Geometric methods for optimal sensor design

M.-A. Belabbas

Abstract

An observer is an estimator of the state of a dynamical system from noisy sensor measurements. The need for observers is ubiquitous, with applications in fields ranging from engineering to biology to economics. The most widely used observer is the Kalman filter, which is known to be the optimal estimator of the state when the noise is additive and Gaussian. Because its performance is limited by the sensors to which it is paired, it is natural to seek an optimal sensor for the Kalman filter. The problem is however not convex and, as a consequence, many ad hoc methods have been used over the years to design sensors. We show in this paper how to characterize and obtain the optimal sensor for the Kalman filter. Precisely, we exhibit a positive definite operator which optimal sensors have to commute with. We furthermore provide a gradient flow to find optimal sensors, and prove the convergence of this gradient flow to the unique minimum in a broad range of applications. This optimal sensor yields the lowest possible estimation error for measurements with a fixed signal to noise ratio. The results presented here also apply to the dual problem of optimal actuator design.

Since the early work of Kalman, Bucy [18, 19] and Stratonovich [31], the estimation of linear systems has expanded its range of applications from its engineering roots [29] to fields such as environmental engineering, where for example it is used to estimate sea-level change [14]; financial engineering, where for example it is used to estimate the realized volatility error [5] or to price energy futures [25]; to economics [2], social science [24], process control [27] or even biology [6]. The common thread to these applications is that one cannot observe exactly all internal variables of a system, but instead needs to estimate them from partial, noisy measurements coming from a set of sensors. We address in this paper the optimal design of such sensors.

There are well-developed methods in control theory to design estimators of the state of a dynamical system based on sensor measurements; such estimators

---

*University of Illinois, Urbana-Champaign, USA. Email: belabbas@illinois.edu
are called observers of the system. It stands to reason that as the signal to noise ratio of the measurements increases, the estimation error afforded by an observer decreases. The question of interest is thus to find which measurements with a given signal to noise ratio are optimal for an observer, optimal in the sense that the estimation error is minimized. Because the Kalman filter is the minimum mean square estimator of the state \([1]\), optimal measurements for the Kalman filter yield the lowest estimation error which one can obtain for a given signal to noise ratio. We call the sensor providing such measurements optimal.

The optimal sensor design problem for Kalman filters is almost as old as the Kalman filter itself, and over the years a variety of methods have been proposed, we refer the reader to the recent thesis \([28]\) for a survey. The major obstacle encountered is that the optimization problem, formulated precisely below, defining an optimal sensor is not convex. To sidestep this obstacle, suboptimal solutions obtained by way of convex relaxations or ad-hoc heuristics for specific application are often used \([30, 27]\). Another approach of choice is to focus on a convex performance measure \([13, 10]\) or optimize bounds for the estimation error \([23]\).

There is also a large literature discussing the properties of, and numerical methods for, optimal sensor/actuator placement in infinite dimensional spaces, see \([12, 26]\) and references therein.

In this paper, we provide an exact characterization of the optimal sensors for Kalman filters by exhibiting a positive definite matrix they have to commute with. We furthermore provide a gradient algorithm—in fact, a Lax equation \([21]\)—to find such optimal sensors and prove its convergence to the global optimum in a broad range of situations. Finally, we demonstrate the efficacy of the methods proposed with simulations and provide a rule of thumb for choosing sensors that work best in low signal to noise ratio settings. We also believe that the geometric analysis provided here sheds light on the intrinsic difficulty of the problem, difficulty that arises because the constraints on the number of observation signals and their signal to noise ratio are not convex. The optimal sensor design problem is equivalent to an optimal actuator placement problem, which we discuss in details below.

Closely related problems which might benefit from the point of view presented here include the optimal scheduling and design of the measurements \([16]\), the joint optimal measurement and control design \([4]\) or the control of complex systems \([23]\).

We now describe the problem and our results precisely. We start with a few conventions used throughout the paper.

All square matrices are real \(n \times n\) matrices unless otherwise specified. We
denote by $I_n$ the $n \times n$ identity matrix, by $\Omega_{ij}$ the skew-symmetric matrix with zero entries everywhere except for the $ij$th and $ji$th ones, which are 1 and $-1$ respectively, and by $\Sigma_{ij}$ the symmetric matrix with zero entries everywhere except for the $ij$th and $ji$th ones, which are both one. We simply say norm of a vector to refer to its 2-norm. For $J$ a differential function on a manifold and $X$ a vector field on the same manifold, we denote by $X \cdot J = dJ \cdot X$ the directional derivative of $J$ along $X$. We denote by $\mathbb{R}^+$ the set of strictly positive real numbers. We use the upper-case letters to denote matrices and lower-case letters for vectors and scalars.

**Optimal sensor design.** Consider the linear stochastic differential equation

$$
\begin{cases}
    dx(t) = Ax(t)dt + Gd\omega(t) \\
    dy(t) = cx(t)dt + dv(t)
\end{cases}
$$

where $\omega(t)$ and $v(t)$ are independent Wiener processes and $A \in \mathbb{R}^{n \times n}, G \in \mathbb{R}^{n \times r}, c \in \mathbb{R}^{1 \times n}$ and $d \in \mathbb{R}$. The process $x(t)$ is called the state process and $y(t)$ the observation process. The vector $c$ is the sensing or observation vector of the system. By an estimator for $x$ is meant a dynamical system with input $y(t)$ and whose state, call it $\hat{x}$, is an estimate of $x$.

The Kalman filter is the optimal, in the mean-squared sense, estimator of the state $x(t)$ given past observations $y([0,t])$. Given the matrices $A, G$ and $c$ as above, the Kalman filter in steady state is

$$
d\hat{x}(t) = A\hat{x}(t)dt - Kc^\top(dy(t) - c\hat{x}(t)dt) + bu(t)dt
$$

where the matrix $K$ is the symmetric positive definite solution of the following Riccati equation:

$$
KA^\top + AK - Kc^\top cK + GG^\top = 0.
$$

Not all sensing matrices are equal for the purpose of estimation. In fact, if $c$ is such that the pair $(A, c)$ is not observable [7], one might not be able to design a convergent observer. Even more, it is not too hard to convince oneself that as the norm of $c$ increases, all other things being equal, the estimation error will decrease [32]. Keeping these observations in mind, it is natural to seek the best sensing matrix of a given norm. To make the statement more precise, denote by $\mathbb{E}$ the expectation operator. Recall that the gain matrix $K$ is also the steady-state covariance of the estimation error [11] $K = \mathbb{E}((x - \hat{x})(x - \hat{x})^\top)$ and thus the trace of $K$ is nothing else than the MSE estimation error:

$$
\text{tr}(K) = \sum_i \mathbb{E}((x_i - \hat{x}_i)^2).
$$
We are thus led to the following optimal sensor design problem: minimize the trace of $K$, where $K$ obeys Eq. (2) and the norm of $c$ is fixed.

**Optimal actuator design.** In view of the well-known duality between observability and controllability, it is not surprising that the optimal actuator design problem takes a formulation similar to the optimal sensor design's. To wit, consider the linear time-invariant system

$$\dot{x} = Ax + bu.$$  

(3)

An optimal linear quadratic controller is a controller which minimizes the cost functional

$$J(x) = \int_0^\infty \left( x(t)^TQx(t) + u^2(t) \right) dt,$$

given $x(0) = x$ and for a user-selected positive definite matrix $Q$. The optimal controller is a feedback controller of the form $u = -b^TKx$ where $K$ obeys the Riccati equation

$$A^TK + KA - Kbb^TK + Q = 0.$$

One can show that starting from an initial condition $x_0$, the “cost of return to zero” with the above controller is $J(x) = x^TKx$. Thus, the expected cost of return to zero for an initial condition $x$ distributed according to an arbitrary rotationally invariant distribution with density $g(r)dr$, where $r = \|x\|$, is

$$\mathbb{E}J = \text{tr}(K) \int_0^\infty g(r)dr.$$

The question thus arises of finding the actuator $b$ that minimizes the trace of $K$. This actuator is the one returning the system to its desired state with the least effort on average. As the norm of $b$ increases, the trace of $K$ decreases and we shall thus fix the norm of $b$. By looking at a broader class of optimization functions below, our results also handle non-rotationally invariant distributions on the initial conditions.

**Main results.** Having shown that optimal sensor and actuator design can both be cast as minimizing the trace of the positive definite solution of the Riccati equation, we pose the following optimization problem, which slightly generalizes the statement introduced above. We adopt the point of view of sensor design, and thus look for an optimal sensing matrix $c$. Without loss of generality, we will represent a sensing matrix by $\sqrt{\gamma}c$ where $\|c\| = 1$ and $\gamma > 0$. 


