

Can Transformers Learn Optimal Filtering for Unknown Systems?

Zhe Du[✉], Member, IEEE, Haldun Balim[✉], Member, IEEE, Samet Oymak[✉], Member, IEEE, and Necmiye Ozay[✉], Senior Member, IEEE

Abstract—Transformer models have shown great success in natural language processing; however, their potential remains mostly unexplored for dynamical systems. In this letter, we investigate the optimal output estimation problem using transformers, which generate output predictions using all the past ones. Particularly, we train the transformer using various distinct systems and then evaluate the performance on unseen systems with unknown dynamics. Empirically, the trained transformer adapts exceedingly well to different unseen systems and even matches the optimal performance given by the Kalman filter for linear systems. In more complex settings with non-i.i.d. noise, time-varying dynamics, and nonlinear dynamics like a quadrotor system with unknown parameters, transformers also demonstrate promising results. To support our experimental findings, we provide statistical guarantees that quantify the amount of training data required for the transformer to achieve a desired excess risk. Finally, we point out some limitations by identifying two classes of problems that lead to degraded performance, highlighting the need for caution when using transformers for control and estimation.

Index Terms—Filtering, neural networks, statistical learning.

I. INTRODUCTION

MANY control problems such as model predictive control and safety analysis are built upon predictions of system's future trajectories. This prediction (or estimation) problem is well studied and dates back to the classical

Manuscript received 15 August 2023; revised 16 October 2023; accepted 5 November 2023. Date of publication 22 November 2023; date of current version 11 December 2023. The work of Zhe Du was supported by the Army Research Office under Grant W911NF2110312. The work of Haldun Balim was supported in part by ONR under Grant N00014-21-1-2431 (CLEVR-AI), and in part by NSF under Grant CNS 1931982. The work of Samet Oymak was supported in part by the Army Research Office under Grant W911NF2110312, and in part by MIDAS PODS Grant. The work of Necmiye Ozay was supported in part by ONR under Grant N00014-21-1-2431 (CLEVR-AI); in part by NSF under Grant CNS 1931982; and in part by MIDAS PODS Grant. Recommended by Senior Editor T. Oomen. (Zhe Du and Haldun Balim contributed equally to this work.) (Corresponding author: Zhe Du.)

Zhe Du is with the Department of Mechanical Engineering, University of California, Riverside, CA 92521 USA (e-mail: zhedu@umich.edu).

Haldun Balim is with the Department of Mechanical and Process Engineering, ETH Zürich, 8092 Zürich, Switzerland (e-mail: hbalim@ethz.ch).

Samet Oymak and Necmiye Ozay are with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: oymak@umich.edu; necmiye@umich.edu).

Digital Object Identifier 10.1109/LCSYS.2023.3335318

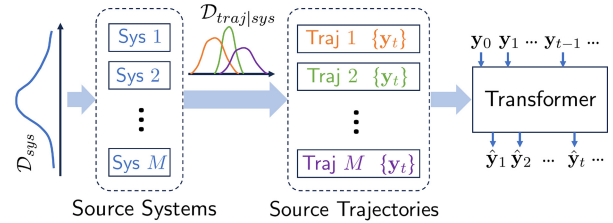


Fig. 1. Training a transformer for dynamical system prediction.

Kalman filter [1], which is optimal for linear systems with Gaussian noise. Methods are also developed for more complex setups, e.g., extended Kalman filter [2] for nonlinear systems, particle filters [3] when system dynamics can be sampled, and adaptive filters and adaptive filters [4] for unknown systems. Existing methods typically require the knowledge of system dynamics, linearity, time-invariance, or Gaussian noise, which, for more challenging and realistic settings, may yield degraded performance.

Prediction, on the other hand, in the domain of natural language processing, has witnessed recent success thanks to the transformer models [5], which are deep learning architectures that can generate text prediction after feeding into an input text sequence. In this letter, we investigate the use of transformers in predicting dynamical system's outputs.

To begin with, we assume a priori access to a collection of M systems drawn from some distribution $\mathcal{D}_{sys}$ and their respective output trajectories $\{y_t\}$. These are referred to as source systems and trajectories respectively. We then train a transformer using the source trajectories so that after feeding into past outputs $y_{0:t-1}$, the transformer is able to produce an estimate $\hat{y}_t$ of the true output y_t (as in Fig. 1). During test-time, given a previously unseen system from the same distribution $\mathcal{D}_{sys}$, we feed its observed trajectory to the trained transformer and evaluate its prediction performance. As discussed in [6], in this setting transformer acts like a data-driven adaptive algorithm: given a system, the transformer is able to automatically adapt to it and make predictions by leveraging past data. In the remainder of this letter, we refer to a transformer trained in this way as meta-output-predictor (MOP).

Contributions: Our first contribution is numerically demonstrating the capabilities of MOP. The experiments show that MOP matches the optimal performance given by the Kalman filter for different unseen linear systems and is able to handle challenging settings such as non-i.i.d. noise, time-varying dynamics, and nonlinear quadrotor systems. Complementing our empirical contributions, we theoretically establish that the

excess risk incurred by MOP decays with rate $\mathcal{O}(1/\sqrt{MT})$ where T is the prediction time horizon, under appropriate assumptions. Motivated by our theoretical analysis, we identify a class of systems with slow mixing properties, for which MOP encounters difficulties in learning the optimal estimator. Our experiments also indicate some limitations of MOP in the presence of distribution shifts.

Related Work: Compared with earlier neural sequence models, transformers [5] incorporate the *attention* mechanism that is able to better keep longer memories thus can handle longer input sequences. As a result, a transformer can be trained to perform a variety of tasks rather than a single task [7], [8], [9], which is known as in-context learning and serves as the foundation of MOP training in our work. Particularly, transformers are shown to be able to in-context learn linear functions [10]; in-context reinforcement learning is studied in [11]. Recent work in [6] studies theoretical properties of transformer-based in-context learning for both i.i.d. data and data with Markovian temporal dependencies (i.e., state trajectories), and provides guarantees in terms of excess risk and transfer risk. Compared with [6], we (i) consider the system output prediction problem with data being non-Markovian, (ii) demonstrate the versatility of MOP through evaluations on several challenging scenarios, and (iii) study scenarios that can lead to degradations in MOP performance.

