsf{K}_0), \quad \|\mathbf{Z}_\mathsf{K}^{-1}\| \leq \frac{1}{\lambda}\mathcal{L}_\lambda(\mathsf{K}_0).$$

Hence, for $\mathcal{K} \in \mathcal{K}_0$ and $C_{\mathtt{PL}}(\mathsf{K})$ as in Corollary G.1

$$\begin{aligned} C_{\mathtt{PL}}(\mathsf{K}) &= \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{Z}_\mathsf{K}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathcal{L}_{\mathtt{OE}}(\mathsf{K})\right) \\ &\leq \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}\right). \end{aligned}$$

In addition, from Proposition 4.3,

$$C_{\mathtt{lyap}}(\mathsf{K}) = \mathrm{poly}_{\mathrm{op}}\left(\mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathbf{Z}_\mathsf{K}^{-1}, \mathbf{C}, \mathbf{W}_1^{-1}, \mathbf{W}_2^{-1}\right) \leq \mathrm{poly}_{\mathrm{op}}\left(\mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathbf{C}, \mathbf{W}_1^{-1}, \mathbf{W}_2^{-1}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}\right).$$

Moreover, from Proposition 4.2

$$\begin{aligned} \max\{\|\mathbf{A}_\mathsf{K}\|_{\mathrm{op}}, \|\mathbf{B}_\mathsf{K}\|_{\mathrm{op}}, \|\mathbf{C}_\mathsf{K}\|_{\mathrm{F}}\} &\leq \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{Z}_\mathsf{K}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathcal{L}_{\mathtt{OE}}(\mathsf{K})\right) \\ &\leq \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}^{-1}, \mathcal{L}_\lambda(\mathsf{K})\right). \end{aligned}$$

Finally, from Proposition G.5, we have for $\mathsf{K} \in \mathcal{K}_0$,

$$\begin{aligned} C_{\Sigma,1}(\mathsf{K}), C_{\mathtt{grad},1}(\mathsf{K}), C_{\mathtt{grad},2}(\mathsf{K}) &= \mathrm{poly}_{\mathrm{op}}(\mathbf{Z}_\mathsf{K}^{-1}, \mathbf{\Sigma}_{22,\mathsf{K}}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2) \\ &\leq \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}) \\ &\leq \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_\mathsf{K}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2, \mathbf{W}_2^{-1}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}) \end{aligned}$$

Hence, in summary,

$$\begin{aligned} &C_{\mathtt{PL}}(\mathsf{K}), C_{\mathtt{lyap}}(\mathsf{K}), C_{\Sigma,1}(\mathsf{K}), C_{\mathtt{grad},1}(\mathsf{K}), C_{\mathtt{grad},2}(\mathsf{K}) \\ &\quad = \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_\mathsf{K}^{-1}, \mathbf{\Sigma}_\mathsf{K}, \mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{W}_2, \mathbf{W}_2^{-1}, \mathbf{W}_1^{-1}, \mathcal{L}_\lambda(\mathsf{K}_0), \frac{1}{\lambda}). \end{aligned}$$

□

G.6 Conditioning of the stationary covariance (Lemma G.7)

Part (a). Recall the block decomposition

$$\mathbf{\Sigma}_{\mathsf{K}} = \begin{bmatrix} \mathbf{\Sigma}_{11,\mathrm{sys}} & \mathbf{\Sigma}_{12,\mathsf{K}} \\ \mathbf{\Sigma}_{12,\mathsf{K}}^\top & \mathbf{\Sigma}_{22,\mathsf{K}} \end{bmatrix},$$

where we note that $\mathbf{\Sigma}_{11,\mathrm{sys}}$ does not depend on K. From the Schur complement test, $\mathbf{\Sigma}_{\mathsf{K}} \succ 0$ if and only if both $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$ and $\mathbf{\Sigma}_{11,\mathrm{sys}} \succ \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top = \mathbf{Z}_{\mathsf{K}}$. The first of these holds for $\mathsf{K} \in \mathcal{K}_{\mathtt{ctrb}}$, and since $\mathbf{Z}_{\mathsf{K}} \preceq \mathbf{Z}_\star$ (for $\mathbf{Z}_\star$ as in Lemma 3.2), the second holds from Lemma 3.3.

Part (b). We invoke Lemma G.9 below to bound

$$\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\| \le 2\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\| + 2\|\mathbf{X}_{\mathsf{K}}^{-1}\| \max\{1, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|\},$$

where $\mathbf{X}_{\mathsf{K}} = \mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top = \mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_{\mathsf{K}}$ is the Schur complement term. Moreover, since $\mathbf{Z}_{\mathsf{K}} \preceq \mathbf{Z}_\star$, $\mathbf{X}_{\mathsf{K}}^{-1} \preceq (\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)^{-1}$, so $\|\mathbf{X}_{\mathsf{K}}^{-1}\| \le \|(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)^{-1}\| = 1/\lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)$. Hence,

$$\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\| \le 2\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\| + 2[\lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)]^{-1} \max\{1, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|\},$$

as needed. By Lemma 3.3, we have $\sigma_\star = \lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}} - \mathbf{Z}_\star)$

Part (c). Invoking Lemma G.9 part (a), we directly obtain $\|\mathbf{\Sigma}_{\mathsf{K}}\| \le 2\max\{\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}\|\}$. By

Now the remaining part is to prove the following Lemma.

Lemma G.9. *Suppose that $\mathbf{\Lambda} \succeq 0$ is positive semidefinite and has block-diagonal decomposition with blocks diagonal blocks $\mathbf{\Lambda}_{11}, \mathbf{\Lambda}_{22}$. Then,*

(a) $\|\mathbf{\Lambda}\| \le 2\max\{\|\mathbf{\Lambda}_{11}\|, \|\mathbf{\Lambda}_{22}\|\}$.

(b) If in addition $\mathbf{\Lambda} \succ 0$, then defining the Schur complement $\mathbf{X} := \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{12}^\top$, we have

$$\|\mathbf{\Lambda}^{-1}\| \le 2\|\mathbf{\Lambda}_{22}^{-1}\| + 2\|\mathbf{X}^{-1}\| \max\{1, \|\mathbf{\Lambda}_{22}^{-1}\|\|\mathbf{\Lambda}_{11}\|\}.$$

Proof. We prove each part in sequence:

Part (a). It suffices to prove that

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{12}^\top & \mathbf{\Lambda}_{22} \end{bmatrix} \preceq 2\bar{\mathbf{\Lambda}}, \quad \text{where } \bar{\mathbf{\Lambda}} := \begin{bmatrix} \mathbf{\Lambda}_{11} & 0 \\ 0 & \mathbf{\Lambda}_{22} \end{bmatrix}$$

To show the above, consider any vector $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$. First, for the modified vector $\widetilde{\mathbf{v}} = (\mathbf{v}_1, -\mathbf{v}_2)$, we compute

$$0 \le \widetilde{\mathbf{v}}^\top \mathbf{\Lambda}\widetilde{\mathbf{v}} = \mathbf{v}_1^\top \mathbf{\Lambda}_{11}\mathbf{v}_1 + \mathbf{v}_2^\top \mathbf{\Lambda}_{22}\mathbf{v}_2 - 2\mathbf{v}_1^\top \mathbf{\Lambda}_{12}\mathbf{v}_2.$$

Hence,

$$\begin{aligned} \mathbf{v}^\top \mathbf{\Lambda}\mathbf{v} &= \mathbf{v}_1^\top \mathbf{\Lambda}_{11}\mathbf{v}_1 + \mathbf{v}_2^\top \mathbf{\Lambda}_{22}\mathbf{v}_2 + 2\mathbf{v}_1^\top \mathbf{\Lambda}_{12}\mathbf{v}_2 \\ &\le 2\mathbf{v}_1^\top \mathbf{\Lambda}_{11}\mathbf{v}_1 + 2\mathbf{v}_2^\top \mathbf{\Lambda}_{22}\mathbf{v}_2 = 2\mathbf{v}^\top \bar{\mathbf{\Lambda}}\mathbf{v}. \end{aligned}$$

Part (b). Introduce the $\mathbf{X} := \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{12}^\top$ as the Schur-complement term. From the block-matrix inversion formula,

$$\mathbf{\Lambda}^{-1} = \begin{bmatrix} \mathbf{X}^{-1} & * \\ * & \mathbf{\Lambda}_{22}^{-1}\left(\mathbf{I} + \mathbf{\Lambda}_{22}^{-1/2}\mathbf{\Lambda}_{12}^\top \mathbf{X}^{-1}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1/2}\right) \end{bmatrix}.$$

From part (a), we then bound

$$\begin{aligned}\|\mathbf{\Lambda}^{-1}\| &\leq 2\max\left\{\|\mathbf{X}^{-1}\|, \|\mathbf{\Lambda}_{22}^{-1}\left(\mathbf{I} + \mathbf{\Lambda}_{22}^{-1/2}\mathbf{\Lambda}_{12}^{\top}\mathbf{X}^{-1}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1/2}\right)\|\right\} \\ &\leq 2\max\left\{\|\mathbf{X}^{-1}\|, \|\mathbf{\Lambda}_{22}^{-1}\|\left(1 + \|\mathbf{X}^{-1}\| \cdot \|\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1/2}\|^2\right)\|\right\}.\end{aligned}$$

The term $\|\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1/2}\|^2 = \|\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{12}^{\top}\| \leq \|\mathbf{\Lambda}_{11}\|$, where we used that $\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{12}^{\top} \preceq \mathbf{\Lambda}_{11}$ by the Schur complement test. Hence, we conclude

$$\begin{aligned}\|\mathbf{\Lambda}^{-1}\| &\leq 2\max\left\{\|\mathbf{X}^{-1}\|, \|\mathbf{\Lambda}_{22}^{-1}\|\left(1 + \|\mathbf{X}^{-1}\| \cdot \|\mathbf{\Lambda}_{11}\|\right)\|\right\} \\ &\leq 2\|\mathbf{\Lambda}_{22}^{-1}\| + 2\|\mathbf{X}^{-1}\|\max\{1, \|\mathbf{\Lambda}_{22}^{-1}\|\|\mathbf{\Lambda}_{11}\|\},\end{aligned}$$

which completes the proof of Lemma G.9. ☐

This finally completes the proof of Lemma G.7. ☐

G.7 Proof of Lemma G.8

To see that $\mathcal{K}_0$ is bounded, we use that $\|\mathbf{A}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}_{\mathsf{K}}\|$ are uniformly bounded on $\mathcal{K}_0$. This is a consequence of the bounds on these parameters in Proposition 4.2, as well as the fact that the various terms in those bounds are in terms of $\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|, \|\mathbf{Z}_{\mathsf{K}}^{-1}\|$ and $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$, all of which are shown to be uniformly bounded on $\mathcal{K}_0$.

To show $\mathcal{K}_0$ is closed, it suffices to show that for any convergent sequence of controllers $\mathsf{K}^{(i)}$ in $\mathcal{K}_0$, its limit is in $\mathcal{K}_0$. In light of the boundness discussion above, this follows directly from the following lemma.

Lemma G.10. *Let $\mathsf{K}^{(i)} \in \mathcal{K}_{\mathtt{info}}$ be a sequence of controllers converge to some K, such that $\|\mathbf{\Sigma}_{\mathsf{K}^{(i)}}\|, \|\mathbf{\Sigma}_{\mathsf{K}^{(i)}}^{-1}\|, \|\mathbf{Z}_{\mathsf{K}^{(i)}}^{-1}\|$, as well as $\|\mathbf{A}_{\mathsf{K}^{(i)}}\|, \|\mathbf{B}_{\mathsf{K}^{(i)}}\|, \|\mathbf{C}_{\mathsf{K}^{(i)}}\|$ remain uniformly bounded. Then, $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$.*

Proof. We prove stability, $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$, and $\mathbf{Z}_{\mathsf{K}} \succ 0$ in succession.

Stability. Let $\mathbf{\Gamma}^{(i)} := \|\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}}, \mathbf{I}_{2n})\|_{\circ}$. Then, $\sup_i \|\mathbf{\Gamma}^{(i)}\| \leq M$ for some $M > 0$. Moreover, for any $\varepsilon > 0$ and $i \geq i_0$ sufficiently large, we have $\|\mathbf{A}_{\mathrm{cl},\mathsf{K}} - \mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}}\| \leq \varepsilon$. Thus, for such $i \geq i_0$,

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Gamma}^{(i)} + \mathbf{\Gamma}^{(i)}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top} \preceq \mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}}\mathbf{\Gamma}^{(i)} + \mathbf{\Gamma}^{(i)}\mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}}^{\top} + 2M\varepsilon\mathbf{I}_{2n} = -\mathbf{I}_{2n}(1 - 2M\varepsilon). \tag{G.8}$$

Hence, for $\varepsilon = 1/4M$, $\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Gamma}^{(i)} + \mathbf{\Gamma}^{(i)}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top} \preceq -\frac{1}{2}\mathbf{I}_{2n}$. Since $\mathbf{\Gamma}^{(i)} \succ 0$, this implies $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is stable.

Controllability. Define the functions $F_i(\mathbf{\Sigma}) := \mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}_{\mathrm{cl},\mathsf{K}^{(i)}} + \mathbf{W}_{\mathrm{cl},\mathsf{K}^{(i)}}$, so that $\mathbf{\Sigma}_{\mathsf{K}^{(i)}}$ is the unique PSD solution to $F_i(\mathbf{\Sigma}_{\mathsf{K}^{(i)}}) = 0$. By Proposition 4.3, $0 \preceq \mathbf{\Sigma}_{\mathsf{K}^{(i)}} \preceq M\mathbf{I}_{2n}$ for some $i \geq 0$. Hence, there is a subsequence i_j such that $\mathbf{\Sigma}_{\mathsf{K}^{(i_j)}}$ converges to a limit $\bar{\mathbf{\Sigma}}$ on the set $\mathcal{X} := \{\mathbf{\Sigma} : 0 \preceq \mathbf{\Sigma} \preceq M\mathbf{I}_{2n}\}$. Since $\|\mathbf{A}_{\mathsf{K}}^{(i)}\|, \|\mathbf{B}_{\mathsf{K}}^{(i)}\|, \|\mathbf{C}_{\mathsf{K}}^{(i)}\|$ remain uniformly bounded, $F_i \to F(\mathbf{\Sigma}) := \mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}_{\mathrm{cl},\mathsf{K}} + \mathbf{W}_{\mathrm{cl},\mathsf{K}}$ uniformly on this set $\mathcal{X}$, and thus, $F(\bar{\mathbf{\Sigma}}) = \lim_{j\to\infty} F_j(\mathbf{\Sigma}_{\mathsf{K}^{(i_j)}}) = \mathbf{0}$. Hence, since $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is stable as established above, $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}_{\mathsf{K}}$. Since this holds for all subsequences, we have $\lim_{i\to\infty}\mathbf{\Sigma}_{\mathsf{K}^{(i)}} = \mathbf{\Sigma}_{\mathsf{K}}$. Hence, $\mathbf{\Sigma}_{\mathsf{K}} \succ 0$, since by assumption $\|\mathbf{\Sigma}_{\mathsf{K}^{(i)}}^{-1}\|$ is uniformly bounded in i. Thus $\mathbf{\Sigma}_{22,\mathsf{K}} \succ 0$, and thus, $\mathsf{K} \in \mathcal{K}_{\mathtt{ctrb}}$.

Informativity. As established above, $\lim_{i\to\infty}\mathbf{\Sigma}_{\mathsf{K}^{(i)}} = \mathbf{\Sigma}_{\mathsf{K}}$. Since the transformation mapping $\mathbf{\Sigma}_{\mathsf{K}^{(i)}} \to \mathbf{Z}_{\mathsf{K}^{(i)}}$ is continuous for $\mathbf{\Sigma}_{\mathsf{K}^{(i)}} \succ 0$, we see that $\lim_{i\to\infty}\mathbf{Z}_{\mathsf{K}^{(i)}} = \mathbf{Z}_{\mathsf{K}}$. Hence, since $\mathbf{Z}_{\mathsf{K}^{(i)}} \succ 0$ and $\mathbf{Z}_{\mathsf{K}^{(i)}}^{-1}$ is uniformly bounded, $\mathbf{Z}_{\mathsf{K}} \succ 0$. Thus, $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$. ☐

H Proofs for `DCL`s and Gradient Descent

H.1 Proof of Fact 1.1

Throughout, we use the notation $\mathsf{dom}_{>}(f) := \{\boldsymbol{x} \in \mathsf{dom}(f) : f(\boldsymbol{x}) > \inf(f)\}$.

Fact 1.1. Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ be a differentiable, possibly nonconvex function such that $\min_{\boldsymbol{x}} \; f(\boldsymbol{x})$ is finite. Suppose there exists a differentiable function $\Psi : \mathbb{R}^{n_\nu} \to \mathbb{R}^{n_x}$ satisfying the following two properties: (i) the mapping Ψ is surjective, i.e. for all $\boldsymbol{x} \in \mathbb{R}^{n_x}$ there exists $\boldsymbol{\nu} \in \mathbb{R}^{n_\nu}$ such that $\boldsymbol{x} = \Psi(\boldsymbol{\nu})$, (ii) under the change of variables the function $f_{\mathrm{cvx}}(\boldsymbol{\nu}) := f(\Psi(\boldsymbol{\nu}))$ is differentiable and *convex*. Then all first-order stationary points, $\boldsymbol{x}$ s.t $\nabla f(\boldsymbol{x}) = 0$, are globally optimal.

Proof. Let us proceed by contradiction. Suppose that $\bar{\boldsymbol{x}}$ is a suboptimal stationary point, i.e. $\nabla f(\bar{\boldsymbol{x}}) = 0$ but $f(\bar{\boldsymbol{x}}) \neq \min_{\boldsymbol{x}} \; f(\boldsymbol{x})$. Let $\bar{\boldsymbol{\nu}}$ be such that $\bar{\boldsymbol{x}} = \Psi(\bar{\boldsymbol{\nu}})$. By surjectivity of Ψ, such a $\bar{\boldsymbol{\nu}}$ always exists. Next, by application of the chain rule to $f_{\mathrm{cvx}}(\boldsymbol{\nu}) := f(\Psi(\boldsymbol{\nu}))$, we have

$$\nabla f_{\mathrm{cvx}}(\boldsymbol{\nu})|_{\boldsymbol{\nu}=\bar{\boldsymbol{\nu}}} = \nabla f(\boldsymbol{x})|_{\boldsymbol{x}=\bar{\boldsymbol{x}}} \cdot \nabla \Psi(\boldsymbol{\nu})|_{\boldsymbol{\nu}=\bar{\boldsymbol{\nu}}}. \tag{H.1}$$

Therefore, by Eq. (H.1), $\nabla f(\bar{\boldsymbol{x}}) = 0$ implies $\nabla f_{\mathrm{cvx}}(\bar{\boldsymbol{\nu}}) = 0$. However,

$$f_{\mathrm{cvx}}(\bar{\boldsymbol{\nu}}) \overset{(a)}{=} f(\Psi(\bar{\boldsymbol{\nu}})) \overset{(b)}{=} f(\bar{\boldsymbol{x}}) \overset{(c)}{\neq} \min_{\boldsymbol{x}} f(\boldsymbol{x}) \overset{(d)}{=} \min_{\boldsymbol{\nu}} f_{\mathrm{cvx}}(\boldsymbol{\nu}), \tag{H.2}$$

where (a) follows by definition of f_{cvx}, (b) follows from $\bar{\boldsymbol{x}} = \Psi(\bar{\boldsymbol{\nu}})$, (c) follows by suboptimality of $\bar{\boldsymbol{x}}$, and (d) follows by the definition of f_{cvx} and surjectivity of Ψ. However, Eq. (H.2) contradicts the fact that f_{cvx} is a convex function, for which all stationary points must be globally optimal. Therefore, no such suboptimal stationary point $\bar{\boldsymbol{x}}$ can exist. $\square$

H.2 Proof of Theorem 3

We prove Theorem 3, which can be thought of as a (considerable) strengthening of Fact 1.1. The theorem pertains `DCL`s, whose definition we recall below.

Definition 4.1. A triplet of functions $(f_{\mathtt{cvx}}, f_{\mathtt{lft}}, \Phi)$ is a `DCL` of a proper function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ if
(1) $f_{\mathtt{cvx}} : \mathbb{R}^{d_z} \to \bar{\mathbb{R}}$ is a proper convex function whose minimum is attained by some $\boldsymbol{z}^\star$.
(2) For some additional number of parameters $d_\xi \geq 0$, $f_{\mathtt{lft}} : \mathbb{R}^{d+d_\xi} \to \bar{\mathbb{R}}$ is related to f via partial minimization: $f(\boldsymbol{x}) = \min_{\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}} f_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi})$.
(3) There is an open set $\mathcal{Y} \supseteq \mathsf{dom}(f_{\mathtt{lft}})$ for which $\Phi : \mathcal{Y} \to \mathsf{dom}(f_{\mathtt{cvx}})$ is $\mathscr{C}^1$ and satisfies $f_{\mathtt{lft}}(\cdot) = f_{\mathtt{cvx}}(\Phi(\cdot))$.

Before beginning the proof, we explain why the following "trivializing" reparametrization is inadequate.

Remark H.1 (Failure of the trivializing reparametrization)**.** Given a `DCL` $(f_{\mathtt{cvx}}, f_{\mathtt{lft}}, \Phi)$, it may seem that one can avoid the dependence on $\sigma_{d_z}(\nabla\Phi)$ with the following *trivializing reparametrization* obtained by (a) augmenting the lifted parameters $(\boldsymbol{x}, \boldsymbol{\xi})$ with the convex parameter $\boldsymbol{z}$ and (b) defining a new candidate `DCL` $(f_{\mathtt{cvx}}, \widetilde{f}_{\mathtt{lft}}, \widetilde{\Phi})$ given by $\widetilde{f}_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) = f_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi})$ and $\widetilde{\Phi}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) = \boldsymbol{z}$. Note then that $\sigma_{d_z}(\widetilde{\Phi}) = 1$ since $\widetilde{\Phi}$ just projects onto the $\boldsymbol{z}$-coordinates, so this would circumvent the dependence on $\sigma_{d_z}(\nabla\Phi)$. In addition, $(f_{\mathtt{cvx}}, \widetilde{f}_{\mathtt{lft}}, \widetilde{\Phi})$ meets the first two `DCL` two criteria: $f_{\mathtt{cvx}}$ is convex and $f(\boldsymbol{x}) = \min_{(\boldsymbol{\xi}, \boldsymbol{z})} \widetilde{f}_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z})$. However, the candidate `DCL` does not meet the third criterion of Definition 4.1 since the value of $\widetilde{f}_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z})$ does not depend on $\boldsymbol{z}$, so $\widetilde{f}_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) \neq f_{\mathtt{cvx}}(\boldsymbol{z}) = f_{\mathtt{cvx}}(\widetilde{\Phi}(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}))$ in general.

We now begin the proof. We first define a notion of descent direction for functions which strictly generalizes the gradient:

Definition H.1 (Cauchy Directions)**.** Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a proper function, and $\boldsymbol{x} \in \mathsf{dom}(f)$.

(a) We say $\boldsymbol{g} \in \mathbb{R}^d$ is an *Cauchy direction* of f at $\boldsymbol{x}$ if there exists constants $\varepsilon_0 > 0$ such, that for all $\varepsilon \in [0, \varepsilon_0]$, $\boldsymbol{x} - \varepsilon \boldsymbol{g} \in \mathsf{dom}(f)$ and $\lim_{\varepsilon \to 0^+} \frac{f(\boldsymbol{x} - \varepsilon \boldsymbol{g}) - f(\boldsymbol{x})}{\varepsilon} \leq -\|\boldsymbol{g}\|^2$

(b) We say $\boldsymbol{g} \in \mathbb{R}^d$ is a *generalized Cauchy direction* of f at $\boldsymbol{x}$ if, for some $\varepsilon_0 > 0$ the exists a $\mathscr{C}^1$ curve $\phi := [0, \varepsilon_0] \to \mathsf{dom}(f)$ such that $\phi(0) = \boldsymbol{x}$, $\phi'(0) = g$, and $\lim_{\varepsilon \to 0^+} \frac{f(\phi(\varepsilon)) - f(\boldsymbol{x})}{\varepsilon} \le -\|\boldsymbol{g}\|^2$.

Observe that if f is $\mathscr{C}^1$ at x, then the standard gradient $\nabla f(\boldsymbol{z})$ is a Cauchy direction at x; indeed, our nomencalture is a tribute to the 1847 article in which Augustin-Louis Cauchy first described gradient descent, justifying its use via the computation $f(\boldsymbol{x} - \eta \nabla f) = f(\boldsymbol{x}) - \eta \|\nabla f\|^2 + o(\eta)$ (for more in depth history, see e.g. Lemaréchal [2012]). The purpose of generalized Cauchy direction is to accommodate functions whose domains may not contain the segment $\{\boldsymbol{x} - \varepsilon \boldsymbol{g}\}$, but may contain a curve ϕ with the same slope.

Cauchy directions for convex functions. At all high level, we show weak-PL by first showing that $f_{\texttt{cvx}}$ has a Cauchy direction at $\boldsymbol{z}$ of magnitude $\approx f_{\texttt{cvx}}(\boldsymbol{z}) - \inf(f_{\texttt{cvx}})$, and then subsequently showing similar Cauchy directions for $f_{\texttt{lft}}$ and f. For convex functions $f_{\texttt{cvx}}$, we can usally construct Cauchy directions using the subgradient at $\boldsymbol{z}$, a vector $\boldsymbol{g}$ such that $f_{\texttt{cvx}}(\boldsymbol{z}) - f_{\texttt{cvx}}(\boldsymbol{z}') \le \boldsymbol{g}^\top (\boldsymbol{z} - \boldsymbol{z}')$ for all other $\boldsymbol{z}' \in \mathsf{dom}(f_{\texttt{cvx}})$. However, in certain pathological cases, the subgradient may not exist.

Hence, we take a more conservative approach by showing considering not the whole domain of $f_{\texttt{cvx}}$, but rather the line segment joining $\boldsymbol{z}$ to any minimizer $\boldsymbol{z}^\star$, defining the function

$$\psi(t) = f_{\texttt{cvx}}(\boldsymbol{z} + t(\boldsymbol{z}^\star - \boldsymbol{z})) \tag{H.3}$$

This approach allows for pathological cases where the subgradient is "infinite" (in the sense of $h = \infty$, in the sense of the proof below.)

Lemma H.1. *Suppose that $f_{\texttt{cvx}}$ is a proper convex function, with $\boldsymbol{z}^\star \in \arg\min_{\boldsymbol{z}} f_{\texttt{cvx}}(\boldsymbol{z})$ attained. Then, for any $\boldsymbol{z} \in \mathsf{dom}_{>}(f_{\texttt{cvx}})$, $f_{\texttt{cvx}}$ admits a Cauchy direction g satisfying*

$$\|\boldsymbol{g}\| \ge \frac{f_{\texttt{cvx}}(\boldsymbol{z}) - \inf(f_{\texttt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|} \tag{H.4}$$

Proof. Recall $\psi(t)$ from Eq. (H.3), and define the secant-approximation function $\phi(t) := \frac{\psi(t) - \psi(0)}{t}$ for $t \in (0, 1]$. From convexity, one can check that $\phi(t)$ is non-increasing on $t \in (0, 1]$. Hence, the limit

$$h = \lim_{t \to 0^+} \phi(t) = \lim_{t \to 0^+} \frac{\psi(t) - \psi(0)}{t}$$

exists, and has $h \in \{-\infty\} \cup (-\infty, \phi(1)]$, where again, since $\phi(t)$ is non-increasing, we note that

$$h \le \phi(1) = -(f_{\texttt{cvx}}(\boldsymbol{z}) - \inf(f_{\texttt{cvx}})) \tag{H.5}$$

Let us first assume $h \neq -\infty$. We now claim that $\boldsymbol{g} = -|h|(\boldsymbol{z}^\star - \boldsymbol{z})/\|\boldsymbol{z} - \boldsymbol{z}^\star\|^2$ is a Cauchy direction of f at x; this will conclude the proof since by Eq. (H.5)

$$\|\boldsymbol{g}\| = \frac{|h|}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|} \ge \frac{|f(\boldsymbol{z}) - f(\boldsymbol{z}^\star)|}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|}.$$

Let us show that $\boldsymbol{g}$ is a Cauchy direction. First, since f is convex, $\mathsf{dom}(f)$ is convex. Thus, since $\boldsymbol{z}, \boldsymbol{z}^\star \in \mathsf{dom}(f)$, the line seqment joining $\boldsymbol{z}, \boldsymbol{z}^\star$ is contained in $\mathsf{dom}(f)$, and hence for ε sufficiently small, $\boldsymbol{z} - \varepsilon \boldsymbol{g}$ lies on this line segment, and is therefore also contained in $\mathsf{dom}(f)$.