Throughout this paper, we use the notation

\[ C := c^\top c. \]

With a slight abuse of language, which will be justified below, we will refer to both \( C \) and \( c \) as observation matrices. Note that the spectral decomposition of \( C \) yields \( c \) unambiguously. We define the cost function

\[ J(\gamma, c) := \text{tr}(LK). \] (4)

where \( K \) satisfies the Riccati equation

\[ AK + KA^\top - \gamma K c^\top c K + Q = 0 \]

and \( L \) and \( Q \) are positive definite matrices. Note that \( J(\gamma, c) = J(\gamma, -c) \). We call an observation vector \( c \) optimal if it is a global minimizer of \( J(\gamma, c) \) for \( \gamma \) fixed, and we call it extremal if it is a singular point of \( J(\gamma, c) \), but not necessarily a minimum. We let \( [A, B] = AB - BA \) be the commutator of two matrices \( A \) and \( B \).

A square matrix is called stable if its eigenvalues have negative real parts. For \( c \in \mathbb{R}^{1 \times n} \), we denote by \( \text{span} c \) the subspace of \( \mathbb{R}^n \) spanned by \( c \). A subspace \( V \) of \( \mathbb{R}^n \) is an invariant subspace of \( M \in \mathbb{R}^{n \times n} \) if \( M V \subset V \). If \( M \) is symmetric positive definite, all its \( p \)-dimensional invariant subspaces are spanned by \( p \) eigenvectors of \( M \). We refer to the invariant subspace of \( M \) spanned by the eigenvector corresponding to the largest (resp. \( p \)-th largest) eigenvalue as the highest (resp. \( p \)-th highest) invariant subspace of \( M \). The main results of the paper are the following:

1. An observation vector \( c \in \mathbb{R}^{1 \times n} \) is extremal if \( \text{span} c \) is an eigenspace of the positive definite matrix

\[ M := KRK \] (5)

where \( K \) and \( R \) are the positive definite solutions of the equations

\[ A^\top K + KA - \gamma K c K + Q = 0 \]

\[ (A - \gamma c K)R + R(A - \gamma c K)^\top + L = 0. \]

2. Generically\(^1\) for \( L, Q \) symmetric positive definite, \( \gamma > 0 \) small and \( A \) stable, there is a unique (up to a sign) optimal observation vector \( c \). It spans the highest invariant subspace of \( M \).

\(^1\)By generically, we mean everywhere except on a set of codimension at least one.
3. With the same assumptions as in item 2, the Lax equation

\[ \frac{d}{dt} C = -[C, [C, M]], \]

with \( K, R \) as above converges from a set of measure one of initial conditions to an optimal observation matrix.

1 Optimal sensor and actuator placement.

On the Riccati equation. The Riccati equation plays a central role in the theory of linear systems, and much has been written about its properties. We only mention here, and without proof, the facts needed to prove our results. A pair \((A, c)\) is called detectable if there exists a matrix \( D \) such that \( A - c^T D \) is stable. If the pair \((A, c)\) is observable \([7]\), it is detectable and if a pair \((A, c)\) is detectable and \( Q \) is positive definite, the Riccati equation \( A^T K + KA - \gamma KCK + Q = 0 \) has a unique positive-definite solution. Moreover, this solution is such that \( A - CK \) is a stable matrix \([7]\). In this paper, we will restrict our attention to stable matrices \( A \), in which case the pair \((A, c)\) is detectable regardless of \( c \). We discuss this assumption in the last section.

We gather the facts needed in the following result, which is essentially \([11]\).

**Lemma 1.** Let \( A \) be a stable matrix and \( Q \) a symmetric positive definite matrix. Let \( C \in \mathbb{R}^{n \times n} \) be symmetric positive definite with \( \|C\| = 1 \) and let \( \gamma \geq 0 \). Then the positive definite solution \( K \) of the Riccati equation \( A^T K + KA - \gamma KCK + Q = 0 \) is analytic with respect to \( C \) and \( \gamma \).

Real projective space and isospectral matrices. Denote by \( SO(n) \) the special orthogonal group, that is the set of matrices \( \Theta \in \mathbb{R}^{n \times n} \) such that \( \Theta^T \Theta = I_n \) and \( \det(\Theta) = 1 \). We denote by \( so(n) = \{ \Omega \in \mathbb{R}^{n \times n} \mid \Omega^T = -\Omega \} \) the Lie algebra of \( SO(n) \) and use the notation

\[ \text{ad}_C A := [C, A] := CA - AC. \]

Let \( \Lambda \) be a diagonal matrix. We denote by \( \text{Sym}(\Lambda) \) the orbit of the special orthogonal group \( SO(n) \) acting on \( \Lambda \) by conjugation; that is

\[ \text{Sym}(\Lambda) = \{ C \in \mathbb{R}^{n \times n} \mid C = \Theta^T \Lambda \Theta \text{ for } \Theta \in SO(n) \}. \]

The set \( \text{Sym}(\Lambda) \) is the set of all real symmetric matrices which can be diagonalized to \( \Lambda \). We call \( \text{Sym}(\Lambda) \) an **isospectral manifold**. A simple computation
shows that its tangent space $T_C \text{Sym}(\Lambda)$ at a point $C$ is the following vector space:

$$T_C \text{Sym}(\Lambda) = \{ [C, \Omega] = \text{ad}_C \Omega \mid \Omega \in \mathfrak{so}(n) \}.$$  

(6)

We will only need the case here of $\Lambda$ having all entries zero except for one entry, which is one. Since we clearly have that $\text{Sym}(\Lambda) = \text{Sym}(\Lambda')$ if and only if $\Lambda$ and $\Lambda'$ are conjugate, we can denote such isospectral manifold by $\text{Sym}(n)$. Note that if $C \in \text{Sym}(n)$, then $C^2 = C$ and $C$ is of rank 1. Thus $\text{Sym}(n)$ can be thought of as the space of rank 1 orthogonal projectors in $\mathbb{R}^n$. In particular, $\text{Sym}(n)$ is homeomorphic to the real projective space $\mathbb{RP}(n-1)$. We thus have $\dim(\text{Sym}(n)) = n - 1$.

We now define the function (see Eq. (4))

$$\bar{J}(\gamma, C) : \mathbb{R}^+ \times \text{Sym}(n) \mapsto \mathbb{R} : (\gamma, C) \mapsto \text{tr}(LK).$$

Note that given $C$, one can obtain $c$ uniquely via spectral decomposition. With a slight abuse of notation, we will omit the bar over $J$ and write $J(\gamma, C)$ as well.

**The normal metric.** The manifold $\text{Sym}(\Lambda)$ possesses a natural metric called the normal metric or Einstein metric. The main idea behind the definition of the normal metric, which has already been used in engineering applications [15, 8, 9], is to embed $\text{Sym}(\Lambda)$ in the Lie algebra $\mathfrak{su}(n)$ and use the so-called Killing form [20] on $\mathfrak{su}(n)$. Note that because $\Lambda$ is not an element of $\mathfrak{so}(n)$, $\text{Sym}(\Lambda)$ is not an adjoint orbit [3] of $SO(n)$. Furthermore, we wish to include the case of $\Lambda$ having repeated entries, which implies that the operator $[C, \cdot]$ (or $\text{ad}_C$, as defined above) acting on $\mathfrak{so}(n)$ is not invertible. We briefly sketch a construction of the normal metric here that emphasizes the properties we shall need below. We refer the reader to the appendix for a more careful construction.

