Duality-Based Stochastic Policy Optimization for Estimation with Unknown Noise Covariances

Shahriar Talebi, Amirhossein Taghvaei, Mehran Mesbahi

***Abstract*—Duality of control and estimation allows mapping recent advances in data-guided control to the estimation setup. This paper formalizes and utilizes such a mapping to consider learning the optimal (steady-state) Kalman gain when process and measurement noise statistics are unknown. Specifically, building on the duality between synthesizing optimal control and estimation gains, the filter design problem is formalized as direct policy learning. In this direction, the duality is used to extend existing theoretical guarantees of direct policy updates for Linear Quadratic Regulator (LQR) to establish global convergence of the Gradient Descent (GD) algorithm for the estimation problem–while addressing subtle differences between the two synthesis problems. Subsequently, a Stochastic Gradient Descent (SGD) approach is adopted to learn the optimal Kalman gain without the knowledge of noise covariances. The results are illustrated via several numerical examples.**

I. Introduction

Duality of control and estimation provides an important relationship between two distinct synthesis problems in system theory [1]–[3]. In fact, duality has served as an effective bridge for developing theoretical and computational techniques in one domain and then "dualized" for use in the other. For instance, the stability proof of the Kalman filter relies on the stabilizing feature of the optimal feedback gain for the dual LQR optimal control problem [4, Ch. 9]. The aim of this paper is to build on this dualization for the purpose of learning the optimal estimation policy via recent advances in data-driven algorithms for optimal control.

The setup that we consider is the estimation problem for a system with known linear dynamics and observation model, but unknown process and measurement noise covariances. The problem is to learn the optimal steady-state Kalman gain using a training data that consists of independent realizations of the observation signal. This problem has a long history in system theory, often examined in the context of adaptive Kalman filtering [5]–[10]. The classical reference [6] includes a comprehensive summary of four solution approaches to this problem: Bayesian inference [11]–[13], Maximum likelihood [14], [15], covariance matching [9], and innovation correlation methods [5], [7]. The Bayesian and maximum likelihood setup are known to be computationally costly and covariance matching admits undesirable biases in practice. For these reasons, the innovation correlation based approaches are more popular and have been subject of more recent research [16]–[18]. The article [19] includes an excellent survey on this topic. Though relying strongly on the statistical assumptions on the model, these approaches do not provide *non-asymptotic* guarantees.

On the optimal control side, there has been a number of recent advances in data-driven synthesis methods. For example, first order methods have been adopted for state-feedback LQR problems [20], [21]. This direct policy optimization perspective has been particularly effective as it has been shown that the LQR cost is *gradient dominant* [22], allowing the adoption and global convergence of first order methods for optimal feedback synthesis despite the non-convexity of the cost, when represented directly in terms of this policy. Since then, Policy Optimization (PO) using first order methods has been investigated for variants of LQR problem, such as Output-feedback Linear Quadratic Regulators (OLQR) [23], model-free setup [24], risk-constrained setup [25], Linear Quadratic Gaussian (LQG) [26], and recently, Riemannian constrained LQR [27].

This paper aims to bring new insights to the classical estimation problem through the lens of control-estimation duality and utilizing recent advances in data-driven optimal control. In particular, we first argue that the optimal mean-squared error estimation problem is "equivalent" to an LQR problem. This in turn, allows representing the problem of finding the optimal Kalman gain as that of optimal policy synthesis for the LQR problem—under conditions distinct from what has been examined in the literature. In particular in this equivalent LQR formulation, the cost parameters–relating to the noise covariances–are unknown and the covariance of initial state is not positive definite. By addressing these technical issues, we show how exploring this relationship leads to computational algorithms for learning optimal Kalman gain with non-asymptotic error guarantees.

The rest of the paper is organized as follows. The estimation problem is formulated in §II, followed by the estimation-control duality relationship in §III. The theoretical analysis on policy optimization for the Kalman gain appears in §IV while the proofs are deferred to [28]. We propose an SGD algorithm in §V with several numerical examples, followed by concluding remarks in §VI.

II. Background and Problem Formulation

Consider the stochastic difference equation,

$$x(t+1) = Ax(t) + \xi(t), \tag{1a}$$

$$y(t) = Hx(t) + \omega(t), \tag{1b}$$

The authors are with the William E. Boeing Department of Aeronautics and Astronautics, University of Washington, Seattle, WA, USA. S. Talebi is also with the Department of Mathematics at the University of Washington. The research of the first and last authors has been supported by AFOSR grant FA9550-20-1-0053 and NSF grant ECCS-2149470. Emails: *shahriar@uw.edu*, *amirtag@uw.edu*, and *mesbahi@uw.edu*.

where $x(t) \in \mathbb{R}^n$ is the state of the system, $y(t) \in \mathbb{R}^m$ is the observation, and $\{\xi(t)\}_{t\in\mathbb{Z}}$ and $\{\omega(t)\}_{t\in\mathbb{Z}}$ are the uncorrelated zero-mean process and measurement noise vectors, respectively, with the following covariances,

$$\mathbb{E}\left[\xi(t)\xi^\intercal(t)\right] = Q \in \mathbb{R}^{n\times n}, \quad \mathbb{E}\left[\omega(t)\omega^\intercal(t)\right] = R \in \mathbb{R}^{m\times m},$$

for some (possibly time-varying) positive (semi-)definite matrices $Q, R \succcurlyeq 0$. Let m_0 and $P_0 \succcurlyeq 0$ denote the mean and covariance of the initial condition x_0.