Next, we compute

$$\begin{aligned} h = \lim_{t \to 0^+} \frac{f(\boldsymbol{z} + t(\boldsymbol{z}^\star - \boldsymbol{z})) - f(\boldsymbol{z})}{t} &= \lim_{t \to 0^+} \frac{f(\boldsymbol{z} - t\|\boldsymbol{z} - \boldsymbol{z}^\star\|^2 \frac{\boldsymbol{g}}{|h|}) - f(\boldsymbol{z})}{t} \\ &= \frac{\|\boldsymbol{z} - \boldsymbol{z}^\star\|^2}{|h|} \cdot \lim_{t \to 0^+} \frac{f(\boldsymbol{z} - t\boldsymbol{g}) - f(\boldsymbol{z})}{t}. \end{aligned}$$

Hence,

$$\lim_{t \to 0^+} \frac{f(\boldsymbol{z} - t\boldsymbol{g}) - f(\boldsymbol{z})}{t} = \frac{h|h|}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|^2} = \frac{-h^2}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|^2} = -\|\boldsymbol{g}\|^2,$$

as needed. Now, consider the case where $h = -\infty$. Then, for any $\eta > 0$, we see that $\lim_{t \to 0^+} \frac{f(\boldsymbol{z} + t \cdot \eta(\boldsymbol{z}^\star - \boldsymbol{z})) - f(\boldsymbol{z})}{t} = -\infty$. Hence, $\boldsymbol{g} = \eta \cdot (\boldsymbol{z}^\star - \boldsymbol{z})$ is Cauchy direction for any $\eta > 0$. In particular, taking $\eta = \frac{f(\boldsymbol{z}) - f(\boldsymbol{z}^\star)}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|}$ satisfies the conclusion of the lemma. □

Smooth transformations preserve Cauchy directions. We show that the existence of a Cauchy direction is preserved under smooth transformations.

Lemma H.2. *Let $\bar{f}$ be a proper function, $\boldsymbol{z} \in \mathsf{dom}(f)$, and $\boldsymbol{g}$ a Cauchy direction of $\bar{f}$ at $\boldsymbol{z}$. Let Ψ be a $\mathscr{C}^1$ mapping from a neighborhood $\mathcal{Z}$ containing $\boldsymbol{z}$ into a domain $\mathcal{Y}$ such that $\sigma_{d_z}(\nabla\Psi(\boldsymbol{z})) > 0$, and let $f_{\mathtt{lft}} : \mathcal{Y} \to \bar{\mathbb{R}}$ satisfy $\bar{f}(\boldsymbol{z}') = f_{\mathtt{lft}}(\Psi(\boldsymbol{z}'))$ for all $\boldsymbol{z}' \in \mathcal{Z}$. Then, $f_{\mathtt{lft}}$ has a generalized Cauchy direction $\widetilde{\boldsymbol{g}}$ at $\boldsymbol{y} = \Psi(\boldsymbol{z})$ of norm*

$$\|\widetilde{\boldsymbol{g}}\| \geq \frac{\|\boldsymbol{g}\|}{\|\nabla\Psi(\boldsymbol{z})\|_{\mathrm{op}}}.$$

In particular, we take $\bar{f}$ to be proper, convex function $f_{\mathtt{cvx}}$ whose minimum is attained at some $\boldsymbol{z}^\star$, we can take

$$\|\widetilde{\boldsymbol{g}}\| \geq \max_{\boldsymbol{z}^\star \in \arg\min f_{\mathtt{cvx}}} \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\| \cdot \|\nabla\Psi(\boldsymbol{z})\|_{\mathrm{op}}}.$$

Proof. We may assume without loss of generality that $\boldsymbol{g} \neq 0$, for otherwise $\widetilde{\boldsymbol{g}} = 0$ and the constant curve $\phi(\varepsilon) = \Psi(\boldsymbol{z}) = \boldsymbol{y}$ satisfies the conclusion of the lemma. Fix a parameter $\eta > 0$ to be chosen at the end of the proof, and define the curve $\phi(\varepsilon) := \Psi(\boldsymbol{z} - \frac{\varepsilon}{\eta}\boldsymbol{g})$. Then, for ε sufficiently small,

$$f_{\mathtt{lft}}(\phi(\varepsilon)) = \bar{f}(\boldsymbol{z} - \frac{\varepsilon}{\eta}\boldsymbol{g}) < \infty,$$

since $\boldsymbol{g}$ is Cauchy direction of $\bar{f}$. Hence, $\phi(\varepsilon) \in \mathsf{dom}(f_{\mathtt{lft}})$ for ε sufficiently small. We compute

$$\begin{aligned}
\lim_{\varepsilon \to 0^+} \frac{f_{\mathtt{lft}}(\phi(\varepsilon)) - f_{\mathtt{lft}}(\boldsymbol{y})}{\varepsilon} &= \lim_{\varepsilon \to 0^+} \frac{f_{\mathtt{lft}}(\Psi(\boldsymbol{z} - \frac{\varepsilon}{\eta}\boldsymbol{g})) - f_{\mathtt{lft}}(\boldsymbol{y})}{\varepsilon} \\
&= \lim_{\varepsilon \to 0^+} \frac{\bar{f}(\boldsymbol{z} - \frac{\varepsilon}{\eta}\boldsymbol{g}) - \bar{f}(\boldsymbol{z})}{\varepsilon} \\
&= \frac{1}{\eta} \lim_{\varepsilon \to 0^+} \frac{f(\boldsymbol{z} - \varepsilon\boldsymbol{g}) - \bar{f}(\boldsymbol{z})}{\varepsilon} \\
&\leq \frac{-\|\boldsymbol{g}\|^2}{\eta}.
\end{aligned} \tag{H.6}$$

Furthermore,

$$\phi'(0) = -\frac{1}{\eta}\nabla\Psi(\boldsymbol{z})\boldsymbol{g}.$$

Since we assume $\boldsymbol{g} \neq 0$ (see above), and since $\sigma_{d_z}(\nabla\Psi(\boldsymbol{z})) \neq 0$ by assumption, we find that $\|\phi'(0)\| > 0$. Thus, continuing Eq. (H.6),

$$\lim_{\varepsilon \to 0^+} \frac{f_{\mathtt{lft}}(\phi(\varepsilon)) - f_{\mathtt{lft}}(\boldsymbol{y})}{\varepsilon} \leq \frac{-\|\boldsymbol{g}\|^2}{\eta} = -\|\phi'(0)\|^2 \cdot \frac{\|\boldsymbol{g}\|^2}{\eta \cdot \|\phi'(0)\|^2} = -\|\phi'(0)\|^2 \cdot \eta \cdot \frac{\|\boldsymbol{g}\|^2}{\|\nabla\Psi(\boldsymbol{z})\boldsymbol{g}\|^2}.$$

In particular, if we set $\eta = \frac{\|\nabla\Psi(\boldsymbol{z})\boldsymbol{g}\|^2}{\|\boldsymbol{g}\|^2}$, we see that $\phi(\cdot)$ is valid for certifying that $\widetilde{\boldsymbol{g}} = \phi'(0)$ is a generalized Cauchy direction. In this case, we have that

$$\|\phi'(0)\| = \frac{\|\nabla\Psi(\boldsymbol{z})\boldsymbol{g}\| \cdot \|\boldsymbol{g}\|^2}{\|\nabla\Psi(\boldsymbol{z})\boldsymbol{g}\|^2} = \|\boldsymbol{g}\| \cdot \frac{\|\boldsymbol{g}\|}{\|\nabla\Psi(\boldsymbol{z})\boldsymbol{g}\|} \geq \frac{\|\boldsymbol{g}\|}{\|\nabla\Psi(\boldsymbol{z})\|_{\mathrm{op}}}.$$

□

Partial minimization preserves Cauchy directions. For our final lemma, recall the set up of `DCLs`. Let $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{\xi})$, and let $f(\boldsymbol{x}) := \min_{\boldsymbol{\xi}} f_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi})$. As shorthand, we say $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{\xi})$ is *admissible* if $\boldsymbol{x} \in \mathsf{dom}(f)$ and $\boldsymbol{\xi} \in \arg\min_{\boldsymbol{\xi}'} f_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}')$. We show that if $f_{\mathtt{lft}}$ has a (generalied) Cauchy direction $\widetilde{\boldsymbol{g}}$ at an admissible $\boldsymbol{y}$, then the norm of the gradient of f must be at least as large as $\|\widetilde{\boldsymbol{g}}\|$.

Lemma H.3. *Suppose that f is proper, and that $f(\boldsymbol{x})$ is $\mathscr{C}^2$ at $\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathsf{dom}(f)$. Let $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{\xi})$ be f-admissible, and suppose that $f_{\mathtt{lft}}$ has a generalized Cauchy direction $\widetilde{\boldsymbol{g}}$ at $\boldsymbol{y}$. Then,*

$$\|\nabla f(\boldsymbol{x})\| \geq \|\widetilde{\boldsymbol{g}}\|.$$

Proof. Let ϕ be a curve which certifies $\widetilde{\boldsymbol{g}}$ as a Cauchy direction of $f_{\mathtt{lft}}$; namely $\phi(0) = \boldsymbol{y}$, $\phi'(0) = \widetilde{\boldsymbol{g}}$, and

$$\lim_{\varepsilon \to 0^+} \frac{f_{\mathtt{lft}}(\phi(\varepsilon)) - f_{\mathtt{lft}}(\boldsymbol{y})}{\varepsilon} \leq -\|\widetilde{\boldsymbol{g}}\|^2.$$

We write $\phi(\varepsilon) = (\phi_1(\varepsilon), \phi_2(\varepsilon))$ in its $(\boldsymbol{x}, \boldsymbol{\xi})$ components. Then,

$$f(\phi_1(\varepsilon)) = \min_{\boldsymbol{\xi}'} f_{\mathtt{lft}}(\phi_1(\varepsilon), \boldsymbol{\xi}') \leq f_{\mathtt{lft}}(\phi_1(\varepsilon), \phi_2(\varepsilon)) = f_{\mathtt{lft}}(\phi(\varepsilon)).$$

By admissibility of $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{\xi})$, $f(\phi_1(0)) = f(\boldsymbol{x}) = f_{\mathtt{lft}}(\boldsymbol{y})$, so that

$$f(\phi_1(\varepsilon)) - f(\boldsymbol{x}) \leq f_{\mathtt{lft}}(\phi(\varepsilon)) - f_{\mathtt{lft}}(\boldsymbol{y}).$$

Dividing by ε and taking limits,

$$\lim_{\varepsilon \to 0^+} \frac{f(\phi_1(\varepsilon)) - f(\boldsymbol{x})}{\varepsilon} \leq \lim_{\varepsilon \to 0^+} \frac{f_{\mathtt{lft}}(\phi(\varepsilon)) - f_{\mathtt{lft}}(\boldsymbol{y})}{\varepsilon} = -\|\widetilde{\boldsymbol{g}}\|^2.$$

On the other hand, since f and ϕ_1 are both differentiable,

$$\lim_{\varepsilon \to 0^+} \frac{f(\phi_1(\varepsilon)) - f(\boldsymbol{x})}{\varepsilon} = \langle \nabla f(\phi_1(0)), \phi_1'(0) \rangle = \langle (\nabla f(\boldsymbol{x}), \mathbf{0}), \widetilde{\boldsymbol{g}} \rangle,$$

where above $(\nabla f(\boldsymbol{x}), \mathbf{0}) \in \mathbb{R}^{d+d_y}$ has a 0 in the remaining d_y coordinates. Therefore,

$$\langle (\nabla f(\boldsymbol{x}), \mathbf{0}), \widetilde{\boldsymbol{g}} \rangle \leq -\|\widetilde{\boldsymbol{g}}\|^2,$$

which requires $\|\nabla f(\boldsymbol{x})\| = \|(\nabla f(\boldsymbol{x}), \mathbf{0})\| \geq \|\widetilde{\boldsymbol{g}}\|$. □

Concluding the proof of Theorem 3.

Proof of Theorem 3. Given $\boldsymbol{x} \in \operatorname{dom}(f)$, pick any $\boldsymbol{z}^\star \in \arg\min(f_{\mathtt{cvx}})$, and any $\boldsymbol{\xi} \in \arg\min_{\boldsymbol{\xi}'} f_{\mathtt{lft}}(\boldsymbol{x}, \boldsymbol{\xi}')$. Set $\boldsymbol{z} = \Phi(\boldsymbol{y})$, and note that $f(\boldsymbol{x}) = f_{\mathtt{lft}}(\boldsymbol{y}) = f_{\mathtt{cvx}}(\Phi(\boldsymbol{y})) = f_{\mathtt{cvx}}(\boldsymbol{z})$, so $\boldsymbol{z} \in \operatorname{dom}(f_{\mathtt{cvx}})$ and $\boldsymbol{y} \in \operatorname{dom}(f_{\mathtt{lft}})$. Note that we cannot have $\boldsymbol{z}^\star = \boldsymbol{z}$, since $\boldsymbol{x} \in \operatorname{dom}_{>}(f)$ implies

$$f_{\mathtt{cvx}}(\boldsymbol{z}) = f(\boldsymbol{x}) > \inf_{\boldsymbol{x}'} f(\boldsymbol{x}') = \inf_{\boldsymbol{y}} f_{\mathtt{lft}}(\boldsymbol{y}) = \inf_{\boldsymbol{y}} f_{\mathtt{cvx}}(\Phi(\boldsymbol{y})) \geq f_{\mathtt{cvx}}(\boldsymbol{z}^\star).$$

By Lemma H.1, $f_{\mathtt{cvx}}$ has a Cauchy direction $\boldsymbol{g}$ at $\boldsymbol{z}$ satisfying

$$\|g\| \geq \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|}.$$

Next, from the $\mathtt{DCL}$, the mapping $\Phi : \mathcal{Y} \to \mathcal{Z}$ is $\mathscr{C}^1$ on an open neighborhoods containing $\boldsymbol{y}$. Hence, $\nabla\Phi(\boldsymbol{y})$ is defined. We now claim that $f_{\mathtt{lft}}$ has a generalized generalized Cauchy direction $\widetilde{\boldsymbol{g}}$ of norm

$$\|\widetilde{\boldsymbol{g}}\| \geq \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|} \cdot \sigma_{d_z}(\nabla\Phi(\boldsymbol{y})). \tag{H.7}$$

Indeed, if $\sigma_{d_z}(\nabla\Phi(\boldsymbol{y})) = 0$, $\widetilde{\boldsymbol{g}} = 0$ suffices (the zero vector is always a generalized Cauchy direction). Otherwise, if $\sigma_{d_z}(\nabla\Phi(\boldsymbol{y})) > 0$, the fact that $d_y \geq d_z$ and the implicit function theorem implies that Φ admits a $\mathscr{C}^1$ right inverse Ψ satisfying $\Phi \circ \Psi(\boldsymbol{z}') = \boldsymbol{z}'$ and $\Psi(\boldsymbol{z}) = \boldsymbol{y}$ on a neighborhood of $\boldsymbol{z}$. This mapping must satisfy $\nabla\Psi(\boldsymbol{z}) = \nabla\Phi(\boldsymbol{y})^\dagger$, so that in particular, $\|\nabla\Psi(\boldsymbol{z})\|_{\mathrm{op}}^{-1} = \sigma_{d_z}(\nabla\Phi(\boldsymbol{y}))$ and $\sigma_{d_z}(\nabla\Psi(\boldsymbol{z})) > 0$. Hence, Lemma H.2 implies that

$$\|\widetilde{\boldsymbol{g}}\| \geq \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|} \cdot \frac{1}{\|\nabla\Psi(\boldsymbol{z})\|_{\mathrm{op}}} = \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\|\boldsymbol{z} - \boldsymbol{z}^\star\|} \cdot \sigma_{d_z}(\nabla\Phi(\boldsymbol{y})),$$

verifying Eq. (H.7). Finally, by Lemma H.3,

$$\|\nabla f(\boldsymbol{x})\| \geq \|\widetilde{\boldsymbol{g}}\| \geq \frac{f_{\mathtt{cvx}}(\boldsymbol{z}) - \inf(f_{\mathtt{cvx}})}{\boldsymbol{z}} \cdot \sigma_{d_z}(\nabla\Phi(\boldsymbol{y})).$$

Lastly, using the $\mathtt{DCL}$, we have $f_{\mathtt{cvx}}(\boldsymbol{z}) = f(\boldsymbol{x})$, $\inf(f_{\mathtt{cvx}}) = \inf(f)$. Substituting in $\boldsymbol{z} = \Phi(\boldsymbol{y})$ and $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{\xi})$,

$$\|\nabla f(\boldsymbol{x})\| \geq \|\widetilde{\boldsymbol{g}}\| \geq \frac{f(\boldsymbol{x}) - \inf(f)}{\|\Phi(\boldsymbol{x}, \boldsymbol{\xi}) - x^\star\|} \cdot \sigma_{d_z}(\nabla\Phi(\boldsymbol{x}, \boldsymbol{\xi})).$$

Since the above holds for any $x^\star \in \arg\min(f_{\mathtt{cvx}})$ and any $\boldsymbol{\xi} \in \arg\min f_{\mathtt{lft}}(\boldsymbol{x}, \cdot)$, Theorem 3 follows. □

H.3 Analysis of gradient descent and reconditioning under weak-PL

H.3.1 Proof of Proposition G.2

The first step of the proof is to ensure sufficiently small step sizes remain in the set $\mathcal{K}$ for which our regularity conditions.

Claim H.4. *Suppose that $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}$. Then,*

$$f(\widetilde{\boldsymbol{x}}_k - \eta_k \nabla f(\widetilde{\boldsymbol{x}}_k)) \le f(\widetilde{\boldsymbol{x}}_k) - \frac{\eta_k}{2}\|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2 = f(\boldsymbol{x}_k) - \frac{\eta_k}{2}\|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2. \tag{H.8}$$

In addition, for all $t \in [0, \eta_k]$, $\widetilde{\boldsymbol{x}}_k - t\nabla f(\widetilde{\boldsymbol{x}}_k) \in \mathcal{K}$.

The proof of Claim H.4 is somewhat elementary, and deferred to the end of the broader argument. We now argue recursively that $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}$ for all k. We argue inductively, noting $f(\widetilde{\boldsymbol{x}}_0) = f(\boldsymbol{x}_0)$ and $\boldsymbol{\Lambda}(\widetilde{\boldsymbol{x}}_0) = \mathbf{I}_n$ ensures the base case $\widetilde{\boldsymbol{x}}_0 \in \mathcal{K}$. Now, if $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}$,

$$f(\boldsymbol{x}_{k+1}) \overset{(i)}{\le} f(\widetilde{\boldsymbol{x}}_k - \eta_k \nabla f(\widetilde{\boldsymbol{x}}_k)) \le f(\boldsymbol{x}_k) - \frac{\eta_k}{2}\|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2, \tag{H.9}$$

where (i) is an equality under Eq. (G.2), but may be an inequality under Eq. (G.3). Hence, $f(\widetilde{\boldsymbol{x}}_{k+1}) = f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) = f(\widetilde{\boldsymbol{x}}_k) \le f(\boldsymbol{x}_0)$ (since $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}$). Hence, since $\widetilde{\boldsymbol{x}}_{k+1}$ is reconditioned, $\widetilde{\boldsymbol{x}}_{k+1} \in \mathcal{K}$ as well.

Subtracting $\inf(f)$ from both sides of Eq. (H.9) and invokingthe $\alpha_{\boldsymbol{x}_0}$-weak PL property of f, the suboptimality gaps $\delta_k := f(\boldsymbol{x}_k) - \inf(f)$ and minimal step $\eta := \min_k\{\eta_k\}$satisfy

$$\delta_{k+1} \le \delta_k - \frac{\eta_k \alpha_{\boldsymbol{x}_0}^2}{2}\delta_k^2 \le \delta_k - \frac{\eta \alpha_{\boldsymbol{x}_0}^2}{2}\delta_k^2. \tag{H.10}$$

We solve this recursion following an argument described in Section 3.2 of Bubeck [2014]. Setting $\omega = \eta \cdot \alpha_{\boldsymbol{x}_0}^2/2$, we have $\delta_k \ge \omega\delta_k^2 + \delta_{k+1}$, or equivalently, $\frac{1}{\delta_{k+1}} \ge \omega\frac{\delta_k}{\delta_{k+1}} + \frac{1}{\delta_k}$. Since $\delta_k \ge \delta_{k+1}$, this implies that $\frac{1}{\delta_{k+1}} \ge \omega + \frac{1}{\delta_k}$. Hence, we find

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \ge \omega.$$

Telescoping, we conclude that $\frac{1}{\delta_{k+1}} \ge \omega(k+1)$, whence

$$f_\lambda(\boldsymbol{x}_k) - \inf(f) = \delta_k \le \frac{1}{\omega k} = \frac{2}{\alpha_{\boldsymbol{x}_0}^2 \eta k}.$$

□

Proof of Claim H.4. Define $\bar{\boldsymbol{x}}(\tau) := \widetilde{\boldsymbol{x}}_k - \tau\nabla f(\widetilde{\boldsymbol{x}}_k)$, noting $\bar{\boldsymbol{x}}(0) = \boldsymbol{x}_k$. The key subtlety in proving the claim is ensuring that the entire line segment $\{\bar{\boldsymbol{x}}(\tau) : \tau \in [0, \eta_k]\}$ lies in the set $\mathcal{K}$ under which relevant regularity conditions on f hold. To start, we may assume without loss of generality that $\nabla f(\widetilde{\boldsymbol{x}}_k) \ne 0$, for otherwise the bound follows trivially. We make two observations

1. since f is $\beta_{\boldsymbol{x}_0}$-upper-smooth on $\mathcal{K}$, there is an open set containing $\widetilde{\boldsymbol{x}}_k$ on which f is $\mathscr{C}^2$. Then, there exist some τ_1 such that, for all $\tau \in [0, \tau_1]$, $f(\bar{\boldsymbol{x}}(\tau)) = f(\boldsymbol{x}_k) - \tau\|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2 + o(\tau) < \phi(0)$. Note further that $= f(\mathsf{recond}_{\boldsymbol{\Lambda}}(\boldsymbol{x}_k)) = f(\boldsymbol{x}_k) \le f(\boldsymbol{x}_0)$ (since $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}$ by assumption.)
2. Since $\boldsymbol{\Lambda}(\widetilde{\boldsymbol{x}}) = I$ and $\boldsymbol{\Lambda}$ is L-Lipschitz on $\mathcal{K}$, there exist some τ_2 such that, for all $\tau \in [0, \tau_2]$, $\|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau)) - \mathbf{I}_n\|_{\mathrm{op}} \le \frac{1}{2}$.

Let us choose τ_0 as the largest real satisfying the above two constraints:

$$\tau_0 := \sup\left\{\tau \le \eta_k : \forall\tau' \in [0,\tau),\ f(\bar{\boldsymbol{x}}(\tau')) \le f(\widetilde{\boldsymbol{x}}_k) \text{ and } \|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau')) - \mathbf{I}_n\|_{\mathrm{op}} \le \frac{1}{2}\right\},$$

and observe that $\bar{\boldsymbol{x}}(\tau) \in \mathcal{K}$ for all $\tau \in [0, \tau_0]$ by construction.

First, we show that $\tau_0 > 0$. Indeed, by assumption, there is any open set containing $\mathcal{K}$ on which f is $\mathscr{C}^2$, and hence, on this open set f is finite. In particular, there is an open set $\mathcal{U} \subset \mathsf{dom}(f)$ with $\widetilde{\boldsymbol{x}}_k \in \mathcal{U}$, f is $\mathscr{C}^2$ on $\mathcal{U}$. Since f is $\mathscr{C}^2$ on $\mathcal{U}$ and $\nabla f(\widetilde{\boldsymbol{x}}_k) \neq 0$, there exists some $\tau_0 > 0$ for such that, for all $\tau' \in [0, \tau_0)$, $f(\bar{\boldsymbol{x}}_k(\tau')) = f(\widetilde{\boldsymbol{x}}_k - \tau' \nabla f(\widetilde{\boldsymbol{x}}_k)) < f(\widetilde{\boldsymbol{x}}_k)$, and since $\boldsymbol{\Lambda}$ is continuous on $\mathsf{dom}(f) \supset \mathcal{U}$ and $\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}_k) = \mathbf{I}_n$, we can shrink τ_0 if necessary to ensure that $\|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau')) - \mathbf{I}_n\|_{\mathrm{op}} = \|\boldsymbol{\Lambda}(\widetilde{\boldsymbol{x}}_k - \tau' \nabla f(\widetilde{\boldsymbol{x}}_k))) - \mathbf{I}_n\|_{\mathrm{op}} \le \frac{1}{2}$.

Further, observe that by $\beta_{\boldsymbol{x}_0}$-smoothness of f on $\mathcal{K}$, a Taylor expansion along the segment parameterized by $\bar{\boldsymbol{x}}(\tau)$ yields

$$
\begin{aligned}
f(\bar{\boldsymbol{x}}(\tau)) &\le f(\widetilde{\boldsymbol{x}}_k) - \left(\tau - \frac{\tau^2 \beta_{\boldsymbol{x}_0}}{2}\right) \cdot \|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2 \quad \forall \tau \in [0, \tau_0] \\
&\le f(\widetilde{\boldsymbol{x}}_k) - \frac{\tau}{2} \cdot \|\nabla f(\widetilde{\boldsymbol{x}}_k)\|^2 \quad \forall \tau \in \left[0, \min\left\{\tau_0, \frac{1}{\beta_{\boldsymbol{x}_0}}\right\}\right]. \qquad \text{(H.11)}
\end{aligned}
$$

To conclude, it suffices to show $\tau_0 \ge \eta_k$. For the sake of contradiction, suppose instead that $\tau_0 < \eta_k \le \min\{\frac{1}{\beta_{\boldsymbol{x}_0}}, \frac{1}{2L_{f,\boldsymbol{x}_0} L_{\mathrm{cond},\boldsymbol{x}_0}}\}$. By (a) continuity of f and $\boldsymbol{\Lambda}$ on $\mathcal{K}$ (b) continuity of $\tau \mapsto \bar{\boldsymbol{x}}(\tau) \in \mathcal{K}$, and (c) the assumption that $\mathcal{K}$ is closed, it must be the case that, either (a) $f(\bar{\boldsymbol{x}}(\tau_0)) = f(\widetilde{\boldsymbol{x}}_k)$ or (b) $\|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau)) - \mathbf{I}_n\|_{\mathrm{op}} = \frac{1}{2}$. To see that (a) cannot hold, we have that $\tau_0 \le \eta_k \le \frac{1}{\beta}$ and Eq. (H.11) implies that $f(\bar{\boldsymbol{x}}(\tau)) < f(\widetilde{\boldsymbol{x}}_k)$. To see that (b) cannot hold, we use $\boldsymbol{\Lambda}(\widetilde{\boldsymbol{x}}_k) = \mathbf{I}_n$ and L_{cond}-Lipschitzness of $\boldsymbol{\Lambda}$ in the $\|\cdot\|_2 \to \|\cdot\|_{\mathrm{op}}$ norm, $L_{f,\boldsymbol{x}_0}$ Lipschitzness of f, and the bound $\tau_0 \le \eta_k \le \frac{1}{2L_{f,\boldsymbol{x}_0} L_{\mathrm{cond},\boldsymbol{x}_0}}$ to attain

$$
\begin{aligned}
\|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau_0)) - \mathbf{I}_n\|_{\mathrm{op}} &= \|\boldsymbol{\Lambda}(\bar{\boldsymbol{x}}(\tau_0)) - \boldsymbol{\Lambda}(\widetilde{\boldsymbol{x}}_k)\|_{\mathrm{op}} \\
&\le L_{\mathrm{cond},\boldsymbol{x}_0} \|\bar{\boldsymbol{x}}(\tau_0) - \widetilde{\boldsymbol{x}}_k\| \le L_{\mathrm{cond},\boldsymbol{x}_0} L_{f,\boldsymbol{x}_0} \tau_0 \le \frac{1}{2}.
\end{aligned}
$$

□

H.3.2 Proof of Proposition G.4

We assume without loss of generality that $\boldsymbol{x}_0 \notin \arg\min(f)$.

