Maximizing the Value of Predictions in Control: Accuracy Is Not Enough

Yiheng Lin
California Institute of Technology
Pasadena, CA, USA
yihengl@caltech.edu

Christopher Yeh
California Institute of Technology
Pasadena, CA, USA
cyeh@caltech.edu

Zaiwei Chen
Purdue University
West Lafayette, IN, USA
chen5252@purdue.edu

Adam Wierman
California Institute of Technology
Pasadena, CA, USA
adamw@caltech.edu

Abstract

We study the value of stochastic predictions in online optimal control with random disturbances. Prior work provides performance guarantees based on prediction error but ignores the stochastic dependence between predictions and disturbances. We introduce a general framework modeling their joint distribution and define "prediction power" as the control cost improvement from the optimal use of predictions compared to ignoring the predictions. In the time-varying Linear Quadratic Regulator (LQR) setting, we derive a closed-form expression for prediction power and discuss its mismatch with prediction accuracy and connection with online policy optimization. To extend beyond LQR, we study general dynamics and costs. We establish a lower bound on prediction power under two sufficient conditions that generalize the properties of the LQR setting, characterizing the fundamental benefit of incorporating stochastic predictions. We apply this lower bound to non-quadratic costs and show that even weakly dependent predictions yield significant performance gains.

1 Introduction

Understanding the benefits of predictions in control has received significant attention recently [1, 2, 3, 4, 5]. In this work, we study a class of discrete-time online optimal control problems in general time-varying systems, where random disturbances W_t affect state transitions. The agent leverages a prediction vector containing information about future disturbances to minimize the expected total cost over a finite horizon T. To study the impact of using predictions, a fundamental question is how to model disturbances and their relationship to predictions. Prior works adopt different modeling approaches [1, 5, 6], each with distinct strengths and limitations.

A common paradigm assumes perfect predictions over a finite horizon k, yielding an elegant characterization of "prediction power" that improves with larger k. Under this model, predictions exactly reveal future disturbances $W_t, \ldots, W_{t+k-1}$. [1] shows how prediction power grows with k in the LQR setting, and subsequent work extends this result to time-varying systems [3]. As a result, the marginal benefit of one additional prediction decays exponentially with k, offering insight into how to select k. However, longer-horizon predictions are more costly and less accurate, and real-world predictions are rarely perfect [7], making this idealized setting challenging in practice.

A natural extension of accurate predictions is to consider bounded prediction errors, which better captures practical challenges. Specifically, prediction errors measure the distance between the predicted and actual disturbances, and the resulting cost bounds depend on these errors [2, 5, 4]. This extension recovers the perfect predictions setting when errors shrink to zero. However, it can be overly pessimistic because, for any predictor, the same performance bound must also apply to an adversary that generates the worst prediction sequence to penalize the predictive policy subject to the same error bound. It overlooks stochastic dependencies between predictions and disturbances that can be valuable for improving control costs.

In this work, we propose a general stochastic model that captures the distributional dependencies between predictions and disturbances, without restricting prediction targets, horizon length, or requiring strict error bounds. Compared with previous stochastic methods [6, 8], our approach further relaxes problem-specific assumptions and directly focuses on the incremental benefit of predictions. Such benefits can be subtle—often overlooked by classical metrics like regret or competitive ratio. To capture them, we define *prediction power* as the improvement in expected total cost when predictions are fully exploited, which builds on and generalizes the notion from [1]. Our framework thus characterizes when and why predictions significantly boost online control performance.

Contributions. We introduce a general stochastic model (Definition 2.1) that describes how disturbances relate to all candidate predictors. We then define *prediction power* (Definition 2.4) which quantifies the incremental control-cost improvement gained by fully leveraging these predictions. To illustrate this concept, we derive an exact expression for prediction power in the benchmark setting of time-varying linear quadratic regulator (LQR) control (Theorem 3.2). Using this closed-form formula, we provide examples (e.g., Example 3.3) that illustrate why analyzing prediction accuracy is insufficient—improving prediction accuracy may not always improve prediction power. Finally, we demonstrate the connection between prediction power and online policy optimization (Example 3.4), highlighting how practical algorithms can attain (a portion of) the maximum potential.

We extend our analysis of prediction power beyond the LQR setting. This generalization poses significant challenges due to the lack of closed-form expressions for the optimal policy. Building on insights from the LQR analysis, we identify two key structural conditions: a quadratic growth condition on the optimal Q-function (Condition 4.1) and a positive semi-definite covariance condition on the optimal policy's actions (Condition 4.2). These conditions are sufficient to derive a general lower bound on prediction power, formalized in Theorem 4.3. We apply this result to the setting of time-varying linear dynamics with non-quadratic cost functions. Under assumptions on costs and on the joint distribution of predictions and disturbances, we establish a lower bound on prediction power (Theorem 4.8), demonstrating that even weak predictions can yield strict performance gains.

Related Literature. Our work is closely related to the line of works on using predictions in online control. Our prediction power is inspired by [1], which defines the prediction power as the maximum control cost improvement enabled by k steps of accurate predictions in the time-invariant LQR setting. Compared with [1], we extend the notion of prediction power to allow general dependencies between predictions and disturbances, and we consider more general dynamics/costs (Section 4). Rather than focusing on the prediction power, many works study the power of a certain policy class such as MPC [2, 3, 5, 4], Averaging Fixed Horizon Control [6, 8], Receding Horizon Gradient Descent [9, 10], and others [11]. While one can say the power of (generalized) MPC equals the prediction power in the LQR setting [1] (Section 3), they are not the same in general (see Appendix C.2).

Our work is, in part, motivated by both empirical and theoretical findings in the decision-focused learning (DFL) (also referred to as "predict-then-optimize") literature that prediction models with the same prediction accuracy may have very different control costs (see [12] for a recent survey). Research on DFL typically considers predictions given as point estimates of some uncertain input to decision-makers modeled as optimization problems, such as stochastic optimization [13], linear programs [14], or model predictive control [15], although more recent works have started exploring other forms of predictions such as prediction sets [16, 17]. In contrast, our work does not require any particular form of decision-maker; instead, our main result characterizes the benefit of optimally leveraging predictions, for whatever form an optimal controller may take. Whereas DFL aims to design procedures for training prediction models that reduce downstream control costs, our work studies a more fundamental question about how much performance gain is achievable with better predictions. Another major difference between our work and typical DFL literature is that our controller must decide control actions sequentially in a dynamical system, where the predictions are

revealed in an online process. The controller can also build its knowledge from past disturbances and predictions to infer future disturbances or the optimal control action. Thus, our online setting presents unique challenges compared with making a one-shot decision for a classic optimization problem.

2   Problem Setting

We consider a finite-horizon discrete-time optimal control problem with time-varying dynamics and cost functions, where state transitions are subject to random disturbances:

Control dynamics: $X_{t+1} = f_t(X_t, U_t; W_t)$, $0 \leq t < T$, with the initial state $X_0 = x_0$;

Stage cost: $h_t(X_t, U_t)$, $0 \leq t < T$,   and terminal cost: $h_T(X_T)$.   (1)

At each time step t, we let X_t denote the system state and U_t denote the control action chosen by an agent. The function $f_t : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^n$ defines how the next state X_{t+1} depends on the current state X_t, the control action U_t, and the random disturbance W_t. The agent incurs a stage cost $h_t(X_t, U_t)$ at each time step $t < T$ and a terminal cost $h_T(X_T)$ at the final time step T. At each time step t, the controller observes the past disturbance W_{t-1} and a (possibly random) prediction vector $V_t(\theta) \in \mathbb{R}^d$ before selecting a control action U_t, where θ is a parameter of the predictor generating the prediction. We formally define the concept of *predictions* and the parameter θ in the following.

Definition 2.1 (Predictions). *At each time step t, the predictor with parameter $\theta \in \Theta$ provides a prediction $V_t(\theta)$, where Θ denotes the set of all possible predictor parameters. The predictions $\{V_{0:T-1}(\theta)\}_{\theta \in \Theta}$ and the disturbances $W_{0:T-1}$ live in the same probability space.*

We do not require the prediction $V_t(\theta)$ to have any specific form as a function of θ. The parameter θ is primarily used for distinguishing different predictor candidates.

Compared with previous works [3, 18] that assume predictions targeting specific disturbances, Definition 2.1 focuses on the stochastic relationship between predictions and system uncertainties, yielding a unified framework for comparing different forms of prediction based on their effectiveness for control—even if their precise nature is unknown. Because predictions and disturbances share the same probability space, we can compare prediction sequences $V_{0:T-1}(\theta)$ and $V_{0:T-1}(\theta')$, generated by different predictors with parameters θ and θ'.

Observe that the disturbances $W_{0:T-1}$ and predictions in Definition 2.1 do not depend on the current state or past trajectory, reflecting their exogenous nature. For example, consider the problem of quadcopter control in windy conditions [19]. In this case, the wind disturbances are not influenced by the quadcopter's state or control inputs. Under this causal relationship, we define the *problem instance* as $\Xi = (W_{0:T-1}, \{V_{0:T-1}(\theta)\}_{\theta \in \Theta})$, and make the following assumption.

Assumption 2.2. *The problem instance Ξ is sampled from the distribution of problem instances before the control process starts, i.e., it will not be affected by the controller's states/actions.*

Let $\xi = (w_{0:T-1}, \{v_{0:T-1}(\theta)\}_{\theta \in \Theta})$ denote a realization of the problem instance, including disturbances and all parameterized predictions. Under Assumption 2.2, Ξ is viewed as realized to ξ before control begins, although the agent observes each disturbance and prediction step by step. Similar assumptions about oblivious environments or predictions appear in online optimization [20, 21], ensuring that future disturbances or predictions will not be affected by past states or actions. Hence, for a fixed predictor parameter θ, we define a *predictive policy* as a mapping from the current state and past disturbances and predictions to a control action.

Definition 2.3 (Predictive policy). *Consider a fixed predictor parameter θ. For each time step t, let $I_t(\theta) := (W_{0:t-1}, V_{0:t}(\theta))$ denote the history of past disturbances and predictions, and let $\mathcal{F}_t(\theta) := \sigma(I_t(\theta))$[1]. A* predictive policy *that applies to the predictor with parameter θ is a sequence of functions $\pi_{0:T-1}$, where π_t maps a state/history pair to a control action.*

Given a fixed predictive policy sequence $\pi = \pi_{0:T-1}$ for a predictor parameter θ, we evaluate its performance via the expected total cost over Ξ: $J^\pi(\theta) := \mathbb{E}[\sum_{t=0}^{T-1} h_t(X_t, U_t) + h_T(X_T)]$, where $X_0 = x_0$, $X_{t+1} = f_t(X_t, U_t; W_t)$, $U_t = \pi_t(X_t; I_t(\theta))$, for $t = 0, \ldots, T-1$. The optimal cost under θ is defined as $J^*(\theta) = \min_\pi J^\pi(\theta)$, where the minimum is over all predictive policies that use the predictor parameter θ.