Now, let us fix a time horizon $T > 0$ and define an estimation policy, denoted by $\mathcal{P}$, as a map that takes a history of the observation signal $\mathcal{Y}_T = \{y(0), y(1), \ldots, y(T-1)\}$ as an input and outputs an estimate of the state $x(T)$, denoted by $\hat{x}_{\mathcal{P}}(T)$. The filtering problem of interest is finding the estimation policy $\mathcal{P}$ that minimizes the mean-squared error,

$$\mathbb{E}\left[\|x(T) - \hat{x}_{\mathcal{P}}(T)\|^2\right]. \tag{2}$$

We make the following assumptions in our problem setup: 1) The matrices A and H are known, but the process and the measurement noise covariance matrices, Q and R, are *not* available. 2) We have access to a training data-set that consists of independent realizations of the observation signal $\{y(t)\}_{t=0}^{T}$. However, ground-truth measurements of $x(T)$ is *not* available.[1]

It is not possible to directly minimize (2) as the ground-truth measurement $x(T)$ is not available. Instead, we propose to minimize the mean-squared error in predicting the observation $y(T)$ as a surrogate objective function. In particular, let us first define $\hat{y}_{\mathcal{P}}(T) = H\hat{x}_{\mathcal{P}}(T)$ as the prediction for the observation $y(T)$. This is indeed a prediction since the estimate $\hat{x}_{\mathcal{P}}(T)$ depends only on the observations up to time $T-1$. The optimization problem is now finding the estimation policy $\mathcal{P}$ that minimizes the mean-squared prediction error,

$$\mathcal{J}_T^{\text{est}}(\mathcal{P}) := \mathbb{E}\left[\|y(T) - \hat{y}_{\mathcal{P}}(T)\|^2\right]. \tag{3}$$

1) Kalman filter: Indeed, when Q and R are known, the solution is given by the celebrated Kalman filter algorithm [2]. The algorithm involves an iterative procedure to update the estimate $\hat{x}(t)$ according to

$$\hat{x}(t+1) = A\hat{x}(t) + L(t)(y(t) - H\hat{x}(t)),\ \hat{x}(0) = m_0, \tag{4}$$

where $L(t) := AP(t)H^\intercal(HP(t)H^\intercal + R)^{-1}$ is the Kalman gain, and $P(t) := \mathbb{E}[(x(t)-\hat{x}(t))(x(t)-\hat{x}(t))^\intercal]$ is the error covariance matrix that satisfies the Ricatti equation,

$$P(t+1) = (A - L(t)H)P(t)A^\intercal + Q, \quad P(t_0) = P_0.$$

Note that the update law presented here combines the information and dynamic update steps of the Kalman filter.

It is known that $P(t)$ converges to an steady-state value P_∞ when the pair (A, H) is observable and the pair $(A, Q^{\frac{1}{2}})$ is controllable [29], [30]. In such a case, the gain converges to $L_\infty := AP_\infty H^\intercal(HP_\infty H^\intercal + R)^{-1}$, the so-called steady-state Kalman gain. It is a common practice to evaluate the steady-state Kalman gain L_∞ offline and use it, instead of $L(t)$, to update the estimate in real-time.

2) Learning the optimal Kalman gain: Inspired by the structure of the Kalman filter, we consider restriction of the estimation policies $\mathcal{P}$ to those realized with a constant gain. In particular, we define the estimate $\hat{x}_L(T)$ as one given by the Kalman filter at time T realized by the constant gain L. Rolling out the update law (4) for $t = 0$ to $t = T-1$, and replacing $L(t)$ with L, leads to the following expression for the estimate $\hat{x}_L(T)$ as a function of L,

$$\hat{x}_L(T) = A_L^T m_0 + \textstyle\sum_{t=0}^{T-1} A_L^{T-t-1} L y(t), \tag{5}$$

where $A_L := A - LH$. Note that this estimate does not require knowledge of the matrices Q or R. By considering $\hat{y}_L(T) := H\hat{x}_L(T)$, the problem is now finding the optimal gain L that minimizes the mean-squared prediction error

$$J_T^{\text{est}}(L) := \mathbb{E}\left[\|y(T) - \hat{y}_L(T)\|^2\right]. \tag{6}$$

Numerically, this problem falls into the realm of stochastic optimization and can be solved by algorithms such as Stochastic Gradient Descent (SGD). Such an algorithm would require accessing independent realizations of the observation signal. An algorithm that utilizes such realizations is presented in §V. Theoretically, however, it is not yet clear if this optimization problem is well-posed and admits a unique minimizer. This is the subject of §IV, where certain properties of the objective function, such as its gradient dominance and smoothness, are established. These theoretical results are then used to analyze first-order optimization algorithms and provide stability guarantees of the estimation policy iterates. The results are based on the duality relationship between estimation and control that is presented next.