Denote by $\text{co ad}_C$ the codomain of $\text{ad}_C$ and by $\ker \text{ad}_C \subset \mathfrak{so}(n)$ the kernel of $\text{ad}_C$. Note that $\text{co ad}_C = T_C \text{Sym}(\Lambda)$. The bilinear operator

$$\kappa : \mathfrak{so}(n) \times \mathfrak{so}(n) : (\Omega_1, \Omega_2) \mapsto -\text{tr}(\Omega_1 \Omega_2)$$

is symmetric and positive definite. It can thus be used to define the orthogonal complement $(\ker \text{ad}_C)^{\perp}$ of $\ker \text{ad}_C$ in $\mathfrak{so}(n)$, which we identify with $\mathfrak{so}(n)/\ker \text{ad}_C$. With the above, we can define the invertible map

$$\tilde{\text{ad}}_C : \mathfrak{so}(n)/\ker \text{ad}_C \mapsto \text{co ad}_C.$$

The normal metric $\kappa_n$ is defined as
\[ \kappa_n(X, Y) := -\text{tr}(\text{ad}_C^{-1} X \text{ad}_C^{-1} Y) \] (7)
for \( X, Y \in T_C \text{Sym}(\Lambda) \). One can show that the normal metric is positive definite and non-degenerate. Moreover, we have that
\[ \text{ad}_C \text{ad}_C^{-1} X = X. \] (8)

Another property we shall need below is the \textit{ad-invariance} of the trace, which refers to the following relation:
\[ \text{tr}((\text{ad}_C \Omega_1) \Omega_2) = -\text{tr}(\Omega_1 \text{ad}_C \Omega_2). \] (9)

### 1.1 Gradient flow for optimal sensor design

We now evaluate the gradient flow of \( J = \text{tr}(LK) \) with respect to the normal metric. Fix \( C \in \text{Sym}(n) \) such that \((A, C)\) is detectable and let \( K \) be the corresponding solution of the Riccati equation. Recall that the gradient of \( J \) evaluated at \( C \), denoted by \( \nabla J(C) \), obeys the relation [17]
\[ \kappa_n(\nabla J(C), X) = dJ \cdot X, \text{ for all } X \in T_C \text{Sym}(\Lambda). \] (10)

Let \( C(t) \), be a differentiable curve in \( \text{Sym}(\Lambda) \) defined for \(|t| < \varepsilon\) small and such that \( C(0) = C \) and \( \frac{d}{dt} \bigg|_{t=0} C(t) = X \). We can choose \( \varepsilon \) small enough so that \((A, C(t))\) is detectable for all \(|t| < \varepsilon\). From Lemma [1] we conclude that for all such \( t \), there exists a unique positive definite solution \( K(t) \) to the algebraic Riccati equation \( A^\top K(t) + K(t)A - \gamma K(t)C(t)K(t) + Q = 0 \) and that the curve \( K(t) \) is differentiable in \( t \). Then
\[ dJ \cdot X = \left. \frac{d}{dt} \right|_{t=0} J(C(t)) = \text{tr}(L \left. \frac{d}{dt} \right|_{t=0} K(t)) \] (11)

Differentiating the Riccati equation, and writing \( \dot{K} \) for \( \frac{d}{dt} \bigg|_{t=0} K \), we obtain
\[ A^\top \dot{K} + \dot{K}A - \gamma \dot{K}CK - \gamma K\dot{K}C - \gamma KC\dot{K} = 0. \]
The above equation is a Lyapunov equation [7], which we can write as
\[ (A - \gamma CK)^\top \dot{K} + \dot{K}(A - \gamma CK) - \gamma KXX = 0 \] (12)
and whose solution is
\[ \dot{K} = -\gamma \int_0^\infty e^{(A - \gamma CK)t} KXXe^{(A - \gamma CK)t} \, dt. \] (13)
Using Eqs. (10), (11) and (13) we obtain
\[
\text{tr}(\tilde{\text{ad}}^{-1}_C \nabla J \tilde{\text{ad}}^{-1}_C X) = \gamma \text{tr}(L \int_0^\infty e^{(A-\gamma CK)^\top} K X e^{(A-\gamma CK)t} dt)
\]

From (6), we can write \(X = \text{ad}_C \Omega\) for some \(\Omega \in (\ker \text{ad}_C)^\perp\). Using the cyclic and ad-invariance properties of the trace, the last equation can be rewritten as
\[
\text{tr}(\tilde{\text{ad}}^{-1}_C (\nabla J) \Omega) = \gamma \text{tr}(\text{ad}_C \left[ K \int_0^\infty e^{(A-\gamma CK)t} L e^{(A-\gamma CK)^\top} dt K \right] \Omega).
\]
The above equation holds for all \(\Omega \in (\ker \text{ad}_C)^\perp\) and thus
\[
\tilde{\text{ad}}^{-1}_C \nabla J = \gamma \text{ad}_C \left[ K \int_0^\infty e^{(A-\gamma CK)t} L e^{(A-\gamma CK)^\top} dt K \right] + \Delta
\]
for some \(\Delta \in ((\ker \text{ad}_C)^\perp)^\perp = \ker \text{ad}_C\). Taking \(\text{ad}_C\) on both sides of the last relation, we obtain \(\nabla J = \gamma \text{ad}_C \text{ad}_C K R K R\) where \(R\) is the solution of the Lyapunov equation \((A - \gamma CK) R + R(A - \gamma CK)^\top + L = 0\).

**Theorem 1.** The gradient of the function \(J(\gamma, C) = \text{tr}(LK)\) with respect to the normal metric and for \(\gamma > 0\) fixed is
\[
\nabla J = \gamma [C, [C, M]]
\]
where \(M = KRK\) and \(K\), \(R\) obey the equations
\[
A^\top K + KA + Q - \gamma KCK = 0 \\
(A - \gamma CK) R + R(A - \gamma CK)^\top + L = 0.
\]

Moreover, an observation matrix \(C \in \text{Sym}(n)\) is extremal if it is an orthogonal projection onto an invariant subspace of \(M\). Equivalently, an observation vector \(c \in \mathbb{R}^{1 \times n}\) of norm \(\gamma\) is extremal if \(\text{span} \ c\) is an eigenspace of \(M\).

**Proof.** The first part of the statement was proven above. We thus focus on the second part. Recall that extremal points of \(J\) are zeros of its gradient, and thus \(C\) is extremal if and only if \([C, M] = 0\). Because the positive definite solution of the Riccati equation is such that the matrix \((A - \gamma CK)\) is stable and because \(L\) is symmetric positive definite, we have that \(R\) is symmetric positive definite and thus so is the product \(M\). The result is now a consequence of the fact that symmetric matrices commute if and only if they have the same invariant subspaces.

\[\square\]
Remark 1. It is tempting to conjecture that if $c_1$ is an extremal observation vector, and $K_1$ and $R_1$ are the corresponding solutions of Eq. (14) above, then any eigenvector of $K_1R_1K_1$ is also extremal. This however is not the case.

Remark 2. Note that both $R$ and $K$ depend on the parameter $\gamma$.

1.2 The extremal points of $J$

We have derived in the previous section the gradient of $J$. Because $J$ is a lower-bounded function defined on a compact domain, it is clear that the gradient flow will converge to the set of extremal points of $J$. However, $J$ is not convex and thus we do not have, a priori, convergence to the global minimum of $J$. We show that for $\gamma$ small $J$ has a unique minimum, a unique maximum and that the other extremal points are finite in number and saddle points. This shows that, in that regime, the gradient flow will essentially converge to the global minimum. We will discuss in the last section how small $\gamma$ needs to be in practice.

To this end, we need to evaluate the Hessian of $J$. Let $\langle \cdot, \cdot \rangle$ be a Riemannian metric on $\text{Sym}(n)$. Recall that the Levi-Civita connection is the unique connection that is compatible with the metric and torsion free [17]. We denote by $d^2J$ the Hessian of the function $J$. It is a symmetric, bilinear form on $T\text{Sym}(n)$ and for the vector fields $X$ and $Y$, it is given by [17]

$$d^2J(X,Y) = X \cdot Y \cdot J - \nabla_X Y \cdot J. \tag{15}$$

The Hessian thus depends on the choice of a particular connection. We call the Hessian with respect to the normal metric the bilinear form (15) where $\nabla$ is the Levi-Civita connection of the normal metric.

Remark 3. The choice of connection does not affect the type of extremal points of $J$ of course, but it is convenient to fix a connection for the perturbation argument that will be used below. Also note that the Hessian can be used to accelerate gradient flows or algebraic equation solvers [33].

Let $C \in \text{Sym}(n)$ and $X, Y \in T_C\text{Sym}(n)$ such that $X = \text{ad}_C \Omega_x, \ Y = \text{ad}_C \Omega_y$. We write $d^2J(\Omega_x, \Omega_y)$ for the Hessian of $J$ at $C$ evaluated at $(X, Y)$. We can show (see Proposition 1 in the appendix) that the Hessian of $J$ with respect to the normal metric is

$$d^2J(\Omega_1, \Omega_2) = \gamma \text{tr} \left\{ [C, \Omega_1][M, \Omega_2] + [C, VRK + KWK + KRV]\Omega_2 - \frac{1}{2} [C, M][\Omega_1, \Omega_2] \right\} \tag{16}$$

10
where \( K \) and \( R \) are as in the statement of Theorem 1 and \( V \) and \( W \) are defined in Eqs. (26) and (27).