Claim H.5. *Fix $\eta > 0$ consider the iterates $\boldsymbol{x}_k$ and $\widetilde{\boldsymbol{x}}_k$ produced by the updates in Eq. (G.2) with $\eta_k = \eta$, η satisfies the step-size conditions of Proposition G.2. Then $\boldsymbol{x}_0$ is in the same connected component of $\mathsf{dom}(f)$ as $\widetilde{\boldsymbol{x}}_k$ for all k.*

Proof. Since $\boldsymbol{\Lambda}(\cdot)$ is a connected reconditiong operator, each $\boldsymbol{x}_k$ and $\widetilde{\boldsymbol{x}}_k$ lie in the same path-connected component of $\mathsf{dom}(f)$ for all k. Moreover, by Claim H.4, the line segment $\widetilde{\boldsymbol{x}}_k - t\nabla f(\widetilde{\boldsymbol{x}}_k), t \in [0, \eta]$ lies entirely in $\mathcal{K}$, so $\widetilde{\boldsymbol{x}}_k$ and $\boldsymbol{x}_{k+1} = \widetilde{\boldsymbol{x}}_k - \eta \nabla f(\widetilde{\boldsymbol{x}}_k)$ lie in the same connected component of $\mathsf{dom}(f)$. Since path-connectedness is an equivalence relation, the result follows. □

Now, since $\mathcal{K}(\boldsymbol{x}_0)$ is compact, and $\widetilde{\boldsymbol{x}}_k \in \mathcal{K}(\boldsymbol{x}_0)$ for all $k \ge 0$, there exists a convergent subsequence $\widetilde{\boldsymbol{x}}_{k_i} \to \bar{\boldsymbol{x}} \in \mathcal{K}(\boldsymbol{x}_0)$. Since f is continuous on $\mathcal{K}(\boldsymbol{x}_0)$, $\lim_{i \to \infty} f(\widetilde{\boldsymbol{x}}_{k_i}) = f(\bar{\boldsymbol{x}})$, so by Proposition G.2, $f(\bar{\boldsymbol{x}}) = \inf(f)$, i.e. $\bar{\boldsymbol{x}} \in \arg\min(f) \cap \mathcal{K}(\boldsymbol{x}_0)$. Since $\bar{\boldsymbol{x}} \in \mathcal{K}(\boldsymbol{x}_0)$ is contained in an open set $\mathcal{U}$, which is in turn contained in $\mathsf{dom}(f)$, there is an open ball of radius r, $\mathcal{B}_r(\bar{\boldsymbol{x}})$, contained in $\mathsf{dom}(f)$. Since for some $i_\star$ sufficiently large, $\widetilde{\boldsymbol{x}}_{k_\star} \in \mathcal{B}_r(\bar{\boldsymbol{x}})$, $\widetilde{\boldsymbol{x}}_{k_\star}$ is in the same path-connected component as $\bar{\boldsymbol{x}}$. But by Claim H.5, it is also in the same path-connected component as $\boldsymbol{x}_0$. Since path-connectedness is an equivalence relation, the result follows.

I DCL for Output Estimation (Proposition 4.2)

In this section, we establish the weak-PL property of our regularized loss function $\mathcal{L}_\lambda(\cdot) = \mathcal{L}_{\texttt{OE}}(\cdot) + \lambda \mathcal{R}_{\texttt{info}}(\cdot)$. Our strategy is to show that $\mathcal{L}_\lambda(\cdot)$ admits a $\mathtt{DCL}$, which leads to a weak-PL constant $\alpha(\mathsf{K})$ for each K, whose parameters are themselves bounded in terms of $\mathcal{L}_\lambda(\cdot)$. Before continuing, we recall that n denotes the dimension of the system state $\mathbf{x}$ (and internal state $\hat{\mathbf{x}}$), m of the observation $\mathbf{y}$, and p the output $\mathbf{z}$, and that $\mathrm{poly}_{\mathrm{op}}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \kappa)$ denote a (universal) polynomial function of operator norm of matrix, arguments $\|\mathbf{X}_1\|, \|\mathbf{X}_2\|, \ldots$, and a polynomial in scalar argument κ. We use $\mathbb{I}_\infty$ to denote the 1-∞ indicator, i.e. for some event $\mathcal{E}$, $\mathbb{I}_\infty\{\mathcal{E}\} = 1$ if $\mathcal{E}$ is true, and $\mathbb{I}_\infty\{\mathcal{E}\} = +\infty$ otherwise.

All proofs of the lemmas that follow are deferred to Appendix I.1. To proceed, we need to invoke Theorem 3 by specifying the `DCL` of the function

$$\mathcal{L}_\lambda(\mathsf{K}) = \mathcal{L}_{\mathtt{OE}}(\mathsf{K}) + \lambda \mathcal{R}_{\mathtt{info}}(\mathsf{K}) = \lim_{t\to\infty} \mathbb{E}[\|\mathbf{z}(t) - \hat{\mathbf{z}}(t)\|^2] + \lambda \mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}].$$

Throughout, given a matrix $\mathbf{\Sigma} \succ 0$ partitioned in 2×2 blocks, we more generally define

$$\mathbf{Z}(\mathbf{\Sigma}) := \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^\top. \tag{I.1}$$

With the above notation, we can express

$$\begin{aligned}
\mathcal{L}_\lambda(\mathsf{K}) &= \lim_{t\to\infty} \mathbb{E}\left[\|\mathbf{G}\mathbf{x}(t) - \mathbf{C}_{\mathsf{K}}\hat{\mathbf{x}}(t)\|^2\right] + \lambda \cdot \mathrm{tr}\left[\mathbf{Z}_{\mathsf{K}}^{-1}\right] \\
&= \lim_{t\to\infty} \mathrm{tr}\left[\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix} \mathbb{E}\left[\begin{bmatrix}\mathbf{x}(t) \\ \hat{\mathbf{x}}(t)\end{bmatrix}\begin{bmatrix}\mathbf{x}(t) \\ \hat{\mathbf{x}}(t)\end{bmatrix}^\top\right]\begin{bmatrix}\mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top\end{bmatrix}\right] + \lambda \mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}})^{-1}\right] \\
&= \mathrm{tr}\left[\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix} \mathbf{\Sigma}_{\mathsf{K}} \begin{bmatrix}\mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top\end{bmatrix}\right] + \lambda \mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}})^{-1}\right].
\end{aligned} \tag{I.2}$$

This leads to the following notion of the lifted function.

Definition I.1 (The lifted function)**.** We define the lifted function on the space of parameters $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{K}_{\mathtt{info}} \times \mathbb{S}^{2n}$ as follows

$$f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma}) = \left(\mathrm{tr}\left[\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix} \mathbf{\Sigma} \begin{bmatrix}\mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top\end{bmatrix} + \lambda \cdot \mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma})^{-1}\right]\right]\right) \cdot \mathbb{I}_\infty\{(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\mathtt{lft}},\} \tag{I.3a}$$

$$\mathcal{C}_{\mathtt{lft}} := \left\{ (\mathsf{K}, \mathbf{\Sigma}) : \begin{array}{c} (i)\ \mathbf{\Sigma} \succ 0,\ \mathbf{Z}(\mathbf{\Sigma}) \succ 0 \quad (ii)\ \mathbf{A}\mathbf{\Sigma}_{11} + \mathbf{\Sigma}_{11}\mathbf{A}^\top + \mathbf{W}_1 = 0 \\ (iii)\ \begin{pmatrix}\mathbf{A} & 0 \\ \mathbf{B}_{\mathsf{K}}\mathbf{C} & \mathbf{A}_{\mathsf{K}}\end{pmatrix}\mathbf{\Sigma} + \mathbf{\Sigma}\begin{pmatrix}\mathbf{A} & 0 \\ \mathbf{B}_{\mathsf{K}}\mathbf{C} & \mathbf{A}_{\mathsf{K}}\end{pmatrix}^\top + \begin{pmatrix}\mathbf{W}_1 & 0 \\ 0 & \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top\end{pmatrix} \preceq 0 \end{array} \right\}. \tag{I.3b}$$

We extend $f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma})$ to the space of all (unconstrained, even possible unstable) filters $\mathsf{K} = (\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}})$ by setting the lifted function to be infinte when $\mathsf{K} \notin \mathcal{K}_{\mathtt{info}}$: $f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma}) = f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma})\mathbb{I}_\infty\{\mathsf{K} \in \mathcal{K}_{\mathtt{info}}\}$.[7]

Step 1. Verifying the lifting. We first verify that $f_{\mathtt{lft}}$ is indeed a lifted function of $\mathcal{L}_\lambda$.

Lemma I.1. *For any feasible* $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$,

$$\mathcal{L}_\lambda(\mathsf{K}) = \min_{\mathbf{\Sigma} \in \mathbb{S}^{2n}} f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma}),$$

and this minimum is attained for $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathsf{K}}$.

Step 2. Convex reparametrization. Next, we introduce the transformation Φ:

Definition I.2. We define the convex parameter $\boldsymbol{\nu} := (\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{M}_1, \mathbf{M}_2)$ and the transformation

$$\boldsymbol{\nu}^\top = \Phi(\mathsf{K}, \mathbf{\Sigma}) := \begin{pmatrix} \mathbf{U}(\mathbf{A}_{\mathsf{K}}\mathbf{V}^\top + \mathbf{B}_{\mathsf{K}}\mathbf{C}(\mathbf{\Sigma})_{11}) + (\mathbf{\Sigma}^{-1})_{11}\mathbf{A}(\mathbf{\Sigma})_{11} \\ \mathbf{U}\mathbf{B}_{\mathsf{K}} \\ \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top \\ (\mathbf{\Sigma}^{-1})_{11} \\ (\mathbf{\Sigma})_{11} \end{pmatrix}, \tag{I.4a}$$

$$\text{where } \begin{pmatrix}\mathbf{U} \\ \mathbf{V}\end{pmatrix} := \begin{pmatrix}(\mathbf{\Sigma}^{-1})_{12} \\ (\mathbf{\Sigma})_{12}\end{pmatrix}. \tag{I.4b}$$

---

[7] This formalism is just to accomodate for the fact that we encode constraints on domains in the function in general `DCL` framework.

We let d_ν denote the dimension of the parameter $\boldsymbol{\nu}$ and let d_y denote the dimension of the parameters $(\mathsf{K}, \boldsymbol{\Sigma})$, both as Euclidean vectors. One can then verify that $d_\nu \le d_y$; that is, the lifted function indeed has more parameters than the convex one. The following shows that there exists a convex function $f_{\texttt{cvx}}$, which completes the `DCL`:

Lemma I.2. *There exists a convex function* $f_{\texttt{cvx}} : \mathbb{R}^{d_\nu} \to \bar{\mathbb{R}}$ *such that*

$$f_{\texttt{lft}}(\mathsf{K}, \boldsymbol{\Sigma}) = f_{\texttt{cvx}}(\Phi(\mathsf{K}, \boldsymbol{\Sigma})). \tag{I.5}$$

The transformation Φ and associated convex function $f_{\texttt{cvx}}$ was first developed by Scherer et al. [1997], cf. also Masubuchi et al. [1998] for contemporaneous independent work.

Step 3. Controlling the weak-PL constant. Lastly, we show that the `DCL` lends itself to a bounded PL constant by invoking Theorem 3. To do this, we need to show that the image of $\Phi(\mathsf{K}, \boldsymbol{\Sigma})$ is not too large, and that $\nabla\Phi(\cdot)$ has rank at least d_ν. We establish both in sequence. Let $\mathbf{U}_{\mathsf{K}}$ and $\mathbf{V}_{\mathsf{K}}$ be corresponding to Eq. (I.4b) with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathsf{K}}$, i.e.

$$\mathbf{U}_{\mathsf{K}} = (\boldsymbol{\Sigma}_{\mathsf{K}}^{-1})_{12}, \quad \mathbf{V}_{\mathsf{K}} = (\boldsymbol{\Sigma}_{\mathsf{K}})_{12}. \tag{I.6}$$

Lemma I.3 (Parameter compactness)**.** *Consider* $(\mathsf{K}, \boldsymbol{\Sigma}_{\mathsf{K}})$*, where* $\boldsymbol{\Sigma}_{\mathsf{K}}$ *is the stationary covariance associated with* K*. Then,*

$$\|\Phi(\mathsf{K}, \boldsymbol{\Sigma})\|_{\ell_2} \le (\max\{n, \sqrt{nm}\} + \sqrt{\mathcal{L}_{\texttt{OE}}(\mathsf{K})}) \cdot \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}), \tag{I.7}$$

where $\|\boldsymbol{\nu}\|_{\ell_2} := \sqrt{\sum_{i=1}^{3} \|\mathbf{L}_i\|_{\mathrm{F}}^2 + \sum_{j=1}^{2} \|\mathbf{M}_i\|_{\mathrm{F}}^2}$ *denotes the Euclidean norm of the parameter* $\boldsymbol{\nu}$*. Moreover, if* $\mathbf{U}_{\mathsf{K}}$ *and* $\mathbf{V}_{\mathsf{K}}$ *are invertible, then the filter parameters are bounded by*

$$\max\{\|\mathbf{A}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|\} \le \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{U}_{\mathsf{K}}^{-1}, \mathbf{V}_{\mathsf{K}}^{-1}), \quad \|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}} \le \sqrt{\mathcal{L}_{\texttt{OE}}(\mathsf{K})/\|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|}.$$

Lemma I.4 (Conditioning of $\nabla\Phi$)**.** *Suppose that* $\mathsf{K} \in \mathcal{K}_{\texttt{info}}$*. Then,* Φ *is differentiable in an open neighborhood of* $(\mathsf{K}, \boldsymbol{\Sigma}_{\mathsf{K}})$*, and if* $\mathbf{U}_{\mathsf{K}}$ *and* $\mathbf{V}_{\mathsf{K}}$ *are invertible,*

$$\begin{aligned}\frac{1}{\sigma_{d_\nu}(\nabla\Phi(\mathsf{K}, \boldsymbol{\Sigma}_{\mathsf{K}}))} &\le \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}, \mathbf{U}_{\mathsf{K}}^{-1}, \mathbf{V}_{\mathsf{K}}^{-1}\right) \\ &\le \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{U}_{\mathsf{K}}^{-1}, \mathbf{V}_{\mathsf{K}}^{-1}, \mathcal{L}_{\texttt{OE}}(\mathsf{K})\right),\end{aligned}$$

where the last line is a consequence of Lemma I.3.

To conclude, we eliminate dependencies on $\mathbf{U}_{\mathsf{K}}$ and $\mathbf{V}_{\mathsf{K}}$:

Lemma I.5. *If* $\mathbf{Z} = \mathbf{Z}(\boldsymbol{\Sigma})$ *is invertible, the matrices* $\mathbf{U} = (\boldsymbol{\Sigma}^{-1})_{12}$ *and* $\mathbf{V} = \boldsymbol{\Sigma}_{12}$ *are invertible, and their inverses are bounded in operator norm as*

$$\|\mathbf{U}^{-1}\| \le \sqrt{\|\mathbf{Z}^{-1}\|\|\boldsymbol{\Sigma}^{-1}\|}, \quad \|\mathbf{V}^{-1}\| \le \|\boldsymbol{\Sigma}\|\sqrt{\|\boldsymbol{\Sigma}^{-1}\|^3\|\|\mathbf{Z}^{-1}\|}.$$

As a consequence of Lemmas I.3 and I.4,

$$\max\left\{\|\mathbf{A}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \frac{1}{\sigma_{d_y}(\nabla\Phi(\mathsf{K}, \boldsymbol{\Sigma}_{\mathsf{K}}))}\right\} \le \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{Z}_{\mathsf{K}}^{-1}, \mathcal{L}_{\texttt{OE}}(\mathsf{K})\right). \tag{I.8}$$

The conclusion of Eq. (I.8) and the bound $\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}} \le \sqrt{\mathcal{L}_{\texttt{OE}}(\mathsf{K})/\|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|}$ from Lemma I.3 are precisely the conclusions of Proposition 4.2. □

I.1 Supporting proofs for Proposition 4.2

I.2 Proof of Lemma I.1

Recall from Eq. (I.2) that

$$\mathcal{L}_\lambda(\mathsf{K}) = [\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}]\, \boldsymbol{\Sigma}_{\mathsf{K}} \begin{bmatrix} \mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top \end{bmatrix} + \lambda \mathrm{tr}\left[\mathbf{Z}(\boldsymbol{\Sigma}_{\mathsf{K}})^{-1}\right].$$

Since $\mathbf{\Sigma}_{\mathsf{K}}$ satisfies the constraint in Eq. (I.3b) with equality, and since $\mathbf{\Sigma}_{\mathsf{K}} \succ 0$ and $\mathbf{Z}_{\mathsf{K}} \succ 0$ for any $\mathbf{\Sigma}_{\mathsf{K}} \in \mathcal{K}_{\texttt{info}}$ by Lemma G.7, we see that $(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}}) \in \mathcal{C}_{\texttt{lft}}$, and therefore

$$\mathcal{L}_{\lambda}(\mathsf{K}) = \left\{ [\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}]\, \mathbf{\Sigma}_{\mathsf{K}} \begin{bmatrix} \mathbf{G}^{\top} \\ -\mathbf{C}_{\mathsf{K}}^{\top} \end{bmatrix} + \lambda \mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}})^{-1}\right] \right\} \mathbb{I}_{\infty}\{(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}}) \in \mathcal{C}_{\texttt{lft}}\} := f_{\texttt{lft}}(\mathsf{K}, \mathbf{\Sigma}_{\mathsf{K}}).$$

Next, let $\mathbf{\Sigma}$ be any other matrix such that $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\texttt{lft}}$. Examining $f_{\texttt{lft}}$, it suffices to show that

$$\text{(a) } \mathbf{\Sigma} \succeq \mathbf{\Sigma}_{\mathsf{K}} \qquad \text{and} \qquad \text{(b) } \mathbf{Z}(\mathbf{\Sigma}) \preceq \mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}}).$$

We show (a) and (b) hold as follows.

Proof of point (a). Recall the matix $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ and $\mathbf{W}_{\mathrm{cl},\mathsf{K}}$

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}} := \begin{bmatrix} \mathbf{A} & 0 \\ \mathbf{B}_{\mathsf{K}}\mathbf{C} & \mathbf{A}_{\mathsf{K}} \end{bmatrix}, \quad \mathbf{W}_{\mathrm{cl},\mathsf{K}} := \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^{\top} \end{bmatrix} \succeq 0.$$

Then, $\mathbf{\Sigma}_{\mathsf{K}}$ is the solution to the Lyapunov equation

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}_{\mathsf{K}} + \mathbf{\Sigma}_{\mathsf{K}}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top} + \mathbf{W}_{\mathrm{cl},\mathsf{K}} = 0. \tag{I.9}$$

Since $\mathbf{\Sigma} \in \mathcal{C}_{\texttt{lft}}$, Eq. (I.3b) part (iii) implies

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top} + \mathbf{W}_{\mathrm{cl},\mathsf{K}} \preceq 0. \tag{I.10}$$

Subtracting these equations gives

$$0 \succeq \mathbf{A}_{\mathrm{cl},\mathsf{K}}(\mathbf{\Sigma} - \mathbf{\Sigma}_{K}) + (\mathbf{\Sigma} - \mathbf{\Sigma}_{K})\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top}.$$

In other words, there exists a matrix $\mathcal{Q} \succeq 0$ such that

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top}(\mathbf{\Sigma} - \mathbf{\Sigma}_{K}) + (\mathbf{\Sigma} - \mathbf{\Sigma}_{K})\mathbf{A}_{\mathrm{cl},\mathsf{K}}^{\top} + \mathcal{Q} = 0. \tag{I.11}$$

Since it is assumed $\mathsf{K} \in \mathcal{K}_{\texttt{info}} \subset \mathcal{K}_{\texttt{stab}}$, then $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is Hurwitz. Therefore, the uniqueness of solutions to Lyapunov equations with stable matrices shows that the unique solution to Eq. (I.11) is some matrix $\mathbf{\Sigma} - \mathbf{\Sigma}_{K} = \widetilde{\mathbf{\Sigma}} \succeq 0$, as needed.

Proof of point (b). We build on $\mathbf{\Sigma} \succeq \mathbf{\Sigma}_{\mathsf{K}}$. Recall that $\mathbf{\Sigma}_{\mathsf{K}} \succ 0$ as noted above, so that we can invert $\mathbf{\Sigma}^{-1} \preceq \mathbf{\Sigma}_{\mathsf{K}}^{-1}$. Taking the bottom-right block and using the block inversion formula,

$$(\mathbf{\Sigma}_{11} - \mathbf{Z}(\mathbf{\Sigma}))^{-1} \preceq (\mathbf{\Sigma}_{11} - \mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}}))^{-1},$$

which is equivalent after inversion to

$$\mathbf{\Sigma}_{11} - \mathbf{Z}(\mathbf{\Sigma}) \succeq \mathbf{\Sigma}_{11,\mathsf{K}} - \mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}}). \tag{I.12}$$

Next, observe that since $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{11,\mathsf{K}} = \mathbf{\Sigma}_{11,\mathrm{sys}}$ for $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\texttt{lft}}$ (this follows from the uniqueness of solutions to Lyapunov equations with Hurwitz matrices and constraint (ii) of Eq. (I.3b)). Therefore, Eq. (I.12) simplifies to $\mathbf{Z}(\mathbf{\Sigma}_{\mathsf{K}}) \succeq \mathbf{Z}(\mathbf{\Sigma})$, as needed. □

I.3 Proof of Lemma I.2

Consider the parametrization $\boldsymbol{\nu} = (\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{M}_1, \mathbf{M}_2) = \Phi(\mathsf{K}, \mathbf{\Sigma})$. We can then write

$$f_{\texttt{lft}}(\mathsf{K}, \mathbf{\Sigma}) = \left(\underbrace{\mathrm{tr}\left[[\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}]\, \mathbf{\Sigma} \begin{bmatrix} \mathbf{G}^{\top} \\ -\mathbf{C}_{\mathsf{K}}^{\top} \end{bmatrix} \right]}_{\widetilde{f}_1(\mathsf{K}, \mathbf{\Sigma})} + \lambda \cdot \underbrace{\mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma})^{-1}\right]}_{\widetilde{f}_2(\mathsf{K}, \mathbf{\Sigma})} \right) \cdot \mathbb{I}_{\infty}\{(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\texttt{lft}}\}.$$

We show that

(a) Whenever $(\mathsf{K}, \mathbf{\Sigma}) \in \mathsf{dom}(f_{\texttt{lft}})$ (that is, $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\texttt{lft}}$), then there are affine matrix-valued functions $\widetilde{\mathbf{C}}(\cdot) \in \mathbb{R}^{p \times 2n}$ and $\widetilde{\mathbf{X}}(\cdot) \in \mathbb{S}^{2n}$ with $\widetilde{\mathbf{X}}(\nu) \succ 0$ of $\boldsymbol{\nu}$ such that

$$\widetilde{f}_1(\mathsf{K}, \mathbf{\Sigma}) = \mathrm{tr}[\widetilde{\mathbf{C}}(\boldsymbol{\nu})^{\top}\widetilde{\mathbf{X}}(\boldsymbol{\nu})^{-1}\widetilde{\mathbf{C}}(\boldsymbol{\nu})].$$

(b) Whenever $(\mathsf{K}, \mathbf{\Sigma}) \in \mathsf{dom}(f_{\mathtt{lft}})$ (that is, $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\mathtt{lft}}$), then $\mathbf{M}_1 \succ 0$ and $\mathbf{M}_2 \succ \mathbf{M}_1^{-1}$. One can further express $\widetilde{f}_2(\mathsf{K}, \mathbf{\Sigma}) = \mathrm{tr}[(\mathbf{M}_2 - \mathbf{M}_1^{-1})^{-1}]$.

(c) There exists a convex set $\mathcal{C}_{\mathtt{cvx}}$ such that $(\mathsf{K}, \mathbf{\Sigma}) \in \mathcal{C}_{\mathtt{lft}}$ if and only if $\boldsymbol{\nu} \in \mathcal{C}_{\mathtt{cvx}}$.

We turn to the verification of points (a)-(c) momentarily. Presently, let us conclude the proof. Points (a)-(c) directly imply that $f_{\mathtt{lft}}(\mathsf{K}, \mathbf{\Sigma}) = f_{\mathtt{cvx}}(\Phi(\mathsf{K}, \mathbf{\Sigma}))$, where

$$f_{\mathtt{cvx}}(\boldsymbol{\nu}) := \left(\mathrm{tr}[\widetilde{\mathbf{C}}(\boldsymbol{\nu})^\top \widetilde{\mathbf{X}}(\cdot)^{-1} \widetilde{\mathbf{C}}(\boldsymbol{\nu})] + \lambda \cdot \mathrm{tr}[(\mathbf{M}_2 - \mathbf{M}_1^{-1})^{-1}]\right) \mathbb{I}_\infty\{\boldsymbol{\nu} \in \mathcal{C}_{\mathtt{cvx}}\}.$$

To conclude, it remains to show that $f_{\mathtt{cvx}}(\cdot)$ is convex. Since $\mathcal{C}_{\mathtt{cvx}}$ is convex by point (c), it suffices to show that the functions (i) $\boldsymbol{\nu} \mapsto \widetilde{\mathbf{C}}(\boldsymbol{\nu})^\top \widetilde{\mathbf{X}}(\boldsymbol{\nu})^{-1} \widetilde{\mathbf{C}}(\boldsymbol{\nu})$ and (ii) that $(\mathbf{M}_1, \mathbf{M}_2) \mapsto \mathrm{tr}[(\mathbf{M}_2 - \mathbf{M}_1^{-1})^{-1}]$ are both convex. Since $\widetilde{\mathbf{C}}(\cdot)$ and $\widetilde{\mathbf{X}}(\cdot)$ are affine in $\boldsymbol{\nu}$ (and affine composition preserves convexity), point (i) follows from the following lemma:

Lemma I.6. *The function* $g(\widetilde{\mathbf{C}}, \widetilde{\mathbf{X}}) = \mathrm{tr}[\widetilde{\mathbf{C}}^\top \widetilde{\mathbf{X}}^{-1} \widetilde{\mathbf{C}}]$ *is convex on the domain* $(\widetilde{\mathbf{C}}, \widetilde{\mathbf{X}}) \in \mathbb{R}^{\widetilde{p} \times \widetilde{n}} \times \mathbb{S}^{\widetilde{n}}_{++}$.

Point (ii) follows from the following lemma:

Lemma I.7. *The function* $h(\mathbf{M}_1, \mathbf{M}_2) = \mathrm{tr}[(\mathbf{M}_2 - \mathbf{M}_1^{-1})^{-1}]$ *is convex on the domain* $\{(\mathbf{M}_1, \mathbf{M}_2) \in \mathbb{S}^n_{++} \times \mathbb{S}^n_{++} : \mathbf{M}_2 \succ \mathbf{M}_1^{-1}\}$.

The proof of these lemmas is defered to Appendix I.7.

Proof of point (a). Introduce

$$\widetilde{\mathbf{C}}(\boldsymbol{\nu}) := [\mathbf{G}\mathbf{M}_2 - \mathbf{L}_3 \quad \mathbf{G}]^\top = \left[\mathbf{G}\mathbf{M}_2 - \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top \quad \mathbf{G}\right]^\top, \quad \widetilde{\mathbf{X}}(\boldsymbol{\nu}) := \begin{pmatrix} \mathbf{M}_2 & \mathbf{I} \\ \mathbf{I} & \mathbf{M}_1 \end{pmatrix}.$$

It is shown in the proof of part (c) below that $\widetilde{\mathbf{X}}(\boldsymbol{\nu}) \succ 0$. We compute (noting that $\mathbf{M}_2$ is symmetric) that

$$\begin{aligned}
\widetilde{\mathbf{C}}(\boldsymbol{\nu})^\top \widetilde{\mathbf{X}}(\boldsymbol{\nu})^{-1} \widetilde{\mathbf{C}}(\boldsymbol{\nu}) &= \left[\mathbf{G}\mathbf{M}_2 - \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top \quad \mathbf{G}\right] \begin{bmatrix} \mathbf{M}_2 & \mathbf{I} \\ \mathbf{I} & \mathbf{M}_1 \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{G}\mathbf{M}_2 - \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top)^\top \\ \mathbf{G}^\top \end{bmatrix} \\
&= [\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}] \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{I} & 0 \end{bmatrix}^\top \begin{bmatrix} \mathbf{M}_2 & \mathbf{I} \\ \mathbf{I} & \mathbf{M}_1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{G} \\ -\mathbf{C}_{\mathsf{K}} \end{bmatrix} \\
&\overset{(i)}{=} [\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}] \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{V}^\top & -\mathbf{U}^{-1}\mathbf{M}_1\mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top \end{bmatrix} \\
&\overset{(ii)}{=} [\mathbf{G} \quad -\mathbf{C}_{\mathsf{K}}]\, \mathbf{\Sigma} \begin{bmatrix} \mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top \end{bmatrix} = \widetilde{f}_1(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}, \mathbf{\Sigma}).
\end{aligned}$$

Here, equality (i) uses the block matrix inversion formula, and the facts that $\mathbf{M}_1, \mathbf{M}_2$ are invertible, and $\mathbf{I} = \mathbf{M}_1\mathbf{M}_2 + \mathbf{U}\mathbf{V}^\top$ as to be shown in Claim I.8; Equality (ii) is given by the following calculation, whose steps follow from Claim I.8.

$$\begin{align*}
\mathbf{\Sigma} &= \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{M}_1 & \mathbf{U} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{I} & 0 \end{bmatrix} \tag*{Eq. (I.13a)} \\
&= \begin{bmatrix} \mathbf{I} & 0 \\ -\mathbf{U}^{-1}\mathbf{M}_1 & \mathbf{U}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{I} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ -\mathbf{U}^{-1}(\mathbf{M}_1\mathbf{M}_2 - \mathbf{I}) & -\mathbf{U}^{-1}\mathbf{M}_1\mathbf{V} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{V}^\top & -\mathbf{U}^{-1}\mathbf{M}_1\mathbf{V} \end{bmatrix}. \tag*{Eq. (I.13c)}
\end{align*}$$

Proof of point (b). To show point (b), we have

$$\begin{aligned}
\mathrm{tr}\left[\mathbf{Z}(\mathbf{\Sigma})^{-1}\right] &= \mathrm{tr}[(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^\top)^{-1}] \\
&= \mathrm{tr}[\left(-(\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^\top) + \mathbf{\Sigma}_{11}\right)^{-1}] \\
&= \mathrm{tr}[\left(-(\mathbf{\Sigma}^{-1})_{11}^{-1} + \mathbf{\Sigma}_{11}\right)^{-1}] = \mathrm{tr}[(\mathbf{M}_2 - \mathbf{M}_1^{-1})^{-1}].
\end{aligned}$$

In addition, $\mathbf{M}_1 = -(\mathbf{\Sigma}^{-1})_{11} \succ 0$ since $\mathbf{\Sigma} \succ 0$ (see the definition of the constraint set $\mathcal{C}_{\mathtt{lft}}$ in Eq. (I.3b)). Lastly, since $\mathbf{Z}(\mathbf{\Sigma}) \succ 0$ from the definition of $\mathcal{C}_{\mathtt{lft}}$, it must be the case that $\mathbf{M}_2 \succ \mathbf{M}_1^{-1}$.