III. Estimation-Control Duality Relationship

We use the duality framework, as described in [31, Ch.7.5], to relate the problem of learning the optimal estimation policy to that of learning the optimal control policy for an LQR problem. In order to do so, we introduce the adjoint system:

$$z(t) = A^\intercal z(t+1) - H^\intercal u(t+1), \tag{7}$$

where $z(t) \in \mathbb{R}^n$ is the adjoint state and $\mathcal{U}_T := \{u(1), \ldots, u(T)\} \in \mathbb{R}^{mT}$ are the control variables (dual to the observation signal $\mathcal{Y}_T$). The adjoint state is initialized at $z(T) = a \in \mathbb{R}^n$ and simulated *backward in time* starting with $t = T-1$. We now formalize a relationship between estimation policies for the system (1) and control policies for the adjoint system (7). Consider estimation policies that are linear functions of the observation history $\mathcal{Y}_T \in \mathbb{R}^{mT}$ and the initial mean vector $m_0 \in \mathbb{R}^n$. We characterize such policies with a linear map $\mathcal{L} : \mathbb{R}^{mT+n} \to \mathbb{R}^n$ and let the estimate $\hat{x}_{\mathcal{L}}(T) := \mathcal{L}(m_0, \mathcal{Y}_T)$. The adjoint of this linear map, denoted by $\mathcal{L}^\dagger : \mathbb{R}^n \to \mathbb{R}^{mT+n}$, is used

[1]This setting arises in various applications, such as aircraft wing dynamics, when approximate or reduced-order models are employed, and the effect of unmodelled dynamics and disturbances are captured by the process noise.

to define a control policy for the adjoint system (7). In particular, the adjoint map takes $a \in \mathbb{R}^n$ as input and outputs $\mathcal{L}^\dagger(a) = \{b, u(1), \dots, u(T)\} \in \mathbb{R}^{mT+n}$. This relationship can be depicted as,

$$\begin{array}{ccc} \{m_0, y(0), \dots, y(T-1)\} & \xrightarrow{\mathcal{L}} & \hat{x}_{\mathcal{L}}(T) \\ \{b, u(1), \dots, u(T)\} & \xleftarrow{\mathcal{L}^\dagger} & a \end{array}$$

Note that $\langle a, \mathcal{L}(m_0, \mathcal{Y}_T)\rangle_{\mathbb{R}^n} = \langle \mathcal{L}^\dagger(a), (m_0, \mathcal{Y}_T)\rangle_{\mathbb{R}^{mT+n}}$, so

$$b^\intercal m_0 + \textstyle\sum_{t=0}^{T-1} u(t+1)^\intercal y(t) = a^\intercal \hat{x}_{\mathcal{L}}(T). \tag{8}$$

The following proposition relates the mean-squared error for a linear estimation policy, to the following LQR cost:

$$\begin{aligned} &\mathcal{J}_T^{\text{LQR}}(a, \{b, \mathcal{U}_T\}) := [z^\intercal(0) m_0 - b^\intercal m_0]^2 \\ &+ z^\intercal(0) P_0 z(0) + \textstyle\sum_{t=1}^{T} \left[z^\intercal(t) Q z(t) + u^\intercal(t) R u(t)\right]. \end{aligned} \tag{9}$$

Proposition 1. *Consider the estimation problem for the system* (1) *and the LQR problem* (9) *subject to the adjoint dynamics* (7)*. For each estimation policy $\hat{x}_{\mathcal{L}}(T) = \mathcal{L}(m_0, \mathcal{Y}_T)$, with a linear map $\mathcal{L}$, and for any $a \in \mathbb{R}^n$ we have the identity*

$$\mathbb{E}\left[|a^\intercal x(T) - a^\intercal \hat{x}_{\mathcal{L}}(T)|^2\right] = \mathcal{J}_T^{LQR}(a, \mathcal{L}^\dagger(a)).$$

Furthermore, the prediction error as in (6) satisfies

$$J_T^{est}(L) = \sum_{i=1}^{m} \mathcal{J}_T^{LQR}(H_i, \mathcal{L}^\dagger(H_i)) + \operatorname{tr}[R],$$

where $\hat{y}_{\mathcal{L}}(T) := H\hat{x}_{\mathcal{L}}(T)$ and $H_i^\intercal \in \mathbb{R}^n$ is the i-th row of the $m \times n$ matrix H for $i = 1, \dots, m$.

Remark 1. The duality is also true in the continuous-time setting where the estimation problem is related to a continuous-time LQR problem. Recent extensions to the nonlinear setting appears in [32] with a comprehensive study in [33]. This duality is different than the maximum likelihood approach which involves an optimal control problem over the original dynamics instead of the adjoint system.

1) Duality in the constant control gain regime: In this section, we use the aforementioned duality relationship to show that the estimation policy with constant gain is dual to the control policy with constant feedback gain. This result is then used to obtain an explicit formula for the objective function (6).