We now show that for \( \gamma \) small, \( J \) has exactly \( n \) extremal points, and only one of them is stable. To this end, denote by \( C_i(\gamma), i \in I(\gamma) \) the zeros of \( \nabla J(\gamma, C) / \gamma \) indexed by the index set \( I(\gamma) \). We now show that for \( \gamma > 0 \) small, and for generic \( Q, L \), the set \( I(\gamma) \) has cardinality \( n \). Recall that the signature of an extremal point \( C \) of \( J \) is a triplet of integers \((n_+, n_-, n_0)\), where \( n_+ \) (resp. \( n_- \) and \( n_0 \)) denotes the number of positive (resp. negative, zero) eigenvalues of the Hessian of \( J \) at \( C \). We have the following result, which covers items 2 and 3 of the main result.

**Theorem 2.** Let \( A \) be a stable matrix. For \( \gamma > 0 \) small and generically for \( Q, L \) positive definite matrices, the function \( J(\gamma, C) : \mathbb{R}^+ \times \text{Sym}(n) \to \mathbb{R} : J(\gamma, C) = \text{tr}(LK) \)

where \( K \) satisfies the Riccati equation (14) has exactly \( n \) extremal points, with signatures \((n - 1, 0, 0), (n - 2, 1, 0), \ldots, (0, n - 1, 0)\) respectively. Moreover, the extremal point of signature \((n - p, p - 1, 0)\) is the orthogonal projection matrix onto the \( p \)-th highest invariant subspace of \( M \).

We prove the result in two steps: first, we show that there are \( n \) extremal points for \( \gamma \) small and then we evaluate their signatures. The proof of the first item goes by studying the parametrized family of vector fields \( F(\gamma, c) \). When \( \gamma > 0 \), \( F \) and \( \nabla J \) clearly have the same zeros. We then show that \( F(0, c) \) has exactly \( n \) zeros and that these zeros persist for \( \gamma > 0 \) small.

**Proof.** We introduce the following rescaled gradient of \( J \):

\[
F \colon [0, \infty) \times \text{Sym}(n) \to T\text{Sym}(n) \colon (\gamma, C) \mapsto [C, M]
\]

where \( M = KRRK \) with \( R \) and \( K \) obeying Eq. (14). One should think of \( F(\gamma, C) \) as a parametrized family of vector fields on \( \text{Sym}(n) \). We denote by \( K_0 \) and \( R_0 \) the solutions of \( A^\top K_0 + K_0 A + Q = 0 \) and \( A R_0 + R_0 A^\top + L = 0 \) respectively. Set

\[
M_0 := K_0 R_0 K_0.
\]

For \( \gamma = 0 \) and generically for \( Q \) and \( L \) positive definite, we conclude from Lemma 3 that \( M_0 \) has distinct eigenvalues. Thus, there are exactly \( n \) matrices of rank one which commute with \( M_0 \). Thus \( I(0) \) contains \( n \) elements, say \( I(0) = \{1, 2, \ldots, n\} \). Denote by \( C_i(0) \) the corresponding zeros of \( F \).

We denote by \( \nabla F \) the covariant derivative of \( F \) where \( \nabla \) is the Levi-Civita connection associated to the normal metric. In order to show that for \( \gamma > 0 \),
small enough, $\mathcal{S}(\gamma) = \mathcal{S}(0)$, it is sufficient to show that the linear map $\nabla F : T_C \text{Sym}(n) \rightarrow T_C \text{Sym}(n) : X \mapsto \nabla_X F(0,C_i)$ is non-degenerate at the $n$ points $(0,C_i(0))$. From Lemma 5 and for $\Omega_x$ such that $X = [C, \Omega_x]$ we have

$$\nabla_X F = -\frac{1}{2} ([M_0, [C, \Omega_x]] + [\Omega_x, [C, M_0]]).$$

When $C$ is an eigenvector of $M_0$, the second term vanishes. We are left with

$$\nabla_X F(0,C_i) = -\frac{1}{2} [M_0, [C, \Omega_x]].$$

We now show that the covariant derivative is non-degenerate. For this, we need the following two facts: first, for any orthogonal matrix $\Theta \in SO(n)$, the conjugation map $\text{Ad}_\Theta : \mathfrak{so}(n) \rightarrow \mathfrak{so}(n) : \Omega \mapsto \Theta \Omega \Theta^{-1}$ has $\text{Ad}_\Theta^{-1}$ for inverse and is consequently surjective onto $\mathfrak{so}(n)$. Second, for arbitrary matrices $\Theta \in SO(n)$ and $A, B \in \mathbb{R}^{n \times n}$,

$$\text{Ad}_\Theta [A, B] = [\text{Ad}_\Theta A, \text{Ad}_\Theta B]. \quad (17)$$

Since $M_0$ has distinct eigenvalues, it has exactly $n$ eigenvectors. Using these facts, $\nabla_X F(0,C_i)$ is non-degenerate if and only if the linear map

$$X \mapsto \text{Ad}_\Theta \nabla_X F(0,C_i) = [\Theta M_0 \Theta^T, [\Theta C \Theta^T, \Omega_x]]$$

is non-degenerate.

Let $\Theta$ be the orthogonal matrix whose columns are the eigenvectors of $M_0$. With this choice of $\Theta$, the previous equation reduces to

$$\text{Ad}_\Theta \nabla_X F(0,C_i) = [D, [E, \Omega_x]] = \text{ad}_D \text{ad}_E \Omega_x,$$

where $E$ is a matrix with zero entries except for one diagonal entry which is equal to 1 and $D$ is a diagonal matrix with the eigenvalues of $M_0$ on its diagonal. Without loss of generality, we can assume that the diagonal entries of $D$ are sorted in decreasing order. A short calculation shows that the commutator of a diagonal matrix with diagonal entries $d_i$ and a matrix $A = (a)_{ij}$ has entry $ij$ equal to $a_{ij}(d_i - d_j)$. We deduce from this calculation that $\text{ad}_D$ is full rank. Thus $\text{Ad}_\Theta \nabla_X F(0,C_i)$ is of full rank. We have thus shown that for $\gamma$ small, the function $J$ has exactly $n$ extremal points.

The proof of the second part of the Theorem relies on the expansion of the Hessian of $J$ described in Corollary 1 in the appendix. Namely, it says that the Hessian of $J$ with respect to the normal metric has the following expansion around $\gamma = 0$:  

12
\[ d^2 J(\Omega_1, \Omega_2) \approx \gamma \text{tr} \left \{ [C, \Omega_1][M_0, \Omega_2] - \frac{1}{2} [C, M_0][\Omega_1, \Omega_2] \right \} + \gamma^2 T(\Omega_1, \Omega_2) \quad (18) \]

where \( \Omega_1, \Omega_2 \in \text{so}(n) \) and the bilinear form \( T \) contains terms of zeroth and higher orders in \( \gamma \).

Recall that for \( X_1, X_2 \in T_C \text{Sym}(n) \), there exists \( \Omega_1, \Omega_2 \in \mathfrak{Z}(C)^\top \) such that \( \text{ad}_C \Omega_i = X_i \) for \( i = 1, 2 \). With this notation, we know from Corollary [1] that the dominating term in the Hessian of \( J \) for \( \gamma > 0 \) small is the following bilinear form on \( T_C \text{Sym}(n) \):

\[ \gamma \text{tr} \left \{ [C, \Omega_1][M_0, \Omega_2] - \frac{1}{2} [C, M_0][\Omega_1, \Omega_2] \right \}. \]

We want to evaluate the signature of the \( n \) Hessians obtained by letting \( C = C_i(\gamma) \) for \( \gamma \) small. We will evaluate the signature of the forms at the points \( C_i(0) \) and observe that no eigenvalues are zero. Whence a standard argument using continuity of the eigenvalues with respect to \( \gamma \) shows that the signatures at the extremal points \( C_i(\gamma) \) for \( \gamma \) small are the same as the ones at the points \( C_i(0) \). This will conclude the proof.

Note that the second term in the expression above vanishes at the \( C_i(0) \), and multiplication by \( \gamma > 0 \) clearly does not affect the signature, thus the signature of the Hessian of \( J \) for \( \gamma \) small is given by the signature of \( \text{tr} \left \{ [C, \Omega_1][M_0, \Omega_2] \right \} \).

We can simplify the problem further as follows: as was done in the first part of the proof, let \( \Theta \) be the orthogonal matrix whose columns contain the eigenvectors of \( M_0 \). The cyclic invariance of the trace, Eq. (17) and the fact that \( \text{Ad}_\Theta \) is an isomorphism on \( \text{so}(n) \) together imply that the bilinear form

\[ H : T_E \text{Sym}(n) \times T_E \text{Sym}(n) \rightarrow \mathbb{R} : (\Omega_1, \Omega_2) \rightarrow \text{tr} \left \{ [E, \Omega_1][D, \Omega_2] \right \}, \quad (19) \]

where \( D \) is a diagonal matrix with the eigenvalues of \( M_0 \) on the diagonal and \( E \) is a diagonal matrix with one entry equal to 1 and \( n - 1 \) equal to 0, has the same signature as the Hessian of \( J \) at extremal points.