Proof of point (c). We first remark that, as in the specification of the lifted constraint set $\mathcal{C}_{\texttt{lft}}$, specification of the convex constraint set $\mathcal{C}_{\texttt{cvx}}$ does not invole the parameter $\mathbf{C}_\mathsf{K}$ at all. We first show that $\boldsymbol{\nu} = \Phi(\mathsf{K}, \boldsymbol{\Sigma})$ satisfies some useful identities, using the convex parameterization in [Scherer et al., 1997, Masubuchi et al., 1998].

Claim I.8. $\boldsymbol{\nu} = \Phi(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \boldsymbol{\Sigma})$ *satisfies the identities*

$$\mathbf{X} = \begin{pmatrix}\mathbf{M}_2 & \mathbf{V}\\ \mathbf{I} & 0\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{I} & 0\\ \mathbf{M}_1 & \mathbf{U}\end{pmatrix} \qquad \boldsymbol{\Sigma} = \mathbf{X}^{-1} = \begin{pmatrix}\mathbf{I} & 0\\ \mathbf{M}_1 & \mathbf{U}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{M}_2 & \mathbf{V}\\ \mathbf{I} & 0\end{pmatrix} \tag{I.13a}$$

$$(\mathbf{A}_\mathsf{K} \quad \mathbf{B}_\mathsf{K}) = (\mathbf{U}^{-1} \quad 0)\begin{pmatrix}\mathbf{L}_1 - \mathbf{M}_1\mathbf{A}\mathbf{M}_2 & \mathbf{L}_2\\ 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{V}^\top & 0\\ \mathbf{C}\mathbf{M}_2 & \mathbf{I}\end{pmatrix}^{-1}, \tag{I.13b}$$

$$\mathbf{I} = \mathbf{M}_1\mathbf{M}_2 + \mathbf{U}\mathbf{V}^\top. \tag{I.13c}$$

Proof of Claim I.8. To satisfy Eqs. (I.13a) and (I.13c), one uses the variables (written in terms of $\mathbf{X}$)

$$\begin{pmatrix}\mathbf{M}_1 & \mathbf{M}_2\\ \mathbf{U} & \mathbf{V}\end{pmatrix} = \begin{pmatrix}(\mathbf{X})_{11} & (\mathbf{X}^{-1})_{11}\\ (\mathbf{X})_{12} & (\mathbf{X}^{-1})_{12}\end{pmatrix} = \begin{pmatrix}(\boldsymbol{\Sigma}^{-1})_{11} & (\boldsymbol{\Sigma})_{11}\\ (\boldsymbol{\Sigma}^{-1})_{12} & (\boldsymbol{\Sigma})_{12}\end{pmatrix}. \tag{I.14}$$

Next, by Eq. (I.13b) we have

$$\begin{pmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{pmatrix} = \begin{pmatrix}\mathbf{U}(\mathbf{A}_\mathsf{K}\mathbf{V}^\top + \mathbf{B}_\mathsf{K}\mathbf{C}\mathbf{M}_2)\\ \mathbf{U}\mathbf{B}_\mathsf{K}\end{pmatrix} + \begin{pmatrix}\mathbf{M}_1\mathbf{A}\mathbf{M}_2\\ 0\end{pmatrix}.$$

Hence, combining with Eq. (I.14) and setting $\mathbf{L}_3 = \mathbf{C}_\mathsf{K}\mathbf{V}^\top$, these identities are satisfied for

$$\boldsymbol{\nu}^\top = \begin{pmatrix}\mathbf{L}_1\\ \mathbf{L}_2\\ \mathbf{L}_3\\ \mathbf{M}_1\\ \mathbf{M}_2\end{pmatrix} = \begin{pmatrix}\mathbf{U}(\mathbf{A}_\mathsf{K}\mathbf{V}^\top + \mathbf{B}_\mathsf{K}\mathbf{C}(\boldsymbol{\Sigma})_{11}) + (\boldsymbol{\Sigma}^{-1})_{11}\mathbf{A}(\boldsymbol{\Sigma})_{11}\\ \mathbf{U}\mathbf{B}_\mathsf{K}\\ \mathbf{C}_\mathsf{K}\mathbf{V}^\top\\ (\boldsymbol{\Sigma}^{-1})_{11}\\ (\boldsymbol{\Sigma})_{11}\end{pmatrix}, \quad \text{where } \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} := \begin{pmatrix}(\boldsymbol{\Sigma}^{-1})_{12}\\ (\boldsymbol{\Sigma})_{12}\end{pmatrix}.$$

□

To conclude, we use Claim I.8 to check that $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \boldsymbol{\Sigma}_\mathsf{K}) \in \mathcal{C}_{\texttt{lft}}$ if and only if $\Phi(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \boldsymbol{\Sigma}_\mathsf{K}) \in \mathcal{C}_{\texttt{cvx}}$ for some convex constraint set $\mathcal{C}_{\texttt{cvx}}$. Recall the definition of $\mathcal{C}_{\texttt{lft}}$ in Eq. (I.3b). Via a Schur complement argument, we can express $(\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K}, \boldsymbol{\Sigma})$ in $\mathcal{C}_{\texttt{lft}}$ if and only if $\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}$ and $\mathbf{X} := \boldsymbol{\Sigma}^{-1}$ satisfy

$$\begin{bmatrix}\mathbf{X}\begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_\mathsf{K}\mathbf{C} & \mathbf{A}_\mathsf{K}\end{bmatrix} + \begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_\mathsf{K}\mathbf{C} & \mathbf{A}_\mathsf{K}\end{bmatrix}^\top\mathbf{X} & \mathbf{X}\begin{bmatrix}\mathbf{I} & 0\\ 0 & \mathbf{B}_\mathsf{K}\end{bmatrix}\\ \begin{bmatrix}\mathbf{I} & 0\\ 0 & \mathbf{B}_\mathsf{K}\end{bmatrix}^\top\mathbf{X} & -\begin{bmatrix}\mathbf{W}_1^{-1} & 0\\ 0 & \mathbf{W}_2^{-1}\end{bmatrix}\end{bmatrix} \preceq 0 \tag{I.15a}$$

$$\mathbf{X} \succ 0 \tag{I.15b}$$

$$\mathbf{A}(\mathbf{X}^{-1})_{11} + (\mathbf{X}^{-1})_{11}\mathbf{A}^\top + \mathbf{W}_1 = 0. \tag{I.15c}$$

Substituting Eq. (I.13a)-Eq. (I.13b), we see that Eqs. (I.15a) to (I.15c) are (respectively) equivalent to the following constraints

$$\begin{bmatrix}\widetilde{\mathbf{A}}(\boldsymbol{\nu})^\top + \widetilde{\mathbf{A}}(\boldsymbol{\nu}) & \widetilde{\mathbf{B}}(\boldsymbol{\nu})\\ \widetilde{\mathbf{B}}(\boldsymbol{\nu})^\top & -\begin{bmatrix}\mathbf{W}_1^{-1} & 0\\ 0 & \mathbf{W}_2^{-1}\end{bmatrix}\end{bmatrix} \preceq 0, \tag{I.16a}$$

$$\text{where } \widetilde{\mathbf{A}}(\boldsymbol{\nu}) := \begin{pmatrix}\mathbf{A}\mathbf{M}_2 & \mathbf{A}\\ \mathbf{L}_1 & \mathbf{M}_1\mathbf{A} + \mathbf{L}_2\mathbf{C}\end{pmatrix}, \quad \widetilde{\mathbf{B}}(\boldsymbol{\nu}) := \begin{pmatrix}\mathbf{I} & 0\\ \mathbf{M}_1 & \mathbf{L}_2\end{pmatrix},$$

$$\widetilde{\mathbf{X}}(\boldsymbol{\nu}) \succ 0, \text{ where } \widetilde{\mathbf{X}}(\boldsymbol{\nu}) := \begin{pmatrix}\mathbf{M}_2 & \mathbf{I}\\ \mathbf{I} & \mathbf{M}_1\end{pmatrix} \tag{I.16b}$$

$$\mathbf{A}\mathbf{M}_2 + \mathbf{M}_2\mathbf{A}^\top + \mathbf{W}_1 = 0. \tag{I.16c}$$

Here, the equivalence between Eq. (I.15a) and Eq. (I.16a) invokes the following identity, derived similarly to the expression for $\mathbf{\Sigma}$ derived in part (a) above:

$$\mathbf{X} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{M}_2 & \mathbf{V} \\ \mathbf{I} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{M}_1 & \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{U} \\ \mathbf{U}^\top & -\mathbf{V}^{-1}\mathbf{M}_2\mathbf{U} \end{bmatrix}.$$

The equivalence between Eq. (I.15b) and Eq. (I.16b) can be verified via the Schur complement. It is clear that Eqs. (I.16a) to (I.16c) determine a convex constraint set. □

I.4 Proof of Lemma I.3

Fix $(\mathsf{K}, \mathbf{\Sigma}_\mathsf{K})$ for $\mathsf{K} \in \mathcal{K}_{\texttt{info}}$, and $\boldsymbol{\nu} = (\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{M}_1, \mathbf{M}_2)$ be the associated convex parameter, $\boldsymbol{\nu} = \Phi(\mathsf{K}, \mathbf{\Sigma}_\mathsf{K})$ defines the matrix $\mathbf{\Lambda}$ as in Eq. (I.16a)

$$\mathbf{\Lambda} := \begin{bmatrix} \widetilde{\mathbf{A}}^\top + \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{B}}^\top & -\begin{bmatrix} \mathbf{W}_1^{-1} & 0 \\ 0 & \mathbf{W}_2^{-1} \end{bmatrix} \end{bmatrix}, \quad \text{where } \widetilde{\mathbf{A}} := \begin{pmatrix} \mathbf{A}\mathbf{M}_2 & \mathbf{A} \\ \mathbf{L}_1 & \mathbf{M}_1\mathbf{A} + \mathbf{L}_2\mathbf{C} \end{pmatrix}, \quad \widetilde{\mathbf{B}} := \begin{pmatrix} \mathbf{I} & 0 \\ \mathbf{M}_1 & \mathbf{L}_2 \end{pmatrix}.$$

Since $(\mathsf{K}, \mathbf{\Sigma}_\mathsf{K}) \in \mathsf{dom}(f_{\texttt{lft}})$, $\boldsymbol{\nu} \in \mathsf{dom}(f_{\texttt{cvx}})$, and hence Eq. (I.16a) implies $\mathbf{\Lambda} \preceq 0$.

We begin our argument by bounding the operator norms of the matrices $\mathbf{L}_1$ and $\mathbf{L}_2$, which we ultimately translate into bounds on $\mathbf{A}_\mathsf{K}$ and $\mathbf{B}_\mathsf{K}$. Our arguments use the following Schur complement test for negative semidefinite matrices:

Lemma I.9. *Let* $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{12}^\top & \mathbf{X}_{22} \end{bmatrix}$ *satisfy* $\mathbf{X} \preceq 0$ *and* $\mathbf{X}_{22} \prec 0$. *Then,* $\|\mathbf{X}_{12}\|^2 / \|\mathbf{X}_{22}\| \le \|\mathbf{X}_{11}\|$.

Proof. Since $\mathbf{X} \preceq 0$, $-\mathbf{X} \succeq 0$. By the PSD Schur complement test applied to $-\mathbf{X}$,

$$0 \preceq -\mathbf{X}_{11} - (-\mathbf{X}_{12})(-\mathbf{X}_{22}^{-1})(-\mathbf{X}_{12})^\top = -\mathbf{X}_{11} + (\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{12}^\top).$$

Hence, $-\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{12}^\top \preceq -\mathbf{X}_{11}$. Now, observe that $\mathbf{X} \preceq 0$ implies that $-\mathbf{X}_{11}$, $-\mathbf{X}_{22}^{-1}$, $-(\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{12}^\top)$ are all PSD. Thus, $\|\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{12}^\top\| \le \|\mathbf{X}_{11}\|$, so that $\|\mathbf{X}_{12}\|^2 \sigma_{\min}(\mathbf{X}_{22}^{-1}) \le \|\mathbf{X}_{11}\|$ (where $\sigma_{\min}$ denotes minimal singular value). Noting $\sigma_{\min}(\mathbf{X}_{22}^{-1}) = 1/\|\mathbf{X}_{22}\|$ concludes. □

We begin bounding $\|\mathbf{L}_2\|$.

Claim I.10. *We have the bound*

$$\|\mathbf{L}_2\| \le 2\|\mathbf{C}\|\|\mathbf{W}_2^{-1}\| + \sqrt{2\|\mathbf{M}_1\|\|\mathbf{A}\|\|\mathbf{W}_2^{-1}\|}.$$

Proof. Let $\mathbf{\Lambda}_{(2,4)}$ denote the submatrix of $\mathbf{\Lambda}$ corresponding to the 2nd and 4th rows/columns:

$$\mathbf{\Lambda}_{(2,4)} := \begin{bmatrix} \mathbf{L}_2\mathbf{C} + (\mathbf{L}_2\mathbf{C})^\top + \mathbf{M}_1\mathbf{A} + (\mathbf{M}_1\mathbf{A})^\top & \mathbf{L}_2 \\ \mathbf{L}_2^\top & -\mathbf{W}_2^{-1} \end{bmatrix}.$$

Since $\mathbf{\Lambda} \preceq 0$, $\mathbf{\Lambda}_{(2,4)} \preceq 0$. Lemma I.9 gives

$$\|\mathbf{L}_2\|^2 / \|\mathbf{W}_2^{-1}\| \le 2\|\mathbf{L}_2\|\|\mathbf{C}\| + 2\|\mathbf{M}_1\|\|\mathbf{A}\|.$$

Hence, $x := \|\mathbf{L}_2\|^2$ satisfies a quadratic inequality $ax^2 - bx - c \le 0$, $a = 1/\|\mathbf{W}_2^{-1}\|$, $b = 2\|\mathbf{C}\|$ and $c = 2\|\mathbf{M}_1\|\|\mathbf{A}\|$. Solving the quadratic equation, using $a, b, c \ge 0$ and taking the positive root, and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$,

$$x \le \frac{b + \sqrt{b^2 + 4ac}}{2a} \le \frac{2b + 2\sqrt{ac}}{2a} \le \frac{b}{a} + \sqrt{c/a},$$

that is,

$$\|\mathbf{L}_2\| \le 2\|\mathbf{C}\|\|\mathbf{W}_2^{-1}\| + \sqrt{2\|\mathbf{M}_1\|\|\mathbf{A}\|\|\mathbf{W}_2^{-1}\|}.$$

□

Next, we bound $\|\mathbf{L}_1\|$ in terms of $\|\mathbf{L}_2\|$:

Claim I.11.

$$\|\mathbf{L}_1\| \leq 2\sqrt{\|\mathbf{A}\|\|\mathbf{M}_2\|\left(\|\|\mathbf{A}\|\|\mathbf{M}_1\| + \|\mathbf{C}\|\|\mathbf{L}_2\|\right)} + \|\mathbf{A}\|.$$

Proof. Observe that $\mathbf{\Lambda} \preceq 0$ implies $\widetilde{\mathbf{A}} + \widetilde{\mathbf{A}}^\top \preceq 0$. That is,

$$\begin{pmatrix} \mathbf{A}\mathbf{M}_2 + (\mathbf{A}\mathbf{M}_2)^\top & \mathbf{A} + \mathbf{L}_1^\top \\ \mathbf{L}_1 + \mathbf{A}^\top & \mathbf{W}_3 \end{pmatrix} \preceq 0, \quad \text{where } \mathbf{W}_3 := \mathbf{M}_1\mathbf{A} + (\mathbf{M}_1\mathbf{A})^\top + \mathbf{L}_2\mathbf{C} + (\mathbf{L}_2\mathbf{C})^\top.$$

Now, we know that $\mathbf{W}_3 \preceq 0$, but to invoke a Schur complement, we need strict inequality. To this end, for some $\lambda > 0$ to be choosen larger, we know that $\mathbf{W}_3 - \lambda\mathbf{I} \prec 0$, and

$$\begin{pmatrix} \mathbf{A}\mathbf{M}_2 + (\mathbf{A}\mathbf{M}_2)^\top & \mathbf{A} + \mathbf{L}_1^\top \\ \mathbf{L}_1 + \mathbf{A}^\top & \mathbf{W}_3 - \lambda\mathbf{I} \end{pmatrix} \preceq 0.$$

Using Lemma I.9

$$\mathbf{A}\mathbf{M}_2 + (\mathbf{A}\mathbf{M}_2)^\top - (\mathbf{A} + \mathbf{L}_1^\top)(\mathbf{W}_3 - \lambda\mathbf{I})^{-1}(\mathbf{L}_1 + \mathbf{A}^\top) \preceq 0,$$

$$(\mathbf{A} + \mathbf{L}_1^\top)^\top(\lambda\mathbf{I} - \mathbf{W}_3)^{-1}(\mathbf{L}_1 + \mathbf{A}^\top) \preceq -\big(\mathbf{A}\mathbf{M}_2 + (\mathbf{A}\mathbf{M}_2)^\top\big),$$

and hence

$$\frac{\|\mathbf{A} + \mathbf{L}_1^\top\|^2}{\|\mathbf{W}_3 - \lambda\mathbf{I}\|} \leq 2\|\mathbf{A}\|\|\mathbf{M}_2\|.$$

Since $\mathbf{W}_3 \preceq 0$, $\|\mathbf{W}_3 - \lambda\mathbf{I}\| = \lambda + \|\mathbf{W}\|$. Hence,

$$\|\mathbf{A} + \mathbf{L}_1^\top\|^2 \leq 2\|\mathbf{A}\|\|\mathbf{M}_2\| \leq (\lambda + \|\mathbf{W}_3\|) \cdot 2\|\mathbf{A}\|\|\mathbf{M}_2\|.$$

Since this is irrespective of $\lambda > 0$,

$$\begin{aligned} \|\mathbf{A} + \mathbf{L}_1^\top\|^2 &\leq 2\|\mathbf{A}\|\|\mathbf{M}_2\|\|\mathbf{W}_3\| \\ &\leq 4\|\mathbf{A}\|\|\mathbf{M}_2\|\left(\|\mathbf{A}\|\|\mathbf{M}_1\| + \|\mathbf{C}\|\|\mathbf{L}_2\|\right). \end{aligned}$$

Hence,

$$\|\mathbf{L}_1\| \leq 2\sqrt{\|\mathbf{A}\|\|\mathbf{M}_2\|\left(\|\mathbf{A}\|\|\mathbf{M}_1\| + \|\mathbf{C}\|\|\mathbf{L}_2\|\right)} + \|\mathbf{A}\|.$$

$\square$

Lastly, let us bound $\mathbf{L}_3$.

Claim I.12. *We have that* $\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}} \leq \sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})/\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|}$ *and* $\|\mathbf{L}_3\|_{\mathrm{F}} \leq \|\mathbf{\Sigma}_{\mathsf{K}}\|\sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})/\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|}$.

Proof. As follows from the proof of Lemma I.1,

$$\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) = \mathrm{tr}\left(\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix}\mathbf{\Sigma}_{\mathsf{K}}\begin{bmatrix}\mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top\end{bmatrix}\right) \geq \lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}})(\|\mathbf{G}\|_{\mathrm{F}}^2 + \|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}}^2) \geq \lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}})\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}}^2,$$

which gives the desired bound on $\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}}$. Since $\mathbf{L}_3 = \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top$, and since $\mathbf{V}$ is a submatrix of $\mathbf{\Sigma}_{\mathsf{K}}$,

$$\|\mathbf{L}_3\|_{\mathrm{F}} \leq \|\mathbf{V}\|\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}} \leq \|\mathbf{\Sigma}_{\mathsf{K}}\|\|\mathbf{C}_{\mathsf{K}}\|_{\mathrm{F}}.$$

The lemma follows. $\square$

Summarizing the previous three claims,

$$\begin{aligned} \|\mathbf{L}_2\| &\leq \|\mathbf{L}_2\| \leq 2\|\mathbf{C}\|\|\mathbf{W}_2^{-1}\| + \sqrt{2\|\mathbf{M}_1\|\|\mathbf{A}\|\|\mathbf{W}_2^{-1}\|} = \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1, \mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}) \\ \|\mathbf{L}_1\| &\leq 2\sqrt{\|\mathbf{A}\|\|\mathbf{M}_2\|\left(\|\|\mathbf{A}\|\|\mathbf{M}_1\| + \|\mathbf{C}\|\|\mathbf{L}_2\|\right)} + \|\mathbf{A}\| \\ &\quad = \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1, \mathbf{M}_2, \mathbf{A}, \mathbf{C}, \mathbf{L}_2) = \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1, \mathbf{M}_2, \mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}) \\ \|\mathbf{L}_3\|_{\mathrm{F}} &\leq \|\mathbf{\Sigma}_{\mathsf{K}}\|\sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})/\|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|} = \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{\Sigma}_{\mathsf{K}})\sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})}. \end{aligned}$$

This suffices to bound $\|\boldsymbol{\nu}\|_{\ell_2}$:

$$\begin{aligned}
\|\boldsymbol{\nu}\|_{\ell_2} &= \sqrt{\sum_{i=1}^{2}\left(\|\mathbf{M}_i\|_{\mathrm{F}}^2 + \|\mathbf{L}_i\|_{\mathrm{F}}^2\right) + \|\mathbf{L}_3\|_{\mathrm{F}}} \\
&\le \|\mathbf{L}_3\|_{\mathrm{F}} + \sum_{i=1}^{2}\left(\|\mathbf{M}_i\|_{\mathrm{F}} + \|\mathbf{L}_i\|_{\mathrm{F}}\right) \\
&\overset{(i)}{\le} \max\{n, \sqrt{nm}\}\left(\sum_{i=1}^{2}\left(\|\mathbf{M}_i\| + \|\mathbf{L}_i\|\right)\right) + \|\mathbf{L}_3\|_{\mathrm{F}} \\
&\overset{(ii)}{\le} \max\{n, \sqrt{nm}\} \cdot \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{M}_1, \mathbf{M}_2) + \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1})\sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})} \\
&\overset{(iii)}{=} \max\{n, \sqrt{nm}\} \cdot \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}) + \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}})\sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})} \\
&= \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1})(\max\{n, \sqrt{nm}\} + \sqrt{\mathcal{L}_{\mathtt{OE}}(\mathsf{K})}).
\end{aligned}$$

Above, (i) uses $\mathbf{M}_1, \mathbf{M}_2, \mathbf{L}_1 \in \mathbb{R}^{n\times n}$, and $\mathbf{L}_2 \in \mathbb{R}^{n\times m}$, (ii) uses the bounds on $\|\mathbf{L}_i\|$ developed above, and (iii) uses $\|\mathbf{M}_1\| = \|(\boldsymbol{\Sigma}_{\mathsf{K}}^{-1})_{11}\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|$ and similarly, $\|\mathbf{M}_2\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}\|$.

Next, we bound $\|\mathbf{A}_{\mathsf{K}}\|$ and $\|\mathbf{B}_{\mathsf{K}}\|$. From the definition of the transformation Φ, and recalling $\mathbf{U}_{\mathsf{K}} = (\boldsymbol{\Sigma}_{\mathsf{K}}^{-1})_{12}$ and $\mathbf{V}_{\mathsf{K}} = (\boldsymbol{\Sigma}_{\mathsf{K}})_{12}$, we have

$$\mathbf{L}_1 = \mathbf{U}_{\mathsf{K}}(\mathbf{A}_{\mathsf{K}}\mathbf{V}_{\mathsf{K}}^\top + \mathbf{B}_{\mathsf{K}}\mathbf{C}(\boldsymbol{\Sigma})_{11}) + \underbrace{(\boldsymbol{\Sigma}^{-1})_{11}}_{=\mathbf{M}_1}\mathbf{A}\underbrace{(\boldsymbol{\Sigma})_{11}}_{=\|\mathbf{M}_2\|}, \quad \mathbf{L}_2 = \mathbf{U}_{\mathsf{K}}\mathbf{B}_{\mathsf{K}}. \tag{I.17}$$

Hence, if $\mathbf{U}_{\mathsf{K}}$ and $\mathbf{V}_{\mathsf{K}}$ are invertible,

$$\begin{aligned}
\|\mathbf{B}_{\mathsf{K}}\| &\le \|\mathbf{U}_{\mathsf{K}}^{-1}\|\|\mathbf{L}_2\| = \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1, \mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{U}_{\mathsf{K}}^{-1}), \\
\|\mathbf{A}_{\mathsf{K}}\| &\le \|\mathbf{V}_{\mathsf{K}}^{-1}\|\|\mathbf{U}_{\mathsf{K}}^{-1}\|\left(\|\mathbf{M}_1\|\|\mathbf{A}\|\|\mathbf{M}_2\| + \|\mathbf{L}_1\|\right) + \|\mathbf{B}_{\mathsf{K}}\|\|\mathbf{C}\|\|\mathbf{V}_{\mathsf{K}}^{-1}\|\|\mathbf{U}_{\mathsf{K}}\| \\
&= \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1, \mathbf{M}_2, \mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \mathbf{U}_{\mathsf{K}}, \mathbf{V}_{\mathsf{K}}^{-1}, \mathbf{U}_{\mathsf{K}}^{-1})
\end{aligned}$$

Again, we note that $\|\mathbf{M}_1\| = \|(\boldsymbol{\Sigma}_{\mathsf{K}}^{-1})_{11}\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|$ and similarly, $\|\mathbf{M}_2\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}\|$. Similarly, $\|\mathbf{U}_{\mathsf{K}}\| = \|(\boldsymbol{\Sigma}_{\mathsf{K}}^{-1})_{12}\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|$, hence, we conclude

$$\max\left\{\|\mathbf{A}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|\right\} \le \mathrm{poly}_{\mathrm{op}}(\mathbf{A}, \mathbf{C}, \mathbf{W}_2^{-1}, \boldsymbol{\Sigma}_{\mathsf{K}}, \boldsymbol{\Sigma}_{\mathsf{K}}^{-1}, \mathbf{U}_{\mathsf{K}}^{-1}, \mathbf{V}_{\mathsf{K}}^{-1}), \tag{I.18}$$

as needed. □

I.5 Proof of Lemma I.4

We establish the differentiability and conditioning of Φ for any $(\mathsf{K}, \boldsymbol{\Sigma}) \in \mathcal{C}_{\mathtt{lft}}$; the lemma corresponds to the special case when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathsf{K}}$. We let $\boldsymbol{\nu} = \Phi(\mathsf{K}, \boldsymbol{\Sigma}_{\mathsf{K}})$, where we recall $\boldsymbol{\nu} = (\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{M}_1, \mathbf{M}_2)$ is given by

$$\begin{pmatrix}\mathbf{L}_1\\ \mathbf{L}_2\\ \mathbf{L}_3\\ \mathbf{M}_1\\ \mathbf{M}_2\end{pmatrix} := \begin{pmatrix}\mathbf{U}(\mathbf{A}_{\mathsf{K}}\mathbf{V}^\top + \mathbf{B}_{\mathsf{K}}\mathbf{C}(\boldsymbol{\Sigma})_{11}) + (\boldsymbol{\Sigma}^{-1})_{11}\mathbf{A}(\boldsymbol{\Sigma})_{11}\\ \mathbf{U}\mathbf{B}_{\mathsf{K}}\\ \mathbf{C}_{\mathsf{K}}\mathbf{V}^\top\\ (\boldsymbol{\Sigma}^{-1})_{11}\\ (\boldsymbol{\Sigma})_{11}\end{pmatrix}, \tag{I.19a}$$

$$\text{where } \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} := \begin{pmatrix}(\boldsymbol{\Sigma}^{-1})_{12}\\ (\boldsymbol{\Sigma})_{12}\end{pmatrix}. \tag{I.19b}$$

To see that Φ is differentiable, we see that Phi is a polynomial function in $\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}$ and $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$, and is therefore differentiable in an open neighborhood of any $(\mathsf{K}, \boldsymbol{\Sigma})$ for which $\boldsymbol{\Sigma}$ is invertible.