In terms of filtering/prediction for dynamical systems, there have been many recent advances. When the system dynamics is known, observer design for deterministic systems is studied in [12] through contraction analysis. On the other hand, data-driven adaptive methods have received growing attention. For nonlinear system, techniques such as kernel methods [13] and nonlinear splines [14] are studied. Linear system setups allow for more principled methods such as online optimization [15], explicit [16] or implicit [17], [18] system identification, and policy optimization [19]. Given a class of systems, existing works typically propose algorithms, through which a predictor/filter is learned from data for a specific system. This is in contrast to the framework in our work: Training MOP with various source systems in a class empower MOP the generalizability to the whole class. In other words, the learned MOP is not a specific filter, but a prediction algorithm that can filter any system in the class. And as long as the source systems are representative for the system class, the transformer performance is guaranteed, which is no longer confined by common prerequisites such as dynamics linearity, noise Gaussianity, etc.

II. PROBLEM SETUP

To solve the output estimation problem for an unknown system, we will train a transformer model with data trajectories generated by the following M source systems $\{\mathcal{S}_i\}_{i=1}^M$ drawn from the same distribution $\mathcal{D}_{\text{sys}}$:

$$\mathcal{S}_i : \begin{cases} \mathbf{x}_{i,t+1} = f_i(\mathbf{x}_{i,t}) + \mathbf{w}_{i,t+1} \\ \mathbf{y}_{i,t} = g_i(\mathbf{x}_{i,t}) + \mathbf{v}_{i,t}, \end{cases} \quad (1)$$

where $\mathbf{x}_{i,t} \in \mathbb{R}^n$ and $\mathbf{y}_{i,t} \in \mathbb{R}^m$ are the state and output at time t in the i th system; $f_i(\cdot)$ and $g_i(\cdot)$ are the state dynamics and output functions with $f_i(0) = 0$ and $g_i(0) = 0$; $\mathbf{w}_{i,t} \sim \mathcal{N}(0, \sigma_{\mathbf{w},i}^2 \mathbf{I}_n)$ and $\mathbf{v}_{i,t} \sim \mathcal{N}(0, \sigma_{\mathbf{v},i}^2 \mathbf{I}_m)$ are the process and output noise, which are mutually independent for all i and t . For simplicity, it is assumed that the initial state $\mathbf{x}_{i,0} = 0$. These source systems may be obtained through pre-existing

datasets or simulation environments. The target system under evaluation is denoted by $\mathcal{S}_0$, which is drawn from the same distribution $\mathcal{D}_{\text{sys}}$ and does not have to be contained within the source systems.

We assume that there exists a constant $L_g > 0$ such that for any i , $\mathbf{x}, \mathbf{x}'$, $\|g_i(\mathbf{x}) - g_i(\mathbf{x}')\| \leq L_g \|\mathbf{x} - \mathbf{x}'\|$. Let $\sigma_{\mathbf{w}} := \max_i \sigma_{\mathbf{w},i}$ and $\sigma_{\mathbf{v}} := \max_i \sigma_{\mathbf{v},i}$. Furthermore, we assume these systems satisfy the following stability condition.

Assumption 1 (Stability): Let $f_i^{(t)}(\cdot, \cdot)$ denote the t -step state evolution function such that $f_i^{(t)}(\mathbf{x}_{i,\tau}, \mathbf{w}_{i,\tau+1:\tau+t}) = \mathbf{x}_{i,\tau+t}$ for all $\tau \in \mathbb{N}$. Then, there exists constants $\rho \in [0, 1)$ and $C_\rho > 0$ such that for any system i and time step t , for any $\mathbf{x}, \mathbf{x}'$ and noise sequence $\mathcal{W} := \{\mathbf{w}_{i,\tau}\}_{\tau \in [t]}$, we have

$$\|f_i^{(t)}(\mathbf{x}, \mathcal{W}) - f_i^{(t)}(\mathbf{x}', \mathcal{W})\| \leq C_\rho \rho^t \|\mathbf{x} - \mathbf{x}'\|. \quad (2)$$

We define the notation $L_\rho := \frac{C_\rho}{1-\rho}$. When the class of dynamical systems we are sampling from consists of linear systems with $f_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x}$, then Assumption 1 is satisfied when the spectral radius $\rho(\mathbf{A}_i) < 1$ for all i . It is also satisfied by systems that are contracting [20] or exponentially incrementally input-to-state stable [21] with input $\mathbf{w}$.

In this letter, we seek to predict system output using a transformer [5], which is a deep sequence model $\text{TF}_\theta(\cdot)$ that maps system output sequences $\mathcal{Y}_t := \mathbf{y}_{0:t}$ to $\hat{\mathbf{y}}_{t+1} := \text{TF}_\theta(\mathcal{Y}_t)$, an estimation of the true output $\mathbf{y}_{t+1}$ at time $t+1$. The trainable parameters of the transformer are denoted by $\theta \in \Theta$ for some parameter set Θ . The transformer structure allows the sequence length t to be varying.

Assuming the access to M length- T output trajectories $\{\mathbf{y}_{i,0:T}\}_{i=1}^M$ generated by each of the M source systems, the goal in this letter is to train a transformer model that, at each time t , can predict the output $\mathbf{y}_{0,t+1}$ of the target system $\mathcal{S}_0$ only using the past outputs $\mathbf{y}_{0,0:t}$. Let $\mathcal{Y}_{i,t} := \mathbf{y}_{i,0:t}$ denote the outputs up to time t , which is also known as the *prompt* (to predict $\mathbf{y}_{i,t+1}$), then the transformer is trained by solving

$$\widehat{\text{TF}} = \arg \min_{\text{TF} \in \mathcal{A}} \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T \ell(\mathbf{y}_{i,t}, \text{TF}(\mathcal{Y}_{i,t-1})), \quad (3)$$

where $\mathcal{A} := \{\text{TF}_\theta : \theta \in \Theta\}$ and $\ell(\cdot, \cdot) \geq 0$ is the loss function. To apply $\widehat{\text{TF}}(\cdot)$ to the target systems $\mathcal{S}_0$, we simply take $\widehat{\text{TF}}(\mathcal{Y}_{0,t})$ as the prediction for $\mathbf{y}_{0,t+1}$.

Training a model as in (3) where the data comes from a diversity of sources is also known as in-context learning. As a result of the training diversity, the transformer can achieve good performance on any of the source systems as well as demonstrate generalization ability for the unseen target system $\mathcal{S}_0$. Hence, we refer to the obtained transformer $\widehat{\text{TF}}$ as meta-output-predictor (MOP).

In what follows, we first empirically demonstrate the performance of MOP in Section III under various setups, which is followed by theoretical analysis in Section IV.