---

[1] For any random variable Y, we use $\sigma(Y)$ to denote the σ-algebra it generates.

Following [1], we define *prediction power* by comparing against a baseline that provides minimal information (e.g., no prediction). Without loss of generality, let $\mathbf{0} \in \Theta$ be the baseline predictor parameter so that any $\theta \neq \mathbf{0}$ provides at least as much information as $\mathbf{0}$, i.e., $\mathcal{F}_t(\theta) \supseteq \mathcal{F}_t(\mathbf{0})$. Based on this baseline, we define *prediction power* as the maximum possible cost improvement achieved by using predictions under θ relative to the baseline, formally stated in Definition 2.4.

Definition 2.4 (Prediction power). *For a predictor with parameter θ, its prediction power in the optimal control problem* (1) *is $P(\theta) := J^*(\mathbf{0}) - J^*(\theta)$.*

Our definition of prediction power is based on the optimal control policy under a given predictor parameter and, therefore, is independent of any specific policy class. Many previous works have considered prediction-enabled improvement within a specific policy class [9, 10, 6], where they focus on changes in $J^\pi(\theta)$ rather than $J^*(\theta)$. In other works, policies include parameters that can be tuned to perform optimally under a specific predictor; that is, $\min_{\pi \in \text{a policy class}} J^\pi(\theta)$. While these approaches are useful in specific application scenarios, our definition, based on the general optimal policy, is more universal because: (1) imposing policy class constraints may lead to performance loss, and (2) the extent of improvement can depend on policy design and parameterization, which shifts the focus away from valuing predictions themselves.

When the baseline predictor $\mathbf{0}$ is no prediction, the prediction power is guaranteed to be non-negative. This is because the optimal policy for the no-prediction case is also a predictive policy when predictions are available, i.e., the controller can simply ignore the predictions. But the prediction power could be zero, for example, when the predictions are independent of the disturbances. Further, the prediction power does not increase if one replaces the original prediction $V_t(\theta)$ by any function of the history $I_t(\theta)$, because this additional step cannot increase the information available at time step t.

Throughout this paper, we use $\bar{\pi} = \bar{\pi}_{0:T-1}$ and $\pi^\theta = \pi^\theta_{0:T-1}$ to denote the optimal policy for the predictor with parameter $\mathbf{0}$ and θ, respectively. In other words, $J^{\bar{\pi}}(\mathbf{0}) = J^*(\mathbf{0})$ and $J^{\pi^\theta}(\theta) = J^*(\theta)$. To compare the policies π^θ and $\bar{\pi}$, we introduce the *instance-dependent Q function*, inspired by the Q function in the study of Markov decision processes (MDPs). For a given state-action pair (x, u) and problem instance ξ, the instance-dependent Q function for a policy π evaluates the remaining cost incurred by taking action u from state x and then following policy π for all future time steps. Recall that the history $I_\tau(\theta)$ contains all past disturbances and predictions that are observed until a time step τ. Using $\iota_\tau(\theta)$ to denote the realization of $I_\tau(\theta)$, the instance-dependent Q function is defined as

$$Q_t^{\pi^\theta}(x, u; \xi) = \sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T), \quad \text{where } x_t = x, \ u_t = u, \tag{2}$$

subject to the constraints that $x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_\tau)$ for $t \leq \tau < T$ and $u_\tau = \pi^\theta_\tau(x_\tau; \iota_\tau(\theta))$ for $t < \tau < T$. Recall that ξ is the realization of problem instance Ξ that contains all disturbances and predictions for all time steps. Thus, the disturbance w_τ and the history $\iota_\tau(\theta)$ in (2) are decided by the problem instance ξ, which is an input to $Q_t^{\pi^\theta}$. Similarly, we define $Q_t^{\bar{\pi}}(x, u; \xi)$ by replacing θ with $\mathbf{0}$ and π^θ with $\bar{\pi}$ in (2). Importantly, our instance-dependent Q function is different from the classical definition of the Q function for MDPs or reinforcement learning (RL), where it is the *expectation* of the cost to go. The instance-dependent Q function denotes the actual remaining cost, which is a $\sigma(\Xi)$-measurable *random variable*. The classic definition of the Q function can be recovered by taking the conditional expectation, i.e., $\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$. It is worth noting that our instance-dependent Q function is about the *cost* instead of the *reward*, so lower values are better.

With this definition of the instance-dependent Q function, the policies $\bar{\pi}$ and π^θ can be expressed as recursively minimizing the corresponding expected Q functions conditioned on the available history. Starting with $C_T^{\pi^\theta}(x; \xi) = h_T(x)$, for time step $t = T-1, \ldots, 0$, we have

$$Q_t^{\pi^\theta}(x, u; \xi) := h_t(x, u) + C_{t+1}^{\pi^\theta}(f_t(x, u; w_t); \xi), \quad \text{for } x \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \text{ and problem instance } \xi;$$

$$\pi_t^\theta(x; \iota_t(\theta)) := \arg\min_{u \in \mathbb{R}^m} \mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right], \quad \text{for } x \in \mathbb{R}^n \text{ and history } \iota_t(\theta);$$

$$C_t^{\pi^\theta}(x; \xi) := Q_t^{\pi^\theta}(x, \pi_t^\theta(x; \iota_t(\theta)); \xi), \quad \text{for } x \in \mathbb{R}^n \text{ and problem instance } \xi. \tag{3}$$

Similar recursive relationships also define the optimal policy $\bar{\pi}$ for the baseline predictions, and we only need to replace θ with $\mathbf{0}$ and π^θ with $\bar{\pi}$ in the above equations. The recursive equations in (3) can be viewed as a generalization of the classical Bellman optimality equation for general MDPs.

3 LTV Dynamics with Quadratic Costs

We first characterize the prediction power (Definition 2.4) in a linear time-varying (LTV) dynamical system with quadratic costs, where the dynamics and costs are given by:

Control dynamics: $X_{t+1} = A_t X_t + B_t U_t + W_t$, for $0 \le t < T$;

stage cost: $X_t^\top Q_t X_t + U_t^\top R_t U_t$, for $0 \le t < T$; and terminal cost: $X_T^\top P_T X_T$, $\quad$ (4)

where $Q_{0:T-1}, R_{0:T-1}$, and P_T are symmetric positive definite. The classic linear quadratic regulator (LQR) problem, along with its time-varying variant that we consider, has been used widely as a benchmark setting in the learning-for-control literature. It also serves as a good approximation of nonlinear systems near equilibrium points, making it amenable to standard analytical tools. We begin by defining key quantities that will be useful for stating the main results in this section. For $t = T-1, \ldots, 0$, we define the matrices H_t, P_t, and K_t recursively according to

$$H_t = B_t (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top, \quad P_t = Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} H_t P_{t+1} A_t, \text{ and}$$
$$K_t = (R_t + B_t^\top P_{t+1} B_t)^{-1} (B_t^\top P_{t+1} A_t). \tag{5}$$

Moreover, we define the transition matrix Φ_{t_2, t_1} as $\Phi_{t_2, t_1} = I$ if $t_2 \le t_1$ and

$$\Phi_{t_2, t_1} = (A_{t_2-1} - B_{t_2-1} K_{t_2-1})(A_{t_2-2} - B_{t_2-2} K_{t_2-2}) \cdots (A_{t_1} - B_{t_1} K_{t_1}), \text{ if } t_2 > t_1. \tag{6}$$

The matrix K_t is the feedback gain matrix in the optimal policy, and P_t is the matrix that defines the quadratic term in the optimal cost-to-go function. To simplify notation, we define the shorthands $W_{\tau|t}^\theta := \mathbb{E}[W_\tau \mid I_t(\theta)]$ and $w_{\tau|t}^\theta := \mathbb{E}[W_\tau \mid I_t(\theta) = \iota_t(\theta)]$.

Proposition 3.1. *In the case of LTV dynamics with quadratic costs, the conditional expectation of the optimal Q function* $\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$ *can be expressed as*

$$\left(u + K_t x - \bar{u}_t^\theta(\iota_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t) \left(u + K_t x - \bar{u}_t^\theta(\iota_t(\theta))\right) + \psi_t^{\pi^\theta}(x; \iota_t(\theta)),$$

where $\psi_t^{\pi^\theta}(x; \iota_t(\theta))$ *is a function of the state x and the history $\iota_t(\theta)$ that does not depend on the control action u. Here,* $\bar{u}_t^\theta(\iota_t(\theta)) := -(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} w_{\tau|t}^\theta$. *And the optimal policy can be expressed as* $\pi_t^\theta(x; \iota_t(\theta)) = -K_t x + \bar{u}_t^\theta(\iota_t(\theta))$.

We derive the closed-form expressions in Proposition 3.1 by induction following the backward recursive equations in (3); the full proof is deferred to Appendix A.1. With these expressions, we obtain a closed-form expression of the prediction power. We defer its proof to Appendix A.2.

Theorem 3.2. *In the case of LTV dynamics with quadratic costs, the prediction power of the predictor with parameter θ is* $P(\theta) = \sum_{t=0}^{T-1} \text{Tr}\left\{(R_t + B_t^\top P_{t+1} B_t) \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right]\right\}$.

While the optimal policy in Proposition 3.1 is restricted to the LQR case, we can interpret the optimal policy as planning according to the conditional expectation following the idea of model predictive control (MPC) [1], which is easier to generalize. The agent needs to solve an optimization problem and re-plan at every time step. At time step t, the agent solves

$$\arg\min_{u_{t:T-1}} \quad \mathbb{E}\left[\sum_{\tau=t}^{T-1} h_\tau(X_\tau, u_\tau) + h_T(X_T) \,\Big|\, I_t(\theta) = \iota_t(\theta)\right] \tag{7}$$

subject to the constraints that $X_{\tau+1} = f_\tau(X_\tau, u_\tau; W_\tau)$ for $\tau \ge t$ and $X_t = x$. Then, the agent commits to the first entry $u_{t|t}$ of the optimal solution as $\pi_t^\theta(x; \iota_t(\theta))$. In the LQR setting, we can further simplify it to be *planning according to* $w_{\tau|t}^\theta$, i.e.,

$$\arg\min_{u_{t:T-1}} \quad \sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T) \tag{8}$$

subject to the constraints that $x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_{\tau|t}^\theta)$ for $\tau \ge t$ and $x_t = x$. We defer a detailed discussion and proof to Appendix A.3.