Let's turn to the condition of $\nabla\Phi$. We then fix a target perturbation $\boldsymbol{\Delta}_{\mathtt{cvx}} := (\boldsymbol{\Delta}_{\mathbf{L}_1}, \boldsymbol{\Delta}_{\mathbf{L}_2}, \boldsymbol{\Delta}_{\mathbf{L}_3}, \boldsymbol{\Delta}_{\mathbf{M}_1}, \boldsymbol{\Delta}_{\mathbf{M}_2})$ such that its ℓ_2-norm as an Euclidean vector (equivalently, the sum of Frobenius norms of its parameters) is

$$\|\boldsymbol{\Delta}_{\mathtt{cvx}}\|_{\ell_2}^2 = \sum_{i=1}^{3} \|\boldsymbol{\Delta}_{\mathbf{L}_i}\|_{\mathrm{F}}^2 + \sum_{j=1}^{2} \|\boldsymbol{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}}^2 = 1.$$

Our strategy is to compute a perturbation $\boldsymbol{\Delta}_{\mathtt{lft}} = (\boldsymbol{\Delta}_{\mathbf{A}}, \boldsymbol{\Delta}_{\mathbf{B}}, \boldsymbol{\Delta}_{\mathbf{C}}, \boldsymbol{\Delta}_{\boldsymbol{\Sigma}})$ of the parameters $(\mathsf{K}, \boldsymbol{\Sigma})$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi((\mathsf{K}, \boldsymbol{\Sigma}) + t\boldsymbol{\Delta}_{\mathtt{lft}})\big|_{t=0} = \boldsymbol{\Delta}_{\mathtt{cvx}}. \tag{I.20}$$

Noting the identity

$$\nabla\Phi(\boldsymbol{y}) \cdot \boldsymbol{\Delta}_{\mathtt{lft}} = \boldsymbol{\Delta}_{\mathtt{cvx}},$$

it thus suffices to compute uniform upper bound on $\|\boldsymbol{\Delta}_{\mathtt{lft}}\|_{\ell_2}^2 = \|\boldsymbol{\Delta}_{\mathbf{A}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_{\mathbf{B}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_{\mathbf{C}}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\|_{\mathrm{F}}^2$ for which Eq. (I.20) holds. For convenience, let $j \in \{1, 2\}$ (resp $i \in \{1, 2, 3\}$) $\Phi_{\mathbf{M}_j}$ (resp $\Phi_{\mathbf{L}_i}$) denote the restriction of Φ's image to the $\mathbf{M}_j$ (resp. $\mathbf{L}_i$) coordinate.

Handling the M_j-blocks. We proceed to choose $\boldsymbol{\Delta}_{\mathtt{lft}}$ by first ensuring $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{M}_j}((\mathsf{K}, \boldsymbol{\Sigma}) + t\boldsymbol{\Delta}_{\mathtt{lft}})\big|_{t=0} = \boldsymbol{\Delta}_{\mathbf{M}_j}$ for $j \in \{1, 2\}$, and then continue to show the same for the $\mathbf{L}_i$-coordinates. Since $\Phi_{\mathbf{M}_j}$ are functions of $\boldsymbol{\Sigma}$, it suffices for now to choose perturbations of $\boldsymbol{\Sigma}$; abusing notation, we shall simply write $\Phi_{\mathbf{M}_j}(\boldsymbol{\Sigma})$ to express this fact. We consider a perturbation of the form

$$\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}, \quad \text{where } \boldsymbol{\Delta}_{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Delta}_{11} & \boldsymbol{\Delta}_{12} \\ \boldsymbol{\Delta}_{12}^\top & 0 \end{bmatrix}. \tag{I.21}$$

Since $\Phi_{\mathbf{M}_2}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}_{11}$, we have $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{M}_2}(\boldsymbol{\Sigma} + t\boldsymbol{\Delta}_{\boldsymbol{\Sigma}})\big|_{t=0} = \boldsymbol{\Delta}_{11}$, so it suffices to choose

$$\boldsymbol{\Delta}_{11} = \boldsymbol{\Delta}_{\mathbf{M}_2}. \tag{I.22}$$

Next, we consider the $\mathbf{M}_1$-block. For convenience, we define the curve $\bar{\boldsymbol{\Sigma}}(t) = \boldsymbol{\Sigma} + t\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}$. Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{M}_1}(\boldsymbol{\Sigma} + t\boldsymbol{\Delta}_{\boldsymbol{\Sigma}})\big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{M}_1}(\bar{\boldsymbol{\Sigma}}(t))\big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}(\bar{\boldsymbol{\Sigma}}(t)^{-1})_{11}\big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}(\bar{\boldsymbol{\Sigma}}_{11} - \bar{\boldsymbol{\Sigma}}_{12}\bar{\boldsymbol{\Sigma}}_{22}^{-1}\bar{\boldsymbol{\Sigma}}_{12}^\top)^{-1}\big|_{t=0} \\
&= \underbrace{(\bar{\boldsymbol{\Sigma}}_{11} - \bar{\boldsymbol{\Sigma}}_{12}\bar{\boldsymbol{\Sigma}}_{22}^{-1}\bar{\boldsymbol{\Sigma}}_{12}^\top)^{-1}\big|_{t=0}}_{=\mathbf{M}_1} \cdot \left(\frac{\mathrm{d}}{\mathrm{d}t}(\bar{\boldsymbol{\Sigma}}_{11} - \bar{\boldsymbol{\Sigma}}_{12}\bar{\boldsymbol{\Sigma}}_{22}^{-1}\bar{\boldsymbol{\Sigma}}_{12}^\top)\big|_{t=0}\right) \cdot \underbrace{\dots}_{=\mathbf{M}_1} \\
&= \mathbf{M}_1\left(\boldsymbol{\Delta}_{11} - \boldsymbol{\Delta}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12} - (\boldsymbol{\Delta}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12})^\top\right)\mathbf{M}_1.
\end{aligned}$$

Hence, we can take

$$\begin{aligned}
\boldsymbol{\Delta}_{12} &= \frac{1}{2}\left(\boldsymbol{\Delta}_{\mathbf{M}_2} - \mathbf{M}_1^{-1}\boldsymbol{\Delta}_{\mathbf{M}_1}\boldsymbol{\Delta}_{11}^{-1}\right)\boldsymbol{\Sigma}_{12}^{-1}\boldsymbol{\Sigma}_{22} \\
&= \frac{1}{2}\left(\boldsymbol{\Delta}_{\mathbf{M}_2} - \mathbf{M}_1^{-1}\boldsymbol{\Delta}_{11}\boldsymbol{\Delta}_{11}^{-1}\right)\mathbf{V}^{-1}\boldsymbol{\Sigma}_{22} \qquad \text{(I.23)}
\end{aligned}$$

Some directional derivatives. To handle the $\mathbf{L}_i$ blocks, we extend the "bar" notation to the variables $\bar{\mathbf{M}}_1(t), \bar{\mathbf{M}}_2(t), \bar{\mathbf{U}}(t), \mathbf{V}(t)$ to denote the matrices corresponding to $\bar{\boldsymbol{\Sigma}}(t) = \boldsymbol{\Sigma} + t\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}$, i.e.

$$\bar{\mathbf{M}}_1(t) = (\bar{\boldsymbol{\Sigma}}^{-1}(t))_{11}, \quad \bar{\mathbf{M}}_2(t) = \bar{\boldsymbol{\Sigma}}_{11}(t), \quad \bar{\mathbf{V}} = \bar{\boldsymbol{\Sigma}}_{12}(t), \quad \bar{\mathbf{U}} = (\bar{\boldsymbol{\Sigma}}(t)^{-1})_{12}.$$

Since $\bar{\boldsymbol{\Sigma}}(0) = \boldsymbol{\Sigma}$, the above matrices are evaluated to their "non-barred" counterparts when $t = 0$. Moreover, by choice of $\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}$, we have

$$\bar{\mathbf{M}}_1'(0) = \boldsymbol{\Delta}_{\mathbf{M}_1}, \quad \bar{\mathbf{M}}_2'(0) = \boldsymbol{\Delta}_{\mathbf{M}_2}, \quad \bar{\mathbf{V}}'(0) = \boldsymbol{\Delta}_{12}.$$

Using the block matrix inversion formula, we have

$$\bar{\mathbf{U}} = (\bar{\boldsymbol{\Sigma}})_{12}^{-1} = -(\bar{\boldsymbol{\Sigma}}^{-1})_{11}\bar{\boldsymbol{\Sigma}}_{12}\bar{\boldsymbol{\Sigma}}_{22}^{-1} = -\bar{\mathbf{M}}_1\bar{\boldsymbol{\Sigma}}_{12}\boldsymbol{\Sigma}_{22}^{-1},$$

where above we use $\bar{\mathbf{M}}_1 = (\bar{\boldsymbol{\Sigma}}^{-1})_{11}$ and $\bar{\boldsymbol{\Sigma}}_{22}^{-1}(t) = \boldsymbol{\Sigma}_{22}^{-1}$ is constant for all t. Therefore,

$$\begin{aligned}
\bar{\mathbf{U}}'(0) &= -\bar{\mathbf{M}}_1'(0)\bar{\boldsymbol{\Sigma}}_{12}(0)\boldsymbol{\Sigma}_{22}^{-1} - \bar{\mathbf{M}}_1(0)\bar{\boldsymbol{\Sigma}}_{12}'(0)\boldsymbol{\Sigma}_{22}^{-1} \\
&= -\boldsymbol{\Delta}_{\mathbf{M}_1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} - \mathbf{M}_1\boldsymbol{\Delta}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\
&= -\boldsymbol{\Delta}_{\mathbf{M}_1}\mathbf{V}(\boldsymbol{\Sigma}_{22})^{-1} - \mathbf{M}_1\boldsymbol{\Delta}_{12}\boldsymbol{\Sigma}_{22}^{-1}. \qquad \text{(I.24)}
\end{aligned}$$

Handling the $\mathbf{L}_i$-blocks. Let us also define $\bar{\mathbf{B}}_{\mathsf{K}}(t) = \mathbf{B} + t\boldsymbol{\Delta}_{\mathbf{B}}$ and $\bar{\mathbf{A}}_{\mathsf{K}}(t) = \mathbf{A} + t\boldsymbol{\Delta}_{\mathbf{A}}$. Using the "bar"-notation, we can compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{L}_2}(\boldsymbol{\nu} + t\boldsymbol{\Delta}_{\mathtt{cvx}})\big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}(\bar{\mathbf{B}}_{\mathsf{K}}\bar{\mathbf{U}})\big|_{t=0} = \bar{\mathbf{B}}_{\mathsf{K}}'(0)\mathbf{U} + \mathbf{B}_{\mathsf{K}}\bar{\mathbf{U}}'(0) = \boldsymbol{\Delta}_{\mathbf{B}}\mathbf{U} + \mathbf{B}_{\mathsf{K}}\bar{\mathbf{U}}'(0).$$

Hence, we set

$$\boldsymbol{\Delta}_{\mathbf{B}} = (\boldsymbol{\Delta}_{\mathbf{L}_2} - \mathbf{B}_{\mathsf{K}}\bar{\mathbf{U}}'(0))\mathbf{U}^{-1} \tag{I.25}$$

Similarly, we can select

$$\boldsymbol{\Delta}_{\mathbf{C}} = \mathbf{V}^{-\top}(\boldsymbol{\Delta}_{\mathbf{L}_3} - \mathbf{C}_{\mathsf{K}}(\bar{\mathbf{V}}'(0))^{\top}). \tag{I.26}$$

Finally, we compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{\mathbf{L}_1}(\boldsymbol{\nu} + t\boldsymbol{\Delta}_{\mathtt{cvx}})\big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\bar{\mathbf{U}}(\bar{\mathbf{A}}_{\mathsf{K}}\bar{\mathbf{V}}^{\top} + \bar{\mathbf{B}}_{\mathsf{K}}\mathbf{C}\bar{\mathbf{M}}_2) + \bar{\mathbf{M}}_1\mathbf{A}\bar{\mathbf{M}}_2\right)\big|_{t=0} \\
&= \mathbf{U}\bar{\mathbf{A}}_{\mathsf{K}}'(0)\mathbf{V} + \bar{\mathbf{U}}'(0)(\mathbf{A}_{\mathsf{K}}\mathbf{V} + \mathbf{K}_{\mathsf{K}}\mathbf{C}\mathbf{M}_2) + \mathbf{U}\left(\mathbf{A}\bar{\mathbf{V}}'(0) + \bar{\mathbf{B}}_{\mathsf{K}}'(0)\mathbf{C}\mathbf{M}_2 + \mathbf{B}_{\mathsf{K}}\mathbf{C}\bar{\mathbf{M}}_2'(0)\right) \\
&\quad + \bar{\mathbf{M}}_1'(0)\mathbf{A}\mathbf{M}_2 + \mathbf{M}_1(0)\mathbf{A}\bar{\mathbf{M}}_2'(0) \\
&= \mathbf{U}\boldsymbol{\Delta}_{\mathbf{A}}\mathbf{V} + \bar{\mathbf{U}}'(0)(\mathbf{A}_{\mathsf{K}}\mathbf{V} + \mathbf{B}_{\mathsf{K}}\mathbf{C}\mathbf{M}_2) + \mathbf{U}\left(\mathbf{A}\boldsymbol{\Delta}_{12} + \boldsymbol{\Delta}_{\mathbf{B}}\mathbf{C}\mathbf{M}_2 + \mathbf{B}_{\mathsf{K}}\mathbf{C}\boldsymbol{\Delta}_{\mathbf{M}_2}\right) \\
&\quad + \boldsymbol{\Delta}_{\mathbf{M}_1}\mathbf{A}\mathbf{M}_2 + \mathbf{M}_1\mathbf{A}\boldsymbol{\Delta}_{\mathbf{M}_2}.
\end{aligned}$$

Hence, we select

$$\begin{aligned}
\boldsymbol{\Delta}_{\mathbf{A}} &= \mathbf{U}^{-1}\boldsymbol{\Delta}_{\mathbf{L}_1}\mathbf{V}^{-1} - \mathbf{U}^{-1}\bar{\mathbf{U}}'(0)\left(\mathbf{A}_{\mathsf{K}} + \mathbf{B}_{\mathsf{K}}\mathbf{C}\mathbf{M}_2\right)\mathbf{V}^{-1} \\
&\quad - \left(\mathbf{A}\boldsymbol{\Delta}_{12} + \boldsymbol{\Delta}_{\mathbf{B}}\mathbf{C}\mathbf{M}_2 + \mathbf{B}_{\mathsf{K}}\mathbf{C}\boldsymbol{\Delta}_{\mathbf{M}_2}\right)\mathbf{V}^{-1} - \mathbf{U}^{-1}\left(\boldsymbol{\Delta}_{\mathbf{M}_1}\mathbf{A}\mathbf{M}_2 + \mathbf{M}_1\mathbf{A}\boldsymbol{\Delta}_{\mathbf{M}_2}\right)\mathbf{V}^{-1}.
\end{aligned} \tag{I.27}$$

Bounding the norm of $\boldsymbol{\Delta}_{\mathtt{cvx}}$. We begin with some useful bounds:

$$\max\{\|\mathbf{M}_1\|, \|\mathbf{U}\|, \|(\boldsymbol{\Sigma}_{22})^{-1}\|, \|\mathbf{M}_2^{-1}\|\} \le \|\boldsymbol{\Sigma}^{-1}\| \tag{I.28a}$$

$$\max\{\|\boldsymbol{\Sigma}_{22}\|, \|\mathbf{M}_1^{-1}\|, \|\mathbf{M}_2\|, \|\mathbf{V}\|\} \le \|\boldsymbol{\Sigma}\|. \tag{I.28b}$$

Eq. (I.28a) uses the fact that $\mathbf{M}_1$ and $\mathbf{U}$ are submatrices of $\boldsymbol{\Sigma}^{-1}$, and the fact that for any positive-definite matrix, $\|\boldsymbol{\Sigma}_{11}^{-1}\|$ (which is just $\|\mathbf{M}_2^{-1}\|$) and $\|\boldsymbol{\Sigma}_{22}^{-1}\|$ are both at most $\|\boldsymbol{\Sigma}^{-1}\|$ (as can be verified by the block-matrix inverse formula). Finally, Eq. (I.28b) follows from similar reasoning.

Notational aside. In what follows, we apply our $\mathrm{poly}_{\mathrm{op}}(\cdot)$ notation, which denotes a universal polynomial in the operator norms of its matrix arguments, and in the values of its scalar arguments. We let $\mathrm{poly}_{\mathrm{op}}(\cdot)$ include universal constant terms (e.g. $1 + \|\mathbf{X}\|$ is $\mathrm{poly}_{\mathrm{op}}(\mathbf{X})$).

From Eq. (I.23), we can bound

$$\|\boldsymbol{\Delta}_{12}\|_{\mathrm{F}} = \mathrm{poly}_{\mathrm{op}}(\mathbf{M}_1^{-1}, \boldsymbol{\Sigma}_{22}, \mathbf{V}^{-1}) \cdot (\|\boldsymbol{\Delta}_{\mathbf{M}_1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{\mathbf{M}_2}\|_{\mathrm{F}}).$$

In light of Eq. (I.28b),

$$\|\boldsymbol{\Delta}_{12}\|_{\mathrm{F}} = \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}, \mathbf{V}^{-1}) \cdot (\|\boldsymbol{\Delta}_{\mathbf{M}_1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{\mathbf{M}_2}\|_{\mathrm{F}}), \tag{I.29}$$

which means that

$$\|\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\|_{\mathrm{F}} \overset{(i)}{=} \sqrt{\|\boldsymbol{\Delta}_{\mathbf{M}_2}\|_{\mathrm{F}}^2 + 2\|\boldsymbol{\Delta}_{12}\|_{\mathrm{F}}^2} \le \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}, \mathbf{V}^{-1}) \cdot (\|\boldsymbol{\Delta}_{\mathbf{M}_1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{\mathbf{M}_2}\|_{\mathrm{F}}),$$

where (i) uses Eq. (I.21) and Eq. (I.22), and the second inequality calls Eq. (I.29).

Next, from Eq. (I.24) and Eqs. (I.28a) and (I.28b),

$$\begin{aligned}
\|\bar{\mathbf{U}}'(0)\|_{\mathrm{F}} &\le \|\boldsymbol{\Delta}_{\mathbf{M}_1}\|_{\mathrm{F}}\|\mathbf{V}\|\|\boldsymbol{\Sigma}_{22}^{-1}\| - \|\mathbf{M}_1\|\|\boldsymbol{\Delta}_{12}\|\|(\boldsymbol{\Sigma}_{22})^{-1}\| \\
&= \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1})\left(\|\boldsymbol{\Delta}_{\mathbf{M}_1}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{12}\|_{\mathrm{F}}\right) \\
&= \mathrm{poly}_{\mathrm{op}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1}, \mathbf{V}^{-1})\left(\sum_{j=1}^{2}\|\boldsymbol{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}}\right).
\end{aligned}$$

Continuing,

$$\begin{aligned}\|\mathbf{\Delta}_{\mathbf{B}}\|_{\mathrm{F}} &= \|\mathbf{U}^{-1}\| \left(\|\mathbf{\Delta}_{\mathbf{L}_2}\|_{\mathrm{F}} - \|\mathbf{B}_{\mathsf{K}}\|\|\bar{\mathbf{U}}'(0)\|_{\mathrm{F}}\right) \\ &\le \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}^{-1}, \mathbf{\Sigma}, \mathbf{V}^{-1}, \mathbf{U}^{-1}, \mathbf{B}_{\mathsf{K}}) \left(\sum_{j=1}^{2} \|\mathbf{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{L}_2}\|_{\mathrm{F}}\right),\end{aligned}$$

and similarly, since $\bar{\mathbf{V}}'(0) = \mathbf{\Delta}_{12}$ bounded as in Eq. (I.29),

$$\begin{aligned}\|\mathbf{\Delta}_{\mathbf{C}}\|_{\mathrm{F}} &= \|\mathbf{V}^{-1}\| \left(\|\mathbf{\Delta}_{\mathbf{L}_3}\|_{\mathrm{F}} - \|\mathbf{C}_{\mathsf{K}}\|\|\bar{\mathbf{V}}'(0)\|_{\mathrm{F}}\right) \\ &\le \mathrm{poly}_{\mathrm{op}}(\mathbf{\Sigma}^{-1}, \mathbf{\Sigma}, \mathbf{V}^{-1}, \mathbf{U}^{-1}, \mathbf{C}_{\mathsf{K}}) \left(\sum_{j=1}^{2} \|\mathbf{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{L}_3}\|_{\mathrm{F}}\right).\end{aligned}$$

Finally,

$$\begin{aligned}\|\mathbf{\Delta}_{\mathbf{A}}\|_{\mathrm{F}} &= \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{U}, \mathbf{V}, \mathbf{M}_1, \mathbf{M}_2, \mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{U}^{-1}, \mathbf{V}^{-1}\right) \\ &\quad \times \left(\|\bar{\mathbf{U}}'(0)\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{L}_1}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{B}}\|_{\mathrm{F}} + \sum_{j=1}^{2} \|\mathbf{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}}\right) \\ &= \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{\Sigma}, \mathbf{\Sigma}^{-1}, \mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{U}^{-1}, \mathbf{V}^{-1}\right) \left(\sum_{j=1}^{2} \|\mathbf{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{L}_1}\|_{\mathrm{F}}\right).\end{aligned}$$

In sum,

$$\begin{aligned}\|\mathbf{\Delta}_{\mathtt{lft}}\|_{\ell_2}^2 &= \|\mathbf{\Delta}_{\mathbf{A}}\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_{\mathbf{C}}\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_{\mathbf{C}}\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_{\mathbf{\Sigma}}\|_{\mathrm{F}}^2 \\ &= \mathrm{poly}_{\mathrm{op}}\left(\mathbf{A}, \mathbf{C}, \mathbf{\Sigma}, \mathbf{\Sigma}^{-1}, \mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}, \mathbf{C}_{\mathsf{K}}, \mathbf{U}^{-1}, \mathbf{V}^{-1}\right) \cdot \left(\sum_{j=1}^{2} \|\mathbf{\Delta}_{\mathbf{M}_j}\|_{\mathrm{F}}^2 + \sum_{i=1}^{3} \|\mathbf{\Delta}_{\mathbf{L}_i}\|_{\mathrm{F}}^2\right).\end{aligned}$$

The bound follows. □

I.6 Proof of Lemma I.5

Let $\mathbf{Z} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^{\top}$. Then,

$$\begin{aligned}\|\mathbf{Z}^{-1}\| = \|\mathbf{\Sigma}_{12}^{-\top}\mathbf{\Sigma}_{22}\mathbf{\Sigma}_{12}^{-1}\| &\ge \|\mathbf{\Sigma}_{12}^{-1}\|^2 \lambda_{\min}(\mathbf{\Sigma}_{22}) \\ &\ge \|\mathbf{\Sigma}_{12}^{-1}\|^2 \lambda_{\min}(\mathbf{\Sigma}_{22}) \\ &\ge \|\mathbf{\Sigma}_{12}^{-1}\|^2 \lambda_{\min}(\mathbf{\Sigma}).\end{aligned}$$

Hence,

$$\|\mathbf{V}^{-1}\| = \|\mathbf{\Sigma}_{12}^{-1}\| \le \sqrt{\frac{\|\mathbf{Z}^{-1}\|}{\lambda_{\min}(\mathbf{\Sigma})}} = \sqrt{\|\mathbf{Z}^{-1}\|\|\mathbf{\Sigma}^{-1}\|}.$$

Next, from the block matrix inversion identity, we have

$$\mathbf{U} := (\mathbf{\Sigma}^{-1})_{12} = -(\mathbf{\Sigma}^{-1})_{11}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} = -(\mathbf{\Sigma}^{-1})_{11}\mathbf{V}\mathbf{\Sigma}_{22}^{-1}.$$

Hence,

$$\begin{aligned}\|\mathbf{U}^{-1}\| = \|\mathbf{\Sigma}_{22}\mathbf{V}^{-1}(\mathbf{\Sigma}^{-1})_{11}^{-1}\| &\le \frac{\|\mathbf{\Sigma}_{22}\|}{\lambda_{\min}((\mathbf{\Sigma}^{-1})_{11})}\|\mathbf{V}^{-1}\| \\ &\le \frac{\|\mathbf{\Sigma}\|}{\lambda_{\min}(\mathbf{\Sigma}^{-1})}\|\mathbf{V}^{-1}\| = \|\mathbf{\Sigma}\|\|\mathbf{\Sigma}^{-1}\|\|\mathbf{V}^{-1}\| \\ &\le \|\mathbf{\Sigma}\|\sqrt{\|\mathbf{\Sigma}^{-1}\|^3\|\|\mathbf{Z}^{-1}\|}.\end{aligned}$$

The conclusion invokes Lemma I.4. □

I.7 Proof of convexity lemmas

Here we prove Lemmas I.6 and I.7, restated below for convenience. Both proofs use the fact that convexity is preserved under partial minimization.

Fact I.1 (Chapter 3.2.5 of Boyd et al. [2004])**.** Let $\widetilde{\phi}(\boldsymbol{x},\boldsymbol{y})$ be a convex function in two arguments such that $\phi(\boldsymbol{x}) := \min_{\boldsymbol{y}} \widetilde{\phi}(\boldsymbol{x},\boldsymbol{y})$ is finite and attained for each $\boldsymbol{x}$. Then $\phi(\boldsymbol{x})$ is convex.

Lemma I.6. *The function* $g(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}}) = \mathrm{tr}[\widetilde{\mathbf{C}}^\top\widetilde{\mathbf{X}}^{-1}\widetilde{\mathbf{C}}]$ *is convex on the domain* $(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}}) \in \mathbb{R}^{\widetilde{p}\times\widetilde{n}} \times \mathbb{S}^{\widetilde{n}}_{++}$.

Proof. Observe that we can express

$$g(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}}) = \min_{\mathbf{E}\in\mathbb{S}^n} \widetilde{g}(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}},\mathbf{E}), \quad \widetilde{g}(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}},\mathbf{E}) = \left(\mathrm{tr}(\mathbf{E})\cdot\mathbb{I}_\infty\left\{\mathbf{E}\succeq 0, \quad \begin{bmatrix}\mathbf{E} & \widetilde{\mathbf{C}}\\ \widetilde{\mathbf{C}}^\top & \widetilde{\mathbf{X}}\end{bmatrix}\succeq 0\right\}\right). \tag{I.30}$$

Indeed, since $\widetilde{\mathbf{X}}\succ 0$ on the domain of g, the Schur complement test implies that $\begin{bmatrix}\mathbf{E} & \widetilde{\mathbf{C}}^\top\\ \widetilde{\mathbf{C}} & \widetilde{\mathbf{X}}\end{bmatrix}\succeq 0$ if and only if $\mathbf{E}\succeq\widetilde{\mathbf{C}}\widetilde{\mathbf{X}}^{-1}\widetilde{\mathbf{C}}^\top$. Hence, the minimal value of $\mathrm{tr}(\mathbf{E})$ is attained with $\mathbf{E} = \widetilde{\mathbf{C}}\widetilde{\mathbf{X}}^{-1}\widetilde{\mathbf{C}}^\top$. Observing that $\widetilde{g}(\widetilde{\mathbf{C}},\widetilde{\mathbf{X}},\mathbf{E})$ is convex (affine function with a convex constraint), Fact I.1 implies that its partial minimization g is convex. $\square$

Lemma I.7. *The function* $h(\mathbf{M}_1,\mathbf{M}_2) = \mathrm{tr}[(\mathbf{M}_2-\mathbf{M}_1^{-1})^{-1}]$ *is convex on the domain* $\{(\mathbf{M}_1,\mathbf{M}_2)\in\mathbb{S}^n_{++}\times\mathbb{S}^n_{++} : \mathbf{M}_2\succ\mathbf{M}_1^{-1}\}$.

Proof. Introduce the function

$$\widetilde{h}(\mathbf{M}_1,\mathbf{M}_2,\mathbf{E}) = \mathrm{tr}[\mathbf{E}^{-1}]\cdot\mathbb{I}_\infty\left\{\begin{bmatrix}\mathbf{M}_2-\mathbf{E} & \mathbf{I}_n\\ \mathbf{I}_n & \mathbf{M}_1\end{bmatrix}\succeq 0, \quad \mathbf{E}\succ 0, \quad \mathbf{M}_1\succ 0\right\}.$$

Since the function $\mathbf{E}\mapsto\mathrm{tr}[\mathbf{E}^{-1}]$ is convex for $\mathbf{E}\succ 0$, the function $\widetilde{h}$ is also convex. The constraint in $\widetilde{h}$ is equivalent to $\mathbf{M}_1\succ 0$, $\mathbf{E}\succ 0$, and $\mathbf{M}_2-\mathbf{E}-\mathbf{M}_1^{-1}\succeq 0$. Rearranging that is $\mathbf{E}\preceq\mathbf{M}_2-\mathbf{M}_1^{-1}$, or equivalently, $\mathbf{E}^{-1}\succeq(\mathbf{M}_2-\mathbf{M}_1^{-1})^{-1}$. Hence, $\widetilde{h}$ can be written as

$$\widetilde{h}(\mathbf{M}_1,\mathbf{M}_2,\mathbf{E}) = \mathrm{tr}[\mathbf{E}^{-1}]\cdot\mathbb{I}_\infty\left\{\mathbf{E}^{-1}\succeq(\mathbf{M}_2-\mathbf{M}_1^{-1})^{-1}, \quad \mathbf{E}\succ 0, \quad \mathbf{M}_1\succ 0\right\}.$$

From the above form, it is clear that $\min_{\mathbf{E}}\widetilde{h}(\mathbf{M}_1,\mathbf{M}_2,\mathbf{E}) = \widetilde{h}(\mathbf{M}_1,\mathbf{M}_2)$, which is finite and attained by $\mathbf{E}=\mathbf{M}_2-\mathbf{M}_1^{-1}$ on the domain of h. $\square$

J Bounds on Solutions to Closed-Loop Lyapunov Equations (Proposition 4.3)

The following proposition gives a more granular statement of Proposition 4.3 in the main text.