III. EXPERIMENTS

In this section, we present the experimental results for the transformer-based MOP in different scenarios. In each scenario, during the training, we fix the number of source systems $M = 20000$ and training trajectory length $T = 50$. To evaluate the performance of MOP on different unseen test systems, for each experimental setup, we randomly generate

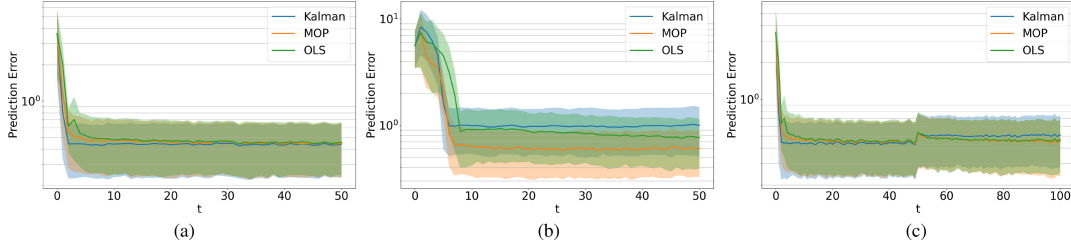


Fig. 2. Output predictions for linear systems: (a) with i.i.d. Gaussian noise; (b) with colored (non-i.i.d.) noise; (c) with dynamics changes at $t = T/2$. MOP performs as well as or better than Kalman filter even though it does not have access to the system dynamics.

$N = 1000$ systems and record the prediction error $\|\hat{\mathbf{y}}_t - \mathbf{y}_t\|$ over trajectories each with length $T = 50$, where $\hat{\mathbf{y}}_t$ denotes the prediction for $\mathbf{y}_t$. We use GPT-2 [22] architecture with 12 layers, 8 attention heads and 256 embedding dimensions. In each experimental setup, the transformer model is trained for 10000 training steps with batch size 64. The ℓ_2 -norm is selected as the training loss function. The code we use to produce the figures and execute our algorithm can be accessed at <https://github.com/haldunbalim/Meta-Output-Predictor>.

A. Linear Systems

We first consider the simplest setting with linear systems and i.i.d. Gaussian noise, i.e., $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $g(\mathbf{x}) = \mathbf{C}\mathbf{x}$ in (1). The state dimension is $n = 10$ and the output dimension is $m = 5$. For each source and test system, we generate matrix $\mathbf{A}$ with entries sampled uniformly between $[-1, 1]$, which is then followed by scaling so that the largest eigenvalue is 0.95. The $\mathbf{C}$ matrix is generated with entries sampled uniformly between $[0, 1]$. The noise covariance are $\sigma_w^2 = 0.01$ and $\sigma_v^2 = 0.01$. Kalman filter and linear autoregressive predictor are used as baselines, where the latter is given by $\hat{\mathbf{y}}_{t+1} = \alpha_1 \mathbf{y}_t + \alpha_2 \mathbf{y}_{t-1}$ and the matrix parameters (α_1, α_2) are updated in an online fashion using the ordinary least squares (OLS). The results are presented in Fig. 2(a). We see that after some burn-in time (~ 20 steps), MOP eventually matches the performance of Kalman filter. This is because the transformer needs to collect certain amount ($\mathcal{O}(n+m)$) of data to implicitly learn the system dynamics, while Kalman filter, designed with the exact system knowledge, reaches optimality immediately.

In the next experiment, we consider the case where the noise process $\{\mathbf{w}_t\}$ is non-i.i.d. Specifically, we let $\mathbf{w}_t = \sum_{t'=t-4}^t \boldsymbol{\eta}_{t'}$ and $\mathbf{v}_t = \sum_{t'=t-4}^t \boldsymbol{\gamma}_{t'}$ where $\boldsymbol{\eta}_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$ and $\boldsymbol{\gamma}_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. When applying the Kalman filter in this case, we disregard the fact that $\mathbf{w}_t$ and $\mathbf{v}_t$ each are temporally correlated and simply use the variances of $\mathbf{w}_t$ and $\mathbf{v}_t$ for prediction. We note that for non-i.i.d. noise Kalman filter is no longer optimal. Fig. 2(b) shows the results for this case. We can observe the advantage of MOP over Kalman filter as Kalman filter has lost its optimality whereas MOP has learned the non-i.i.d noise prior during training.

Next, we evaluate the ability of MOP to adapt to run-time changes in the dynamics. Specifically, when generating the test trajectories, we change the underlying dynamics to a randomly generated new one at time $t = T/2 = 50$. The results are presented in Fig. 2(c). We see that when dynamics changes occur, there are sudden jumps in prediction error for both MOP and the Kalman filter; as we collect more data from the new dynamics, MOP quickly adapts, and achieves the same performance as before at around $t = 100$. The convergence of MOP after dynamics changes is much slower than the one at

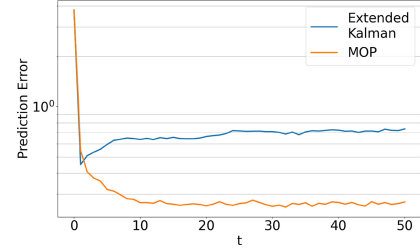


Fig. 3. Output predictions for planar quadrotor systems.

the beginning because the prompt always contains data from the original system.

B. Planar Quadrotor Systems

We consider the underactuated 6D planar quadrotor systems as in [23] with the following discrete-time dynamics:

$$\underbrace{\begin{bmatrix} x_{t+1} \\ z_{t+1} \\ \phi_{t+1} \\ \dot{x}_{t+1} \\ \dot{z}_{t+1} \\ \dot{\phi}_{t+1} \end{bmatrix}}_{=\mathbf{x}_{t+1}} = \begin{bmatrix} x_t + (\dot{x}_t \cos(\phi_t) - \dot{z}_t \sin(\phi_t))\tau \\ z_t + (\dot{x}_t \sin(\phi_t) + \dot{z}_t \cos(\phi_t))\tau \\ \phi_t + \dot{\phi}_t \tau \\ \dot{x}_t + (\dot{z}_t \phi_t - g \sin(\phi_t))\tau \\ \dot{z}_t + (-\dot{x}_t \phi_t - g \cos(\phi_t) + (u_{0t} + u_{1t})/m)\tau \\ (u_{0t} - u_{1t})l\tau/J \end{bmatrix} + \mathbf{w}_t$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{v}_t.$$

The mass, length and moment of inertia parameters (m, l, J) are chosen uniformly from $[0.5, 2]$, g is set to be constant 10. For each system a trajectory is generated by randomly sampled actions. The noise $\mathbf{w}, \mathbf{v}$ are sampled from $\mathcal{N}(0, 0.01)$. The discretization time $\tau = 0.1$. The matrix $\mathbf{C} \in \mathbb{R}^{3 \times 6}$ has elements uniformly sampled in $[0, 1]$. The results are provided in Fig. 3. We see that MOP significantly outperforms the extended Kalman filter.