The MPC forms of the optimal policy in (7) extends the result in [1], which shows that MPC is the optimal predictive policy under the accurate prediction model in time-invariant LQR. When

the predictions are inaccurate, and the system is time-varying, MPC is still optimal if we solve the predictive optimal control problem in expectation (7).

Evaluation. One can follow the expressions in Theorem 3.2 to evaluate the prediction power, but it requires taking the conditional covariance on the top of conditional expectations ($\bar{u}_t^\theta(\iota_t(\theta))$ in Proposition 3.1). To avoid this recursive structure, an alternative way is to first construct the *surrogate optimal action*, which is defined as

$$\bar{u}_t^*(\Xi) := -(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} W_\tau. \tag{9}$$

We call $\bar{u}_t^*(\Xi)$ the surrogate-optimal action, because it is the optimal action that an agent should take with the oracle knowledge of all future disturbances at time t. The prediction power in Theorem 3.2 can be expressed as $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\mathbf{0})\right]\right] - \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta)\right]\right]$. Following this decomposition, we propose an evaluation approach that constructs $\bar{u}_t^*(\Xi)$ before estimating its conditional covariance with respect to $I_t(\theta)$ and $I_t(\mathbf{0})$ separately. We defer the details to Appendix A.4.

3.1 Prediction Power $\neq$ Accuracy

As Proposition 3.1 suggests, one way to implement the optimal policy is to predict each of the future disturbances $W_{t:T-1}$ and generate the estimations $w_{(t:T-1)|t}^\theta$ (i.e., the conditional expectation of future disturbances given the history $\iota_t(\theta)$ at time step t) in deciding the action at time step t. However, two controllers with the same estimation error (as measured by mean squared error (MSE)) can have very different control costs. Because of this reason, the control cost bounds depend on the estimation errors in previous works [5, 2, 4] must be loose, so one cannot rely on them to infer or compare the values of different predictors.

To illustrate this point, we provide an example where the prediction power can change significantly when the prediction accuracy does not change.

Example 3.3. *Consider the time-invariant LQR setting, i.e., assume $A_t = A, B_t = B, Q_t = Q, R_t = R$ for all t and $P_T = P$ is the solution to the Discrete-time Riccati Equation (DARE) in (4). Suppose the disturbance is sampled $W_t \overset{i.i.d.}{\sim} N(0, \mathbb{I})$ at every time step t, where we use the notation $\mathbb{I}$ to denote the identity matrix. Let $\rho \in [0, \frac{\sqrt{2}}{2}]$ be a fixed coefficient. We construct a class of predictors from the disturbances $\{W_t\}$ by applying the affine transformation $V_t(\theta) := \rho\theta W_t + \epsilon_t(\rho, \theta)$ for $\theta \in \mathbb{R}^{2\times 2}$ that satisfies $\theta\theta^\top \preceq \frac{1}{2}\mathbb{I}$, where the random noise $\epsilon_t(\rho, \theta)$ is independently sampled from a Gaussian distribution $N(0, \mathbb{I} - \rho^2\theta\theta^\top)$.*

We can construct θ such that $V_t(\theta)$ and $V_t(\mathbb{I})$ achieve the same mean-square error (MSE) when predicting each individual entry of W_t, yet $P(\theta) > P(\mathbb{I})$. To construct θ, note that $(W_t, V_t(\theta))$ satisfies $\mathbb{E}[W_t \mid V_t(\theta)] = \rho\theta^\top V_t$ and $\mathbf{Cov}[W_t \mid V_t(\theta)] = \mathbb{I} - \rho^2\theta^\top\theta$. Thus, we can change θ without affecting the MSE of predicting each individual entry as long as the diagonal entries of $\theta^\top\theta$ remain the same. However, by Theorem 3.2, we know the prediction power is equal to $\rho^2 T \cdot \mathrm{Tr}\{\theta^\top\theta PHP\}$, where $H = B(R + B^\top PB)^{-1}B^\top$. Thus, the off-diagonal entries of $\theta^\top\theta$ can also affect the value of $\mathrm{Tr}\{\theta^\top\theta PHP\}$. We instantiate this example with a 2-D double-integrator dynamical system in Appendix A.5.1: the predictors with parameters $\mathbb{I}$ and θ share the same MSE but their prediction powers are significantly different.[2]

Example 3.3 shows how prediction power can vary even when the accuracy of predicting each entry of the disturbance W_t remains the same, where the construction leverages the covariance between the predictions for different entries of W_t. While the construction in Example 3.3 requires $n \geq 2$, we also provide an example with $n = 1$ and multiple steps of predictions in Appendix A.5. From these examples, it is clear that one should not use the MSEs of predicting future disturbances to infer the prediction power. The intuition behind Example 3.3 does not require a very specific choice of the function $V_t(\theta)$: The underlying idea is that the MSEs of estimating each individual entry of the multi-dimensional disturbance W_t are insufficient to decide the prediction power. The off-diagonal entries of the covariance matrix $\mathbf{Cov}[W_t \mid V_t(\theta)]$ also matter, and their impact on the prediction power depends on the dynamics (A, B, Q, R). Therefore, a general accuracy metric (like MSE) that is unaware of (A, B, Q, R) can be misaligned with the prediction power. The mismatch also relates to the findings in the decision-focused learning literature discussed in the related work section.

---

[2]The simulation code for all examples (Examples 3.3, 3.4, and A.2) can be found at https://github.com/yihenglin97/Prediction-Power.
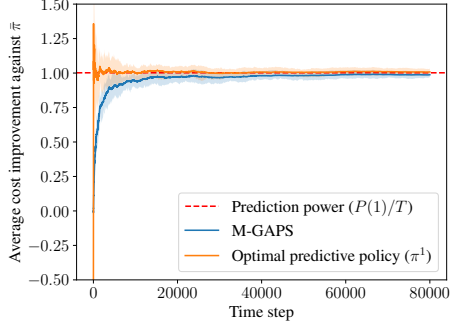
Figure 1: Example 3.4: Prediction $V_t(1)$ is available. Candidate policy: $u_t = -Kx_t + \Upsilon_t v_t(1)$.
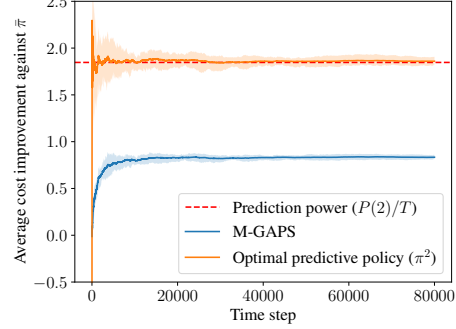

Figure 2: Example 3.4: Prediction $V_t(2)$ is available. Candidate policy: $u_t = -Kx_t + \Upsilon_t v_t(2)$.

3.2 Prediction Power and Online Policy Optimization

The closed-form expression of the prediction power, presented in Theorem 3.2, characterizes the maximum potential of using a given prediction sequence $V_{0:T-1}(\theta)$. Here, we draw a connection between prediction power and online policy optimization [22, 23], which aims to learn and adapt the optimal control policy within a certain policy class over time: the prediction power serves as an improvement upper bound of applying online policy optimization to predictive policies, although it is generally unattainable. In the following example, we demonstrate this bound using M-GAPS [24], a state-of-the-art online policy optimization algorithm.

Example 3.4. *We construct two scenarios under the same setting as Example 3.3. First, when the prediction is $V_t(1) \coloneqq \rho W_t + \epsilon_t(\rho, I)$, we let M-GAPS adapt within the candidate policy class $u_t = -Kx_t + \Upsilon_t v_t(1)$, where $\Upsilon_t \in \mathbb{R}^{1\times2}$ is the policy parameter.[3] Here, the optimal predictive policy π^1 is contained in the candidate policy class. We plot the average cost improvement of M-GAPS and π^1 compared against the optimal no-prediction policy $\bar{\pi}$ in Figure 1. From the initialization $\Upsilon_0 = \mathbf{0}$, M-GAPS tunes Υ_t to improve the average cost over time, and the average cost improvement against $\bar{\pi}$ converges towards the averaged prediction power $P(1)/T$.*

In the second scenario, we change the prediction to apply M-GAPS to $V_t(2) \coloneqq V_{t+1}(1)$ (i.e., the same prediction as before is made available 1-step ahead). We let M-GAPS adapt within the same candidate policy class $u_t = -Kx_t + \Upsilon_t v_t(2)$, where the policy parameter is still $\Upsilon_t \in \mathbb{R}^{1\times2}$. Unlike the first scenario, the optimal predictive policy π^2 is not *contained in the candidate policy class, because π^2 uses both $v_t(2)$ and $v_{t-1}(2)$ to decide the action. As a result, M-GAPS cannot achieve an improvement that is close to the averaged prediction power $P(2)/T$, which is achievable by π^2 (see Figure 2).*

The details of Example 3.4 are provided in Appendix A.6. It demonstrates how prediction power serves as an upper bound for the cost improvement achieved by online policy optimization. Conversely, online policy optimization offers practical tools to achieve (part of) the potential benefit of using predictions without requiring explicit knowledge or estimation of the joint distribution between predictions and true disturbances.

4 Characterizing Prediction Power: A General-Purpose Theorem

In this section, we provide a theorem to characterize the prediction power $P(\theta)$ within the general problem setting introduced in Section 2. Our result relies on two conditions about a growth property of the expected Q function under π^θ and the covariance of the optimal policy's action when conditioned on the σ-algebra $\mathcal{F}_t(\mathbf{0})$ of the baseline. We state these conditions and provide intuitive explanations.

---

[3]Intuitively, M-GAPS works by taking the gradient of the cost function with respect to the policy parameters at every time step, and it takes gradient steps to update Υ_t, allowing it to converge towards the optimal policy parameters. Our goal is to highlight the connection between prediction power and online policy optimization, so the specific online policy optimization algorithms and their proofs are not the primary focus here.

Condition 4.1. *For a sequence of positive semi-definite matrices $M_{0:T-1}$, the following inequality holds for all time steps $0 \leq t < T$: For any $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and history $\iota_t(\theta)$,*

$$\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) - C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta) \right] \geq (u - \pi_t^\theta(x; \iota_t(\theta)))^\top M_t (u - \pi_t^\theta(x; \iota_t(\theta))). \quad (10)$$

The LQR setting (Section 3) satisfies Condition 4.1 with $M_t = R_t + B_t^\top P_{t+1} B_t$. But Condition 4.1 is applicable beyond the LQR setting. For example, we show it holds under non-quadratic cost functions in Section 4.1.