Proposition J.1. *Let* $\|\cdot\|_\circ$ *denote either the operator, Frobenius, or nuclear norm, and let* clyap *denote the integral in Eq. (J.1), which corresponds to the continuous Lyapunov opertor when its argument is Hurwitz. Then, for any matrix* $\mathbf{Y}\in\mathbb{S}^{2n}$*,*

$$\|\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}},\mathbf{Y})\|_\circ \le C_{\mathtt{lyap}}(\mathsf{K})\cdot\|\mathbf{Y}\|_\circ,$$

where $C_{\mathtt{lyap}}(\mathsf{K}) = \mathrm{poly}\left(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}^{-1}\|, \|\mathbf{Z}_{\mathsf{K}}^{-1}\|, \|\mathbf{W}_1^{-1}\|, \|\mathbf{W}_2^{-1}\|, \|\mathbf{C}\|\right)$. *More precisely,*

$$C_{\mathtt{lyap}}(\mathsf{K}) := \frac{8t_\star(\mathsf{K})^2\|\mathbf{\Sigma}_{\mathsf{K}}\|^2\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|^2\|\mathbf{C}\|^2}{\lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}})\lambda_{\min}(\mathbf{W}_2)\lambda_{\min}(\mathbf{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}\cdot\max\left\{1, \frac{4\|\mathbf{\Sigma}_{22,\mathsf{K}}\|}{\lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}})}\right\}, \quad \textit{where}$$

$$t_\star(\mathsf{K}) := \frac{\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_1)}\log\left(\frac{\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|^2}{\lambda_{\min}(\mathbf{W}_1)}\max\left\{\frac{2}{\lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}})}, \frac{4\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{\Sigma}_{11,\mathrm{sys}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}\right\}\right).$$

The following corollary is also useful for establishing compact level sets.

J.1 Preliminaries on Lyapunov solutions

As a shorthand, we let clyap denote the following limit, if it converges:[8]

$$\mathsf{clyap}(\mathbf{X},\mathbf{Y}) = \lim_{t\to\infty}\int_0^t \exp(s\mathbf{X})\mathbf{Y}\exp(s\mathbf{X})^\top \mathrm{d}s = \int_0^\infty \exp(s\mathbf{X})\mathbf{Y}\exp(s\mathbf{X})^\top \mathrm{d}s. \tag{J.1}$$

We also define a "finite-time version", which is defined for all $\mathbf{X}\in\mathbb{R}^{d\times d}$ and $\mathbf{Y}\in\mathbb{S}^d$:

$$\mathsf{clyap}^{[t]}(\mathbf{X},\mathbf{Y}) = \int_0^t \exp(s\mathbf{X})\mathbf{Y}\exp(s\mathbf{X})^\top \mathrm{d}s, \quad \mathsf{clyap}^{[>t]}(\mathbf{X},\mathbf{Y}) = \int_t^\infty \exp(s\mathbf{X})\mathbf{Y}\exp(s\mathbf{X})^\top \mathrm{d}s.$$

The name clyap is short for "continuous Lyapunov", and is motivated by the following lemma:

Lemma J.2. *Suppose that $\mathbf{X}$ is Hurwitz stable. Then, $\mathbf{\Gamma} = \mathsf{clyap}(\mathbf{X},\mathbf{Y})$ exists and is the unique solution to the Lyapunov equation*

$$\mathbf{X}\mathbf{\Gamma} + \mathbf{\Gamma}\mathbf{X}^\top + \mathbf{Y} = 0.$$

In addition, if there exists a sequence $t_1, t_2, \ldots$ such that $\lim_{k\to\infty}\|\exp(t_k\mathbf{X})\|_{\mathrm{F}} \to 0$, then $\mathbf{X}$ is Hurwitz stable.

J.2 Proof of Proposition J.1

J.2.1 Setup.

Recall that $\mathbf{\Sigma}_{\mathsf{K}} = \mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{W}_{\mathrm{cl},\mathsf{K}})$ and $\mathbf{\Sigma}_{11,\mathrm{sys}} = \mathsf{clyap}(\mathbf{A},\mathbf{W}_1)$ solve the equations

$$\underbrace{\begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_{\mathsf{K}}\mathbf{C} & \mathbf{A}_{\mathsf{K}}\end{bmatrix}}_{=\mathbf{A}_{\mathrm{cl},\mathsf{K}}}\mathbf{\Sigma}_{\mathsf{K}} + \mathbf{\Sigma}_{\mathsf{K}}\begin{bmatrix}\mathbf{A} & 0\\ \mathbf{B}_{\mathsf{K}}\mathbf{C} & \mathbf{A}_{\mathsf{K}}\end{bmatrix}^\top + \underbrace{\begin{bmatrix}\mathbf{W}_1 & 0\\ 0 & \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top\end{bmatrix}}_{\mathbf{W}_{\mathrm{cl},\mathsf{K}}} = 0, \quad \mathbf{A}\mathbf{\Sigma}_{11,\mathrm{sys}} + \mathbf{\Sigma}_{11,\mathrm{sys}}\mathbf{A}^\top + \mathbf{W}_1 = 0.$$

Define the matrix $\mathbf{\Sigma}_w = \mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{W}_0)$ and $\mathbf{\Sigma}_v = \mathsf{clyap}(\mathbf{A}_{\mathsf{K}}, \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top)$ as the solutions to the Lyapunov equations

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}_w + \mathbf{\Sigma}_w\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top + \underbrace{\begin{bmatrix}\mathbf{W}_1 & 0\\ 0 & 0\end{bmatrix}}_{:=\mathbf{W}_0} = 0, \quad \mathbf{A}_{\mathsf{K}}\mathbf{\Sigma}_v + \mathbf{\Sigma}_v\mathbf{A}_{\mathsf{K}}^\top + \mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top = 0. \tag{J.2}$$

We recall the following closed-form expression for the solution to Lyapunov equations.

In particular,

$$\begin{aligned}
\mathbf{\Sigma}_{\mathsf{K}} &= \int_0^\infty \exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_{\mathrm{cl},\mathsf{K}}\exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}\tau\\
\mathbf{\Sigma}_w &= \int_0^\infty \exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_0\exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}\tau\\
\mathbf{\Sigma}_v &= \int_0^\infty \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top \mathrm{d}\tau\\
\mathbf{\Sigma}_{11,\mathrm{sys}} &= \int_0^\infty \exp(\tau\mathbf{A})\mathbf{W}_1\exp(\tau\mathbf{A})^\top \mathrm{d}\tau.
\end{aligned}$$

Throughout, we use the following decompositions

$$\mathbf{\Sigma}_{\mathsf{K}} = \mathbf{\Sigma}_{\mathsf{K}}^{[t]} + \mathbf{\Sigma}_{\mathsf{K}}^{[>t]}, \quad \mathbf{\Sigma}_w = \mathbf{\Sigma}_w^{[t]} + \mathbf{\Sigma}_w^{[>t]}, \quad \mathbf{\Sigma}_v = \mathbf{\Sigma}_v^{[t]} + \mathbf{\Sigma}_v^{[>t]}, \quad \mathbf{\Sigma}_{11,\mathrm{sys}} = \mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]} + \mathbf{\Sigma}_{11,\mathrm{sys}}^{[>t]},$$

where we define

$$\mathbf{\Sigma}_{\mathsf{K}}^{[t]} := \int_0^t \exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_{\mathrm{cl},\mathsf{K}}\exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}\tau, \quad \mathbf{\Sigma}_{\mathsf{K}}^{[>t]} := \int_t^\infty \exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{W}_{\mathrm{cl},\mathsf{K}}\exp(\tau\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}\tau,$$

and where $\mathbf{\Sigma}_w^{[t]}, \mathbf{\Sigma}_w^{[>t]}, \mathbf{\Sigma}_v^{[t]}, \mathbf{\Sigma}_v^{[>t]}, \mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]}, \mathbf{\Sigma}_{11,\mathrm{sys}}^{[>t]}$ are all defined analogously. The following computations are useful.

---

[8] That is, if $\lim_{t=0}^\infty \|\exp(s\mathbf{X})\mathbf{Y}\exp(s\mathbf{X})^\top\|_{\mathrm{F}}\,\mathrm{d}s$ is finite.

Lemma J.3 (Computations of exponentials)**.** *The following characterizes the exponentials of* $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$*:*

(a) Defining $\mathbf{M}(t) = \int_0^t \exp((t-s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s$*, one has*

$$\exp(t\mathbf{A}_{\mathrm{cl},\mathsf{K}}) = \begin{bmatrix}\exp(t\mathbf{A}) & 0 \\ \mathbf{M}(t) & \exp(t\mathbf{A}_{\mathsf{K}})\end{bmatrix}.$$

(b) The following computation holds

$$\exp(t\mathbf{A}_{\mathrm{cl},\mathsf{K}})\begin{bmatrix}\mathbf{W}_1 & 0 \\ 0 & 0\end{bmatrix}\exp(t\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top = \begin{bmatrix}\exp(t\mathbf{A})\mathbf{W}_1\exp(t\mathbf{A})^\top & \exp(t\mathbf{A})\mathbf{W}_1\mathbf{M}(t)^\top \\ \mathbf{M}(t)\mathbf{W}_1\exp(t\mathbf{A})^\top & \mathbf{M}(t)\mathbf{W}_1\mathbf{M}(t)^\top\end{bmatrix}.$$

Proof. Part (b) follows directly from part (a) and a straightforward computation. To prove part (a), we observe that the desired identity holds at time $t = 0$. To prove it holds for all t, it suffices to equate derivatives. First, we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{M}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^t \exp((t-s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s \\ &= \exp((t-s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\big|_{s=t} + \int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\exp((t-s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s \\ &= \mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(t\mathbf{A}) + \int_0^t \mathbf{A}_{\mathsf{K}}\exp((t-s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s = \mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(t\mathbf{A}) + \mathbf{A}_{\mathsf{K}}\mathbf{M}(t).\end{aligned}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\exp(t\mathbf{A}) & 0 \\ \mathbf{M}(t) & \exp(t\mathbf{A}_{\mathsf{K}})\end{bmatrix} = \begin{bmatrix}\mathbf{A}\exp(t\mathbf{A}) & 0 \\ \mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(t\mathbf{A}) + \mathbf{A}_{\mathsf{K}}\mathbf{M}(t) & \mathbf{A}_{\mathsf{K}}\exp(t\mathbf{A}_{\mathsf{K}})\end{bmatrix} = \mathbf{A}_{\mathrm{cl},\mathsf{K}}\begin{bmatrix}\exp(t\mathbf{A}) & 0 \\ \mathbf{M}(t) & \exp(t\mathbf{A}_{\mathsf{K}})\end{bmatrix}.$$

Similarly, $\frac{\mathrm{d}}{\mathrm{d}t}\exp(t\mathbf{A}_{\mathrm{cl},\mathsf{K}}) = \mathbf{A}_{\mathrm{cl},\mathsf{K}}\exp(t\mathbf{A}_{\mathrm{cl},\mathsf{K}})$. The identity follows from uniqueness of solutions to ODEs. ☐

The following lemma is straightforward to verify using the previous one.

Lemma J.4 (Useful identitities)**.** *The following identities hold:*

(a) One has the decompositions

$$\mathbf{\Sigma}_{\mathsf{K}} = \mathbf{\Sigma}_w + \begin{bmatrix}0 & 0 \\ 0 & \mathbf{\Sigma}_v\end{bmatrix}, \quad \mathbf{\Sigma}_{\mathsf{K}}^{[t]} = \mathbf{\Sigma}_w^{[t]} + \begin{bmatrix}0 & 0 \\ 0 & \mathbf{\Sigma}_v^{[t]}\end{bmatrix} \tag{J.3}$$

(b) $\mathbf{\Sigma}_{11,\mathsf{K}}^{[t]} = \mathbf{\Sigma}_{11,w}^{[t]} = \mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]}$ *and similarly,* $\mathbf{\Sigma}_{11,\mathsf{K}}^{[>t]} = \mathbf{\Sigma}_{11,w}^{[>t]} = \mathbf{\Sigma}_{11,\mathrm{sys}}^{[>t]}$*, and* $\mathbf{\Sigma}_{11,\mathsf{K}} = \mathbf{\Sigma}_{11,w} = \mathbf{\Sigma}_{11,\mathrm{sys}}$.

As a consequence, we find that $\mathbf{\Sigma}_{11,w}^{[t]}$ is invertible for all t. Lastly, we show $\mathbf{\Sigma}_{11,w}^{[t]} \succ 0$.

Lemma J.5. $\mathbf{\Sigma}_{11,\mathsf{K}}^{[t]} = \mathbf{\Sigma}_{11,w}^{[t]} = \mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]} \succ 0$ *for all* $t > 0$.

Proof. The equivalence $\mathbf{\Sigma}_{11,\mathsf{K}}^{[t]} = \mathbf{\Sigma}_{11,w}^{[t]} = \mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]}$ is given by Lemma J.4. Using the formula $\mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]} = \int_0^t \exp(\tau\mathbf{A})\mathbf{W}_1\exp(\tau\mathbf{A})^\top\mathrm{d}\tau$, we see that we can $\mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]} = \int_0^t \mathbf{N}(\tau)\mathrm{d}\tau$, where $\mathbf{N}(\cdot)$ is a continuous matrix valued function with $\mathbf{N}(0) = \mathbf{W}_1 \succ 0$. Hence, for all vectors $\mathbf{v} \neq 0$, the function $f(\cdot;\mathbf{v}) = \mathbf{v}^\top\mathbf{N}(\cdot)\mathbf{v}$ is continous and has $f(0;\mathbf{v}) = 0$. Thus, $\mathbf{v}^\top\mathbf{\Sigma}_{11,\mathrm{sys}}^{[t]}\mathbf{v} = \int_0^t f(\tau;\mathbf{v})\mathrm{d}\tau > 0$ for all nonzero $\mathbf{v}$. ☐

J.2.2 A Lyapunov argument

In this section, we show that if there is a finite t for which $\lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}}^{[t]})$ is strictly positive, then one can bound the solutions to $\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{Y})$ in terms of this t and other problem-dependent quantities. We begin with a general lemma that bounds the decay of matrix exponentials, with their finite-time Gramians.

Lemma J.6. *Fix a matrix* $\mathbf{X} \in \mathbb{R}^{d\times d}$*, and matrix* $\mathbf{Y}_0 \in \mathbb{S}^d$*, and suppose that the solution* $\mathbf{\Gamma}_0 = \mathsf{clyap}(\mathbf{X}, \mathbf{Y}_0)$ *exists. Define* $\mathbf{\Gamma}_0^{[t]} = \mathsf{clyap}^{[t]}(\mathbf{X}, \mathbf{Y}_0)$*, and* $\mathbf{\Gamma}_0^{[>t]}$ *analogously. Then, for all* $s, t \geq 0$*,*

- *(a)* $\mathbf{P}_0(t) = \mathbf{\Gamma}_0^{[>t]}$*, where* $\mathbf{P}_0(t) := \exp(t\mathbf{X})\mathbf{\Gamma}_0 \exp(t\mathbf{X})^\top$.
- *(b)* $\mathbf{P}_0(s+t) \preceq \rho_0(s) \cdot \mathbf{P}_0(t)$*, where* $\rho_0(s) := 1 - \frac{\lambda_{\min}(\mathbf{\Gamma}_0^{[s]})}{\|\mathbf{\Gamma}_0\|}$.
- *(c) In particular, if* $\mathbf{\Gamma}_0 \succ 0$*, then* $\mathbf{X}$ *is Hurwitz stable.*

Proof. **Part (a).** We see that

$$\begin{aligned}
\mathbf{\Gamma}_0^{[>t]} &:= \int_t^\infty \exp(\tau\mathbf{X})\mathbf{Y}_0 \exp(\tau\mathbf{X})^\top \mathrm{d}\tau \\
&= \exp(t\mathbf{X}) \left(\int_t^\infty \exp((\tau - t)\mathbf{X})\mathbf{Y}_0 \exp((\tau-t)\mathbf{X})^\top \mathrm{d}\tau\right) \exp(t\mathbf{X})^\top \\
&= \exp(t\mathbf{X}) \left(\int_0^\infty \exp(\tau\mathbf{X})\mathbf{Y}_0 \exp(\tau\mathbf{X})^\top \mathrm{d}\tau\right) \exp(t\mathbf{X})^\top \\
&= \exp(t\mathbf{X})\mathbf{\Gamma}_0 \exp(t\mathbf{X})^\top := \mathbf{P}_0(t).
\end{aligned}$$

Part (b). We use the decomposition

$$\mathbf{\Gamma}_0 = \mathbf{\Gamma}_0^{[t]} + \mathbf{\Gamma}_0^{[>t]} = \mathbf{\Gamma}_0^{[t]} + \mathbf{P}_0(t).$$

For a fixed t and $s \geq 0$, we have

$$\begin{aligned}
\mathbf{P}_0(s+t) &= \exp(t\mathbf{X}) \cdot \exp(s\mathbf{X})\mathbf{\Gamma}_0 \exp(s\mathbf{X})^\top \cdot \exp(t\mathbf{X})^\top \\
&= \exp(t\mathbf{X}) \cdot (\mathbf{\Gamma}_0 - \mathbf{\Gamma}_0^{[s]}) \cdot \exp(t\mathbf{X})^\top \\
&= \exp(t\mathbf{X}) \cdot \mathbf{\Gamma}_0^{1/2}(\mathbf{I}_n - \mathbf{\Gamma}_0^{-1/2}\mathbf{\Gamma}_0^{[s]}\mathbf{\Gamma}_0^{-1/2})\mathbf{\Gamma}_0^{1/2} \cdot \exp(t\mathbf{X})^\top \\
&\preceq \lambda_{\max}(\mathbf{I}_n - \mathbf{\Gamma}_0^{-1/2}\mathbf{\Gamma}_0^{[s]}\mathbf{\Gamma}_0^{-1/2}) \cdot \underbrace{\exp(t\mathbf{X}) \cdot \mathbf{\Gamma}_0^{1/2}\mathbf{\Gamma}_0^{1/2} \cdot \exp(t\mathbf{X})^\top}_{=\mathbf{P}_0(t)} \\
&\preceq \underbrace{\left(1 - \frac{\lambda_{\min}(\mathbf{\Gamma}_0^{[s]})}{\|\mathbf{\Gamma}_0\|}\right)}_{=\rho_0(s)} \cdot \mathbf{P}_0(t).
\end{aligned}$$

Part (c). Suppose that $\mathbf{\Gamma}_0 \succ 0$. Then, since $\mathbf{\Gamma}_0^{[s]}$ is monotone, there exists a finite s such that $\mathbf{\Gamma}_0^{[s]} \succ 0$. Thus, $\rho_0(s) < 1$. Then, by iterating part (b), we have that for any finite $k \in \mathbb{N}$

$$\mathbf{P}_0(ks+t) \preceq \rho_0(s)^k \cdot \mathbf{P}_0(t),$$

so that $\lim_{k\to\infty} \mathbf{P}_0(ks+t) = \lim_{k\to\infty} \exp((ks+t)\mathbf{X})\mathbf{\Gamma}_0 \exp((ks+t)\mathbf{X})^\top = 0$. Since $\mathbf{\Gamma}_0 \succ 0$, this implies that $\lim_{k\to\infty} \|\exp((ks+t)\mathbf{X})\| = 0$. By Lemma J.2, this can only occur if $\mathbf{X}$ is Hurwitz stable. □

By integrating Lemma J.6, we bound $\|\mathsf{clyap}(\mathbf{X}, \mathbf{Y})\|_\circ$ in terms of $\lambda_{\min}(\mathbf{\Gamma}_0^{[t]})$.

Lemma J.7. *Consider the setup of Lemma J.6, and suppose that* $t > 0$ *is such that* $\lambda_{\min}(\mathbf{\Gamma}_0^{[t]}) > 0$*. Then, for any* $\mathbf{Y} \in \mathbb{S}^d$*, and for* $\|\cdot\|_\circ$ *denoting either operator, Frobenius, or nuclear norm,*

$$\|\mathsf{clyap}(\mathbf{X}, \mathbf{Y})\|_\circ \leq \frac{t\|\mathbf{\Gamma}_0\|^2}{\lambda_{\min}(\mathbf{\Gamma}_0)\lambda_{\min}(\mathbf{\Gamma}_0^{[t]})} \cdot \|\mathbf{Y}\|_\circ.$$

Proof. Using Lemma J.2, we write $\bar{\mathbf{\Sigma}}$ explicitly and bound it as follows

$$\begin{aligned}
\|\mathsf{clyap}(\mathbf{X},\mathbf{Y})\|_\circ &= \left\|\int_0^\infty \exp(\tau\mathbf{X})\mathbf{Y}\exp(\tau\mathbf{X})^\top \mathrm{d}\tau\right\|_\circ \\
&\le \int_0^\infty \left\|\exp(\tau\mathbf{X})\mathbf{Y}\exp(\tau\mathbf{X})^\top\right\|_\circ \mathrm{d}\tau \\
&\overset{(i)}{\le} \|\mathbf{Y}\|_\circ \int_0^\infty \|\exp(\tau\mathbf{X})\|^2 \,\mathrm{d}\tau \\
&= \|\mathbf{Y}\|_\circ \int_0^\infty \left\|\exp(\tau\mathbf{X})\exp(\tau\mathbf{X})^\top\right\| \mathrm{d}\tau \\
&\le \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot \int_0^\infty \|\exp(\tau\mathbf{X})\mathbf{\Gamma}_0\exp(\tau\mathbf{X})^\top\|\mathrm{d}\tau \\
&= \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot \sum_{k=0}^\infty \int_{tk}^{t(k+1)} \|\exp(\tau\mathbf{X})\mathbf{\Gamma}_0\exp(\tau\mathbf{X})^\top\|\mathrm{d}\tau \\
&= \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot \sum_{k=0}^\infty \int_{tk}^{t(k+1)} \|\mathbf{P}_0(\tau)\|\mathrm{d}\tau.
\end{aligned}$$

Here, (i) uses that $\|\mathbf{X}_1\mathbf{X}_2\| \le \min\{\|\mathbf{X}_1\|\|\mathbf{X}_2\|_\circ, \|\mathbf{X}_1\|_\circ\|\mathbf{X}_2\|\}$ for any $\circ$ denoting either the operator, Frobenius, or trace norms (or more generally, any Schatten norm). From Lemma J.6, $\mathbf{P}_0(\tau)$ is non-increasing in the PSD order and $\|\mathbf{P}_0(tk)\| \le \|\mathbf{P}_0(0)\|\rho_0(t)^k$. Hence, noting that $\mathbf{P}_0(0) = \mathbf{\Gamma}_0$,

$$\int_{tk}^{t(k+1)} \|\mathbf{P}_0(\tau)\|\mathrm{d}\tau \;\le\; t\|\mathbf{P}_0(tk)\| \le t\|\mathbf{P}_0(0)\|\rho_0(t)^k = t\|\mathbf{\Gamma}_0\|\rho_0(t)^k.$$

Thus, if $\lambda_{\min}(\mathbf{\Gamma}_0^{[t]}) > 0$, then $\rho_0(t) < 1$, so we can sum

$$\begin{aligned}
\|\mathsf{clyap}(\mathbf{X},\mathbf{Y})\|_\circ &\le \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot t\|\mathbf{\Gamma}_0\| \cdot \sum_{k=0}^\infty \rho_0(t)^k \\
&= \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot t\|\mathbf{\Gamma}_0\| \cdot \frac{1}{1-\rho_0(t)} \\
&= \|\mathbf{Y}\|_\circ \|\mathbf{\Gamma}_0^{-1}\| \cdot t\|\mathbf{\Gamma}_0\| \cdot \frac{\|\mathbf{\Gamma}_0\|}{\lambda_{\min}(\mathbf{\Gamma}_0^{[t]})}.
\end{aligned}$$

Hence,

$$\|\mathsf{clyap}(\mathbf{X},\mathbf{Y})\|_\circ \le \|\mathbf{Y}\|_\circ \cdot t\frac{\|\mathbf{\Gamma}_0^{-1}\|\|\mathbf{\Gamma}_0\|^2}{\lambda_{\min}(\mathbf{\Gamma}_0^{[t]})} = \|\mathbf{Y}\|_\circ \cdot \frac{t\|\mathbf{\Gamma}_0\|^2}{\lambda_{\min}(\mathbf{\Gamma}_0)\lambda_{\min}(\mathbf{\Gamma}_0^{[t]})}.$$

□

Specializing with $\mathbf{X} \leftarrow \mathbf{A}_{\mathrm{cl},\mathsf{K}}$, $\mathbf{Y}_0 \leftarrow \mathbf{W}_{\mathrm{cl},\mathsf{K}}$, $\mathbf{\Gamma}_0 \leftarrow \mathbf{\Sigma}_{\mathsf{K}}$, we arrive at the following lemma:

Lemma J.8. *Suppose that $t > 0$ is such that $\lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}}^{[t]}) > 0$. Then, for any $\mathbf{Y} \in \mathbb{S}^{2n}$, and for $\|\cdot\|_\circ$ denoting either operator, Frobenius, or nuclear norm,*

$$\|\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}},\mathbf{Y})\|_\circ \le C_{[t]}(\mathsf{K}) \cdot \|\mathbf{Y}\|_\circ, \quad \textit{where } C_{[t]}(\mathsf{K}) := \frac{t\|\mathbf{\Sigma}_{\mathsf{K}}\|^2}{\lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}})\lambda_{\min}(\mathbf{\Sigma}_{\mathsf{K}}^{[t]})}.$$

Thus, it remains to show that, for any appropriate choice of t

$$C_{[t]}(\mathsf{K}) \le C_{\mathtt{lyap}}(\mathsf{K}). \tag{J.4}$$

J.2.3 Lower bounding finite-time covariance in terms of diagonal blocks

In order to upper bound $C_{[t]}(\mathsf{K})$ from Lemma J.8, we must lower bound $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]})$. Recall that from Lemma J.4,

$$\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]} = \boldsymbol{\Sigma}_{w}^{[t]} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{v}^{[t]} \end{bmatrix}. \tag{J.5}$$

Leveraging this form, we show that it suffices to lower bound $\lambda_{\min}(\boldsymbol{\Sigma}_{v}^{[t]})$, that is, the finite-time covariance introduced by the observation noise into the policy.

Lemma J.9. *Suppose that t is large enough such that $\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[<t]}\| \le \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})$, then*

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]}\right) \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{v}^{[t]}) \min\left\{1, \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})}{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}\right\}.$$

Proof. Applying Lemma J.14 to the decomposition Eq. (J.5), we have

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]}\right) \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{v}^{[t]}) \min\left\{1, \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{11,w}^{[t]})}{2\|\boldsymbol{\Sigma}_{22,w}^{[t]} + \boldsymbol{\Sigma}_{v}^{[t]}\|}\right\}.$$

Substituing in $\boldsymbol{\Sigma}_{11,w}^{[t]} = \boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[t]}$, and bounding $\|\boldsymbol{\Sigma}_{22,w}^{[t]} + \boldsymbol{\Sigma}_{v}^{[t]}\| = \|\boldsymbol{\Sigma}_{22,\mathsf{K}}^{[t]}\| \le \|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|$ in view of Lemma J.4. Moreover, we have $\lambda_{\min}(\boldsymbol{\Sigma}_{11,w}^{[t]}) \ge \lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}}) - \|\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[>t]}\| \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})$, where the last step holds under the assumption on t in the lemma. With these simplifications, we arrive at the desired bound:

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]}\right) \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{v}^{[t]}) \min\left\{1, \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{11,w})}{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}\right\}.$$

□

J.2.4 Lower bounding the contribution of output noise

From Lemma J.9, we have to lower bound $\lambda_{\min}\left(\boldsymbol{\Sigma}_{v}^{[t]}\right)$. This step involves the two most original insights of the proof:

- First, $\boldsymbol{\Sigma}_{v}^{[t]} \succeq \frac{C}{t}\boldsymbol{\Sigma}_{22,w}^{[t]}$, for some system-dependent constant C. Here, $\boldsymbol{\Sigma}_{v}^{[t]}$ represents the part of the internal-state covariance excited by the full-rank observation noise $\mathbf{W}_2$, and $\boldsymbol{\Sigma}_{22,w}^{[t]}$ the part of the covariance excited by the observations $\tilde{\boldsymbol{y}}(t) = \mathbf{C}\boldsymbol{x}(t)$. Essentially, we argue that the covariance excited by any stochastic process $\tilde{\boldsymbol{y}}(t)$ cannot be much greater than the excitation by Gaussian noise $\mathbf{v}(t)$.
- Second, if $\mathbf{Z}_{\mathsf{K}} \succ 0$ and if $\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[>t]}$ is small, then we can lower bound $\lambda_{\min}(\boldsymbol{\Sigma}_{22,w}^{[t]})$ in terms of $\mathbf{Z}_{\mathsf{K}}$. Intuitively, $\boldsymbol{\Sigma}_{22,w}^{[t]}$ describes how much of the process noise $\mathbf{w}(t)$ excites the internal filter state $\hat{\mathbf{x}}(t)$, and $\mathbf{Z}_{\mathsf{K}}$ measures the correlation between $\hat{\mathbf{x}}(t)$ and $\mathbf{x}(t)$. This argument therefore uses the insight that, if $\hat{\mathbf{x}}(t)$ and $\mathbf{x}(t)$ have nontrivial correlation, some of the process noise $\mathbf{w}(t)$ must be exciting the filter state $\hat{\mathbf{x}}(t)$.