IV. THEORETICAL GUARANTEES

Before analyzing the performance of MOP $\widehat{\text{TF}}$, we first introduce a few notions and assumptions. The analysis in this section generalizes that in [6], which studies a special case where state is observed (i.e., g is known and equal to the identity map and there is no measurement noise).

A. Preliminaries

Definition 1 (Covering Number): Consider a set $\mathcal{Q}$ and a distance metric $d(\cdot, \cdot)$ on $\mathcal{Q}$. For a set $\mathcal{Q}_N := \{q_1, \dots, q_N\}$, we say it is an ϵ -cover of $\mathcal{Q}$ if for any $q \in \mathcal{Q}$, there exists $q_i \in \mathcal{Q}_N$ such that $d(q, q_i) \leq \epsilon$. The number $\mathcal{N}(\mathcal{Q}, d, \epsilon)$ is the smallest $N \in \mathbb{N}$ such that $\mathcal{Q}_N$ is an ϵ -cover of $\mathcal{Q}$.

We will analyze the set $\mathcal{A}$ through its ϵ -cover. To do so, we define the following distance on $\mathcal{A}$.

Definition 2 (Distance Metric): Let $\mathcal{Y}_T := \mathbf{y}_{0:T}$ denote a trajectory of some system i under the noise sequence $\{\mathbf{w}_{0:T}, \mathbf{v}_{0:T}\}$. For any two transformers $\text{TF}, \text{TF}' \in \mathcal{A}$, define the distance metric $\mu(\text{TF}, \text{TF}') := \sup_{t \leq T} \sup_{\mathbf{w}_{0:t}, \mathbf{v}_{0:t}} \frac{\|\text{TF}(\mathcal{Y}_{t-1}) - \text{TF}'(\mathcal{Y}_{t-1})\|}{\max_{\tau \leq t} \|\mathbf{w}_{\tau-1}\| + \max_{\tau \leq t} \|\mathbf{v}_{\tau-1}\|}$.

Though this metric is regarding the transformers TF, TF' in the transformer space $\mathcal{A}$, it can be viewed as a metric between their respective parameters θ, θ' in the parameter set Θ . When the noise is bounded, the distance can also be defined without using the normalization factor in the denominator, e.g., [6]. Next, we quantify the robustness of the transformer in terms of how much the prediction changes with respect to the perturbations of its prompt. This will help us establish generalization bounds for the transformer trained in (3).

Assumption 2 (Transformer Robustness): Consider a trajectory $\{\mathbf{y}_{0:t}\}$ generated by some system i under the noise sequence $\{\mathbf{w}_{0:t}, \mathbf{v}_{0:t}\}$. Let $\{\mathbf{y}'_{0:t}\}$ denote another trajectory under the same noise except that at time τ , $\{\mathbf{w}_\tau, \mathbf{v}_\tau\}$ is replaced by $\{\mathbf{w}'_\tau, \mathbf{v}'_\tau\}$. Let $\mathcal{Y}_t := \mathbf{y}_{0:t}$ and $\mathcal{Y}'_t := \mathbf{y}'_{0:t}$. Suppose the loss function $\ell(\mathbf{y}, \cdot)$ is L_ℓ -Lipschitz. Let $\mathcal{B} := \cap_{j=0}^t \{\|\mathbf{w}_j\| \leq \bar{w}, \|\mathbf{v}_j\| \leq \bar{v}\}$ for some $\bar{w}, \bar{v} \geq 0$ and $\mathbb{E}[\cdot] := \mathbb{E}[\cdot | \mathcal{B}]$. Then, there exist constants $K \geq 0$ such that for any system i , any t and $\{\mathbf{w}_{0:t-1}, \mathbf{v}_{0:t-1}\}$, any $\tau < t$ and $\{\mathbf{w}'_\tau, \mathbf{v}'_\tau\}$, and any $\text{TF} \in \mathcal{A}$, we have

$$\begin{aligned} & \mathbb{E}'_{\mathbf{w}_t, \mathbf{v}_t} [|\ell(\mathbf{y}_t, \text{TF}(\mathcal{Y}_{t-1})) - \ell(\mathbf{y}_t, \text{TF}(\mathcal{Y}'_{t-1}))|] \\ & \leq \frac{KL_\ell}{t-\tau} \sum_{j=\tau}^{t-1} \|\mathbf{y}_j - \mathbf{y}'_j\|. \end{aligned}$$

In this assumption, trajectories $\mathbf{y}'_j = \mathbf{y}_j$ for $j < \tau$ and possibly differ afterward due to the perturbation at time τ , which explains the summation term in the upper bound. It is shown in [6, Lemma B.5] that this assumption holds for a wide class of transformers.

B. Performance Guarantees

For a transformer $\text{TF} \in \mathcal{A}$, we define the following risk to evaluate its performance on the target system S_0 over the time horizon T

$$\mathcal{L}(\text{TF}) := \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\ell(\mathbf{y}_{0:t}, \text{TF}(\mathcal{Y}_{0,t-1}))], \quad (4)$$

where the expectation is over the target system S_0 and noise terms $\{\mathbf{w}_{0:t}, \mathbf{v}_{0:t}\}$. Let $\text{TF}^* \in \mathcal{A}$ denote an optimal transformer that minimizes $\mathcal{L}(\text{TF})$. Define the excess risk for $\widehat{\text{TF}}$ obtained via minimizing the loss in (3) as

$$\text{Risk}(\widehat{\text{TF}}) := \mathcal{L}(\widehat{\text{TF}}) - \mathcal{L}(\text{TF}^*). \quad (5)$$

Then, we have the following performance guarantees on $\widehat{\text{TF}}$.

Theorem 1: Suppose Assumptions 1 and 2 hold, and the loss function $\ell(\mathbf{y}, \cdot)$ is L_ℓ -Lipschitz and $\ell(\cdot, \cdot) \leq B$ for some $B \geq 0$. Then, when $MT \geq 3 \max(\sqrt{n}, \sqrt{m})$, for all $\epsilon > 0$, with probability at least $1 - \delta$,

$$\text{Risk}(\widehat{\text{TF}}) \leq 12B\delta + 4L_\ell\epsilon + \bar{B} \sqrt{\frac{\log(4\mathcal{N}(\mathcal{A}, \mu, \epsilon')/\delta)}{cMT}},$$

where $\bar{B} := 2B + 7KL_\ell(L_g L_\rho \sigma_{\mathbf{w}} + \sigma_{\mathbf{v}}) \sqrt{\log(4MT/\delta) \log(T)}$; $\epsilon' := \epsilon / ((\sigma_{\mathbf{w}} + \sigma_{\mathbf{v}}) \sqrt{\log(4MT/\delta)})$; c is some absolute constant;

$\mathcal{N}(\cdot, \cdot, \cdot)$ and μ are the covering number and distance metric from Definitions 1 and 2.