Consider the adjoint system (7) with the linear feedback law $u(t) = L^\intercal z(t)$. Then,

$$z(t) = (A_L^\intercal)^{T-t} a, \quad \text{for} \quad t = 0, 1, \dots, T. \tag{10}$$

Therefore, as a function of a, $u(t) = L^\intercal (A_L^\intercal)^{T-t} a$. Moreover, for this choice of control, the optimal $b = z(0) = (A_L^\intercal)^T a$. These relationships are used to identify the control policy $\mathcal{L}^\dagger(a) = ((A_L^\intercal)^T a, L^\intercal (A_L^\intercal)^{T-1} a, \dots, L^\intercal a)$. This control policy corresponds to an estimation policy by the adjoint relationship (8):

$$a^\intercal \hat{x}_{\mathcal{L}}(T) = a^\intercal A_L^T m_0 + \textstyle\sum_{t=0}^{T-1} a^\intercal A_L^{T-t-1} L y(t), \quad \forall a \in \mathbb{R}^n.$$

As this relationship holds for all $a \in \mathbb{R}^n$, we have,

$$\hat{x}_{\mathcal{L}}(T) = A_L^T m_0 + \textstyle\sum_{t=0}^{T-1} A_L^{T-t-1} L y(t),$$

that coincides with the Kalman filter estimate with constant gain L given by the formula (5). Therefore, the adjoint relationship (8) relates the control policy with constant gain $L^\intercal$ to the Kalman filter with the constant gain L.

Next, we use this relationship to evaluate the mean-squared prediction error (6). Denote by $J_T^{\text{LQR}}(a, L^\intercal)$ as the LQR cost (9) associated with the control policy with constant gain $L^\intercal$ and $b = z(0)$. Then, from the explicit formula for $z(t)$ and $u(t)$ above, we have,

$$J_T^{\text{LQR}}(a, L^\intercal) = a^\intercal X_T(L) a,$$

where

$$X_T(L) := A_L^T P_0 (A_L^\intercal)^T + \sum_{t=1}^{T} A_L^{T-t} (Q + LRL^\intercal)(A_L^\intercal)^{T-t}.$$

Therefore, by the second claim in Proposition 1, the mean-squared prediction error (6) becomes,

$$J_T^{\text{est}}(L) - \operatorname{tr}[R] = \sum_{i=1}^{m} J_T^{\text{LQR}}(H_i, L^\intercal) = \operatorname{tr}\left[X_T(L) H^\intercal H\right],$$

where we have used the cyclic permutation property of the trace and the identity $H^\intercal H = \sum_{i=1}^m H_i H_i^\intercal$.

2) Duality in steady-state regime: Define the set of Schur stabilizing gains

$$\mathcal{S} := \{L \in \mathbb{R}^{n \times m} : \rho(A - LH) < 1\}.$$

For any $L \in \mathcal{S}$, in the steady-state limit as $T \to \infty$: $X_T(L) \to X_\infty(L) := \sum_{t=0}^{\infty} (A_L)^t (Q + LRL^\intercal) (A_L^\intercal)^t$. The limit coincides with the unique solution X of the discrete Lyapunov equation $X = A_L X A_L^\intercal + Q + LRL^\intercal$, which exists as $\rho(A_L) < 1$. Therefore, the steady-state limit of the mean-squared prediction error assumes the form,

$$J(L) := \lim_{T \to \infty} J_T^{\text{est}}(L) = \operatorname{tr}\left[X_\infty(L) H^\intercal H\right] + \operatorname{tr}[R].$$

Given the steady-state limit, we formally analyze the following constrained optimization problem:

$$\begin{aligned} \min_{L \in \mathcal{S}} \ &\leftarrow J(L) = \operatorname{tr}\left[X_{(L)} H^\intercal H\right] + \operatorname{tr}[R], \\ &\text{s.t.} \quad X_{(L)} = A_L X_{(L)} A_L^\intercal + Q + LRL^\intercal. \end{aligned} \tag{11}$$

Remark 2. Note that the latter problem is technically the dual of the optimal LQR problem as formulated in [20] by relating $A \leftrightarrow A^\intercal$, $-H \leftrightarrow B^\intercal$, $L \leftrightarrow K^\intercal$, and $H^\intercal H \leftrightarrow \Sigma$. However, one main difference here is that the matrices Q and R are unknown, and the $H^\intercal H$ may *not* be positive definite, for example, due to rank deficiency in H specially whenever $m < n$. Thus, in general, the cost function $J(L)$ is not necessarily coercive in L, which can drastcially effect the optimization landscape. For the same reason, in contrast to the LQR case [20], [22], the gradient dominant property of $J(L)$ is not clear in the filtering setup. In the next section, we show that such issues can be avoided as long as the pair (A, H) is observable.

IV. Theoretical analysis

In this section, we provide theoretical analysis of the proposed optimization problem (11). The following lemma is useful for our subsequent analysis which is a direct consequence of duality described in Remark 2, Lemmas 3.5 and 3.6 in [20], and the fact that the spectrum of a matrix remains unchanged under the transpose operation.