Lemma J.10. *For all t, we have*

$$\boldsymbol{\Sigma}_{22,w}^{[t]} \preceq \frac{t\|\mathbf{C}\|^2\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_2)} \cdot \boldsymbol{\Sigma}_{v}^{[t]}.$$

In particular, $\boldsymbol{\Sigma}_{22,\mathsf{K}}^{[t]} \succ 0$ if and only if $\boldsymbol{\Sigma}_{v}^{[t]} \succ 0$.[9]

---

[9] Note that this lemma and its conclusion only requires Assumption 2.3. It does not even require stability of $\mathbf{A}_{\mathsf{K}}$.

Proof. From Lemma J.3, we have that

$$
\begin{aligned}
&\mathbf{\Sigma}^{[t]}_{22,w}\\
&= \int_0^t \left(\mathbf{M}(\tau)\mathbf{W}_1\mathbf{M}(\tau)^\top\right) \mathrm{d}\tau\\
&= \int_0^t \left(\int_0^\tau \exp((\tau - s_1)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s_1\right)\mathbf{W}_1\left(\int_0^\tau \exp((\tau - s_2)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathrm{d}s_2\right)^\top \mathrm{d}\tau\\
&= \int_0^t \tau^2\left(\frac{1}{\tau}\int_0^\tau \exp((\tau - s_1)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1^{1/2}\mathrm{d}s_1\right)\left(\frac{1}{\tau}\int_0^\tau \exp((\tau - s_2)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1^{1/2}\mathrm{d}s_2\right)^\top \mathrm{d}\tau\\
&\overset{(i)}{\preceq} \int_0^t \tau^2\cdot\frac{1}{\tau}\int_0^\tau \left(\exp((\tau - s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1^{1/2}\right)\left(\exp((\tau - s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1^{1/2}\right)^\top \mathrm{d}s\mathrm{d}\tau\\
&= \int_0^t\int_0^\tau \tau\exp((\tau - s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1\exp(s\mathbf{A})^\top\mathbf{C}^\top\mathbf{B}_{\mathsf{K}}^\top\exp((\tau - s)\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}s\mathrm{d}\tau\\
&\preceq t\int_0^t\int_0^\tau \exp((\tau - s)\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1\exp(s\mathbf{A})^\top\mathbf{C}^\top\mathbf{B}_{\mathsf{K}}^\top\exp((\tau - s)\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}s\mathrm{d}\tau.
\end{aligned}
$$

Here, inequality (i) invoked Lemma J.15. Using the integral re-arrangment

$$
\begin{aligned}
\int_{\tau=0}^t\int_{s=0}^\tau \mathbf{N}(\tau - s, s)\mathrm{d}s\mathrm{d}\tau &= \int_{s=0}^t\int_{\tau=s}^t \mathbf{N}(\tau - s, s)\mathrm{d}\tau\mathrm{d}s\\
&= \int_{s=0}^t\int_{\tau=0}^{t-s} \mathbf{N}(\tau, s)\mathrm{d}\tau\mathrm{d}s\\
&\preceq \int_{s=0}^t\int_{\tau=0}^{t} \mathbf{N}(\tau, s)\mathrm{d}\tau\mathrm{d}s
\end{aligned}
$$

for any PSD-matrix valued function $\mathbf{N}(\cdot,\cdot) : [0,t]^2 \to \mathbb{S}^n_+$, we obtain that

$$
\begin{aligned}
\mathbf{\Sigma}^{[t]}_{22,w} &\preceq t\int_0^t\int_0^t \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s\mathbf{A})\mathbf{W}_1\exp(s\mathbf{A})^\top\mathbf{C}^\top\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}s\mathrm{d}\tau\\
&= t\int_0^t \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\left(\mathbf{C}\left(\int_0^t \exp(s\mathbf{A})\mathbf{W}_1\exp(s\mathbf{A})^\top\mathrm{d}s\right)\mathbf{C}^\top\right)\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}\\
&= t\int_0^t \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\left(\mathbf{C}\mathbf{\Sigma}^{[t]}_{11,\mathrm{sys}}\mathbf{C}^\top\right)\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}\tau\\
&= t\int_0^t \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2^{1/2}\left(\mathbf{W}_2^{-1/2}\mathbf{C}\mathbf{\Sigma}^{[t]}_{11,\mathrm{sys}}\mathbf{C}^\top\mathbf{W}_2^{-1/2}\right)\mathbf{W}_2^{1/2}\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}\tau.
\end{aligned}
$$

We render the above integral as

$$
t\int_{s_2=0}^t\int_{s_1=0}^t \exp(s_1\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{C}\exp(s_2\mathbf{A})\mathbf{W}_1\exp(s_2\mathbf{A})^\top\mathbf{C}^\top\mathbf{B}_{\mathsf{K}}^\top\exp(s_1\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}s_1\mathrm{d}s_2.
$$

Bounding $\|\mathbf{W}_2^{-1/2}\mathbf{C}\mathbf{\Sigma}^{[t]}_{11,\mathrm{sys}}\mathbf{C}^\top\mathbf{W}_2^{-1/2}\| \le \frac{\|\mathbf{C}\|^2\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_2)} \le \frac{\|\mathbf{C}\|^2\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_2)}$, we have

$$
\begin{aligned}
\mathbf{\Sigma}^{[t]}_{22,w} &\preceq \frac{t\|\mathbf{C}\|^2\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_2)}\cdot\int_0^t \exp(\tau\mathbf{A}_{\mathsf{K}})\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{B}_{\mathsf{K}}^\top\exp(\tau\mathbf{A}_{\mathsf{K}})^\top\mathrm{d}\tau\\
&= \frac{t\|\mathbf{C}\|^2\|\mathbf{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_2)}\cdot\mathbf{\Sigma}^{[t]}_v.
\end{aligned}
$$

The last point follows from $\mathbf{\Sigma}^{[t]}_{22,\mathsf{K}} = \mathbf{\Sigma}^{[t]}_{22,w} + \mathbf{\Sigma}^{[t]}_v$ by Lemma J.4. □

Lemma J.11. *Suppose that t is sufficiently large such that*

$$
\|\mathbf{\Sigma}^{[>t]}_{11,\mathrm{sys}}\| \le \frac{\lambda_{\min}(\mathbf{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4\|\mathbf{\Sigma}_{22,\mathsf{K}}\|}.
$$

Then, it holds that

$$\lambda_{\min}\left(\boldsymbol{\Sigma}^{[t]}_{22,w}\right) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}.$$

Therefore, in view of Lemma J.10,

$$\lambda_{\min}\left(\boldsymbol{\Sigma}^{[t]}_{v}\right) \geq \frac{\lambda_{\min}(\mathbf{W}_2)\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4t\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2\|\mathbf{C}\|^2}.$$

Proof. We assume that $\lambda_{\min}(\mathbf{Z}_{\mathsf{K}}) \geq 0$ for otherwise the lemma is vacuous. We compute

$$\sigma_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}})^2 = \lambda_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top) \geq \frac{1}{\|\boldsymbol{\Sigma}_{22,\mathsf{K}}^{-1}\|}\lambda_{\min}\left(\boldsymbol{\Sigma}_{12,\mathsf{K}}\boldsymbol{\Sigma}_{22,\mathsf{K}}^{-1}\boldsymbol{\Sigma}_{12,\mathsf{K}}^\top\right) = \lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}}), \tag{J.6}$$

and take t sufficiently large that

$$\|\boldsymbol{\Sigma}^{[>t]}_{12,w}\| \leq \frac{1}{2}\sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})} \leq \frac{1}{2}\sigma_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}}). \tag{J.7}$$

Now invoking Eq. (J.3) on the $(1,2)$-block of $\boldsymbol{\Sigma}_{\mathsf{K}}$, then if Eq. (J.7) holds,

$$\boldsymbol{\Sigma}_{12,\mathsf{K}} = \boldsymbol{\Sigma}_{12,w} = \boldsymbol{\Sigma}^{[t]}_{12,w} + \boldsymbol{\Sigma}^{[>t]}_{12,w}, \quad \text{so that } \sigma_{\min}\left(\boldsymbol{\Sigma}^{[t]}_{12,w}\right) \geq \sigma_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}}) - \|\boldsymbol{\Sigma}^{[>t]}_{12,w}\| \geq \frac{1}{2}\sigma_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}}).$$

Next, since $\boldsymbol{\Sigma}^{[t]}_{w} \succeq 0$, the Schur complement test implies that

$$\begin{aligned}
\boldsymbol{\Sigma}^{[t]}_{22,w} &\succeq \boldsymbol{\Sigma}^{[t]\top}_{12,w}\left(\boldsymbol{\Sigma}^{[t]}_{11,w}\right)^{-1}\boldsymbol{\Sigma}^{[t]}_{12,w} \\
&\succeq \frac{1}{\|\boldsymbol{\Sigma}^{[t]}_{11,w}\|}\boldsymbol{\Sigma}^{[t]\top}_{12,w}\boldsymbol{\Sigma}^{[t]}_{12,w} \\
&\succeq \frac{1}{\|\boldsymbol{\Sigma}_{11,w}\|}\boldsymbol{\Sigma}^{[t]\top}_{12,w}\boldsymbol{\Sigma}^{[t]}_{12,w} = \frac{1}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}\boldsymbol{\Sigma}^{[t]\top}_{12,w}\boldsymbol{\Sigma}^{[t]}_{12,w},
\end{aligned}$$

where above we use $\boldsymbol{\Sigma}^{[t]}_{11,w} \preceq \boldsymbol{\Sigma}_{11,w}$, and $\boldsymbol{\Sigma}_{11,w} = \boldsymbol{\Sigma}_{11,\mathrm{sys}}$ in view of Eq. (J.3). Here, invertibility of $\boldsymbol{\Sigma}^{[t]}_{11,w}$ is guaranteed by Lemma J.5. Therefore, if Eq. (J.7) holds,

$$\lambda_{\min}\left(\boldsymbol{\Sigma}^{[t]}_{22,w}\right) \geq \frac{1}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}\sigma_{\min}(\boldsymbol{\Sigma}^{[t]}_{12,w})^2 \geq \frac{1}{4\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}\sigma_{\min}(\boldsymbol{\Sigma}_{12,\mathsf{K}})^2 \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|},$$

where the last inequality applies Eq. (J.6). Lastly, we simplify the condition $\|\boldsymbol{\Sigma}^{[>t]}_{12,w}\| \leq \frac{1}{2}\sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}$ in Eq. (J.7). We have

$$\|\boldsymbol{\Sigma}^{[>t]}_{12,w}\|^2 \overset{(i)}{\leq} \|\boldsymbol{\Sigma}^{[>t]}_{11,w}\| \cdot \|\boldsymbol{\Sigma}^{[>t]}_{22,w}\| \overset{(ii)}{\leq} \|\boldsymbol{\Sigma}^{[>t]}_{11,w}\| \cdot \|\boldsymbol{\Sigma}_{22,w}\| \overset{(iii)}{\leq} \|\boldsymbol{\Sigma}^{[>t]}_{11,\mathrm{sys}}\| \cdot \|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|,$$

where (i) uses Lemma J.13, (ii) uses $\boldsymbol{\Sigma}^{[>t]}_{22,w} \preceq \boldsymbol{\Sigma}_{22,w}$, and (iii) uses both that $\boldsymbol{\Sigma}_{22,w} \preceq \boldsymbol{\Sigma}_{22,\mathsf{K}}$ by Eq. (J.3) and that $\boldsymbol{\Sigma}^{[>t]}_{11,w} = \boldsymbol{\Sigma}^{[>t]}_{11,\mathrm{sys}}$ by Lemma J.4 part (b). Therefore, the condition $\|\boldsymbol{\Sigma}^{[>t]}_{12,w}\| \leq \frac{1}{2}\sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}$ is met as soon as

$$\|\boldsymbol{\Sigma}^{[>t]}_{11,\mathrm{sys}}\| \leq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}.$$

□

J.2.5 Bounding the decay of the true system

Recall that both Lemma J.9 and Lemma J.11 require us to bound the decay of $\|\boldsymbol{\Sigma}^{[>t]}_{11,\mathrm{sys}}\|$. This is achieved in the following lemma.

Lemma J.12. *For any* $t > 0$*, we have*

$$\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[>t]}\| \le e^{\frac{-t\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}} \cdot \frac{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2}{\lambda_{\min}(\mathbf{W}_1)}.$$

Proof. Define $\mathbf{N}(s) = \exp(t\mathbf{A})\boldsymbol{\Sigma}_{11,\mathrm{sys}}\exp(t\mathbf{A})^\top$. We compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{N}(s) &= \exp(t\mathbf{A})\mathbf{A}\boldsymbol{\Sigma}_{11,\mathrm{sys}}\exp(t\mathbf{A})^\top + \exp(t\mathbf{A})\boldsymbol{\Sigma}_{11,\mathrm{sys}}\mathbf{A}^\top\exp(t\mathbf{A})^\top \\
&= \exp(t\mathbf{A})\left(\mathbf{A}\boldsymbol{\Sigma}_{11,\mathrm{sys}} + \boldsymbol{\Sigma}_{11,\mathrm{sys}}\mathbf{A}^\top\right)\exp(t\mathbf{A})^\top \\
&= \exp(t\mathbf{A})\left(-\mathbf{W}_1\right)\exp(t\mathbf{A})^\top \\
&\preceq \frac{-\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|} \cdot \exp(t\mathbf{A})\boldsymbol{\Sigma}_{11,\mathrm{sys}}\exp(t\mathbf{A})^\top = \frac{-\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|} \cdot \mathbf{N}(s).
\end{aligned}$$

Applying Lemma J.16,

$$\exp(t\mathbf{A})\boldsymbol{\Sigma}_{11,\mathrm{sys}}\exp(t\mathbf{A})^\top = \mathbf{N}(t) \preceq \mathbf{N}(0) \cdot e^{\frac{-t\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}} = \boldsymbol{\Sigma}_{11,\mathrm{sys}} \cdot e^{\frac{-t\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}}.$$

As a consequence,

$$\begin{aligned}
\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[>t]} &= \int_{s=t}^{\infty} \exp(s\mathbf{A})\boldsymbol{\Sigma}_{11,\mathrm{sys}}\exp(s\mathbf{A})^\top \mathrm{d}s \preceq \int_{s=t}^{\infty}\left(\boldsymbol{\Sigma}_{11,\mathrm{sys}} \cdot e^{\frac{-s\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}}\right)\mathrm{d}s \\
&= \boldsymbol{\Sigma}_{11,\mathrm{sys}} \cdot \frac{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_1)} \cdot e^{\frac{-t\lambda_{\min}(\mathbf{W}_1)}{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}}.
\end{aligned}$$

The bound follows by taking the operator norm of both sides. □

J.2.6 Concluding the argument

It remains to bound $C_{[t]}(\mathsf{K}) \le C_{\mathtt{lyap}}(\mathsf{K})$, where $C_{[t]}(\mathsf{K}) := \frac{t\|\boldsymbol{\Sigma}_{\mathsf{K}}\|^2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})\lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]})}$ was given in Lemma J.8. Consolidating Lemma J.9 and Lemma J.11, we have that if

$$\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}^{[<t]}\| \le \min\left\{\frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}}), \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{4\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}\right\}. \tag{J.8}$$

Then,

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]}\right) \ge \frac{\lambda_{\min}(\mathbf{W}_2)\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}{8t\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2\|\mathbf{C}\|^2} \cdot \min\left\{1, \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})}{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}\right\}.$$

Or by inverting,

$$\frac{1}{\lambda_{\min}\left(\boldsymbol{\Sigma}_{\mathsf{K}}^{[t]}\right)} \le \frac{8t\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2\|\mathbf{C}\|^2}{\lambda_{\min}(\mathbf{W}_2)\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})} \cdot \max\left\{1, \frac{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})}\right\}.$$

Hence, if t satisfies Eq. (J.8), then

$$C_{[t]}(\mathsf{K}) \le \frac{8t^2\|\boldsymbol{\Sigma}_{\mathsf{K}}\|^2\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2\|\mathbf{C}\|^2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})\lambda_{\min}(\mathbf{W}_2)\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})} \cdot \max\left\{1, \frac{4\|\boldsymbol{\Sigma}_{22,\mathsf{K}}\|}{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})}\right\}. \tag{J.9}$$

It remains to select t large enough to satisfy Eq. (J.8). From Lemma J.12, it is enough to take

$$t = t_\star(\mathsf{K}) = \frac{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\mathbf{W}_1)}\log\left(\frac{\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|^2}{\lambda_{\min}(\mathbf{W}_1)}\max\left\{\frac{2}{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})}, \frac{4\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|}{\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})\lambda_{\min}(\mathbf{Z}_{\mathsf{K}})}\right\}\right).$$

Substituting $t_\star(\mathsf{K})$ into Eq. (J.9) yields the desired upper bound $C_{\mathtt{lyap}}(\mathsf{K})$. To see that $C_{\mathtt{lyap}}(\mathsf{K})$ is at most polynomial in $\left(\|\boldsymbol{\Sigma}_{\mathsf{K}}\|, \|\boldsymbol{\Sigma}_{\mathsf{K}}^{-1}\|, \|\mathbf{Z}_{\mathsf{K}}^{-1}\|, \|\mathbf{W}_1^{-1}\|, \|\mathbf{W}_2^{-1}\|, \|\mathbf{C}\|\right)$, we observe that $C_{[t]}(\mathsf{K})$ and $t_\star(\mathsf{K})$ are at most polynomial in

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|, \lambda_{\min}(\mathbf{W}_1)^{-1}, \|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|, \lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}})^{-1}, \|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|, \\
&\lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}})^{-1}, \lambda_{\min}(\mathbf{Z}_{\mathsf{K}})^{-1}, \|\boldsymbol{\Sigma}_{\mathsf{K}}\|, \|\mathbf{C}\|, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})^{-1}, \lambda_{\min}(\mathbf{W}_2)^{-1}.
\end{aligned} \tag{J.10}$$

Noting that $\boldsymbol{\Sigma}_{11,\mathrm{sys}}$ and $\boldsymbol{\Sigma}_{22,\mathsf{K}}$ are submatrices of $\boldsymbol{\Sigma}_{\mathsf{K}}$, so that $\lambda_{\min}(\boldsymbol{\Sigma}_{11,\mathrm{sys}}), \lambda_{\min}(\boldsymbol{\Sigma}_{22,\mathsf{K}}) \ge \lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})$ and $\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|, \|\boldsymbol{\Sigma}_{22,\mathsf{K}}\| \le \|\boldsymbol{\Sigma}_{\mathsf{K}}\|$, $C_{[t]}(\mathsf{K})$ and $t_\star(\mathsf{K})$ are polynomial in

$$\|\boldsymbol{\Sigma}_{11,\mathrm{sys}}\|, \lambda_{\min}(\mathbf{W}_1)^{-1}, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})^{-1}, \lambda_{\min}(\mathbf{Z}_{\mathsf{K}})^{-1}, \|\mathbf{C}\|, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{K}})^{-1}, \lambda_{\min}(\mathbf{W}_2)^{-1},$$

Replacing $\lambda_{\min}(\cdot)^{-1}$ with $\|(\cdot)^{-1}\|$ verifies the simplification.

J.3 Supporting technical tools

We first begin with two linear algebra matrices, both of which pertain to partitioned matrices

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{12}^\top & \mathbf{\Lambda}_{22} \end{bmatrix} \in \mathbb{S}_+^{2n}. \tag{J.11}$$

Lemma J.13. *Let* $\mathbf{\Lambda} \in \mathbb{S}_+^{2n}$ *be PSD be partioned as in Eq. (J.11). Then* $\|\mathbf{\Lambda}_{12}\| \le \sqrt{\|\mathbf{\Lambda}_{11}\|\|\mathbf{\Lambda}_{22}\|}$.

Proof. Let $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in \mathbb{R}^{2n}$, then

$$\mathbf{v}^\top \mathbf{\Lambda} \mathbf{v} = \mathbf{v}_1^\top \mathbf{\Lambda}_{11} \mathbf{v}_1 + \mathbf{v}_2^\top \mathbf{\Lambda}_{22} \mathbf{v}_2 + 2\mathbf{v}_1^\top \mathbf{\Lambda}_{12} \mathbf{v}_2 \ge 0.$$

By considering the same inequality with the vector $\widetilde{\mathbf{v}} = (\mathbf{v}_1, -\mathbf{v}_2)$, we have that for all $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in \mathbb{R}^{2n}$,

$$|\mathbf{v}_1^\top \mathbf{\Lambda}_{12} \mathbf{v}_2| \le \frac{1}{2}\left(\mathbf{v}_1^\top \mathbf{\Lambda}_{11} \mathbf{v}_1 + \mathbf{v}_2^\top \mathbf{\Lambda}_{22} \mathbf{v}_2\right).$$

By considering scalings $\mathbf{v}_\alpha := (\alpha^{1/2}\mathbf{v}_1, \alpha^{-1/2}\mathbf{v}_2) \in \mathbb{R}^{2n}$ for $\alpha > 0$, we have

$$\begin{aligned} |\mathbf{v}_1^\top \mathbf{\Lambda}_{12} \mathbf{v}_2| &\le \frac{1}{2} \inf_{\alpha > 0} \left(\alpha \mathbf{v}_1^\top \mathbf{\Lambda}_{11} \mathbf{v}_1 + \alpha^{-1} \mathbf{v}_2^\top \mathbf{\Lambda}_{22} \mathbf{v}_2\right) \\ &= \sqrt{\mathbf{v}_1^\top \mathbf{\Lambda}_{11} \mathbf{v}_1 \cdot \mathbf{v}_2^\top \mathbf{\Lambda}_{22} \mathbf{v}_2} \\ &\le \sqrt{\|\mathbf{\Lambda}_{11}\|\|\mathbf{\Lambda}_{22}\|}\|\mathbf{v}_1\|\|\mathbf{v}_2\|, \end{aligned}$$

which completes the proof. □

Lemma J.14. *Let* $\mathbf{\Lambda} \in \mathbb{S}_+^{2n}$ *be PSD and be partitioned as in Eq. (J.11). Then any given* $\mathbf{\Lambda}_0 \in \mathbb{S}_+^n$*, we have*

$$\lambda_{\min}\left(\mathbf{\Lambda} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Lambda}_0 \end{bmatrix}\right) \ge \frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_0) \min\left\{1, \frac{\lambda_{\min}(\mathbf{\Lambda}_{11})}{2\|\mathbf{\Lambda}_{22} + \mathbf{\Lambda}_0\|}\right\}.$$

Proof. Without loss of generality, may assume $\mathbf{\Lambda}_0, \mathbf{\Lambda}_{11} \succ 0$ since otherwise the lemma is vacuous. For compactness, denote

$$\bar{\mathbf{\Lambda}} := \mathbf{\Lambda} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Lambda}_0 \end{bmatrix}.$$

It suffices to exhibit λ such that $\lambda_{\min}(\bar{\mathbf{\Lambda}}) \ge \lambda$. From the Schur complement test, we have that $\lambda_{\min}(\bar{\mathbf{\Lambda}}) \ge \lambda$ as long as $\bar{\mathbf{\Lambda}}_{11} \succeq \lambda \mathbf{I}_n$ and

$$\bar{\mathbf{\Lambda}}_{22} - \lambda \mathbf{I}_n \succeq \bar{\mathbf{\Lambda}}_{12}^\top (\bar{\mathbf{\Lambda}}_{11} - \lambda \mathbf{I}_n)^{-1} \bar{\mathbf{\Lambda}}_{12},$$

so, substituing in the form of $\bar{\mathbf{\Lambda}}$, we have $\mathbf{\Lambda}_{11} \succeq \lambda \mathbf{I}_n$ and

$$\mathbf{\Lambda}_{22} + \mathbf{\Lambda}_0 - \lambda \mathbf{I}_n \succeq \mathbf{\Lambda}_{12}^\top (\mathbf{\Lambda}_{11} - \lambda \mathbf{I}_n)^{-1} \mathbf{\Lambda}_{12}.$$

If we take $\lambda = \alpha \lambda_{\min}(\mathbf{\Lambda})$ for some $\alpha < 1$, then

$$\mathbf{\Lambda}_{12}^\top (\mathbf{\Lambda}_{11} - \lambda \mathbf{I}_n)^{-1} \mathbf{\Lambda}_{12} \preceq (1-\alpha)^{-1} \mathbf{\Lambda}_{12}^\top (\mathbf{\Lambda}_{11})^{-1} \mathbf{\Lambda}_{12}^\top \preceq (1-\alpha)^{-1} \mathbf{\Lambda}_{22},$$

where the last step applies the Schur complement test to $\mathbf{\Lambda}$. Thus, it is enough

$$\mathbf{\Lambda}_{22} + \mathbf{\Lambda}_0 - \lambda \mathbf{I}_n \succeq (1-\alpha)^{-1} \mathbf{\Lambda}_{22},$$

so that, with rearranging and substituing in the definition of λ, it suffices to choose $\alpha \le 1/2$ and

$$\mathbf{\Lambda}_0 \succeq \frac{\alpha}{1-\alpha} \mathbf{\Lambda}_{22} + \alpha \lambda_{\min}(\mathbf{\Lambda}_{11}).$$

Thus, it is enough that $\alpha \ge 0$ satisfies

$$\lambda_{\min}(\mathbf{\Lambda}_0) \succeq \alpha\left(2\|\mathbf{\Lambda}_{22}\| + \lambda_{\min}(\mathbf{\Lambda}_{11})\right), \quad \alpha \le 1/2.$$

Hence, choosing the maximal α which satisfies the above display,

$$\begin{aligned}
\lambda_{\min}(\bar{\mathbf{\Lambda}}) \geq \alpha\lambda_{\min}(\mathbf{\Lambda}_{11}) &= \min\left\{\frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_{11}), \frac{\lambda_{\min}(\mathbf{\Lambda}_0)\lambda_{\min}(\mathbf{\Lambda}_{11})}{(2\|\mathbf{\Lambda}_{22}\| + \lambda_{\min}(\mathbf{\Lambda}_{11}))}\right\} \\
&\overset{(i)}{\geq} \min\left\{\frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_{11}), \frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_0), \frac{\lambda_{\min}(\mathbf{\Lambda}_0)\lambda_{\min}(\mathbf{\Lambda}_{11})}{4\|\mathbf{\Lambda}_{22}\|}\right\} \\
&\overset{(ii)}{\geq} \min\left\{\frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_{11}), \frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_0), \frac{\lambda_{\min}(\mathbf{\Lambda}_0)\lambda_{\min}(\mathbf{\Lambda}_{11})}{4\|\mathbf{\Lambda}_{22} + \mathbf{\Lambda}_0\|}\right\} \\
&\overset{(iii)}{=} \min\left\{\frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_0), \frac{\lambda_{\min}(\mathbf{\Lambda}_0)\lambda_{\min}(\mathbf{\Lambda}_{11})}{4\|\mathbf{\Lambda}_{22} + \mathbf{\Lambda}_0\|}\right\}.
\end{aligned}$$

Here (i) used that $\frac{a}{b+c} \geq \min\{\frac{a}{2b}, \frac{a}{2c}\}$, (ii) that since $\mathbf{\Lambda}_0, \mathbf{\Lambda}_{22} \succeq 0$, we can replace $\|\mathbf{\Lambda}_0 + \mathbf{\Lambda}_{22}\| \geq \|\mathbf{\Lambda}_{22}\|$, and (iii) that $\frac{\lambda_{\min}(\mathbf{\Lambda}_0)}{\|\mathbf{\Lambda}_{22}+\mathbf{\Lambda}_0\|} \leq 1$ (again, for $\mathbf{\Lambda}_{22}, \mathbf{\Lambda}_0 \succeq 0$). The bound follows by factoring. $\square$

Lemma J.15. *For any continuous matrix valued function* $\mathbf{X}(s) \in \mathbb{R}^{n\times n}$, $\left(\int_0^1 \mathbf{X}(s_1)\mathrm{d}s_1\right)\left(\int_0^1 \mathbf{X}(s_1)\mathrm{d}s_1\right)^\top \preceq \int_0^1 \mathbf{X}(s)\mathbf{X}(s)^\top \mathrm{d}s$.