For fixed failure probability δ and distance ϵ , the upper bound decays with rate $\mathcal{O}(1/\sqrt{MT})$. When the transformer mapping is Lipschitz, the covering number term can be upper bounded by $\log \mathcal{N}(\mathcal{A}, \mu, \epsilon') \leq \mathcal{O}(n_\Theta \log(\bar{\Theta} \sqrt{n_\Theta}/\epsilon'))$, where n_Θ and $\bar{\Theta}$ are respectively the dimension and magnitude of the transformer parameter set Θ .

C. Proof of the Main Theorem

In this section, we provide the proof for Theorem 1. Extending Assumption 2, the following lemma tells how the noise $\{\mathbf{w}, \mathbf{v}\}$ would affect the loss performance.

Lemma 1: Suppose Assumptions 1 and 2 hold. Under the same setup as in Assumption 2, let $\bar{\mathbf{y}} := L_g L_\rho \bar{\mathbf{w}} + \bar{\mathbf{v}}$. Then,

$$\mathbb{E}'_{\mathbf{w}_t, \mathbf{v}_t} [|\ell(\mathbf{y}_t, \text{TF}(\mathcal{Y}_{t-1})) - \ell(\mathbf{y}_t, \text{TF}(\mathcal{Y}'_{t-1}))|] \leq \frac{2KL_\ell \bar{\mathbf{y}}}{t-\tau}.$$

Proof: From Assumption 2, it only suffices to show $\sum_{j=\tau}^{t-1} \|\mathbf{y}_j - \mathbf{y}'_j\| \leq 2\bar{\mathbf{y}}$. In Assumption 2, as a result of perturbing the noise sequence $\{\mathbf{w}_{0:t}, \mathbf{v}_{0:t}\}$ at time τ , the original sequence $\{\mathbf{x}_{0:t}, \mathbf{y}_{0:t}\}$ and the perturbed sequence $\{\mathbf{x}'_{0:t}, \mathbf{y}'_{0:t}\}$ are the same up to time $\tau - 1$ and possibly differ afterward. For $j = \tau, \dots, t$, according to the stability in Assumption 1, we have $\|\mathbf{x}_j - \mathbf{x}'_j\| \leq C_\rho \rho^{j-\tau} \|\mathbf{w}_\tau - \mathbf{w}'_\tau\|$. Since, for all i , $g_i(\cdot)$ is assumed Lipschitz, for $j = \tau$, we have $\|\mathbf{y}_j - \mathbf{y}'_j\| \leq L_g \|\mathbf{x}_j - \mathbf{x}'_j\| + \|\mathbf{v}_\tau - \mathbf{v}'_\tau\| \leq L_g C_\rho \rho^{j-\tau} \|\mathbf{w}_\tau - \mathbf{w}'_\tau\| + \|\mathbf{v}_\tau - \mathbf{v}'_\tau\|$; similarly for $j > \tau$, we have $\|\mathbf{y}_j - \mathbf{y}'_j\| \leq L_g C_\rho \rho^{j-\tau} \|\mathbf{w}_\tau - \mathbf{w}'_\tau\|$. Taking the summation gives that $\sum_{j=\tau}^{t-1} \|\mathbf{y}_j - \mathbf{y}'_j\| \leq L_g L_\rho \|\mathbf{w}_\tau - \mathbf{w}'_\tau\| + \|\mathbf{v}_\tau - \mathbf{v}'_\tau\| \leq 2(L_g L_\rho \bar{\mathbf{w}} + \bar{\mathbf{v}}) = 2\bar{\mathbf{y}}$. ■

Proof for Theorem 1: To bound the excess risk $\text{Risk}(\widehat{\text{TF}}) := \mathcal{L}(\widehat{\text{TF}}) - \mathcal{L}(\text{TF}^*)$ in (5), we first define the following empirical risk on the M source systems

$$\hat{\mathcal{L}}(\text{TF}) := \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T \underbrace{\ell(\mathbf{y}_{i,t}, \text{TF}(\mathcal{Y}_{i,t-1}))}_{=: \ell_{i,t}}, \quad (6)$$

Noticing that $\widehat{\text{TF}} = \arg \min_{\text{TF} \in \mathcal{A}} \hat{\mathcal{L}}(\text{TF})$, the decomposition $\text{Risk}(\widehat{\text{TF}}) = (\mathcal{L}(\widehat{\text{TF}}) - \hat{\mathcal{L}}(\widehat{\text{TF}})) + (\hat{\mathcal{L}}(\widehat{\text{TF}}) - \hat{\mathcal{L}}(\text{TF}^*)) + (\hat{\mathcal{L}}(\text{TF}^*) - \mathcal{L}(\text{TF}^*))$ becomes $\text{Risk}(\widehat{\text{TF}}) \leq (\mathcal{L}(\widehat{\text{TF}}) - \hat{\mathcal{L}}(\widehat{\text{TF}})) + (\hat{\mathcal{L}}(\text{TF}^*) - \mathcal{L}(\text{TF}^*))$. This further gives

$$\text{Risk}(\widehat{\text{TF}}) \leq 2 \sup_{\text{TF} \in \mathcal{A}} |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|. \quad (7)$$

In the following, we proceed as follows: (i) assume the noise sequence $\{\mathbf{w}_{i,t}, \mathbf{v}_{i,t}\}$ is bounded and show that for any $\text{TF} \in \mathcal{A}$, $|\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|$ is bounded; (ii) use a covering number argument to bound $\sup_{\text{TF} \in \mathcal{A}} |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|$; (iii) show $\{\mathbf{w}_{i,t}, \mathbf{v}_{i,t}\}$ can be bounded with high probability.

Step (i): Upper bound $|\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|$.

Define the following risks for the system $i = 0, 1, \dots, M$

$$\hat{\mathcal{L}}_i(\text{TF}) := T^{-1} \sum_{t=1}^T \ell_{i,t}, \quad \mathcal{L}_i(\text{TF}) := T^{-1} \sum_{t=1}^T \mathbb{E}[\ell_{i,t}].$$

This gives $\hat{\mathcal{L}}(\text{TF}) = M^{-1} \sum_{i=1}^M \hat{\mathcal{L}}_i(\text{TF})$ and $\mathcal{L}(\text{TF}) = \mathcal{L}_0(\text{TF}) = M^{-1} \sum_{i=1}^M \mathcal{L}_i(\text{TF})$ since $\mathcal{L}_i(\text{TF})$ i.i.d. for all i . We then have $|\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})| = |M^{-1} \sum_{i=1}^M (\hat{\mathcal{L}}_i(\text{TF}) - \mathcal{L}_i(\text{TF}))|$. We

will bound each individual $\hat{\mathcal{L}}_i(\text{TF}) - \mathcal{L}_i(\text{TF})$ and then apply concentration result to bound $|\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|$.