Lemma 1. *The set of Schur stabilizing gains $\mathcal{S}$ is regular open, contractible, and unbounded when $m \geq 2$ and the boundary $\partial\mathcal{S}$ coincides with the set $\{L \in \mathbb{R}^{n\times m} : \rho(A - LH) = 1\}$. Furthermore, $J(.)$ is real analytic on $\mathcal{S}$ whenever Q and R are time-independent.*

1) Coercive property: Next, we provide sufficient conditions to recover the coercive property of $J(.)$ which resembles Lemma 3.7 in [20], but extended for the time-varying cost parameters Q and R.

Proposition 2. *Suppose the pair (A, H) is observable, and Q and R are lower bounded uniformly in time with some positive definite matrices. Then, the function $J(.) : \mathcal{S} \to \mathbb{R}$ is coercive, i.e., for any sequence $\{L_k\} \in \mathcal{S}$,*

$$\text{if } \; L_k \to \partial\mathcal{S} \;\; \text{ or } \;\; \|L_k\| \to \infty \;\; \text{ then } \;\; J(L) \to \infty.$$

Furthermore, for any $\alpha > 0$, the sublevel set $\mathcal{S}_\alpha := \{L \in \mathbb{R}^{n\times m} : J(L) \leq \alpha\}$ is compact and contained in $\mathcal{S}$ whenever Q and R are time-independent.

Remark 3. This approach recovers the claimed coercivity also in the control setting with weaker assumptions. In particular, using this result, one can replace the positive definite condition on the covariance of the initial condition in [20], i.e., $\Sigma \succ 0$, with just the controllability of $(A, \Sigma^{1/2})$.

2) Gradient dominance property: Next, we establish the gradient dominance property which resembles Lemma 3.12 in [20]. While our approach utilizes a similar proof technique, this property is not trivial in this case as $H^\intercal H$ may not be positive definite. This, apparently minor issue, hinders establishing the gradient dominated property globally. However, we are able to recover this property on every sublevel sets of $J(L)$ which is sufficient for the subsequent convergence analysis.

Before presenting the result, we compute the gradient of $J(L)$ to characterize its global minimizer and consider the following simplifying assumption for the rest of the analysis.

Assumption 1. Suppose (A, H) is observable and the covariance matrices $Q \succ 0$ and $R \succ 0$ are time-independent.

The explicit gradient formula for J takes the form,

$$\nabla J(L) = 2Y_{(L)} \left[-LR + A_L X_{(L)} H^\intercal\right],$$

where $Y_{(L)}$ is the unique solution of $Y = A_L^\intercal Y A_L + H^\intercal H$. While the derivation appears in [28], note that the expression for the gradient is consistent with Proposition 3.8 in [20] after applying the duality relationship explained in Remark 2.

We also characterize the global minimizer $L^* = \arg\min_{L\in\mathcal{S}} J(L)$. The domain $\mathcal{S}$ is non-empty whenever (A, H) is observable. Thus, by continuity of $L \to J(L)$, there exists some finite $\alpha > 0$ such that the sublevel set $\mathcal{S}_\alpha$ is non-empty and compact. Therefore, the minimizer is an interior point and thus must satisfy the first-order optimality condition $\nabla J(L^*) = 0$. Moreover, by coercivity, the minimizer is stabilizing and unique satisfying,

$$L^* = AX^*H^\intercal \left(R + HX^*H^\intercal\right)^{-1},$$

with X^* being the unique solution of

$$X^* = A_{L^*} X^* A_{L^*}^\intercal + Q + L^* R (L^*)^\intercal. \tag{12}$$

As expected, the global minimizer L^* is equal to the steady-state Kalman gain, but explicitly dependent on the noise covariances Q and R.

Proposition 3. *Let L^* be the unique optimizer of $J(L)$ over $\mathcal{S}$ and consider any non-empty sublevel set $\mathcal{S}_\alpha$ for some $\alpha > 0$. Then, the function $J(.) : \mathcal{S}_\alpha \to \mathbb{R}$ satisfies*

$$c_1[J(L) - J(L^*)] + c_2\|L - L^*\|_F^2 \leq \langle \nabla J(L), \nabla J(L)\rangle,$$
$$c_3\|L - L^*\|_F^2 \leq J(L) - J(L^*),$$

for some positive constants $c_1 = c_1(\alpha) > 0$, $c_2 = c_2(\alpha) > 0$ and $c_3 = c_3(\alpha) > 0$ that are independent of L.

Remark 4. The proposition above implies that $J(.)$ is gradient dominated on $\mathcal{S}_\alpha$, i.e., for any $L \in \mathcal{S}_\alpha$ we have

$$J(L) - J(L^*) \leq \tfrac{1}{c_1(\alpha)} \langle \nabla J(L), \nabla J(L)\rangle.$$

Note that the first inequality characterizes the dominance gap in terms of the iterate error from the optimality. This is useful in obtaining the iterate convergence results in the next section where we analyze first-order methods in order to solve the minimization problem (11).