We thus have to evaluate the signatures of the \( n \) bilinear forms obtained by letting \( E = \Sigma_{jj} \), for \( j = 1, \ldots, n \), in \( H \). With the ordering of the entries of \( D \) chosen, the signature at \( \Sigma_{jj} \) corresponds to the signature at the extremal point \( C_j(\gamma) \), where \( C_j(\gamma) = c_j c_j^\top \) and \( c_j \) is the eigenvector associated to the \( j \)th largest eigenvalue of \( M \). Recall that from Lemma [2] an orthonormal basis of the tangent space of \( \text{Sym}(n) \) at \( E_1 \) is given by the commutators or \( E_1 \) and
the $n-1$ matrices $\frac{1}{\sqrt{2}}\Omega_{12}, \frac{1}{\sqrt{2}}\Omega_{13}, \ldots, \frac{1}{\sqrt{2}}\Omega_{1n}$. Note that since $D$ is diagonal, $[D, \Omega_{1j}] = \Omega_{1j}(d_1 - d_j)$ and thus

$$H_1(\Omega_{1j}, \Omega_{1l}) = (d_1 - d_j) \delta_{jl}$$

where $\delta_{jl} = 1$ if $j = l$ and zero otherwise. Thus this basis diagonalizes $H_1$ and shows that its eigenvalues are $(d_1 - d_j)$ and are all positive. Thus the signature at $C_1$ is $(n-1, 0, 0)$. We now evaluate the signature of $H_j = \text{tr}([\Sigma_{jj}, \Omega_1][D, \Omega_2])$. An orthonormal basis of the tangent space at $\Sigma_{jj}$ is given by the commutators of $\Sigma_{jj}$ and $\frac{1}{2}\Omega_{jl}$ for $j \in \{1, 2, \cdots, j-1, j+1, \ldots, n\}$ where $\hat{j}$ indicates that $j$ is omitted from the set. A similar computation gives that the eigenvalues of $H_j$ are $(d_j - d_l)$, for $l \in \{1, 2, \hat{j}, \ldots, n\}$. Hence $n-j$ eigenvalues are positive and $j-1$ are negative. Thus the signature of $\Sigma_{jj}$ is $(n-j, j-1, 0)$.

2 Discussion

Summary. We posed and solved the problem of finding the sensor minimizing the estimation error afforded by the Kalman filter. The methodology proposed is applicable to actuator design as well. The optimal sensor design problem is a difficult problem in the sense that it is not convex and, so far, convex relaxations have been the approach of choice. We took a different approach in this paper. We cast the problem as an optimization problem on a real projective space and equipped this space with a Riemannian metric, called the normal metric. We then evaluated the gradient and Hessian of the cost function $J$ to be optimized. We have shown that for $\gamma$ small, where $\gamma$ is the norm of the observation vector or sensor, and a stable infinitesimal generator $A$ of the dynamics, the gradient flow converges with probability one to the global minimum. We have restricted the analysis in this paper to the case of scalar, continuous-time observation signals. Similar results hold for the case of vector-valued observation signals. They are more technically involved and we do not elaborate on these here due to space constraints and the fact that most of the main ideas already appear in the present treatment of the scalar case.

We now discuss the role of the assumptions made. The first statement of the main result, which characterizes optimal observation matrices, holds free of the assumptions that $\gamma$ be small and the infinitesimal generator $A$ be stable. The second and third statements, however, relied on these assumptions. From a practitioner’s point of view, how small does $\gamma$ need to be? We can answer this question using Eq. (16) and the proof of Theorem 2.
holds for $\gamma < \gamma^*$ where $\gamma^*$ is the smallest $\gamma$ such that the bilinear form
\[ \text{tr} \left\{ [C, \Omega_1][M, \Omega_2] + [C, VRK + KWK + KRV] \Omega_2 \right\} \]
with $C$ extremal has a zero eigenvalue. Indeed, for $0 \leq \gamma < \gamma^*$, we then know that the above bilinear form has no zero eigenvalues and its signature is the one of the lowest order term. Note that $\gamma^*$ depends on $A$ and $Q$. We show in Fig. 1 simulation results, which show that this assumption holds for rather large $\gamma$ in general. The curves are obtained as follows. We first set $Q = \frac{1}{2}I_4$. We then sampled four batches of $10^4$ real $4 \times 4$ matrices which are stable and whose eigenvalues with largest real parts were, depending on the batch, $-0.1$, $-0.5$, $-1$, or $-3$ (denoted by $\text{Re}\lambda_m$ in the legend.). We obtained the samples by drawing matrices from a Gaussian ensemble and then translated their eigenvalues by adding a multiple of the identity matrix. For each sample matrix, and for $\gamma$ ranging from $10^{-3}$ to 10 we searched for the zeros of the gradient of $J$ numerically and then checked whether the Hessian at that zero had a signature given by the dominating term. The curves represent the proportion of matrices, out of the $10^4$ samples, for which $\gamma < \gamma^*$. For example, about 80% of the matrices with $\text{Re}\lambda_m = -\frac{1}{2}$ were such that $\gamma = 4$ qualifies as small. Unsurprisingly, as the eigenvalues of $A$ are further away from the imaginary axis, $\gamma^*$ increases and the proportion of matrices for which $\gamma < \gamma^*$, for $\gamma$ fixed increases as well. Indeed, for $\text{Re}\lambda_m = -3$, close to 100% of matrices are such that $\gamma = 4$ qualifies as small.

We also assumed that $A$ was stable to reach our conclusions. Note first that Proposition 1 which provide the Hessian of $J$, holds whether $A$ is stable or not.
The assumption was needed for Lemma 1 to hold when \( \gamma = 0 \), which in turn allowed us to analyze the Hessian of \( J \) via an expansion of the product \( M = KRK \) around \( \gamma = 0 \). When \( A \) is not stable, this expansion does not hold. Furthermore, it is easy to see that there exist loci of codimension one or two of observation vectors for which \( J(\gamma, c) \) is unbounded. Loci of unbounded values can evidently not be crossed by a gradient flow. If the loci are all of co-dimension two, then one might nevertheless have almost global convergence. Even more, since the domain \( \mathbb{R}P(n-1) \) is not orientable when \( n \) is odd, a locus of codimension one does not necessarily split the domain in two disconnected parts. Hence, the analysis of the unstable \( A \) case requires a careful analysis of the undetectable modes and the homology class of their eigenspaces.

**A rule of thumb for sensor choice.** From the proof of Theorem 2, we conclude that a good observation vector to use is the largest eigenvector of \( M_0 \) (this matrix is defined in (5)), which we denote by \( \gamma c_0 \), with \( \|c_0\| = 1 \). Indeed, this vector is optimal for \( \gamma = 0 \) and one can hope that it remains close to optimal as \( \gamma \) increases. Note that it is also a good starting point of the gradient flow. In Fig. 2 we present simulation results that show that this is indeed a sensible choice when \( \gamma \) is small. The curves in Fig. 2 were obtained as follows: for each curve, we sampled \( 10^4 \times 6 \) matrices with real eigenvalues as indicated on the legend. We let \( Q = I_6/\sqrt{6} \). Denote by \( c^* \) be the optimal observer obtained for each sample. Each curve represents the average of \( J(\gamma, c_0)/J(\gamma, c^*) \) as a function of \( \gamma \) for different values of \( \text{Re} \lambda_m \). We see that for \( \gamma \) very close to 0, the \( c_0 \) and \( c^* \)'s performances are nearly indistinguishable. As \( \gamma \) increases, the difference becomes more marked as expected. We also plotted the performance of a random observer, denoted by \( c_r \), which we observe performs predictably worse than both \( c^* \) and \( c_0 \).

**References**

[1] B. Anderson and J. Moore, *Optimal filtering*, Prentice-Hall, 1979.

[2] M. Athans, *The importance of Kalman filtering methods for economic systems*, in Annals of Economic and Social Measurement, vol. 3 n. 1, J. F. et al., ed., National Bureau of Economic Research, 1974, pp. 49–64.

[3] M. Atiyah, *Convexity and commuting hamiltonians*, Bulletin of the London Mathematical Society, 14 (1982), pp. 1–15.
Figure 2: Using $\gamma c_0$ as sensor often yields a close-to-optimal performance. A random choice of sensor (top curve, $c_r$) performs noticeably worse.