Define the event $\mathcal{B}_M := \cap_{i=0}^M \cap_{j=0}^T \{\|\mathbf{w}_{i,j}\| \leq \bar{w}, \|\mathbf{v}_{i,j}\| \leq \bar{v}\}$ for some $\bar{w}, \bar{v} \geq 0$, and Let $\mathbb{P}'(\cdot) := \mathbb{P}(\cdot | \mathcal{B}_M)$ and $\mathbb{E}'[\cdot] := \mathbb{E}[\cdot | \mathcal{B}_M]$ denote the probability measure and expectation conditioning on the event $\mathcal{B}_M$. Let $\mathcal{S}_{i,t} := \{\mathbf{w}_{i,0:t}, \mathbf{v}_{i,0:t}\}$ for $t \geq 1$ and $\mathcal{S}_{i,0} := \phi$. Define $X_{i,t} := \mathbb{E}'[\hat{\mathcal{L}}_i(\text{TF}) | \mathcal{S}_{i,t}]$, then the process $\{X_{i,t}\}_{t=0}^T$ forms a Doob's martingale. Particularly, note that $X_{i,T} = \hat{\mathcal{L}}_i(\text{TF})$ and $X_{i,0} = \mathbb{E}'[\hat{\mathcal{L}}_i(\text{TF})]$. Consider the martingale difference $X_{i,\tau} - X_{i,\tau-1}$, we have

$$\begin{aligned} & |X_{i,\tau} - X_{i,\tau-1}| \\ &= \left| T^{-1} \sum_{t=1}^T \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau}] - \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau-1}] \right| \\ &= \left| T^{-1} \sum_{t=\tau}^T \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau}] - \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau-1}] \right| \\ &\leq B/T + T^{-1} \sum_{t=\tau+1}^T \left| \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau}] - \mathbb{E}'[\ell_{i,t} | \mathcal{S}_{i,\tau-1}] \right|, \end{aligned}$$

where the last line used the fact $\ell(\cdot, \cdot) \leq B$. Note that each summand in the above summation can be upper bounded by $\frac{2KL\bar{y}}{T-\tau}$ according to Lemma 1. The equation above then gives $|X_{i,\tau} - X_{i,\tau-1}| \leq T^{-1}(B + 2KL\bar{y}\log(T))$. With this martingale difference bound, applying the Azuma-Hoeffding's inequality to $\{X_{i,t}\}_{t=0}^T$ gives

$$\begin{aligned} \mathbb{P}'(|X_{i,T} - X_{i,0}| \geq \epsilon) &= \mathbb{P}'\left(|\hat{\mathcal{L}}_i(\text{TF}) - \mathcal{L}_i(\text{TF}) - \Delta_i| \geq \epsilon | \mathcal{B}_M\right) \\ &\leq 2e^{-\frac{\epsilon^2}{(B+2KL\bar{y}\log(T))^2}}. \end{aligned}$$

where $\Delta_i := \mathcal{L}_i(\text{TF}) - \mathbb{E}'[\hat{\mathcal{L}}_i(\text{TF})]$. Let $Y_i := \hat{\mathcal{L}}_i(\text{TF}) - \mathcal{L}_i(\text{TF}) - \Delta_i$, then the above equation tells that Y_i is sub-Gaussian conditioning on $\mathcal{B}_M$. Following from sub-Gaussian concentration bound, we have

$$\mathbb{P}(M^{-1} \left| \sum_{i=1}^M Y_i \right| \geq \epsilon | \mathcal{B}_M) \leq e^{-\frac{cMT\epsilon^2}{(B+2KL\bar{y}\log(T))^2}}, \quad (8)$$

for some absolute constant c . This further translates to, conditioning on $\mathcal{B}_M$, with probability at least $1 - \delta$,

$$\left| M^{-1} \sum_{i=1}^M Y_i \right| \leq (B + 2KL\bar{y}\log(T))\sqrt{\log(2/\delta)/(cMT)}.$$

The definition of Y_i gives $|M^{-1} \sum_{i=1}^M Y_i| = |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF}) - M^{-1} \sum_{i=1}^M \Delta_i| \geq |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})| - |M^{-1} \sum_{i=1}^M \Delta_i|$. This implies, conditioning on $\mathcal{B}_M$, with probability at least $1 - \delta$,

$$\begin{aligned} |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})| &\leq \left| M^{-1} \sum_{i=1}^M \Delta_i \right| \\ &+ (B + 2KL\bar{y}\log(T))\sqrt{\log(2/\delta)/(cMT)}. \end{aligned} \quad (9)$$

Step (ii): Upper bound $\sup_{\text{TF} \in \mathcal{A}} |\hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})|$.

Let $h(\text{TF}) := \hat{\mathcal{L}}(\text{TF}) - \mathcal{L}(\text{TF})$, here we seek to upper bound $\sup_{\text{TF} \in \mathcal{A}} |h(\text{TF})|$. For $\epsilon > 0$, let $\epsilon' := \epsilon/(\bar{w} + \bar{v})$ and let $\mathcal{A}_{\epsilon'}$ denote the minimal ϵ' -covering of $\mathcal{A}$, under the distance μ in Definition 2. Note that $|\mathcal{A}_{\epsilon'}| = \mathcal{N}(\mathcal{A}, \mu, \epsilon')$. This gives that,

$$\sup_{\text{TF} \in \mathcal{A}} |h(\text{TF})| \leq \max_{\text{TF} \in \mathcal{A}_{\epsilon'}} |h(\text{TF})| + \sup_{\text{TF} \in \mathcal{A}} \min_{\text{TF}' \in \mathcal{A}_{\epsilon'}} |h(\text{TF}) - h(\text{TF}')|. \quad (10)$$

For the term $\max_{\text{TF} \in \mathcal{A}_{\epsilon'}} |h(\text{TF})|$, applying the union bound to (7) for all $\text{TF} \in \mathcal{A}_{\epsilon'}$, we obtain that conditioning on $\mathcal{B}_M$, with probability at least $1 - \delta$,

$$\begin{aligned} \max_{\text{TF} \in \mathcal{A}_{\epsilon'}} |h(\text{TF})| &\leq \left| M^{-1} \sum_{i=1}^M \Delta_i \right| \\ &+ (B + 2KL\bar{y}\log(T))\sqrt{\log(2\mathcal{N}(\mathcal{A}, \mu, \epsilon')/\delta)/(cMT)}. \end{aligned} \quad (11)$$