Condition 4.1 states that conditioned on any history $\iota_t(\theta)$, the expected Q function of policy π^θ grows at least quadratically as the action u deviates from the optimal policy's action. Note that one can always pick M_t to be the all-zeros matrix to make Condition 4.1 hold, but the choice of M_t will affect the prediction power bound in Theorem 4.3. When $M_t \succ 0$, deviating from the action of policy π^θ causes a non-negligible loss. The loss is characterized by the difference between the resulting Q function value and the cost-to-go function value. When this condition does not hold with any non-zero matrix M_t, one can construct an extreme case when $Q_t^{\pi^\theta}$ is a constant by letting all cost functions $h_{0:T}$ be constants; in this case, the prediction power must be zero because every policy achieves the same total cost no matter what predictions they use.

Condition 4.2. *One of the following holds for the optimal policy π^θ:*

(a) For positive semi-definite matrices $\Sigma_{0:T-1}$, the following holds for all time steps $0 \leq t < T$:

$$\mathbb{E}\left[\mathbf{Cov}\left[\pi_t^\theta(X; I_t(\theta)) \mid I_t(\mathbf{0}) \right] \right] \succeq \Sigma_t, \text{ for any } \mathcal{F}_t(\mathbf{0})\text{-measurable } X. \quad (11)$$

(b) For nonnegative scalars $\sigma_{0:T-1}$, the following holds for all time steps $0 \leq t < T$:

$$\mathbb{E}\left[\mathrm{Tr}\left\{ \mathbf{Cov}\left[\pi_t^\theta(X; I_t(\theta)) \mid I_t(\mathbf{0}) \right] \right\} \right] \geq \sigma_t, \text{ for any } \mathcal{F}_t(\mathbf{0})\text{-measurable } X. \quad (12)$$

Before discussing the details, we note that by setting $\sigma_t = \mathrm{Tr}(\Sigma_t)$, Condition 4.2 (a) implies (and is therefore stronger than) Condition 4.2 (b). Similar to Condition 4.1, one can always pick Σ_t to be all-zeros matrix to satisfy Condition 4.2 (a), but it will affect the prediction power bound. The LQR setting (Section 3) satisfies Condition 4.2 (a) with $\Sigma_t = \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0}) \right] \right]$.

Condition 4.2 (a) states that conditioned on the history $I_t(\mathbf{0})$ from the baseline, the covariance matrix of policy π^θ's action from any $\mathcal{F}_t(\mathbf{0})$-measurable state is positive semi-definite in expectation. Recall that $\mathcal{F}_t(\mathbf{0}) = \sigma(I_t(\mathbf{0}))$. To understand this, suppose that the agent only has access to the baseline information. Then, the agent cannot predict the action that policy π^θ would take. This should usually hold because the action $\pi_t^\theta(X; I_t(\theta))$ is not $\mathcal{F}_t(\mathbf{0})$-measurable, and the lower bound in (11) implies the mean-square prediction error cannot improve below a certain threshold. When this condition does not hold with non-zero matrix Σ_t (or scalar σ_t), one can design a policy $\bar{\pi}'$ that always picks the same action as π^θ but only requires access to the baseline information $I_t(\mathbf{0})$, which implies $P(\theta) = 0$ because $J^*(\mathbf{0}) \leq J^{\bar{\pi}'}(\mathbf{0}) = J^*(\theta)$. This can happen, for example, when $W_{0:T-1}$ are deterministic.

Note it is possible that the optimal action at different states has a positive variance in different directions, but there is no non-trivial lower bound on the covariance matrix as required by Condition 4.2 (a). In this case, Condition 4.2 (b) provides a weaker alternative and would be useful when we can only establish a lower bound on the trace of the optimal action's covariance matrix (e.g., Section 4.1).

Theorem 4.3. *If Conditions 4.1 and 4.2 (a) hold with matrices $M_{0:T-1}$ and $\Sigma_{0:T-1}$, then $P(\theta) \geq \sum_{t=0}^{T-1} \mathrm{Tr}\{M_t \Sigma_t\}$. Alternatively, if Conditions 4.1 and 4.2 (b) hold with matrices $M_{0:T-1}$ and scalars $\sigma_{0:T-1}$, then $P(\theta) \geq \sum_{t=0}^{T-1} \mu_{min}(M_t) \cdot \sigma_t$, where $\mu_{min}(\cdot)$ returns the smallest eigenvalue.*

We defer the proof of Theorem 4.3 to Appendix B. As a remark, in the LQR setting, the first inequality in Theorem 4.3 holds with equality, and it recovers the same expression as Theorem 3.2 in Section 3. There are two main takeaways of Theorem 4.3. First, recall that one can always pick M_t and Σ_t to be the all-zeros matrices to satisfy Conditions 4.1 and 4.2. In this case, Theorem 4.3 states that $P(\theta) \geq 0$, which means that having predictions, no matter how weak they are, does not hurt. Second, to characterize the improvement in having predictions, Conditions 4.1 and 4.2 can establish a lower bound for the prediction power that is strictly positive if $\mathrm{Tr}\{M_t \Sigma_t\} > 0$ or $\mu_{min}(M_t)\sigma_t > 0$. We provide an example to help illustrate how Conditions 4.1 and 4.2 (a) can work together to ensure that the predictions can lead to a strict improvement on the control cost (see Figure 3 for an illustration).
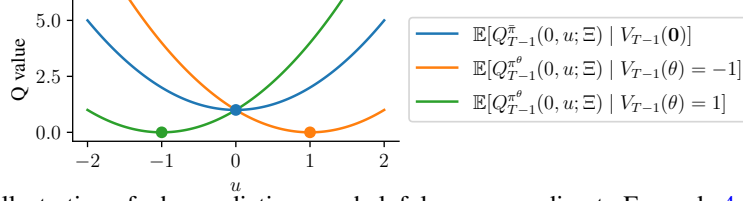
8

Figure 3: An illustration of why predictions are helpful, corresponding to Example 4.4. The expected Q functions with perfect predictions (green and orange lines) have lower minima than the expected Q function with uninformative predictions (blue line).

Example 4.4. *Consider the following optimal control problem*

Dynamics: $X_{t+1} = U_t + W_t$, Stage cost: $h_t(x,u) = x^2$, Terminal cost: $h_T(x) = x^2$,

where each disturbance W_t is sampled independently according to $\mathbb{P}(W_t = -1) = \mathbb{P}(W_t = 1) = \frac{1}{2}$. Suppose that the predictor with parameter θ can predict W_t exactly (i.e., $V_t(\theta) = W_t$), while the baseline predictor is uninformative (e.g., $V_t(\mathbf{0}) = 0$). The Q functions, cumulative cost, and optimal actions under each predictor are

$$Q_t^{\pi^\theta}(x,u;\Xi) = x^2 + (u + V_t(\theta))^2, \qquad Q_t^{\bar\pi}(x,u;\Xi) = x^2 + (u + W_t)^2 + (T - t - 1),$$
$$C_t^{\pi^\theta}(x;\Xi) = x^2, \qquad\qquad C_t^{\bar\pi}(x;\Xi) = x^2 + (T - t),$$
$$\pi_t^\theta(x;I_t(\theta)) = -V_t(\theta) = -W_t, \qquad \bar\pi_t(x;I_t(\mathbf{0})) = 0.$$

The Q function $Q_t^{\pi^\theta}$ is strongly convex in u, with Condition 4.1 holding for any $M_t \in [0,1]$. Furthermore, the optimal action has positive variance, with Condition 4.2 (a) holding for any $\Sigma_t \in [0,1]$. Thus, by Theorem 4.3, the prediction power satisfies $P(\theta) \geq T$. Indeed, by comparing the cumulative cost functions, we see that the predictor with parameter θ incurs a lower cumulative cost by exactly T (as expected by Theorem 3.2).

Figure 3 illustrates the expected Q functions at time $t = T - 1$ and $x = 0$, which the policies $\pi_t^\theta(x;I_t(\theta))$ and $\bar\pi_t(x;I_t(\mathbf{0}))$ seek to minimize. The expected Q functions with perfect predictions have lower minima than the expected Q function with uninformative predictions.

Theorem 4.3 provides a useful tool to characterize the prediction power by reducing the problem of comparing two policies π^θ and $\bar\pi$ over the whole horizon to studying the properties of one policy π^θ at each time step. Our proof of Theorem 4.3 follows the same intuition as the widely-used performance difference lemma in RL (see Lemma 6.1 in [25]), but we adopt novel methods to compare the per-step "advantage" of π^θ along the trajectory of $\bar\pi$ with the conditional covariance of policy π^θ's actions. When only the baseline information is available, the agent must pick a suboptimal action (11) and incur a loss (10) at each step, which accumulates to the total cost difference.

While Theorem 4.3 applies to the general dynamical system and cost functions in (1), the two conditions with their key coefficients M_t and Σ_t (or σ_t) still depend on the optimal Q function and the optimal policy that are implicitly defined through the recursive equations (3). To instantiate Theorem 4.3, we need to derive explicit expressions of M_t and Σ_t under more specific dynamics/costs.

4.1 LTV Dynamics with General Costs

In this section, we consider an online optimal control problem with linear time-varying dynamics and more general cost functions compared with the LQR setting in Section 3.

$$\text{Control dynamics: } X_{t+1} = A_t X_t + B_t U_t + W_t, \text{ for } 0 \leq t < T;$$
$$\text{stage cost: } h_t^x(X_t) + h_t^u(U_t), \text{ for } 0 \leq t < T; \text{ and terminal cost: } h_T^x(X_T). \tag{13}$$

The LTV system with quadratic cost functions studied in Section 3 is a special case of (13). The setting is challenging because the optimal Q function/policy π^θ do not have closed-form expressions like Proposition 3.1. To tackle it, we follow the recursive equations (3) to establish Conditions 4.1 and 4.2 (b). We make the following assumptions about the cost functions and dynamical matrices:

9

Assumption 4.5. *For every time step t, h_t^x is μ_x-strongly convex and ℓ_x-smooth, and h_t^u is μ_u-strongly convex and ℓ_u-smooth. The dynamical matrices satisfy $\mu_A I \preceq A_t^\top A_t \preceq \ell_A I$ and $\mu_B I \preceq B_t^\top B_t \preceq \ell_B I$. Further, we assume $\ell_A < 1$ and $\mu_B > 0$.*

Under Assumption 4.5, we can verify Condition 4.1 and show that the expected cost-to-go functions are well-conditioned. We state this result in Lemma 4.6 and defer its proof to Appendix C.4.