A. Gradient Descent (GD)

Here, we consider the GD policy update:

$$\text{[GD]} \qquad L_{k+1} = L_k - \eta_k \nabla J(L_k),$$

for $k \in \mathbb{Z}$ and a positive stepsize η_k. As a direct consequence of Proposition 3, we can guarantee convergence for the Gradient Flow (GF) algorithm (see [28] for details). But then, establishing convergence for GD relies on carefully choosing the stepsize η_k, and bounding the rate of change of $\nabla J(L)$—at least on each sublevel set. So, the following lemma provides a Lipschitz bound for $\nabla J(L)$ on every sublevel set. This results resembles its "dual" counterpart in [20, Lemma 7.9], however, it is *not* implied directly by the duality argument as $H^\intercal H$ may not be positive definite.

Lemma 2. *Consider any (non-empty) sublevel set $\mathcal{S}_\alpha$ for some $\alpha > 0$. Then,*

$$\|\nabla J(L_1) - \nabla J(L_2)\|_F \leq \ell \, \|L_1 - L_2\|_F, \quad \forall L_1, L_2 \in \mathcal{S}_\alpha,$$

for some positive constant $\ell = \ell(\alpha) > 0$ that is independent of both L_1 and L_2.

In what follows, we establish linear convergence of the GD algorithm. Our convergence result only depends on the

value of α for the initial sublevel set $\mathcal{S}_\alpha$ that contains L_0. Note that our proof technique is distinct from those in [20] and [34]; nonetheless, it involves a similar argument using the gradient dominance property of J.

Theorem 1. *Consider any sublevel set $\mathcal{S}_\alpha$ for some $\alpha > 0$. Then, for any initial policy $L_0 \in \mathcal{S}_\alpha$, the GD updates with any fixed stepsize $\eta_k = \eta \in (0, 1/\ell(\alpha)]$ converges to optimality at a linear rate of $1 - \eta c_1(\alpha)/2$ (in both the function value and the policy iterate). In particular, we have*

$$J(L_k) - J(L^*) \leq [\alpha - J(L^*)](1 - \eta c_1(\alpha)/2)^k,$$

and $\|L_k - L^*\|_F^2 \leq \left[\frac{\alpha - J(L^*)}{c_3(\alpha)}\right](1 - \eta c_1(\alpha)/2)^k$, with $c_1(\alpha)$ and $c_3(\alpha)$ as defined in Proposition 3.

V. Algorithms and Numerical Simulations

In this section, we discuss numerical algorithms in order to solve the minimization problem (11). Note that, it is not possible to implement the gradient-descent algorithm because evaluating the gradient involves the noise covariance matrices Q and R, assumed to be unknown. Instead, here we explore alternative approaches to recover the gradient information from the data at hand.

1) Stochastic Gradient Descent (SGD): Herein, we allow a variable initial time t_0 (instead of just $t_0 = 0$) for the system (1) and use $\mathcal{Y}_{\{t_0:T\}} := \{y(t_0), y(t_0+1), \dots, y(T-1)\}$ to denote the measurement time-span. Using this notation, the statistical steady-state can be equivalently considered as the limit $t_0 \to -\infty$ with fixed T.

Recall that any choice of $L \in \mathcal{S}$ corresponds to a filtering strategy that outputs a prediction $\hat{y}_L(T)$, which with the variable initial time t_0, is given by

$$\hat{y}_L(T) = H A_L^{T-t_0} m_0 + \textstyle\sum_{t=t_0}^{T-1} H A_L^{T-t-1} L y(t).$$

Also, let $e_{\{t_0:T\}}(L) := y(T) - \hat{y}_L(T)$ denote the incurred error corresponding to this filtering strategy and let

$$\varepsilon(L, \mathcal{Y}_{\{t_0:T\}}) := \|e_{\{t_0:T\}}(L)\|^2,$$

denote the squared-norm of the error, where the dependence on the measurement sequence $\mathcal{Y}_{\{t_0:T\}}$ is explicitly specified.

The optimization objective function is then to minimize the expectation of the squared-norm of the error over all possible random measurement sequences:

$$J_{\{t_0:T\}}(L) := \mathbb{E}\left[\varepsilon(L, \mathcal{Y}_{\{t_0:T\}})\right];$$

at the steady-state, we obtain $\lim_{t_0 \to -\infty} J_{\{t_0:T\}}(L) = J(L)$.

The SGD algorithm aims to solve this optimization problem by replacing the gradient, in the GD update, with an unbiased estimate of the gradient in terms of samples from the measurement sequence. In particular, assuming access to an oracle that produces independent realization of the measurement sequence, say M randomly selected measurements $\{\mathcal{Y}^i_{[t_0,T]}\}_{i=1}^M$, the gradient can be approximated according to

$$\nabla J_{\{t_0:T\}}(L) \approx \tfrac{1}{M} \textstyle\sum_{i=1}^M \nabla_L \varepsilon(L, \mathcal{Y}^i_{\{t_0:T\}}).$$

This forms an unbiased estimate of the gradient, i.e.,

$$\mathbb{E}\left[\tfrac{1}{M} \textstyle\sum_{i=1}^M \nabla_L \varepsilon(L, \mathcal{Y}^i_{\{t_0:T\}})\right] = \nabla J_{\{t_0:T\}}(L),$$

with variance that converges to zero with the rate $O(\frac{1}{M})$ as the number of samples increase. The number M is referred to as the batch-size.