Next we consider the term $\sup_{\text{TF} \in \mathcal{A}} \min_{\text{TF}' \in \mathcal{A}_{\epsilon'}} |h(\text{TF}) - h(\text{TF}')|$ in (10). Let $\mathcal{L}'(\text{TF}) := \mathbb{E}[\hat{\mathcal{L}}_0(\text{TF}) | \mathcal{B}_M]$, $\Delta_{h,1} := |\hat{\mathcal{L}}(\text{TF}) - \hat{\mathcal{L}}(\text{TF}')|$, $\Delta_{h,2} := |\mathcal{L}'(\text{TF}) - \mathcal{L}'(\text{TF}')|$, and $\Delta_{h,3} := |\mathcal{L}(\text{TF}) - \mathcal{L}'(\text{TF})| + |\mathcal{L}(\text{TF}') - \mathcal{L}'(\text{TF}')|$. Using the definition of $h(\cdot)$ and the triangular inequality, we have $|h(\text{TF}) - h(\text{TF}')| \leq \Delta_{h,1} + \Delta_{h,2} + \Delta_{h,3}$. By the Lipschitzness of the loss function ℓ and the bound on the distance between TF and TF' , i.e., $\mu(\text{TF}, \text{TF}') \leq \epsilon/(\bar{w} + \bar{v})$, we obtain that, conditioning on $\mathcal{B}_M$, both $\Delta_{h,1}$, $\Delta_{h,2}$ can be upper bounded by $L_\ell \epsilon$. This gives

$$\sup_{\text{TF} \in \mathcal{A}} \min_{\text{TF}' \in \mathcal{A}_{\epsilon'}} |h(\text{TF}) - h(\text{TF}')| \leq 2L_\ell \epsilon + \Delta_{h,3}. \quad (12)$$

Plugging (12) and (11) into (10) followed by invoking $\text{Risk}(\widehat{\text{TF}}) \leq 2 \sup_{\text{TF} \in \mathcal{A}} |h(\text{TF})|$ in (7) gives that, conditioning on $\mathcal{B}_M$, with probability at least $1 - \delta$

$$\begin{aligned} \text{Risk}(\widehat{\text{TF}}) &\leq 4L_\ell \epsilon + 2\Delta_{h,3} + 2 \left| M^{-1} \sum_{i=1}^M \Delta_i \right| \\ &+ 2(B + 2KL\bar{y}\log(T))\sqrt{\log(2\mathcal{N}(\mathcal{A}, \mu, \epsilon')/\delta)/(cMT)}. \end{aligned} \quad (13)$$

Step (iii): Upper bound the noise sequence $\{\mathbf{w}_{i,t}, \mathbf{v}_{i,t}\}$.

Let $\mathcal{E}$ denote the event in (13), then we have $\mathbb{P}(\mathcal{E} | \mathcal{B}_M) \geq 1 - \delta$. In the event $\mathcal{B}_M$, we now set $\bar{w} = \sqrt{3\sigma_w \sqrt{\log(2MT/\delta)}}$ and $\bar{v} = \sqrt{3\sigma_v \sqrt{\log(2MT/\delta)}}$. Using the Gaussian concentration bound and the union bound, we obtain that $\mathbb{P}(\mathcal{B}_M) \geq 1 - \delta$, when $MT \geq 3 \max(\sqrt{n}, \sqrt{m})$. This further yields

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{E}, \mathcal{B}_M) \geq \mathbb{P}(\mathcal{E} | \mathcal{B}_M) \mathbb{P}(\mathcal{B}_M) \geq (1 - \delta)^2 \geq 1 - 2\delta. \quad (14)$$

Now we inspect the term $|M^{-1} \sum_{i=1}^M \Delta_i|$ in the definition of $\mathcal{E}$, i.e., (13). Note that by definition $|\Delta_i| = |\mathbb{E}[\hat{\mathcal{L}}_i(\text{TF})] - \mathbb{E}[\hat{\mathcal{L}}_i(\text{TF}) | \mathcal{B}_M]| \leq |\mathbb{E}[\hat{\mathcal{L}}_i(\text{TF}) | \mathcal{B}_M] (\mathbb{P}(\mathcal{B}_M) - 1)| + |\mathbb{E}[\hat{\mathcal{L}}_i(\text{TF}) | \mathcal{B}_M^c] \mathbb{P}(\mathcal{B}_M^c)| \leq 2B\delta$, where the facts $\ell(\cdot, \cdot) \leq B$ and complement probability $\mathbb{P}(\mathcal{B}_M^c) \leq \delta$ are used. Hence, we have $|M^{-1} \sum_{i=1}^M \Delta_i| \leq 2B\delta$. Similarly, we can show $\Delta_{h,3} \leq 4B\delta$. With these bounds, invoking (14) gives, with probability at least $1 - 2\delta$,

$$\begin{aligned} \text{Risk}(\widehat{\text{TF}}) &\leq +4L_\ell \epsilon + 12B\delta \\ &+ 2(B + 2KL\bar{y}\log(T))\sqrt{\log(2\mathcal{N}(\mathcal{A}, \mu, \epsilon')/\delta)/(cMT)}. \end{aligned}$$

Finally, plugging in the definition $\bar{y} := L_g L_\rho \bar{w} + \bar{v}$ and $\epsilon' := \epsilon/(\bar{w} + \bar{v})$ concludes the proof. ■

V. SYSTEMS THAT ARE HARD TO LEARN IN-CONTEXT

In this section, we investigate two limitations of MOP, one explained by our theoretical guarantees, the other regarding the performance degradation in the face of distribution shifts.

To illustrate the first challenge, consider two distinct classes of linear systems. The first class employs the same generation procedure as described in Section III-A. In the second class, we follow a similar generation procedure, except for the $\mathbf{A}$ matrices, which are generated as upper-triangular matrices.

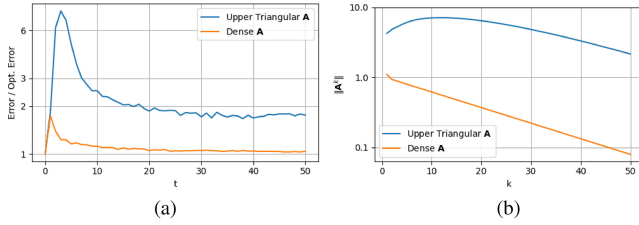


Fig. 4. Comparison between dense and upper-triangular $\mathbf{A}$ matrices: (a) prediction error ratio between MOP and Kalman filter; (b) matrix powers averaged over all source systems.