Lemma 4.6. *Under Assumption 4.5, Condition 4.1 holds with $M_t = \mu_u I$. Further, conditional expectation $\mathbb{E}[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)]$ as a function of x is μ_t-strongly convex and ℓ_t-smooth for any history $\iota_t(\theta)$, where μ_t and ℓ_t are defined as follows: Let $\mu_T = \mu_x$ and $\ell_T = \ell_x$,*

$$\mu_t = \mu_x + \mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}, \text{ and } \ell_t = \ell_x + \ell_A \cdot \ell_{t+1}, \text{ for time } t = T-1, \ldots, 0. \tag{14}$$

To show Lemma 4.6 and (14), we prove properties of infimal convolution (see (35) in Appendix C.1) to preserve strongly convexity, smoothness, and the covariance of the input functions or variables.

To establish the second condition about the covariance of the optimal policy's action, we make the following assumption about the joint distribution of the disturbances and the predictions:

Assumption 4.7. *The disturbances and predictions can be grouped as pairs $\{(W_t, V_t(\theta))\}_{t=0}^{T-1}$, where $(W_t, V_t(\theta))$ is joint Gaussian and independent of $(W_{t'}, V_{t'}(\theta))$ when $t \neq t'$. Further, assume that the baseline is no prediction, i.e., $V_t(\mathbf{0}) = 0$. And for $\theta \in \Theta$, there exists $\lambda_t(\theta) \in \mathbb{R}_{\geq 0}$ such that $\mathbf{Cov}[W_t] - \mathbf{Cov}[W_t \mid V_t(\theta)] \succeq \lambda_t(\theta) I$, for any $0 \leq t < T$.*

With Assumption 4.7 and Lemma 4.6, we can verify that Condition 4.2 (b) holds with

$$\mathrm{Tr}\big\{\mathbf{Cov}\left[\pi_t^\theta(x; I_t(\theta)) \mid \mathcal{F}_t(0)\right]\big\} \geq \sigma_t := \frac{n \lambda_t(\theta) \mu_{t+1}^2 \cdot \mu_B}{2(\ell_u + \ell_{t+1}\sqrt{\ell_B})^2}. \tag{15}$$

Since Conditions 4.1 and 4.2 (b) hold, we apply Theorem 4.3 to obtain the prediction power bound.

Theorem 4.8. *In the case of LTV dynamics with well-conditioned costs, suppose Assumptions 4.5 and 4.7 hold. The prediction power of the predictor with parameter θ is lower bounded by $P(\theta) \geq \sum_{t=0}^{T-1} \mu_u \sigma_t$, where σ_t is defined in (15).*

We provide a more detailed proof outline and the proofs in Appendix C.1. As a remark, the lower bound of the prediction power in Theorem 4.8 shows that even weak predictions (i.e., small but non-zero $\lambda_t(\theta)$ in Assumption 4.7) can help improve the control cost compared with the no-prediction baseline. Although Assumption 4.7 limits $V_t(\theta)$ to be only correlated with W_t, we provide a roadmap towards more general dependencies on all future $W_{t:T-1}$ in Appendix E.

5 Concluding Remarks

In this work, we propose the metric of prediction power and characterize it in the time-varying LQR setting (Theorem 3.2). We extend our analysis to provide a lower bound for the general setting (Theorem 4.3), which is helpful for establishing the incremental value of (weak) predictions beyond LQR (Theorem 4.8). We emphasize that our framework is very broad. For example, if we let the parameter θ represent the dataset that the predictor is trained on, then the prediction power $P(\theta)$ effectively quantifies the value of that particular dataset with respect to the optimal control problem.

We would like to highlight three directions of research inspired by our results. First, while our work establishes prediction power of a predictor with parameter θ relative to a strictly less-informative baseline, it does not immediately enable comparison between two arbitrary parameters θ and θ' when our general lower bounds in Section 4 are not tight. Second, while we discuss how to evaluate prediction power of a given parameter θ, our work does not specify what the optimal θ is. The problem of learning the parameter that maximizes $P(\theta)$ may be interesting future work. Third, a natural future direction is to extend the current analysis to systems with partial observability. This remains challenging, as one must distinguish the disturbances in the system dynamics from the noises in the observation model, and the optimal policy's functional form becomes difficult to characterize due to its dependence on the entire history of partial observations and predictions.

Acknowledgments and Disclosure of Funding

References

[1] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. The power of predictions in online control. *Advances in Neural Information Processing Systems*, 33:1994–2004, 2020.

[2] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Competitive control with delayed imperfect information. In *2022 American Control Conference (ACC)*, pages 2604–2610. IEEE, 2022.

[3] Yiheng Lin, Yang Hu, Guanya Shi, Haoyuan Sun, Guannan Qu, and Adam Wierman. Perturbation-based regret analysis of predictive control in linear time varying systems. *Advances in Neural Information Processing Systems*, 34:5174–5185, 2021.

[4] Yiheng Lin, Yang Hu, Guannan Qu, Tongxin Li, and Adam Wierman. Bounded-regret mpc via perturbation analysis: Prediction error, constraints, and nonlinearity. *Advances in Neural Information Processing Systems*, 35:36174–36187, 2022.

[5] Runyu Zhang, Yingying Li, and Na Li. On the regret analysis of online lqr control with predictions. In *2021 American Control Conference (ACC)*, pages 697–703. IEEE, 2021.

[6] Niangjun Chen, Anish Agarwal, Adam Wierman, Siddharth Barman, and Lachlan LH Andrew. Online convex optimization using predictions. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 191–204, 2015.

[7] Tianyu Chen, Yiheng Lin, Nicolas Christianson, Zahaib Akhtar, Sharath Dharmaji, Mohammad Hajiesmaili, Adam Wierman, and Ramesh K Sitaraman. Soda: An adaptive bitrate controller for consistent high-quality video streaming. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 613–644, 2024.

[8] Niangjun Chen, Joshua Comden, Zhenhua Liu, Anshul Gandhi, and Adam Wierman. Using predictions in online optimization: Looking forward with an eye on the past. *ACM SIGMETRICS Performance Evaluation Review*, 44(1):193–206, 2016.

[9] Yingying Li, Guannan Qu, and Na Li. Using predictions in online optimization with switching costs: A fast algorithm and a fundamental limit. In *2018 Annual American Control Conference (ACC)*, pages 3008–3013. IEEE, 2018.

[10] Yingying Li, Xin Chen, and Na Li. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Yiheng Lin, Gautam Goel, and Adam Wierman. Online optimization with predictions and non-convex losses. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(1):1–32, 2020.

[12] Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, August 2024.

[13] Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based End-to-end Model Learning in Stochastic Optimization. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA, December 2017. Curran Associates, Inc.

[14] Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1):9–26, January 2022.

[15] Brandon Amos, Ivan Jimenez, Jacob Sacks, Byron Boots, and J. Zico Kolter. Differentiable MPC for End-to-end Planning and Control. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[16] Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *Preprint arXiv:2409.20534*, 2024.

[17] Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato. Learning Decision-Focused Uncertainty Sets in Robust Optimization, July 2024.

[18] Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. Robustness and consistency in linear quadratic control with untrusted predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–35, 2022.

[19] Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

[20] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[21] Daan Rutten, Nicolas Christianson, Debankur Mukherjee, and Adam Wierman. Smoothed online optimization with unreliable predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–36, 2023.

[22] Yiheng Lin, James A. Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[23] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[24] Yiheng Lin, James A Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online policy optimization in unknown nonlinear systems. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3475–3522. PMLR, 2024.

[25] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

[26] Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017.

[27] Jerrold E Marsden and Anthony Tromba. *Vector Calculus*. Macmillan, 2003.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims in the abstract are justified by the major contributions, which include pointers to specific theorems and sections in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We highlight some limitations in the last paragraph of Concluding Remarks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All of our theoretical results are based on rigorous assumptions and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the full problem settings and algorithm patameters in all simulations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit the simulation code in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the details in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We discuss why we believe each experiment observation is significant.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We discuss about compute resources for each simulation result.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We check the NeurIPS Code of Ethics and believe our work conforms with every code.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work focuses on theoretical research of online control. We do not see any potential societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use any specific assets beyond standard open-source scientific Python packages such as numpy and matplotlib for running experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Upon paper acceptance, we will release our code on GitHub with a permissive license. Our paper does not introduce any new data or models.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Proofs and Examples for LTV Dynamics with Quadratic Costs

A.1 Proof of Proposition 3.1

Recall that we introduce the shorthand

$$W^\theta_{\tau|t} = \mathbb{E}\left[W_\tau \mid I_t(\theta)\right].$$

We show by induction that

$$\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta)\right]$$
$$= \left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right) + \psi_t^{\pi^\theta}(x; I_t(\theta)),$$
$$\pi_t^\theta(x; I_t(\theta)) = -K_t x + \bar{u}_t^\theta(I_t(\theta)),$$

together with the expression of the optimal cost-to-go function

$$\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right] = x^\top P_t x + 2\left(\sum_{\tau=t}^{T-1} \Phi_{\tau+1,t}^\top P_{\tau+1} W^\theta_{\tau|t}\right)^\top x + \Psi_t(I_t(\theta)), \qquad (16)$$

where recall that for $t_2 > t_1$,

$$\Phi_{t_2,t_1}^\top := (A_{t_1} - B_{t_1} K_{t_1})^\top \cdots (A_{t_2-1} - B_{t_2-1} K_{t_2-1})^\top$$
$$= (A_{t_1}^\top - A_{t_1}^\top P_{t_1+1} H_{t_1}) \cdots (A_{t_2-1}^\top - A_{t_2-1}^\top P_{t_2} H_{t_2-1}).$$

and $\Psi_t(I_t(\theta))$ is a function of the history observations/predictions which does not depend on x. Note that (16) holds when $t = T$ because $C_T^{\pi^\theta}(x; \Xi) = x^\top P_T x$.