Using the stochastic estimation of the gradient, the algorithm proceeds as follows: we let,

$$\text{[SGD]} \qquad L_{k+1} = L_k - \tfrac{\eta_k}{M} \textstyle\sum_{i=1}^M \nabla_L \varepsilon(L, \mathcal{Y}^i_{\{t_0:T\}}),$$

for $k \in \mathbb{Z}$, where $\eta_k > 0$ is the step-size and $\{\mathcal{Y}^i_{\{t_0:T\}}\}$ represent M fresh realizations of the measurement sequence.

Although the convergence of the SGD algorithm is expected to follow similar to the GD algorithm under the gradient dominance condition and Lipschitz property, the analysis becomes complicated due to the possibility of the iterated gain L_k leaving the sub-level sets. It is expected that a convergence guarantee would hold under high-probability due to concentration of the gradient estimate around the true gradient. Complete analysis in this direction will be presented in our subsequent work.

Finally, for implementation purposes, we compute the gradient estimate explicitly in terms of the measurement sequence and the filtering policy L.

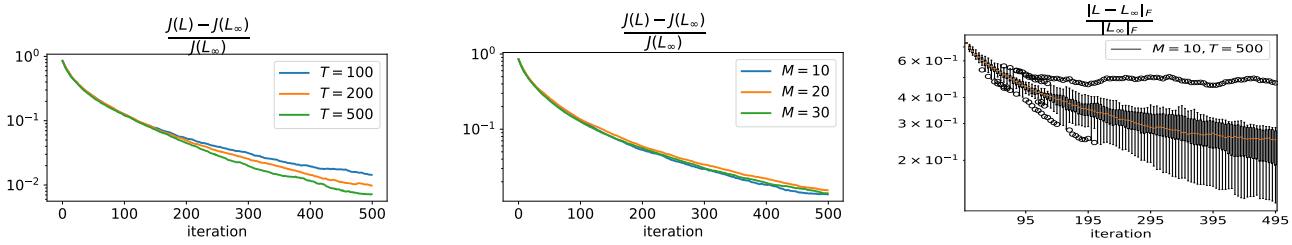
Lemma 3. *Given $L \in \mathcal{S}$ and a sequence of measurements $\mathcal{Y} = \{y(t)\}_{-\infty}^T$, we have,*

$$\nabla_L \varepsilon(L, \mathcal{Y}) = -2 \sum_{t=0}^{\infty} (A_L^\intercal)^t H^\intercal e_T(L) y^\intercal (T-t-1)$$

$$+2 \sum_{t=1}^{\infty} \sum_{k=1}^{t} (A_L^\intercal)^{t-k} H^\intercal e_T(L) y^\intercal (T-t-1) L^\intercal (A_L^\intercal)^{k-1} H^\intercal.$$

Remark 5. Computing the gradient above only requires the knowledge of the system parameters A and H, and does *not* require the noise covariance information Q and R.

2) Numerical Simulations: Herein, we showcase the application of the developed theory for improving the estimation policy for an LTI system. Specifically, we consider an undamped mass-spring system with known parameters (A, H) with $n = 2$ and $m = 1$. In the hindsight, we consider a variance of 0.1 for each state dynamic noise, a state covariance of 0.05 and a variance of 0.1 for the observation noise. Assuming a trajectory of length T at every iteration, the approximate gradient is obtained as in Lemma 3, only requiring an output data sequence collected from the system in (1). Then, the progress of policy updates using the SGD algorithm for different values of trajectory length T and batch size M are depicted in Figure 1 where each figure shows statistics over 20 rounds of simulation. The figure demonstrates a "sublinear rate" of convergence which is expected as every update only relies on an approximation of the gradient—in contrast to the linear convergence established for GD. Finally, Figure 1c demonstrates also the convergence in the Kalman gain as predicted by the properties of J studied in §IV (see Proposition 3).



(a) (b) (c)

Fig. 1: SGD directly from output data and without prior knowledge of the noise covariances or state information. Mean progress of the normalized estimation error over 20 simulations obtained from data trajectories of a) different length T and b) different batch size M; also, c) progress in the Kalman gain with the mean in orange, variance in black line and the outliers in circles.

VI. Conclusions

In this work, we considered the problem of learning the optimal Kalman gain with unknown process and measurement noise covariances. We proposed a direct stochastic PO algorithm with theoretical analysis that are based on the duality between optimal control and estimation. The extension for the other variant of the problem, where the dynamics/observation parameters are also (partially) unknown, is an immediate future direction of this work.

References

[1] R. E. Kalman, "On the general theory of control systems," in *Proceedings First International Conference on Automatic Control, Moscow, USSR*, pp. 481–492, 1960.