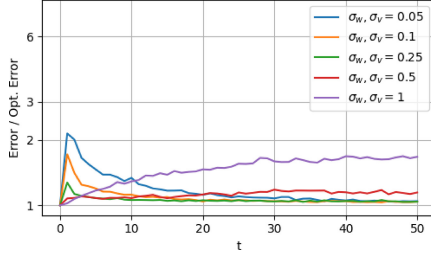


Fig. 5. Performance of MOP compared to Kalman Filter when noise level in test is different than train.

Here, the diagonal entries are sampled from the interval $[-0.95, 0.95]$, while the upper triangular entries are sampled from the range $[-1, 1]$. The experimental results, presented in Fig. 4, demonstrate that, compared with the densely generated $\mathbf{A}$ matrices, the upper-triangular $\mathbf{A}$ matrices make it harder for MOP to learn the optimal Kalman filter. As depicted in Fig. 4, the powers of the upper-triangular $\mathbf{A}$ matrices exhibit a slower decay rate and even initial overshoot in comparison to those of the dense $\mathbf{A}$ matrices. Noticing that $\mathbf{y}_t = \sum_{i=1}^t \mathbf{C}\mathbf{A}^i \mathbf{w}_{t-i} + \mathbf{v}_t$, this implies that upper triangular $\mathbf{A}$ establishes stronger and longer temporal correlation between $\mathbf{y}_t$ and past $\mathbf{y}$'s, i.e., slow mixing. This poses challenges to MOP but can be potentially mitigated by feeding MOP longer prompts, i.e., the time horizon T . Theoretically, the slow decay rate implies larger $L_\rho := C_\rho / (1 - \rho)$, which consequentially gives a looser risk upper bound in Theorem 1.

In our experiments in Section III, the distribution the source and target systems are drawn from is the same. Here we run an experiment to illustrate how MOP behaves if the target distribution is different than the source one. In particular, under the experimental setup of Section III-A, we train the MOP with noise covariances $\sigma_w^2 = \sigma_v^2 = 0.1\mathbf{I}_n$ and test on systems subject to a different noise covariance. As shown in Fig. 5, MOP's performance degrades when the target systems are subject to a different noise distribution, especially when the noise covariance increases.

VI. CONCLUSION

In conclusion, this letter has demonstrated the potential of transformers in addressing prediction problems for dynamical systems. The proposed MOP exhibits remarkable performance by adapting to unseen settings, non-i.i.d. noise, and time-varying dynamics.

This letter motivates new avenues for the application of transformers in continuous control and dynamical systems. Future work could extend the MOP approach to closed-loop control problems to meta-learn policies for problems such as the optimal quadratic control. It is also of interest to explore new training strategies to promote robustness (e.g., against distribution shifts) and safety of this approach in control problems.

REFERENCES

- [1] R. E. Kalman, "A new approach to recursive filtering and prediction problems," *Trans. ASME*, vol. 82, no. 1, pp. 35–45, 1960.
- [2] B. D. Anderson and J. B. Moore, *Optimal Filtering*. North Chelmsford, MA, USA: Courier Corp., 2012.
- [3] P. Del Moral, "Nonlinear filtering: Interacting particle resolution," *Comptes Rendus de l'Académie des Sciences-Ser. I-Math.*, vol. 325, no. 6, pp. 653–658, 1997.
- [4] P. S. R. Diniz, *Adaptive Filtering*. Cham, Switzerland: Springer, 1997.
- [5] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 6000–6010.
- [6] Y. Li, M. E. Ildiz, D. Papailiopoulos, and S. Oymak, "Transformers as algorithms: Generalization and stability in in-context learning," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19565–19594.
- [7] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.
- [8] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" 2021, *arXiv:2101.06804*.
- [9] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *Proc. ICML*, 2021, pp. 12697–12706.
- [10] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant, "What can transformers learn in-context? a case study of simple function classes," in *Proc. NeurIPS*, vol. 35, 2022, pp. 30583–30598.
- [11] M. Laskin et al., "In-context reinforcement learning with algorithm distillation," 2022, *arXiv:2210.14215*.
- [12] B. Yi, R. Wang, and I. R. Manchester, "Reduced-order nonlinear observers via contraction analysis and convex optimization," *IEEE Trans. Autom. Control*, vol. 67, no. 8, pp. 4045–4060, Aug. 2022.
- [13] B. Chen, J. Liang, N. Zheng, and J. C. Príncipe, "Kernel least mean square with adaptive kernel size," *Neurocomputing*, vol. 191, pp. 95–106, May 2016.
- [14] M. Scarpiniti, D. Communiello, R. Parisi, and A. Uncini, "Nonlinear spline adaptive filtering," *Signal Process.*, vol. 93, no. 4, pp. 772–783, 2013.
- [15] O. Anava, E. Hazan, S. Mannor, and O. Shamir, "Online learning for time series prediction," in *Proc. COLT*, 2013, pp. 172–184.
- [16] A. Tsiamis, N. Matni, and G. Pappas, "Sample complexity of Kalman filtering for unknown systems," in *Proc. LADC*, 2020, pp. 435–444.
- [17] M. Kozdoba, J. Marecek, T. Tchakian, and S. Mannor, "On-line learning of linear dynamical systems: Exponential forgetting in Kalman filters," in *Proc. AAAI*, vol. 33, 2019, pp. 4098–4105.
- [18] A. Tsiamis and G. J. Pappas, "Online learning of the Kalman filter with logarithmic regret," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2774–2789, May 2023.
- [19] J. Umenberger, M. Simchowitz, J. Perdomo, K. Zhang, and R. Tedrake, "Globally convergent policy search for output estimation," in *Proc. NeurIPS*, vol. 35, 2022, pp. 22778–22790.
- [20] W. Lohmiller and J.-J. E. Slotine, "On contraction analysis for non-linear systems," *Automatica*, vol. 34, no. 6, pp. 683–696, 1998.
- [21] D. Angeli, "A Lyapunov approach to incremental stability properties," *IEEE Trans. Autom. Control*, vol. 47, no. 3, pp. 410–421, Mar. 2002.
- [22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [23] S. Singh, B. Landry, A. Majumdar, J.-J. Slotine, and M. Pavone, "Robust feedback motion planning via contraction theory," *Int. J. Robot. Res.*, vol. 42, no. 9, pp. 655–688, 2023.