[4] R. Bansal and T. Başar, Simultaneous design of measurement and control strategies for stochastic systems with feedback, Automatica, 25 (1989), pp. 679–694.

[5] O. E. Barndorff-Nielsen and S. Neil, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64 (2002), pp. 253–280.

[6] C. Barnes, D. Silk, and M. Stumpf, Bayesian design strategies for synthetic biology, Interface Focus, 1 (2011), pp. 895–908.

[7] R. Brockett, Finite dimensional linear systems, John Wiley & Sons, 1970.

[8] R. Brockett, Dynamical systems that sort lists, solve linear programming problems and diagonalize symmetric matrices, Linear Algebra Appl, 146 (1991), pp. 79–91.

[9] R. Brockett, Differential geometry and the design of gradient algorithms, in Proceedings of Symposia in Pure Mathematics, American Mathematical Society, 1993, pp. 69–93.

[10] N. Darivandi, K. Morris, and A. Khaejpour, An algorithm for LQ optimal actuator location, Smart materials and structures, 22 (2013), p. 035001.

[11] D. Delchamps, Analytic feedback control and the algebraic Riccati equation, IEEE Transactions on Automatic Control, 29 (1984), pp. 1031–1033.
[12] F. Fahroo and M. A. Demetriou, *Optimal actuator/sensor location for active noise regulator and tracking control problems*, Journal of Computational and Applied Mathematics, 114 (2000), pp. 137–158.

[13] J. Geromel, *Convex analysis and global optimization of joint actuator location and control problems*, IEEE Transactions on Automatic Control, 34 (1989), pp. 711–720.

[14] C. C. Hay, E. Morrow, R. E. Kopf, and J. X. Mitrovica, *Estimating the sources of global sea level rise with data assimilation techniques*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 3692–3699.

[15] U. Helmke and J. B. Moore, *Optimization and dynamical systems*, Springer, London, 1994.

[16] K. Herring and J. Melsa, *Optimum measurements for estimation*, IEEE Transactions on Automatic Control, 19 (1974), pp. 264–266.

[17] J. Jost, *Riemannian Geometry and Geometric Analysis*, Springer, 2011.

[18] R. Kalman, *A new approach to linear filtering and prediction problems*, Journal of Fluids Engineering, 82 (1960), pp. 35–45.

[19] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME, Ser. D, J. Basic Eng., (1961), p. 109.

[20] A. W. Knapp, *Lie groups beyond an introduction*, vol. v. 140 of Progress in mathematics, Birkhäuser, Boston, 2nd ed ed., 2002.

[21] P. D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Communications on pure and applied mathematics, 21 (1968), pp. 467–490.

[22] J. M. Lee, *Manifolds and differential geometry*, vol. v. 107 of Graduate studies in mathematics, American Mathematical Society, Providence, R.I., 2009.

[23] F. Li, M. C. de Oliveira, and R. E. Skelton, *Designing instrumentation for control*, in Model-Based Control, Springer, 2009, pp. 71–88.

[24] Y. Liu, J. Slotine, and A. Barabási, *Observability of complex systems*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 2460–2465.
[25] M. Manoliu and S. Tompaidis, Energy futures prices: term structure models with Kalman filter estimation, Applied Mathematical Finance, 9 (2002), pp. 21–43.

[26] K. Morris, Linear-quadratic optimal actuator location, Automatic Control, IEEE Transactions on, 56 (2011), pp. 113–124.

[27] E. Musulin, C. Benqlilou, M. J. Bagajewicz, and L. Puigjaner, Instrumentation design based on optimal Kalman filtering, Journal of Process Control, 15 (2005), pp. 629–638.

[28] A. C. O’Connor, Optimal Control of Active Sensing Systems, PhD thesis, Harvard University, 2011.

[29] C. R. Rao, A note on Kalman filter, Proceedings of the National Academy of Sciences, 98 (2001), pp. 10557–10559.

[30] S. R. Singiresu, T.-S. Panand, and V. B. Venkayya, Optimal placement of actuators in actively controlled structures using genetic algorithms, AIAA Journal, 29 (1991), pp. 942–943.

[31] R. Stratonovich, Application of the theory of Markov processes for optimum filtration of signals, Radio Eng. Electron. Phys. (USSR), 1 (1960), pp. 1–19.

[32] G. Weredhagen and P. Bélanger, Curvature properties of the algebraic Riccati equation, Systems & Control Letters, 21 (1993), pp. 285–287.

[33] Y.-X. Yuan, Step-sizes for the gradient method, AMS IP studies in Advanced Mathematics, 42 (2008), p. 785.

A Appendix

Normal metric We define here the normal metric on the manifold Sym(Λ). The main idea is to define Sym(Λ) as a submanifold of su(n) and use the Killing form on su(n) to define an inner product. Denote by SU(n) the real group of unitary n × n matrices and by su(n) its Lie algebra. Recall that so(n) is the Lie algebra of skew-symmetric matrices and let Λ(n) be the vector space of traceless, real symmetric matrices. We can decompose su(n) as the direct sum su(n) = so(n) ⊕ Λ(n) where to the subalgebra so(n) of su(n) corresponds the subgroup SO(n) of SU(n).
Because $\mathfrak{su}(n)$ is a compact, semi-simple Lie algebra, the Killing form $\kappa(\cdot,\cdot)$ is a non-degenerate, negative-definite bilinear form and is proportional to $\text{tr}(AB)$ \[20\]. It is furthermore easy to check that

$$\kappa(\operatorname{ad}_C A, B) = -\kappa(A, \operatorname{ad}_C B).$$

The above property is referred to as the *ad-invariance* of the bilinear form $\kappa$. A straightforward computation shows that

$$[\mathfrak{so}(n), i\mathfrak{k}(n)] \subset i\mathfrak{k}(n)$$

and

$$[i\mathfrak{k}(n), i\mathfrak{k}(n)] \subset \mathfrak{so}(n). \tag{20}$$

Finally, note that $\mathfrak{so}(n)$ and $i\mathfrak{k}(n)$ are orthogonal to each other with respect to $\kappa$, i.e.

$$\kappa(\mathfrak{so}(n), i\mathfrak{k}(n)) = 0.$$  

With these facts in mind, we can now construct the normal metric on $\text{Sym}(\Lambda)$. First, note that $\text{Sym}(\Lambda)$ and $\text{Sym}(\Lambda + \alpha I)$ for $I$ the identity matrix and $\alpha \in \mathbb{R}$ are translates of each other. Thus for the purpose of defining a metric, we can assume without loss of generality that $\text{tr} \Lambda = 0$. Notice that $i\text{Sym}(\Lambda) \subset \mathfrak{su}(n)$ and let $iC \in \text{Sym}(\Lambda)$. Denote by $\mathfrak{z}(iC)$ the centralizer of $iC$, that is the subalgebra of elements $\Omega \in \mathfrak{su}(n)$ such that $\operatorname{ad}_{iC} \Omega = 0$, and let $\mathfrak{z}^\perp(iC)$ be its orthogonal complement with respect to the Killing form. Denote by $\co \operatorname{ad}_{iC}$ the codomain of $\operatorname{ad}_{iC}$; the map

$$\tilde{\operatorname{ad}}_{iC} : \mathfrak{z}^\perp(iC) \longrightarrow \co \operatorname{ad}_{iC} : iC \longrightarrow [iC, \Omega]$$

is an isomorphism. Observe that because $iC \in i\mathfrak{k}$, we have using \[20\], and with a slight abuse of notation, that $\tilde{\operatorname{ad}}_{iC} : \mathfrak{z}^\perp(iC) \cap \mathfrak{so}(n) \longrightarrow \co \operatorname{ad}_{iC} \cap i\mathfrak{k}(n)$ is an isomorphism. But we have the equality $\co \operatorname{ad}_{iC} \cap i\mathfrak{k}(n) = iT_C \text{Sym}(\Lambda)$. We drop the factor $i$ and define the normal metric for $X, Y \in T_C \text{Sym}(\Lambda)$ as

$$\kappa_n(X, Y) := -\kappa(\tilde{\operatorname{ad}}_{iC}^{-1} X, \tilde{\operatorname{ad}}_{iC}^{-1} Y). \tag{21}$$

Since $\kappa$ is negative-definite on $\mathfrak{su}(n)$, the normal metric is positive definite on $T_C \text{Sym}(\Lambda)$.