Suppose that (16) holds for $t + 1$. Then, we have

$$\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t; \Xi) \mid I_t(\theta)\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t; \Xi) \mid I_{t+1}(\theta)\right] \mid I_t(\theta)\right]$$
$$= \mathbb{E}\left[(x + W_t)^\top P_{t+1}(x + W_t)\Big| I_t(\theta)\right] + 2\mathbb{E}\left[\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W^\theta_{\tau|t+1}\Bigg| I_t(\theta)\right]^\top x$$
$$+ 2\mathbb{E}\left[\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W^\theta_{\tau|t+1}\right)^\top W_t\Bigg| I_t(\theta)\right] + \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right]$$
$$= x^\top P_{t+1} x + 2\left(P_{t+1} W^\theta_{t|t} + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W^\theta_{\tau|t}\right)^\top x + \text{Tr}\{P_{t+1} \cdot \mathbf{Cov}\left[W_t \mid I_t(\theta)\right]\}$$
$$+ 2\mathbb{E}\left[\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W^\theta_{\tau|t+1}\right)^\top W_t\Bigg| I_t(\theta)\right] + \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right].$$

To simplify the notation, let

$$\bar{\psi}_{t+1}(I_t(\theta)) := \text{Tr}\{P_{t+1} \cdot \mathbf{Cov}\left[W_t \mid I_t(\theta)\right]\} + 2\mathbb{E}\left[\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W^\theta_{\tau|t+1}\right)^\top W_t\Bigg| I_t(\theta)\right]$$
$$+ \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right].$$

We see that the expected Q function is given by

$$\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta)\right]$$
$$= x^\top Q_t x + u^\top R_t u + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid I_t(\theta)\right]$$

$$= x^\top Q_t x + u^\top R_t u + (A_t x + B_t u)^\top P_{t+1}(A_t x + B_t u)$$

$$+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top (A_t x + B_t u) + \bar{\psi}_{t+1}(I_t(\theta))$$

$$= u^\top(R_t + B_t^\top P_{t+1} B_t)u + 2\left(P_{t+1}A_t x + P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t u$$

$$+ x^\top(Q_t + A_t^\top P_{t+1} A_t)x + 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top A_t x + \bar{\psi}_{t+1}(I_t(\theta))$$

$$= \left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right) + \psi_t^{\pi^\theta}(x; I_t(\theta)),$$

where $\psi_t^{\pi^\theta}(x; I_t(\theta))$ is given by

$$x^\top(Q_t + A_t^\top P_{t+1} A_t)x + 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top A_t x + \bar{\psi}_{t+1}(I_t(\theta))$$

$$+ \left(K_t x - \bar{u}_t^\theta(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(K_t x - \bar{u}_t^\theta(I_t(\theta))\right).$$

Using the expected Q function, we know that the optimal policy will pick the action

$$\pi_t(x; I_t(\theta)) = \arg\min_u \mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta)\right] = -K_t x + \bar{u}_t^\theta(I_t(\theta)).$$

Therefore, we see the optimal cost-to-go function at time step t is given by

$$\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]$$

$$= x^\top Q_t x + (K_t x - \bar{u}_t^\theta(I_t(\theta)))^\top R_t(K_t x - \bar{u}_t^\theta(I_t(\theta)))$$

$$+ ((A_t - B_t K_t)x + B_t \bar{u}_t^\theta(I_t(\theta)))^\top P_{t+1}((A_t - B_t K_t)x + B_t \bar{u}_t^\theta(I_t(\theta)))$$

$$+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top ((A_t - B_t K_t)x + B_t \bar{u}_t^\theta(I_t(\theta))) + \bar{\psi}_{t+1}(I_t(\theta))$$

$$= x^\top(Q_t + K_t^\top R_t K_t + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t))x - 2\bar{u}_t^\theta(I_t(\theta))^\top R_t K_t x$$

$$+ 2\bar{u}_t^\theta(I_t(\theta))^\top B_t^\top P_{t+1}(A_t - B_t K_t)x$$

$$+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top (A_t - B_t K_t)x$$

$$+ \bar{u}_t^\theta(I_t(\theta))^\top(R_t + B_t^\top P_{t+1} B_t)\bar{u}_t^\theta(I_t(\theta))$$

$$+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t \bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta)).$$

Note that the term $-2\bar{u}_t^\theta(I_t(\theta))^\top R_t K_t x$ and the term $+2\bar{u}_t^\theta(I_t(\theta))^\top B_t^\top P_{t+1}(A_t - B_t K_t)x$ cancel out because $R_t K_t = B_t^\top P_{t+1}(A_t - B_t K_t)$. We also note that the matrix in the first quadratic term can be simplified to

$$Q_t + K_t^\top R_t K_t + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t)$$

$$= Q_t + K_t^\top B_t^\top P_{t+1}(A_t - B_t K_t) + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t)$$

$$= Q_t + A_t^\top P_{t+1}(A_t - B_t K_t)$$

$$= Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} B_t K_t$$

$$= Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} H_t P_{t+1} A_t$$

$$= P_t,$$

where the last equation follows by the definition of P_t in (5).

Therefore, we obtain that

$$
\mathbb{E}\left[C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta)\right]
$$

$$
= x^\top P_t x + 2\left((A_t^\top - A_t^\top P_{t+1}H_t)(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1}\Phi_{\tau,t+1}^\top P_{\tau+1}W_{\tau|t}^\theta)\right)^\top x
$$

$$
+ \bar{u}_t^\theta(I_t(\theta))^\top (R_t + B_t^\top P_{t+1}B_t)\bar{u}_t^\theta(I_t(\theta))
$$

$$
+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1}\Phi_{\tau,t+1}^\top P_{\tau+1}W_{\tau|t}^\theta\right)^\top B_t\bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta))
$$

$$
= x^\top P_t x + 2\left(\sum_{\tau=t}^{T-1}\Phi_{\tau,t}^\top P_{\tau+1}W_{\tau|t}^\theta\right)^\top x + \bar{\psi}_t(I_t(\theta)),
$$

where the residual term $\bar{\psi}_t(I_t(\theta))$ is given by

$$
\bar{\psi}_t(I_t(\theta)) = \bar{u}_t^\theta(I_t(\theta))^\top (R_t + B_t^\top P_{t+1}B_t)\bar{u}_t^\theta(I_t(\theta))
$$

$$
+ 2\left(P_{t+1}W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1}\Phi_{\tau,t+1}^\top P_{\tau+1}W_{\tau|t}^\theta\right)^\top B_t\bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta)).
$$

Thus, we have shown the statement of Proposition 3.1 and 16 by induction.

A.2 Proof of Theorem 3.2

Note that the cost-to-go function $C_t^{\pi^\theta}(x;\xi)$ can be expressed as $Q_t^{\pi^\theta}(x, \pi_t^\theta(x;\iota_t(\theta);\xi))$. Substituting the expression of $\pi_t^\theta(x;\iota_t(\theta))$ into the expression of $\mathbb{E}\left[Q_t^{\pi^\theta}(x,u;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$ in Proposition 3.1 gives that

$$
\mathbb{E}\left[Q_t^{\pi^\theta}(x,u;\Xi) - C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right]
$$
$$
= (u - \pi_t^\theta(x;\iota_t(\theta)))^\top (R_t + B_t^\top P_{t+1}B_t)(u - \pi_t^\theta(x;\iota_t(\theta))). \tag{17}
$$

Substituting $u = \bar{\pi}_t(x;\iota_t(\mathbf{0}))$ into the above equation gives that

$$
\mathbb{E}\left[Q_t^{\pi^\theta}(x,\bar{\pi}_t(x;\iota_t(\mathbf{0}));\Xi) - C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right]
$$
$$
= (\bar{\pi}_t(x;\iota_t(\mathbf{0})) - \pi_t^\theta(x;\iota_t(\theta)))^\top (R_t + B_t^\top P_{t+1}B_t)(\bar{\pi}_t(x;\iota_t(\mathbf{0})) - \pi_t^\theta(x;\iota_t(\theta)))
$$
$$
= (\bar{u}_t^\theta(\iota_t(\theta)) - \bar{u}_t^{\mathbf{0}}(\iota_t(\mathbf{0})))^\top (R_t + B_t^\top P_{t+1}B_t)(\bar{u}_t^\theta(\iota_t(\theta)) - \bar{u}_t^{\mathbf{0}}(\iota_t(\mathbf{0}))) \tag{18a}
$$
$$
= \mathrm{Tr}\{(R_t + B_t^\top P_{t+1}B_t)(\bar{u}_t^\theta(\iota_t(\theta)) - \bar{u}_t^{\mathbf{0}}(\iota_t(\mathbf{0})))(\bar{u}_t^\theta(\iota_t(\theta)) - \bar{u}_t^{\mathbf{0}}(\iota_t(\mathbf{0})))^\top\}, \tag{18b}
$$

where we use the expression of optimal policies in Proposition 3.1 in (18a) and rearrange the terms in (18b). Note that by Proposition 3.1, we have

$$
\bar{u}_t^{\mathbf{0}}(\iota_t(\mathbf{0})) = \mathbb{E}\left[\bar{u}_t^\theta(I_t(\theta)) \mid I_t(\mathbf{0}) = \iota_t(\mathbf{0})\right].
$$

Therefore, by the tower rule and the definition of conditional covariance, we obtain that

$$
\mathbb{E}\left[Q_t^{\pi^\theta}(x,\bar{\pi}_t(x;\iota_t(\mathbf{0}));\Xi) - C_t^{\pi^\theta}(x;\Xi) \mid I_t(\mathbf{0}) = \iota_t(\mathbf{0})\right]
$$
$$
= \mathrm{Tr}\{(R_t + B_t^\top P_{t+1}B_t)\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid I_t(\mathbf{0}) = \iota_t(\mathbf{0})\right]\}. \tag{19}
$$

Let $\{(\bar{X}_t, \bar{U}_t)\}$ denote the (random) trajectory achieved $\bar{\pi}_{0:T-1}$ under problem instance Ξ. Since $\bar{X}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable, by (19), we obtain that

$$
\mathbb{E}\left[Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t;\Xi) - C_t^{\pi^\theta}(\bar{X}_t;\Xi) \mid \mathcal{F}_t(\mathbf{0})\right] = \mathrm{Tr}\{(R_t + B_t^\top P_{t+1}B_t)\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\},
$$
$$
\tag{20}
$$

where we use $\bar{U}_t = \bar{\pi}_t(\bar{X}_t; I_t(\mathbf{0}))$. Note that we have

$$Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) = h_t(\bar{X}_t, \bar{U}_t) + C_{t+1}^{\pi^\theta}(\bar{X}_{t+1}; \Xi).$$

Substituting this into (20) and taking expectation give that

$$\mathbb{E}\left[h_t(\bar{X}_t, \bar{U}_t) + C_{t+1}^{\pi^\theta}(\bar{X}_{t+1}; \Xi) - C_t^{\pi^\theta}(\bar{X}_t; \Xi)\right]$$
$$= \mathrm{Tr}\{(R_t + B_t^\top P_{t+1} B_t)\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right]\}. \tag{21}$$

Summing (21) over $t = 0, 1, \ldots, T-1$, we obtain that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} h_t(\bar{X}_t, \bar{U}_t) - C_0^{\pi^\theta}(\bar{X}_0; \Xi)\right] = \sum_{t=0}^{T-1} \mathrm{Tr}\{(R_t + B_t^\top P_{t+1} B_t)\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right]\}.$$

Note that the left-hand side equals $P(\theta)$. Thus, we have finished the proof of Theorem 3.2.