Proof. It suffices to show that for any vector $\mathbf{v}_0 \in \mathbb{R}^n$, the function $\mathbf{v}(s) = \mathbf{X}(s)^\top \mathbf{v}_0$ satisfies

$$\left\|\int_0^1 \mathbf{v}(s)\mathrm{d}s\right\|^2 \leq \int_0^1 \|\mathbf{v}(s)\|^2 \mathrm{d}s.$$

We can view both integrals as expectations over a random vector $\tilde{\mathbf{v}} = \mathbf{v}(s)$, where s is drawn uniformly on $[0,1]$. With this interpretation, it suffices that $\|\mathbb{E}[\tilde{\mathbf{v}}]\|^2 \leq \mathbb{E}[\|\tilde{\mathbf{v}}\|^2]$, which is precisely Jensen's inequality. $\square$

Lemma J.16. *Let* $\mathbf{N}(\cdot) : [0,\infty) \to \mathbb{S}_+^n$ *be continuously differentiable PSD-matrix-valued function satisfying*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{N}(t) \preceq -\alpha\mathbf{N}(s), \text{ for some } \alpha > 0.$$

Then, $\mathbf{N}(t) \preceq e^{-\alpha t}\mathbf{N}(0)$ *for all* $t > 0$.

Proof. For fixed $\mathbf{v} \neq 0$, define $f(\cdot;\mathbf{v}) = \mathbf{v}^\top \mathbf{N}(\cdot)\mathbf{v}$. Then, $f(\cdot;\mathbf{v}) \geq 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}f(t;\mathbf{v}) \leq -\alpha f(t;\mathbf{v})$ for all t. Hence, by a scalar ODE comparison inequality, $\mathbf{v}^\top \mathbf{N}(t)\mathbf{v} = f(t;\mathbf{v}) \leq e^{-\alpha t} f(0;\mathbf{v}) = e^{-\alpha t} \cdot \mathbf{v}^\top \mathbf{N}(\cdot)\mathbf{v}$. The lemma follows. $\square$

K Smoothness (Proof of Proposition G.5)

This section bounds the first and second derivatives of $\mathcal{L}_\lambda(\mathsf{K})$, and of $\mathsf{K} \mapsto \mathbf{\Sigma}_\mathsf{K}$, for $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$.

Specification of Derivative Norms. To prove Proposition G.5, we formally define the norms of the relevant derivatives. Let $\Delta_\mathsf{K} = (\mathbf{\Delta_A}, \mathbf{\Delta_B}, \mathbf{\Delta_C})$ denote a perturbation of filter $\mathsf{K} = (\mathbf{A}_\mathsf{K}, \mathbf{B}_\mathsf{K}, \mathbf{C}_\mathsf{K})$, with

$$\|\Delta_\mathsf{K}\|_{\ell_2} = \sqrt{\|\mathbf{\Delta_A}\|_\mathrm{F}^2 + \|\mathbf{\Delta_B}\|_\mathrm{F}^2 + \|\mathbf{\Delta_C}\|_\mathrm{F}^2}.$$

Definition K.1 (Euclidean Norm of Derivatives)**.** We define Euclidean norms of the gradient $\nabla\mathcal{L}_\lambda(\mathsf{K})$, operator-norm of the Hessian $\nabla^2\mathcal{L}_\lambda(\mathsf{K})$, and $\ell_2 \to$ op-norm of the gradients of $\mathbf{\Sigma}_\mathsf{K}$ as

$$\begin{aligned}
\|\nabla^2\mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2\to\ell_2} &:= \sup_{\Delta_\mathsf{K}:\|\Delta_\mathsf{K}\|_{\ell_2}=1} \langle \Delta_\mathsf{K}, \nabla^2\mathcal{L}_\lambda(\mathsf{K})\cdot\Delta_\mathsf{K}\rangle \\
\|\nabla\mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2} &= \sup_{\Delta_\mathsf{K}:\|\Delta_\mathsf{K}\|_{\ell_2}=1} \langle \Delta_\mathsf{K}, \nabla\mathcal{L}_\lambda(\mathsf{K})\rangle \\
\|\nabla\mathbf{\Sigma}_\mathsf{K}\|_{\ell_2\to\mathrm{op}} &:= \sup_{\Delta_\mathsf{K}:\|\Delta_\mathsf{K}\|_{\ell_2}=1} \|\nabla\mathbf{\Sigma}_\mathsf{K}\cdot\Delta_\mathsf{K}\|_\mathrm{op}
\end{aligned}$$

We shall compute these bounds by considering directional derivatives, using that

$$\begin{aligned}
\|\nabla^2 \mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2\to\ell_2} &:= \sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1} \left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_\lambda(\mathsf{K}+t\Delta_{\mathsf{K}})\big|_{t=0}\right| \\
\|\nabla \mathcal{L}_\lambda(\mathsf{K})\|_{\ell_2} &= \sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1} \left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_\lambda(\mathsf{K}+t\Delta_{\mathsf{K}})\big|_{t=0}\right| \\
\|\nabla \mathbf{\Sigma}_{\mathsf{K}}\|_{\ell_2\to\mathrm{op}} &:= \sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1} \left\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Sigma}_{(\mathsf{K}+t\Delta_{\mathsf{K}})}\big|_{t=0}\right\|_{\mathrm{op}}
\end{aligned}$$

Stability preliminaries. For any $\mathsf{K}\in\mathcal{K}_{\texttt{info}}$ (and thus $\mathsf{K}\in\mathcal{K}_{\texttt{stab}}$), $\mathbf{A}_{\mathrm{cl},\mathsf{K}}$ is Hurwitz stable, and the solution to the Lyapunov equation $\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}+\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top+\mathbf{Y}=0$ for any $\mathbf{Y}\in\mathbb{S}_+^{2n}$ can be written as

$$\int_0^\infty \exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})\mathbf{Y}\exp(s\mathbf{A}_{\mathrm{cl},\mathsf{K}})^\top \mathrm{d}s,$$

which recall from Proposition 4.3 that satisfies

$$\|\mathbf{\Sigma}_{\mathsf{K},\mathbf{Y}}\|_\circ \le C_{\texttt{lyap}}(\mathsf{K})\cdot\|\mathbf{Y}\|_\circ,$$

for $\|\cdot\|_\circ$ denoting either operator, Frobenius, or nuclear norm. The explicit form of $C_{\texttt{lyap}}(\mathsf{K})$ is given in Proposition J.1.

Covariance derivatives. We start with derivatives of $\mathbf{\Sigma}_{\mathsf{K}}$. Define

$$\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] := \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Sigma}_{\mathsf{K}+t\Delta_{\mathsf{K}}}\Big|_{t=0}, \quad \mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}] := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{\Sigma}_{\mathsf{K}+t\Delta_{\mathsf{K}}}\Big|_{t=0}.$$

We first compute these derivatives. In what follows, given a symmetric matrix $\mathbf{Y}\in\mathbb{S}^n$, we define its nuclear norm as $\|\mathbf{Y}\|_{\mathrm{nuc}} := \sum_{i=1}^n |\lambda_i(\mathbf{Y})|$.

Lemma K.1 (Bounding derivatives of $\mathbf{\Sigma}_{\mathsf{K}}$)**.** *For any $\mathsf{K}\in\mathcal{K}_{\texttt{stab}}$, we have that $\mathsf{K}\to\mathbf{\Sigma}_{\mathsf{K}}$ is $\mathscr{C}^2$ in a neighorhood containing K, and $\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ and $\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ solve the Lyapunov equations*

$$\mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]+\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top+\mathbf{Y}_1[\Delta_{\mathsf{K}}]=0, \quad \mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]+\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top+\mathbf{Y}_2[\Delta_{\mathsf{K}}]=0,$$

where

$$\begin{aligned}
\mathbf{Y}_1[\Delta_{\mathsf{K}}] &= \begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}\mathbf{\Sigma}_{\mathsf{K}}+\mathbf{\Sigma}_{\mathsf{K}}\begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}^\top+\begin{bmatrix}0 & 0\\ 0 & \mathbf{\Delta_B W}_2\mathbf{B}_{\mathsf{K}}^\top+\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{\Delta_B}^\top\end{bmatrix}\\
\mathbf{Y}_2[\Delta_{\mathsf{K}}] &= \begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]+\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}^\top+\begin{bmatrix}0 & 0\\ 0 & \mathbf{\Delta_B W}_2\mathbf{\Delta_B}^\top\end{bmatrix}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} &\le C_{\texttt{lyap}}(\mathsf{K})\cdot\mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}\\
\|\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} &\le C_{\texttt{lyap}}(\mathsf{K})^2\cdot\mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}^2.
\end{aligned}$$

Proof. The existence of the derivatives $\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ and $\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ in open neighbrhoods is standard (see , e.g. [Tang et al., 2021, Lemma B.1]). We compute the derivatives by implicit differentiation.

$$\mathbf{\Sigma}_{\mathsf{K}+t\Delta_{\mathsf{K}}} = \mathbf{A}_{\mathrm{cl},\mathsf{K}+t\Delta_{\mathsf{K}}}\mathbf{\Sigma}_{\mathsf{K}+t\Delta_{\mathsf{K}}}+\mathbf{\Sigma}_{\mathsf{K}+t\Delta_{\mathsf{K}}}\mathbf{A}_{\mathrm{cl},\mathsf{K}+t\Delta_{\mathsf{K}}}^\top+\begin{bmatrix}\mathbf{W}_1 & 0\\ 0 & (\mathbf{B}_{\mathsf{K}}+t\mathbf{\Delta_B})\mathbf{W}_2(\mathbf{B}_{\mathsf{K}}+t\mathbf{\Delta_B})^\top\end{bmatrix}.$$

Differentiating both sides with respect to t and evaluating at $t=0$, we have

$$\begin{aligned}
\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] &= \mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]+\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top\\
&\quad+\underbrace{\begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}\mathbf{\Sigma}_{\mathsf{K}}+\mathbf{\Sigma}_{\mathsf{K}}\begin{bmatrix}0 & 0\\ \mathbf{\Delta_B C} & \mathbf{\Delta_A}\end{bmatrix}^\top+\begin{bmatrix}0 & 0\\ 0 & \mathbf{\Delta_B W}_2\mathbf{B}_{\mathsf{K}}^\top+\mathbf{B}_{\mathsf{K}}\mathbf{W}_2\mathbf{\Delta_B}^\top\end{bmatrix}}_{:=\mathbf{Y}_1[\Delta_{\mathsf{K}}]}.
\end{aligned}$$

Differentiating twice (and notice that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{A}_{\mathrm{cl},\mathsf{K}+t\Delta_{\mathsf{K}}}\big|_{t=0} = 0$), and evaluating at $t = 0$, we have

$$\begin{aligned}\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}] &= \mathbf{A}_{\mathrm{cl},\mathsf{K}}\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}] + \mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{A}_{\mathrm{cl},\mathsf{K}}^\top \\ &\quad \underbrace{+ \begin{bmatrix} 0 & 0 \\ \mathbf{\Delta_B C} & \mathbf{\Delta_A} \end{bmatrix}\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] + \mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\begin{bmatrix} 0 & 0 \\ \mathbf{\Delta_B C} & \mathbf{\Delta_A} \end{bmatrix}^\top + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Delta_B W}_2\mathbf{\Delta_B}^\top \end{bmatrix}}_{:=\mathbf{Y}_2[\Delta_{\mathsf{K}}]}.\end{aligned}$$

To prove the second part of the lemma, we use Proposition J.1. Since $\mathsf{K} \in \mathcal{K}_{\mathtt{info}}$ and thus $\mathsf{K} \in \mathcal{K}_{\mathtt{stab}}$, we know the solutions to the Lyapunov equations above, i.e., $\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ and $\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$, can be written as $\|\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{Y}_1[\Delta_{\mathsf{K}}])\|_\circ$ and $\|\mathsf{clyap}(\mathbf{A}_{\mathrm{cl},\mathsf{K}}, \mathbf{Y}_2[\Delta_{\mathsf{K}}])\|_\circ$, and can be bounded by

$$\begin{aligned}\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} &\le C_{\mathtt{lyap}}(\mathsf{K})\|\mathbf{Y}_1[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \\ &\le C_{\mathtt{lyap}}(\mathsf{K}) \cdot \left(2\|\mathbf{\Sigma}_{\mathsf{K}}\|\left\|\begin{bmatrix} 0 & 0 \\ \mathbf{C\Delta_B} & \mathbf{\Delta_A}\end{bmatrix}\right\|_{\mathrm{F}} + 2\|\mathbf{\Delta_B}\|_{\mathrm{F}}\|\mathbf{B}_{\mathsf{K}}\|\|\mathbf{W}_2\|\right) \\ &\le C_{\mathtt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}.\end{aligned}$$

Using the above computation and recall that $C_{\mathtt{lyap}}(\mathsf{K}) \ge 1$,

$$\begin{aligned}\|\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} &\le C_{\mathtt{lyap}}(\mathsf{K}) \cdot \|\mathbf{Y}_2[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \\ &\le C_{\mathtt{lyap}}(\mathsf{K}) \cdot \left(2\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}\left\|\begin{bmatrix} 0 & 0 \\ \mathbf{C\Delta_B} & \mathbf{\Delta_A}\end{bmatrix}\right\|_{\mathrm{F}} + 2\|\mathbf{\Delta_B}\|_{\mathrm{F}}^2\|\mathbf{W}_2\|\right) \\ &\le C_{\mathtt{lyap}}(\mathsf{K})(1 + \|\mathbf{C}\|)\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}\|\Delta_{\mathsf{K}}\|_{\ell_2} + C_{\mathtt{lyap}}(\mathsf{K})\|\mathbf{W}_2\|\|\Delta_{\mathsf{K}}\|_{\ell_2}^2 \\ &\le C_{\mathtt{lyap}}(\mathsf{K})^2\mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2 + C_{\mathtt{lyap}}(\mathsf{K})\|\mathbf{W}_2\|\|\Delta_{\mathsf{K}}\|_{\ell_2}^2 \\ &\le C_{\mathtt{lyap}}(\mathsf{K})^2 \cdot \mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2,\end{aligned}$$

which completes the proof. □

Derivatives of $\mathtt{OE}$ loss and regularizer. Next, we compute the derivatives of $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ and $\mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$ in terms of the above derivatives.

Lemma K.2 (Bounding derivatives of $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$)**.** *We have that $\mathcal{L}_{\mathtt{OE}}(\cdot)$ is $\mathscr{C}^2$ in the neighborhood of any $\mathsf{K} \in \mathcal{K}_{\mathtt{stab}}$ and*

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\mathtt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\bigg|_{t=0}\right| &\le \sqrt{n} \cdot C_{\mathtt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2} \\ \left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\mathtt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\bigg|_{t=0}\right| &\le C_{\mathtt{lyap}}(\mathsf{K})^2 \cdot \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2.\end{aligned}$$

Proof. Recall from the computaton in Eq. (I.2) that

$$\mathcal{L}_{\mathtt{OE}}(\mathsf{K}) = \mathrm{tr}\left[\begin{bmatrix}\mathbf{G} & -\mathbf{C}_{\mathsf{K}}\end{bmatrix}\mathbf{\Sigma}_{\mathsf{K}}\begin{bmatrix}\mathbf{G}^\top \\ -\mathbf{C}_{\mathsf{K}}^\top\end{bmatrix}\right] = \mathrm{tr}\left[\begin{bmatrix}\mathbf{G}^\top\mathbf{G} & -\mathbf{G}^\top\mathbf{C}_{\mathsf{K}} \\ -\mathbf{C}_{\mathsf{K}}^\top\mathbf{G} & \mathbf{C}_{\mathsf{K}}^\top\mathbf{C}_{\mathsf{K}}\end{bmatrix} \cdot \mathbf{\Sigma}_{\mathsf{K}}\right].$$

Since Lemma K.1 verifies $\mathsf{K} \mapsto \mathbf{\Sigma}_{\mathsf{K}}$ is $\mathscr{C}^2$ in an open neighborhood around K, we readily see $\mathcal{L}_{\mathtt{OE}}(\mathsf{K})$ is as well. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\mathtt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\bigg|_{t=0} = \mathrm{tr}\left[\begin{bmatrix}\mathbf{G}^\top\mathbf{G} & -\mathbf{G}^\top\mathbf{C}_{\mathsf{K}} \\ -\mathbf{C}_{\mathsf{K}}^\top\mathbf{G} & \mathbf{C}_{\mathsf{K}}^\top\mathbf{C}_{\mathsf{K}}\end{bmatrix} \cdot \mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] + \begin{bmatrix} 0 & -\mathbf{G}^\top\mathbf{\Delta_C} \\ -\mathbf{\Delta_C}^\top\mathbf{G} & \mathbf{\Delta_C}^\top\mathbf{C}_{\mathsf{K}} + \mathbf{C}_{\mathsf{K}}^\top\mathbf{\Delta_C}\end{bmatrix} \cdot \mathbf{\Sigma}_{\mathsf{K}}\right].$$

Thus,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\mathtt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\bigg|_{t=0}\right| &\le \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} + \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\|\Delta_{\mathsf{K}}\|_{\ell_2}\|\mathbf{\Sigma}_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \\ &\le \sqrt{n}\mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\left(\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} + \|\Delta_{\mathsf{K}}\|_{\ell_2}\|\mathbf{\Sigma}_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|\right) \\ &\overset{(i)}{\le} \sqrt{n} \cdot \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\Big(C_{\mathtt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \\ &\qquad \cdot (\|\Delta_{\mathsf{K}}\|_{\ell_2} + \|\Delta_{\mathsf{K}}\|_{\ell_2}\|\mathbf{\Sigma}_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|)\Big) \\ &\overset{(ii)}{\le} \sqrt{n} \cdot C_{\mathtt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}\end{aligned}$$

where (i) uses Lemma K.1, and (ii) uses $C_{\texttt{lyap}}(\mathsf{K}) \geq 1$. Next,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\texttt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\Big|_{t=0} = \mathrm{tr}\Big[\begin{bmatrix} \mathbf{G}^\top\mathbf{G} & -\mathbf{G}^\top\mathbf{C}_{\mathsf{K}} \\ -\mathbf{C}_{\mathsf{K}}^\top\mathbf{G} & \mathbf{C}_{\mathsf{K}}^\top\mathbf{C}_{\mathsf{K}} \end{bmatrix} \cdot \mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$$
$$+ \begin{bmatrix} 0 & -\mathbf{G}^\top\mathbf{\Delta}_{\mathbf{C}} \\ -\mathbf{\Delta}_{\mathbf{C}}^\top\mathbf{G} & \mathbf{\Delta}_{\mathbf{C}}^\top\mathbf{C}_{\mathsf{K}} + \mathbf{C}_{\mathsf{K}}^\top\mathbf{\Delta}_{\mathbf{C}} \end{bmatrix} \cdot \mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Delta}_{\mathbf{C}}^\top\mathbf{\Delta}_{\mathbf{C}} \end{bmatrix} \cdot \mathbf{\Sigma}_{\mathsf{K}}\Big].$$

Using Matrix Holder's inequality, it follows that

$$\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\texttt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\Big|_{t=0}\right| \leq \left(\mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\|\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} + \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|)\|\Delta_{\mathsf{K}}\|_{\ell_2}\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} + \|\mathbf{\Sigma}_{\mathsf{K}}\|\|\mathbf{\Delta}_{\mathbf{C}}\|_{\mathrm{F}}^2\right).$$

Again, invoking Lemma K.1 and appropriate simplifications, we have

$$\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\texttt{OE}}(\mathsf{K} + t\Delta_{\mathsf{K}})\Big|_{t=0}\right| \leq C_{\texttt{lyap}}(\mathsf{K})^2 \cdot \mathrm{poly}(\|\mathbf{G}\|, \|\mathbf{C}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2.$$

□

Next, we turn to controlling the derivatives of the regularizer. Here, we require that $\mathsf{K} \in \mathcal{K}_{\texttt{info}}$, not just $\mathsf{K} \in \mathcal{K}_{\texttt{stab}}$ as above. Introduce $\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Z}_{\mathbf{B}_{\mathsf{K}+t\Delta_{\mathsf{K}}}}\Big|_{t=0}$, and define $\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ analogously.

Lemma K.3 (Bounding derivatives of $\mathbf{Z}_{\mathsf{K}}$)**.** *$\mathbf{Z}_{\mathsf{K}}$ is $\mathscr{C}^2$ in a neighborhood of any $\mathsf{K} \in \mathcal{K}_{\texttt{info}}$, and*

$$\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \leq C_{\texttt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}$$
$$\|\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} \leq C_{\texttt{lyap}}(\mathsf{K})^2 \cdot \mathrm{poly}(\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2.$$

Proof. Using $(\mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top)$ and the facts that (a) $\mathsf{K} \mapsto \mathbf{\Sigma}_{\mathsf{K}}$ is $\mathscr{C}^2$ on some neighborhood, and $\mathbf{X} \mapsto \mathbf{X}^{-1}$ is $\mathscr{C}^2$ on $\mathbb{S}^n_{++}$, we see $\mathbf{Z}_{\mathsf{K}}$ is $\mathscr{C}^2$.

To compute derivatives, let us partition the derivatives $\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ and $\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]$ into two-by-two blocks $\mathbf{\Sigma}'_{ij,\mathsf{K}}[\Delta_{\mathsf{K}}]$ and $\mathbf{\Sigma}''_{ij,\mathsf{K}}[\Delta_{\mathsf{K}}]$ in the obvious way. We have

$$\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}] = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Z}_{\mathbf{B}_{\mathsf{K}+t\Delta_{\mathsf{K}}}}\big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{\Sigma}_{12,\mathsf{K}+t\Delta_{\mathsf{K}}}\mathbf{\Sigma}_{22,\mathsf{K}+t\Delta_{\mathsf{K}}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}+t\Delta_{\mathsf{K}}}^\top)\big|_{t=0}$$
$$= \mathbf{\Sigma}'_{12,\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}} + \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}+t\Delta_{\mathsf{K}}}^{-1}(\mathbf{\Sigma}'_{12,\mathsf{K}}[\Delta_{\mathsf{K}}])^\top$$
$$+ \mathbf{\Sigma}_{12,\mathsf{K}}\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\mathbf{\Sigma}'_{22,\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{\Sigma}_{22,\mathsf{K}+t\Delta_{\mathsf{K}}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}}^\top.$$

Thus,

$$\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \leq \mathrm{poly}(\|\mathbf{\Sigma}_{12,\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|)(\|\mathbf{\Sigma}'_{12,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} + \|\mathbf{\Sigma}'_{22,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}})$$
$$\leq \mathrm{poly}(\|\mathbf{\Sigma}_{12,\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|)\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}$$
$$\leq \mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|)\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}.$$

Thus, the intended bound on $\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}$ follows from Lemma K.1. By the same token, more tedious computations reveal,

$$\|\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} = \|\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathbf{\Sigma}_{12,\mathsf{K}+t\Delta_{\mathsf{K}}}\mathbf{\Sigma}_{22,\mathsf{K}+t\Delta_{\mathsf{K}}}^{-1}\mathbf{\Sigma}_{12,\mathsf{K}+t\Delta_{\mathsf{K}}}^\top)\big|_{t=0}\|_{\mathrm{nuc}}$$
$$= \mathrm{poly}(\|\mathbf{\Sigma}_{12,\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|)\left(\|\mathbf{\Sigma}'_{12,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}^2 + \|\mathbf{\Sigma}'_{12,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}\|\mathbf{\Sigma}'_{22,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} + \|\mathbf{\Sigma}''_{12,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} + \|\mathbf{\Sigma}''_{22,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}}\right)$$
$$\leq \mathrm{poly}(\|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|)\left(\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}^2 + \|\mathbf{\Sigma}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}}\right).$$

Thus, the intended bound on $\|\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}}$ follows from Lemma K.1. □

Lemma K.4 (Bounding derivatives of $\mathcal{R}_{\texttt{info}}(\mathsf{K})$)**.** *Recall $\mathcal{R}_{\texttt{info}}(\mathsf{K}) := \mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$. We have*

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\texttt{info}}(\mathsf{K} + t\Delta_{\mathsf{K}})\Big|_{t=0}\right| \leq \sqrt{n}C_{\texttt{lyap}}(\mathsf{K}) \cdot \mathrm{poly}(\|\mathbf{Z}_{\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}$$
$$\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{R}_{\texttt{info}}(\mathsf{K} + t\Delta_{\mathsf{K}})\Big|_{t=0}\right| \leq C_{\texttt{lyap}}(\mathsf{K})^2 \cdot \mathrm{poly}(\|\mathbf{Z}_{\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|, \|\mathbf{\Sigma}_{\mathsf{K}}\|, \|\mathbf{B}_{\mathsf{K}}\|, \|\mathbf{C}\|, \|\mathbf{W}_2\|) \cdot \|\Delta_{\mathsf{K}}\|_{\ell_2}^2.$$

Proof. We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}[\mathbf{Z}_{\mathsf{K}+t\Delta_{\mathsf{K}}}^{-1}]\Big|_{t=0} = -\mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{Z}_{\mathsf{K}}^{-1}].$$

Thus, invoking Lemma K.3,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\big|_{t=0}\right| &\le \|\mathbf{Z}_{\mathsf{K}}^{-1}\|^2\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}} \le \sqrt{n}\|\mathbf{Z}_{\mathsf{K}}^{-1}\|^2\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}\\ &\le \sqrt{n}C_{\texttt{lyap}}(\mathsf{K})\cdot\mathrm{poly}(\|\mathbf{Z}_{\mathsf{K}}^{-1}\|,\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|,\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}.\end{aligned}$$

Next,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\big|_{t=0} = \mathrm{tr}[2\mathbf{Z}_{\mathsf{K}}^{-1}\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{Z}_{\mathsf{K}}^{-1}\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{Z}_{\mathsf{K}}^{-1} + \mathbf{Z}_{\mathsf{K}}^{-1}\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\mathbf{Z}_{\mathsf{K}}^{-1}],$$

so again applying Lemma K.3,

$$\begin{aligned}\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\big|_{t=0}\right| &\le 2\|\mathbf{Z}_{\mathsf{K}}^{-1}\|^3\|\mathbf{Z}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}}^2 + \|\mathbf{Z}_{\mathsf{K}}^{-1}\|^2\|\mathbf{Z}''_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{nuc}}\\ &\le C_{\texttt{lyap}}(\mathsf{K})^2\cdot\mathrm{poly}(\|\mathbf{Z}_{\mathsf{K}}^{-1}\|,\|\mathbf{\Sigma}_{22,\mathsf{K}}^{-1}\|,\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}^2,\end{aligned}$$

we complete the proof. □

Concluding the proof We now turn to the proof of Proposition G.5.

Proof of Proposition G.5. Combining Lemmas K.2 and K.4,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\lambda}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right| &\le \left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\texttt{OE}}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right| + \lambda\cdot\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right|\\ &\le (1+\lambda)\sqrt{n}\cdot C_{\texttt{lyap}}(\mathsf{K})\cdot\mathrm{poly}(\|\mathbf{G}\|,\|\mathbf{C}_{\mathsf{K}}\|,\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|,\|\mathbf{Z}_{\mathsf{K}}^{-1}\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}\end{aligned}$$

and

$$\begin{aligned}\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\lambda}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right| &\le \left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{L}_{\texttt{OE}}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right| + \lambda\cdot\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{R}_{\texttt{info}}(\mathsf{K}+t\Delta_{\mathsf{K}})\Big|_{t=0}\right|\\ &\le (1+\lambda)C_{\texttt{lyap}}(\mathsf{K})^2\cdot\mathrm{poly}(\|\mathbf{G}\|,\|\mathbf{C}_{\mathsf{K}}\|,\|\mathbf{\Sigma}_{\mathsf{K}}\|,\|\mathbf{B}_{\mathsf{K}}\|,\|\mathbf{C}\|,\|\mathbf{W}_2\|,\|\mathbf{Z}_{\mathsf{K}}^{-1}\|)\cdot\|\Delta_{\mathsf{K}}\|_{\ell_2}^2.\end{aligned}$$

These verify the first two bounds of the proposition. The derivative bound for $\mathbf{\Sigma}_{\mathsf{K}}$ is proven in Lemma K.1,noting that

$$\sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1}\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{F}} \ge \sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1}\|\mathbf{\Sigma}'_{\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{op}} \ge \sup_{\Delta_{\mathsf{K}}:\|\Delta_{\mathsf{K}}\|_{\ell_2}=1}\|\mathbf{\Sigma}'_{22,\mathsf{K}}[\Delta_{\mathsf{K}}]\|_{\mathrm{op}} = \|\nabla\mathbf{\Sigma}_{22,\mathsf{K}}\|_{\ell_2\to\mathrm{op}}.$$

Lastly, we have shown above that $\mathcal{L}_{\texttt{OE}}(\mathsf{K})$ and $\mathbf{Z}_{\mathsf{K}}$ is $\mathscr{C}^2$ in a neighborhood of any $\mathsf{K}\in\mathcal{K}_{\texttt{info}}$. Since $\mathbf{Z}_{\mathsf{K}}$ is invertible on $\mathsf{K}\in\mathcal{K}_{\texttt{info}}$, this implies that $\mathcal{L}_{\lambda} = \mathcal{L}_{\texttt{OE}}(\mathsf{K}) + \lambda\mathrm{tr}[\mathbf{Z}_{\mathsf{K}}^{-1}]$ is $\mathscr{C}^2$ in a neighborhood of any $\mathsf{K}\in\mathcal{K}_{\texttt{info}}$. □