We now provide an orthonormal basis for $T_C \text{Sym}(\Lambda)$:

**Lemma 2.** Let $1 \leq m \leq n$. Let $E \in \text{Sym}(n)$ be the matrix with zero entries except for the diagonal entry $(m, m)$ which is one, that is $E = \Sigma_{mm}$. Then the matrices $\frac{1}{\sqrt{2}} \operatorname{ad}_E \Omega_{mj}$ for and $j \neq m$ form an orthonormal basis of $T_E \text{Sym}(n)$.
Proof. Recall that the tangent space at $E$ is spanned by the matrices $\text{ad}_E \Omega$ for $\Omega \in \mathfrak{so}(n)$. Note that the $\Omega_{ij}$, for $i \neq j$ span $\mathfrak{so}(n)$. Whence, to show that the $\Omega_{ij}$ with $i = m, j \neq m$ span the tangent space at $E$, it is sufficient to show that $\Omega_{ij} \in \mathfrak{z}(E)$ if the conditions $i = m$ and $j \neq m$ are not satisfied. But a short calculation shows that

$$[\Sigma_{kk}, \Omega_{ij}] = \begin{cases} 
\Sigma_{ij} & \text{if } i = k \\
-\Sigma_{ij} & \text{if } j = k \\
0 & \text{otherwise.}
\end{cases}$$

(22)

We conclude from it that $\text{ad}_E \Omega_{ij} \neq 0$ only if either $i = m, j \neq m$ or $i \neq m, j = m$. Since $\Omega_{ij} = -\Omega_{ji}$, the vectors $\text{ad}_E \Omega_{mj}$ with $j \neq m$ span $T_E \text{Sym}(n)$. We now show that these vectors are orthonormal for the normal metric. Again, a simple calculation shows that

$$\kappa_n(\text{ad}_E \Omega_{ij}, \text{ad}_E \Omega_{kl}) = -\text{tr}(\Omega_{ij} \Omega_{kl}) = 2\delta_{ik}\delta_{jl}$$

where $\delta_{ik} = 1$ if $i = k$ and 0 otherwise. This proves the claim.

A.1 The Hessian of $J$ for the normal metric

We first need an explicit expression for the Levi-Civita connection associated to the normal metric. We will derive such an expression for the case of constant vector fields. We call a vector field $X$ in $T\text{Sym}(\Lambda)$ a constant vector field if it is of the form

$$X = [C, \Omega_x]$$

(23)

for a constant $\Omega_x \in \mathfrak{so}(n)$.

Lemma 3. Let $\langle \cdot, \cdot \rangle$ be the normal metric on $\text{Sym}(\Lambda)$ and let $X = [C, \Omega_x], Y = [C, \Omega_y]$ be constant vector fields in $T\text{Sym}(\Lambda)$. Then the Levi-Civita covariant derivative of $Y$ along $X$ is

$$\nabla_X Y = \frac{1}{2}[C, [\Omega_x, \Omega_y]] = \text{ad}_C [\Omega_x, \Omega_y].$$

Proof. Recall that the covariant derivative $\nabla$ obeys the following relation

$$\langle \nabla_X Y, Z \rangle = \frac{1}{2} [X \cdot \langle Y, Z \rangle + Y \cdot \langle Z, X \rangle - Z \cdot \langle X, Y \rangle + \langle [X, Y], Z \rangle - \langle [Y, Z], X \rangle + \langle [Z, X], Y \rangle]$$

(24)
Let $X = [C, \Omega_x], Y = [C, \Omega_y]$ and $Z = [C, \Omega_z]$ be constant vector fields. Note that
\[ X \cdot \langle Y, Z \rangle = X \cdot \text{tr}(\Omega_y \Omega_z) = 0. \]

We thus have
\[ \langle \nabla_X Y, Z \rangle = \frac{1}{2} \left[ \langle \{X, Y\}, Z \rangle - \langle \{Y, Z\}, X \rangle + \langle \{Z, X\}, Y \rangle \right] \]
\[ = \frac{1}{2} \left[ \langle \{C, [\Omega_x, \Omega_y]\}, Z \rangle - \langle \{C, [\Omega_y, \Omega_z]\}, X \rangle + \langle \{C, [\Omega_z, \Omega_x]\}, Y \rangle \right] \]
\[ = \frac{1}{2} \left[ \text{tr}([\Omega_x, \Omega_y] \Omega_z) - \text{tr}([\Omega_y, \Omega_z] \Omega_x) + \text{tr}([\Omega_z, \Omega_x] \Omega_y) \right]. \]

The first two terms cancel each other and we obtain
\[ \langle \nabla_X Y, Z \rangle = \frac{1}{2} \text{tr}([\Omega_x, \Omega_y] \Omega_z). \]

Since the previous equation holds for all $\Omega_z \in \mathfrak{g}(C)^\perp$ we obtain using the same approach as in the derivation of $\nabla J$ for Theorem 1 that
\[ \nabla_X Y = \text{ad}_C [\Omega_x, \Omega_y] \]
as announced. \qedhere

**Proposition 1.** Let $J = \text{tr}(LK)$ be defined as in Theorem 1. The Hessian of $J$ with respect to the normal metric is
\[ d^2 J(\Omega_1, \Omega_2) = \gamma \text{tr} \left\{ [C, \Omega_1] [M, \Omega_2] \right. \]
\[ \left. + [C, VRK + KWK + KRV] \Omega_2 - \frac{1}{2} [C, M] [\Omega_1, \Omega_2] \right\} \] (25)
where $K$ and $R$ are as in the statement of Theorem 1 and $V$ and $W$ are
\[ V = \gamma \int_0^\infty e^{(A-CK)t} K [C, \Omega_1] Ke^{(A-\gamma CK)^\top t} \, dt \] (26)
and
\[ W = \gamma \int_0^\infty e^{(A-\gamma CK)t} \left( -[C, \Omega_1] KR - CVR \right. \]
\[ \left. -RV C - RK [C, \Omega_1] \right) e^{(A-\gamma CK)^\top t} \, dt. \] (27)
Proof. Let $C \in \text{Sym}(n)$ and $X = \text{ad}_C \Omega_x, Y = \text{ad}_C \Omega_y$ be constant vector fields. We start by evaluating the first term in the definition (15) of the Hessian. From the definition of the gradient and the normal metric, we have that
\begin{equation}
Y \cdot J = \gamma \text{tr} \left( [C, M]_{\Omega_y} \right). \tag{28}
\end{equation}

In order to evaluate the differential of the above function along the vector field $X$, we introduce the curve
\begin{equation}
C(t) = e^{t\Omega_x} C e^{-t\Omega_x}.
\end{equation}

We have
\begin{equation}
X \cdot Y \cdot J = \gamma \frac{d}{dt} \bigg|_{t=0} \text{tr} \left( [C(t), K(t)R(t)K(t)]_{\Omega_y} \right).
\end{equation}

We have already given an explicit expression for $\frac{d}{dt} \bigg|_{t=0} K(t)$ in Eqn. (13).

We now derive an expression for $\frac{d}{dt} \bigg|_{t=0} R(t)$. Recall that $R(t)$ obeys the equation
\begin{equation}
(A - \gamma CK)R + R(A - \gamma CK)^\top + L = 0.
\end{equation}

Taking the time derivative of both sides, and using again the short-hand $\frac{d}{dt} \bigg|_{t=0} R = \dot{R}$ and $\frac{d}{dt} \bigg|_{t=0} C = \dot{C} = \text{ad}_C \Omega_x$, we obtain
\begin{equation}
(A - \gamma CK)\dot{R} + \dot{R}(A - \gamma CK)^\top - \gamma(\dot{C}K + C\dot{K})R
- \gamma R(\dot{K}C + K\dot{C}) = 0.
\end{equation}

Setting $S := -(\dot{C}K + C\dot{K})R$, we can write explicitly
\begin{equation}
\dot{R} = \gamma \int_0^\infty e^{(A-CK)t}(S + S^\top) e^{(A-CK)^\top t}. \tag{29}
\end{equation}

Gathering the relations above, we have the following expression for $X \cdot Y \cdot J$:
\begin{equation}
X \cdot Y \cdot J = \gamma \text{tr} \left\{ [[[C, \Omega_x], M]_{\Omega_y} + [C, KRK]_{\Omega_y}
+ [C, KR\dot{K}]_{\Omega_y} + [C, K\dot{R}K]_{\Omega_y}} \right\} \tag{30}
\end{equation}
where $\dot{K}$ and $\dot{R}$ are given explicitly in (13) and (29) respectively. We now focus on the second term in Eqn. (15). From Lemma 3, we now that for constant vector fields $X$ and $Y$,

$$\nabla_X Y = [C, [\Omega_x, \Omega_y]].$$

Let $C(t)$ be the curve in $\text{Sym}(\lambda)$ given by

$$C(t) = e^{t[\Omega_x, \Omega_y]}Ce^{-t[\Omega_x, \Omega_y]}.$$

Using the definition the gradient of $J$ given in (10) and Theorem 1, we get

$$\nabla_X Y \cdot J = \gamma \text{tr}\left( [\Omega_x, \Omega_y]M \right). \quad (31)$$

Using the ad-invariance property of the normal metric, the first term of (30) is equal to $\gamma \text{tr}\left( [C, \Omega_x][M, \Omega_y] \right)$. Now recalling the expression of $d^2J$ given in (15), we obtain the result using (30) and (31).