A.3 Proof of the MPC form

In this section, we show that the MPC policies defined in (7) and (8) are equivalent to the optimal policy in Proposition 3.1.

To simplify the notation, we define the large vectors

$$\vec{x} := \begin{bmatrix} x_t \\ x_{t+1} \\ \vdots \\ x_T \end{bmatrix}, \quad \vec{u} := \begin{bmatrix} u_t \\ u_{t+1} \\ \vdots \\ u_{T-1} \end{bmatrix}, \quad \text{and } \vec{w} := \begin{bmatrix} w_t \\ w_{t+1} \\ \vdots \\ w_{T-1} \end{bmatrix}.$$

Follow the approach of system level thesis, we know the constraints that

$$x_{\tau+1} := A_\tau x_\tau + B_\tau u_\tau + w_\tau, \text{ for } \tau \geq t, \text{ and } x_t = x$$

can be expressed equivalently by the affine relationship

$$\vec{x} := \Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}.$$

Let $\vec{Q} = \mathrm{Diag}(Q_t, \ldots, Q_{T-1}, P_T)$ and $\vec{R} = \mathrm{Diag}(R_t, \ldots, R_{T-1})$. We know the objective function (with equality constraints)

$$\sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T)$$
$$\text{s.t.} \quad x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_t), \text{ for } \tau \geq t, \text{ and } x_t = x, \tag{22}$$

can be written equivalently in the unconstrained form

$$(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w})^\top \vec{Q}(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}) + \vec{u}^\top \vec{R}\vec{u}. \tag{23}$$

We introduce the notations

$$\vec{W} := \begin{bmatrix} W_t \\ W_{t+1} \\ \vdots \\ W_{T-1} \end{bmatrix}, \quad \vec{W}_{\cdot|t}^\theta := \begin{bmatrix} W_{t|t}^\theta \\ W_{t+1|t}^\theta \\ \vdots \\ W_{T-1|t}^\theta \end{bmatrix}, \quad \text{and } \vec{w}_{\cdot|t}^\theta := \begin{bmatrix} w_{t|t}^\theta \\ w_{t+1|t}^\theta \\ \vdots \\ w_{T-1|t}^\theta \end{bmatrix}.$$

The MPC policy in (7) can be expressed as

$$\min_{\vec{u}} \mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^\top \vec{Q}(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^\top \vec{R}\vec{u} \,\Big|\, I_t(\theta) = \iota_t(\theta)\right].$$

Because the objective function can be reduced to

$$\mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^\top \vec{Q}(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^\top \vec{R}\vec{u} \,\Big|\, I_t(\theta) = \iota_t(\theta)\right]$$

$$= (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}^{\theta}_{\cdot|t})^{\top} \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}^{\theta}_{\cdot|t}) + \vec{u}^{\top} \vec{R} \vec{u}$$
$$+ \mathbb{E}\left[(\Phi_w (\vec{W} - \vec{W}^{\theta}_{\cdot|t}))^{\top} \vec{Q} \Phi_w (\vec{W} - \vec{W}^{\theta}_{\cdot|t}) \,\middle|\, I_t(\theta) = \iota_t(\theta) \right],$$

where the last term is independent with x and $\vec{u}$. Thus, the MPC policy in (7) is equivalent to

$$\mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^{\top} \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^{\top} \vec{R} \vec{u} \,\middle|\, I_t(\theta) = \iota_t(\theta) \right],$$

which is the MPC policy in (8).

Now, we show that (8) is equivalent to the optimal policy in Proposition 3.1. For any sequence $w_{t:T-1}$, let $\mathtt{MPC}(x, w_{t:T-1})$ denote the first entry of the solution to

$$\arg\min_{u_{t:T-1}} \sum_{\tau=t}^{T-1} h_{\tau}(x_{\tau}, u_{\tau}) + h_T(x_T)$$
$$\text{s.t.} \quad x_{\tau+1} = f_{\tau}(x_{\tau}, u_{\tau}; w_t), \text{ for } \tau \geq t, \text{ and } x_t = x, \tag{24}$$

To show that (8) is equivalent to the optimal policy in Proposition 3.1, we only need to show that

$$\mathtt{MPC}(x, w_{t:T-1}) = -K_t x - (R_t + B_t^{\top} P_{t+1} B_t)^{-1} B_t^{\top} \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^{\top} P_{\tau+1} w_t \tag{25}$$

holds for any sequence $w_{t:T-1}$. To see this, we consider the case when $w_{t:T-1}$ are deterministic disturbances on and after time step t, i.e., the agent knows $w_{t:T-1}$ exactly at time step t. In this scenario, we know the optimal policy is to follow the planned trajectory according to MPC in (22). On the other hand, by Proposition 3.1, we know the optimal action to take at time t is $-K_t x - (R_t + B_t^{\top} P_{t+1} B_t)^{-1} B_t^{\top} \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^{\top} P_{\tau+1} w_t$. Therefore, the first step planned by MPC must be identical with $-K_t x - (R_t + B_t^{\top} P_{t+1} B_t)^{-1} B_t^{\top} \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^{\top} P_{\tau+1} w_t$. Thus, (25) holds. And replacing $w_{t:(T-1)}$ with $w^{\theta}_{t:(T-1)|t}$ finishes the proof.

A.4 Prediction Power Evaluation

Based on our discussion in Section 3, we propose an algorithm (cf. Algorithm 1) to evaluate the prediction power efficiently given a set of historical problem instances $\{\xi_n\}_{n=1}^{N}$. Recall that we define the surrogate-optimal action as

$$\bar{u}_t^*(\Xi) := -(R_t + B_t^{\top} P_{t+1} B_t)^{-1} B_t^{\top} \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^{\top} P_{\tau+1} W_{\tau}, \tag{26}$$

which is the optimal action that an agent should take with the oracle knowledge of all future disturbances at time t. In the prediction power given by Theorem 3.2, we can express $\bar{u}_t^{\theta}(I_t(\theta))$ as $\mathbb{E}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta) \right]$ by Proposition 3.1, which is the expectation of $\bar{u}_t^*(\Xi)$ condition on the history at time step t.

We design Algorithm 1 as follows. While iterating backward from time step $T - 1$ to 0, the algorithm first constructs a dataset of the surrogate optimal action $\bar{u}_t^*(\Xi)$ as the fitting target. Then, the algorithm estimates the covariance of $\bar{u}_t^*(\Xi)$ when conditioning on $I_t(\mathbf{0})$ and $I_t(\theta)$, respectively, using a subroutine (Algorithm 2). The last step of Algorithm 1 gives the prediction power because $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^{\theta}(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0}) \right] \right]$ can be decomposed as $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\mathbf{0}) \right] \right] - \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta) \right] \right]$, and we prove this result in Lemma A.1. This decomposition is helpful because otherwise, we would need to evaluate the conditional expectation inside another conditional expectation. Specifically, $\bar{u}_t^{\theta}(I_t(\theta))$ needs to be approximated by a learned regressor (say, ϕ) that takes $I_t(\theta)$ as an input. Then, to evaluate $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^{\theta}(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0}) \right] \right]$, we would need to train another regressor to predict the output of ϕ. Our decomposition avoids this hierarchical dependence.

Lemma A.1. *For any random variable X and two σ-algebras $\mathcal{F} \subseteq \mathcal{F}'$, the following equation holds*

$$\mathbb{E}\left[\mathbf{Cov}\left[\mathbb{E}\left[X \mid \mathcal{F}' \right] \mid \mathcal{F} \right] \right] = \mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F} \right] \right] - \mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F}' \right] \right].$$

Proof of Lemma A.1. By the law of total covariance, we see that

$$\mathbf{Cov}\left[X | \mathcal{F} \right] = \mathbf{Cov}\left[\mathbb{E}\left[X \mid \mathcal{F}' \right] \mid \mathcal{F} \right] + \mathbb{E}\left[\mathbf{Cov}\left[X | \mathcal{F}' \right] | \mathcal{F} \right].$$

Algorithm 1 Prediction Power Evaluation

Require: Dataset D of problem instances $\{\xi_n\}_{n=1}^N$.
1: **for** $t = T-1, T-2, \ldots, 0$ **do**
2:     Compute P_t, H_t, K_t and $\{\Phi_{t,t'}\}_{t' \geq t}$ according to (5) and (6).
3:     Compute $M_t = R_t + B_t^\top P_{t+1} B_t$.
4:     **for** $n = 1, 2, \ldots, N$ **do**
5:         Compute $\bar{u}_t^*(\xi_n)$ according to (9) in problem instance ξ_n.
6:     **end for**
7:     Call Algorithm 2 to estimate $\Sigma_t^0 := \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(0)\right]\right]$ using $\{(\bar{u}_t^*(\xi_n), \iota_t^n(0))\}_{n=1}^N$.
8:     Call Algorithm 2 to estimate $\Sigma_t^\theta := \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta)\right]\right]$ using $\{(\bar{u}_t^*(\xi_n), \iota_t^n(\theta))\}_{n=1}^N$.
9: **end for**
10: **return** $P(\theta) = \sum_{t=0}^{T-1} \mathrm{Tr}\{\Sigma_t^0 M_t\} - \sum_{t=0}^{T-1} \mathrm{Tr}\{\Sigma_t^\theta M_t\}$

Taking expectation on both sides gives that

$$\mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F}\right]\right] = \mathbb{E}\left[\mathbf{Cov}\left[\mathbb{E}\left[X \mid \mathcal{F}'\right] \mid \mathcal{F}\right]\right] + \mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F}'\right]\right],$$

which is equivalent to the statement of Lemma A.1. $\qquad\square$

Evaluation of the Expected Conditional Covariance. For two general random variables X and Y, we follow a standard procedure to evaluate the expectation of their conditional covariance $\mathbb{E}\left[\mathbf{Cov}\left[Y \mid X\right]\right]$ using a dataset $\{(x_n, y_n)\}$ that is independently sampled from the joint distribution of (X, Y) (Algorithm 2). The algorithm first trains a regressor ψ that approximates the conditional expectation $\mathbb{E}\left[Y \mid X\right]$, where we use the definition:

$$\mathbb{E}\left[Y \mid X\right] = \min_{\psi \text{ is any function.}} \mathbb{E}\left[\|Y - \psi(X)\|_2^2\right].$$

Then, ψ is used for evaluating the conditional covariance. During training, we split the dataset into train, validation, and test sets in order to prevent overfitting.