[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 03 1960.

[3] J. Pearson, "On the duality between estimation and control," *SIAM Journal on Control*, vol. 4, no. 4, pp. 594–600, 1966.

[4] J. Xiong, *An Introduction to Stochastic Filtering Theory*, vol. 18. OUP Oxford, 2008.

[5] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 15, no. 2, pp. 175–184, 1970.

[6] R. Mehra, "Approaches to adaptive filtering," *IEEE Transactions on Automatic Control*, vol. 17, no. 5, pp. 693–698, 1972.

[7] B. Carew and P. Belanger, "Identification of optimum filter steady-state gain for systems with unknown noise covariances," *IEEE Transactions on Automatic Control*, vol. 18, no. 6, pp. 582–587, 1973.

[8] P. R. Belanger, "Estimation of noise covariance matrices for a linear time-varying stochastic process," *Automatica*, vol. 10, no. 3, pp. 267–275, 1974.

[9] K. Myers and B. Tapley, "Adaptive sequential estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 21, no. 4, pp. 520–523, 1976.

[10] K. Tajima, "Estimation of steady-state Kalman filter gain," *IEEE Transactions on Automatic Control*, vol. 23, no. 5, pp. 944–945, 1978.

[11] D. Magill, "Optimal adaptive estimation of sampled stochastic processes," *IEEE Transactions on Automatic Control*, vol. 10, no. 4, pp. 434–439, 1965.

[12] C. G. Hilborn and D. G. Lainiotis, "Optimal estimation in the presence of unknown parameters," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 1, pp. 38–43, 1969.

[13] P. Matisko and V. Havlena, "Noise covariances estimation for Kalman filter tuning," *IFAC Proceedings Volumes*, vol. 43, no. 10, pp. 31–36, 2010.

[14] R. Kashyap, "Maximum likelihood identification of stochastic linear systems," *IEEE Transactions on Automatic Control*, vol. 15, no. 1, pp. 25–34, 1970.

[15] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[16] B. J. Odelson, M. R. Rajamani, and J. B. Rawlings, "A new autocovariance least-squares method for estimating noise covariances," *Automatica*, vol. 42, no. 2, pp. 303–308, 2006.

[17] B. M. Åkesson, J. B. Jørgensen, N. K. Poulsen, and S. B. Jørgensen, "A generalized autocovariance least-squares method for Kalman filter tuning," *Journal of Process Control*, vol. 18, no. 7-8, pp. 769–779, 2008.

[18] J. Duník, M. Ŝimandl, and O. Straka, "Methods for estimating state and measurement noise covariance matrices: Aspects and comparison," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 372–377, 2009.

[19] L. Zhang, D. Sidoti, A. Bienkowski, K. R. Pattipati, Y. Bar-Shalom, and D. L. Kleinman, "On the identification of noise covariances and adaptive Kalman filtering: A new look at a 50 year-old problem," *IEEE Access*, vol. 8, pp. 59362–59388, 2020.

[20] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[21] J. Bu, A. Mesbahi, and M. Mesbahi, "Policy gradient-based algorithms for continuous-time linear quadratic control," *arXiv preprint arXiv:2006.09178*, 2020.

[22] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 1467–1476, PMLR, 2018.

[23] I. Fatkhullin and B. Polyak, "Optimizing static linear feedback: Gradient method," *SIAM Journal on Control and Optimization*, vol. 59, no. 5, pp. 3887–3911, 2021.

[24] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanovic, "On the linear convergence of random search for discrete-time LQR," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 989–994, 2021.

[25] F. Zhao, K. You, and T. Başar, "Global convergence of policy gradient primal-dual methods for risk-constrained LQRs," *arXiv preprint arXiv:2104.04901*, 2021.

[26] Y. Tang, Y. Zheng, and N. Li, "Analysis of the optimization landscape of linear quadratic gaussian (LQG) control," in *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, vol. 144, pp. 599–610, PMLR, June 2021.

[27] S. Talebi and M. Mesbahi, "Policy optimization over submanifolds for constrained feedback synthesis," *IEEE Transactions on Automatic Control (to appear), arXiv preprint arXiv:2201.11157*, 2022.

[28] S. Talebi, A. Taghvaei, and M. Mesbahi, "Duality-based stochastic policy optimization for estimation with unknown noise covariances," *arXiv preprint arXiv:2210.14878*, 2022.

[29] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, vol. 1072. Wiley-interscience, 1969.

[30] F. Lewis, *Optimal Estimation with an Introduction to Stochastic Control Theory*. New York, Wiley-Interscience, 1986.

[31] K. J. Åström, *Introduction to Stochastic Control Theory*. Courier Corporation, 2012.

[32] J.-W. Kim, P. G. Mehta, and S. P. Meyn, "What is the lagrangian for nonlinear filtering?," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 1607–1614, IEEE, 2019.

[33] J. W. Kim, "Duality for nonlinear filtering," *arXiv preprint arXiv:2207.07709*, 2022.

[34] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2021.

[35] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.