**Corollary 1.** Let $K_0$ and $R_0$ be the solutions of

$$A^T K + KA + Q = 0 \quad (32)$$

and

$$A^T R + RA + L = 0 \quad (33)$$

respectively. Let

$$M_0 := K_0R_0K_0.$$ 

The Hessian of $J$ with respect to the normal metric has the following expansion around $\gamma = 0$:

$$d^2J(\Omega_1, \Omega_2) \approx \gamma \text{tr}\left\{ [C, \Omega_1][M_0, \Omega_2] - \frac{1}{2} [C, M_0][\Omega_1, \Omega_2] \right\} + \gamma^2 T(\Omega_1, \Omega_2) \quad (34)$$

where $\Omega_1, \Omega_2 \in \mathfrak{so}(n)$ and the bilinear form $T$ contains terms of zeroth and higher orders in $\gamma$.

**Proof.** From Lemma 1 we know that for $\gamma$ small, the positive definite solution $K$ of the Riccati equation can be expressed as

$$K = K_0 + \text{h.o.t. in } \gamma,$$

where h.o.t. are higher order terms in $\gamma$. Recall that $R$ obeys the equation $(A - \gamma CK)R - R(A - \gamma CK)^T + L = 0$. This is a linear equation and thus its
solution, when it exists, depends analytically on $\gamma$ and $C$. Whence, similarly as for $K$, we can write for $\gamma$ small

$$R = R_0 + \text{h.o.t. in } \gamma.$$  

We conclude from the above two expansions that we have

$$M \simeq M_0 + \text{h.o.t. in } \gamma.$$  \hspace{1cm} (35)

Now recall the explicit expression of $d^2 J$ at $\gamma C$ derived in Proposition 1. A simple calculation shows that the first and last terms of the right hand side of (25) admit the expansions

$$\text{tr} \left\{ [\gamma C, \Omega_1] [M, \Omega_2] \right\} = \gamma \text{tr} \left\{ [C, \Omega_1] [M_0, \Omega_2] \right\} + \text{h.o.t. in } \gamma$$

and

$$- \text{tr} \left\{ \frac{1}{2} [\gamma C, M] [\Omega_1, \Omega_2] \right\} = - \gamma \frac{1}{2} \text{tr} \left\{ [C, M_0] [\Omega_1, \Omega_2] \right\} + \text{h.o.t. in } \gamma$$

respectively. The second term, since both $V$ and $W$ have order one in $\gamma$, contributes terms of order at least two in $\gamma$. Putting these facts together, we obtain the announced expansion.

Lemma 4. Let $A \in \mathbb{R}^{n \times n}$ be a stable matrix and $Q, L$ be positive definite symmetric matrices. Let $K, R \in \mathbb{R}^{n \times n}$ be the unique positive definite solutions of

$$AK + KA^T + Q = 0$$
$$AR + RA^T + L = 0$$

Then generically for $Q, L$ positive definite, the matrix $M := KRK$ has distinct eigenvalues.

Proof. We first recall that since $A$ is stable, the Lyapunov equations in (36) have each a unique symmetric positive definite solution. We denote them by $\mathcal{L}(Q)$ and $\mathcal{L}(L)$ respectively, i.e.

$$\mathcal{L}(Q) = \int_0^\infty e^{At} Q e^{A^T t} d t.$$

We let $S^+$ be the set of symmetric positive definite matrices of dimension $n$ and define the map $F : S^+ \times S^+ \longrightarrow S^+ : (Q, L) \longmapsto M = KRK$ where $K = \mathcal{L}(Q)$ and
\( R = \mathcal{L}(L) \). The proof of the Lemma goes by showing that for a generic point \((Q, L) \in S^+ \times S^+\), \( F \) is locally surjective, i.e., \( F \) maps small enough neighborhoods of \((Q, L)\) onto neighborhoods of \( F(Q, L) \). From there, the statement of the Lemma follows from a simple contradiction argument. Indeed, assume that \( F \) is locally surjective but that there exists an open set \( V \subset S^+ \times S^+ \) for which \( F(V) \) only contains matrices with non-distinct eigenvalues. The set of matrices in \( S^+ \) which have non-distinct eigenvalues is of codimension 1 and thus for any pair \((Q, L)\) in \( V \), \( F \) is not locally surjective – a contradiction.

The remainder of the proof is dedicated to showing that \( F \) is generically locally surjective (g.l.s.). To this end, note that an open map is clearly g.l.s. and that the composition of generically locally surjective maps is likewise g.l.s. To see that this last statement holds, assume that \( f_1 : M \rightarrow N \) and \( f_2 : N \rightarrow P \) are g.l.s. and let \( f_3 = f_2 \circ f_1 \). Let \( C_1 \subset M \) (resp. \( C_2 \subset N \)) be the set of points where \( f_1 \) is not locally surjective (resp. \( f_2 \)) and let \( D = \{ x \in M \mid f_1(x) \in C_2 \} \). The sets \( C_1 \) and \( C_2 \) are of measure zero by assumption and by the same argument as in the paragraph above, \( D \) is of measure zero in \( M \). Thus for \( x \notin C_1 \cup D \), \( f_3(x) \) is locally surjective and since \( C_1 \cup D \) is of measure zero, \( f_3 \) is g.l.s. .

We now return to the main thread. Let \( G : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n} : G(K, R) = M \). Then we can write \( F \) as the composition \( F = G \circ (\mathcal{L}(Q), \mathcal{L}(L)) \). The operator \( \mathcal{L}^{-1}(X) = AX +XA^T \) is nothing more that the Lyapunov operator. One can show that for \( A \) stable, the Lyapunov operator is of full-rank (observe that its eigenvalues are pairwise sums of eigenvalues of \( A \)). Its inverse \( \mathcal{L} \) is thus a full rank linear map and consequently an open map. By a standard argument, one can show that the map \((Q, L) \rightarrow (\mathcal{L}(Q), \mathcal{L}(L)) \) is also an open map. The map \( G \) is a polynomial map and is clearly surjective. If we can show that \( G \) is g.l.s., then \( F \) is the composition of g.l.s. maps and is thus g.l.s. which proves the Lemma.

It thus remains to show that \( G \) is g.l.s. To see this, first recall that at points \((K, R)\) in the domain of \( G \) where its linearization \( \frac{\partial G}{\partial x} \) is full rank, \( G \) is locally surjective. Now assume that there is an open set \( V \) in the domain of \( G \) where its linearization is nowhere full rank. Then \( \det(\frac{\partial F}{\partial x} \frac{\partial F^T}{\partial x}) = 0 \) on the open set \( V \) and because this determinant is a polynomial function, it is zero everywhere. By Sard Theorem \([22]\), the set \( W \) over which the linearization of \( G \) is not full rank is such that \( G(W) \) has measure zero. But we have just shown that \( W \) is the entire domain of \( G \), which contradicts the fact that \( G \) is surjective. This ends the proof of the Lemma. \( \square \)
Lemma 5. Let $A$ be a stable matrix and $\gamma \geq 0$. Define

$$F : \mathbb{R}^+ \times \text{Sym}(n) \longrightarrow T_c \text{Sym}(n) : (\gamma, C) \longmapsto [C, M]$$

where $K, R$ satisfy Eq. (14) and $M = KRK$. The covariant derivative of $F$ at $(0, C)$ and with respect to its second argument is

$$\nabla_X F = -\frac{1}{2} \left( [M_0, [C, \Omega_x]] + [\Omega_x, [C, M_0]] \right).$$

Proof. We need to evaluate $\nabla_X F(0, C)$ for $X = \text{ad}_C \Omega_x$. Observe that $F(0, C)$ is a constant vector field as defined in (23). From Lemma 3 a short calculation yields

$$\nabla_X F = \frac{1}{2} [C, [\Omega_x, M_0]].$$

Using the Jacobi identity [20], the previous relation can written as

$$\nabla_X F = -\frac{1}{2} \left( [M_0, [C, \Omega_x]] + [\Omega_x, [C, M_0]] \right)$$

as announced. \qed