Algorithm 2 Expected Conditional Covariance Estimator (ECCE)

Require: Dataset D that consists of input/output pairs (x_n, y_n).
1: Split the dataset D to D_{train}, D_{val}, and D_{test}.
2: Initialize a regressor ψ with input x and target output y.
3: Fit ψ to D_{train} with MSE and use D_{val} to prevent overfitting.
4: **return** $\Sigma := \frac{1}{|D_{\text{test}}|} \sum_{n \in D_{\text{test}}} (y_n - \psi(x_n))(y_n - \psi(x_n))^\top$.

A.5   Details of Examples in Section 3.1

In this section, we present the specific instantiation of Example 3.3 in Section 3.1 and another example (Example A.2) for the mismatch between prediction power and prediction accuracy.

A.5.1   Instantiation of Example 3.3

We instantiate Example 3.3 with the following parameters:

$$A = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, \ B = \begin{pmatrix} 0 \\ 0.1 \end{pmatrix}, \ Q = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}, \ R = (1), \text{ and } \theta := \begin{bmatrix} 1 & 0.99 \\ 0 & 0.141 \end{bmatrix}.$$

Under different values of coefficient ρ, we train a linear regressor to predict each entry of W_t from $V_t(\theta)$ (or $V_t(I)$) over a train dataset with 64000 independent samples. We plot in the MSE - ρ curve on a test dataset with 16000 independent samples in Figure 4. From the plot, we see that the predictors $V_t(\theta)$ and $V_t(I)$ achieve the same MSE when predicting each entry of W_t under each $\rho \in \{0, 0.1, \ldots, 0.7\}$.

Then, we use the trained linear regressors as $W_{t|t}^\theta$ and $W_{t|t}^I$ to implement the optimal policy in Proposition 3.1. We plot the averaged total cost over 16000 trajectories with horizon $T = 100$ in
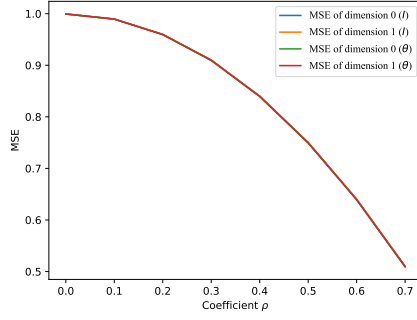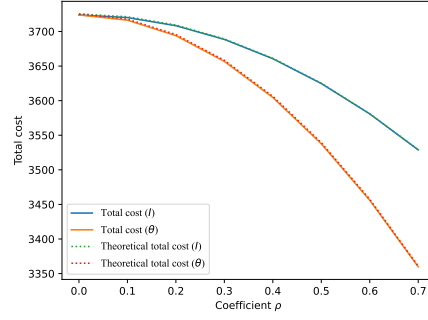
Figure 4: Example 3.3: MSE - ρ curve.

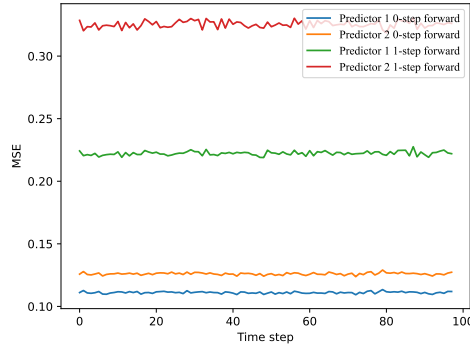Figure 5: Example 3.3: Control cost - ρ curve.

Figure 6: Example A.2: MSE - time curve.

Figure 5. From the plot, we see that the optimal policies under the predictors $V_t(\theta)$ and $V_t(I)$ achieve significantly different control costs when $\rho > 0$. We also plot the theoretical expected control cost in Figure 5 to verify this cost difference. Running this experiment takes about 50 seconds on Apple Mac mini with Apple M1 CPU.

A.5.2 An One-dimension Example

We also provide an example with $n = 1$, where the prediction $V_t(\theta)$ is correlated with two steps of future disturbances W_t and W_{t+1}.

Example A.2. *Suppose the disturbance at each time step can be decomposed as $W_t = \sum_{i=0}^{2} W_t^{(i)}$, where the $\{W_t^{(i)}\}_{i=0}^{2}$ are independently sampled from three mean-zero distributions. We compare two predictors: $V_t(1) = \left(W_t^{(1)}, W_{t+1}^{(0)}\right)$ and $V_t(2) = P\left(W_t^{(0)} + W_t^{(1)}\right) + (A^\top - A^\top PH)PW_{t+1}^{(0)}$. They have the same prediction power when used in the control problem because*

$$\bar{u}_t^2(I_t(2)) = P\left(W_t^{(0)} + W_t^{(1)}\right) + (A^\top - A^\top PH)PW_{t+1}^{(1)} = \bar{u}_t^1(I_t(1)).$$

However, we know that $\mathcal{F}_t(1)$ is a strict super set of $\mathcal{F}_t(2)$, thus $V_t(1)$ can achieve a better MSE than $V_t(2)$ when predicting the disturbances. This is empirically verified in a 1D LQR problem with $A = B = Q = R = (1)$ and $W_t^{(i)} \overset{i.i.d.}{\sim} N(0,1)$, as we plot in Figures 6. In the simulation, we train linear regressors to predict W_t and W_{t+1} with the history $I_t(1)$ or $I_t(2)$ for each time step $t < T = 100$ over a train dataset of size 160000. Then, we plot the MSE - time curve on a test dataset of size 40000. Running this experiment takes about 270 seconds on Apple Mac mini with Apple M1 CPU.

26

A.6 Details of Example 3.4

We instantiate Example 3.4 with the same dynamics and costs as Example 3.3, i.e.,

$$A = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, \ B = \begin{pmatrix} 0 \\ 0.1 \end{pmatrix}, \ Q = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}, \text{ and } R = (1).$$

To build the predictors, we sample the true disturbance $W_t \stackrel{\text{i.i.d.}}{\sim} N(0, I)$ and fix the coefficient $\rho = 0.5$. The online policy optimization starts with the initial policy parameter $\Upsilon_0 = \mathbf{0}$. When implementing M-GAPS in both scenarios, we use the decaying learning rate sequence $\eta_t = (1 + t/1000)^{-0.5}$. The optimal predictive policy for using $V_{0:t-1}(1)$ or $V_{0:t-1}(2)$ are $\pi^1_{0:T-1}$ and $\pi^2_{0:T-1}$, whose closed-form expressions are given by Proposition 3.1. Note that for the history $\iota_t(1)$, the optimal predictive policy π^1_t only depends on $v_t(1)$ because all other entries are independent with future disturbances $W_{t:T-1}$. Similarly, for the history $\iota_t(2)$, the optimal predictive policy π^2_t only depends on $v_{t-1}(2)$ and $v_t(2)$.

In Figures 1 and 2, we compute the average cost improvement of M-GAPS (or the optimal predictive policy) against the optimal no-prediction controller $\bar{\pi}_t(x) = -Kx$. That is, on each problem instance ξ, we plot

$$\frac{1}{t+1} \left(-(\text{cost of M-GAPS until time } t) + (\text{cost of } \bar{\pi} \text{ until time } t) \right)$$

for time $t = 0, 1, \ldots, T - 1$. The prediction power (averaged over time) is given by $P(\theta)/T$. We simulate 30 random trajectories with $T = 80000$ and plot the mean with the 25-th and 75-th percentiles as shaded areas. From the plots, we see that M-GAPS' average cost improvement converges towards the prediction power over time in the first scenario but stays far away with the prediction power in the second scenario. This is as expected, because the optimal predictive policy π^2_t is not in the candidate policy set of M-GAPS in the second scenario. Simulating the first scenario takes about 200 seconds on Apple Mac mini with Apple M1 CPU. The second takes about 210 seconds on the same hardware.

B Proof of Theorem 4.3

Since we assume x_0 is the initial state (deterministic) and π^θ is the optimal policy under the predictor with parameter θ, we have

$$\mathbb{E}\left[C_0^{\pi^\theta}(x_0; \Xi) \right] = J^{\pi^\theta}(\theta) = J^*(\theta).$$

Similarly, we also have that

$$\mathbb{E}\left[C_0^{\bar{\pi}}(x_0; \Xi) \right] = J^{\bar{\pi}}(\mathbf{0}) = J^*(\mathbf{0}).$$

Let $\{\bar{X}_{0:T}, \bar{U}_{0:T-1}\}$ be the trajectory of the baseline controller $\bar{\pi}_{0:T-1}$ under instance Ξ starting from $\bar{X}_0 = x_0$. First, we will prove by backwards induction that the difference in cumulative costs between the optimal controller π^θ and $\bar{\pi}$ has the following decomposition:

$$C_0^{\pi^\theta}(x_0; \Xi) - C_0^{\bar{\pi}}(x_0; \Xi) = \sum_{t=0}^{T-1} \left(C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \right). \tag{27}$$

For the base case at time $T - 1$, we apply the definition of $C_{T-1}^{\bar{\pi}}$ to get

$$C_{T-1}^{\pi^\theta}(\bar{X}_{T-1}; \Xi) - C_{T-1}^{\bar{\pi}}(\bar{X}_{T-1}; \Xi) = C_{T-1}^{\pi^\theta}(\bar{X}_{T-1}; \Xi) - Q_{T-1}^{\pi^\theta}(\bar{X}_{T-1}, \bar{U}_{T-1}; \Xi).$$

For the inductive step, suppose that

$$C_{\tau+1}^{\pi^\theta}(\bar{X}_{\tau+1}; \Xi) - C_{\tau+1}^{\bar{\pi}}(\bar{X}_{\tau+1}; \Xi) = \sum_{t=\tau+1}^{T-1} \left(C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \right).$$

Note that for any $t < T$,

$$Q_t^{\bar{\pi}}(\bar{X}_t, \bar{U}_t; \Xi) = Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) - \left(C_{t+1}^{\pi^\theta}(\bar{X}_{t+1}; \Xi) - C_{t+1}^{\bar{\pi}}(\bar{X}_{t+1}; \Xi) \right).$$

27

Therefore,

$$
\begin{aligned}
& C_\tau^{\pi^\theta}(\bar{X}_\tau; \Xi) - C_\tau^{\bar\pi}(\bar{X}_\tau; \Xi) \\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau; \Xi) - Q_\tau^{\bar\pi}(\bar{X}_\tau, \bar{U}_\tau; \Xi) \\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau; \Xi) - \left[Q_\tau^{\pi^\theta}(\bar{X}_\tau, \bar{U}_\tau; \Xi) - \left(C_{\tau+1}^{\pi^\theta}(\bar{X}_{\tau+1}; \Xi) - C_{\tau+1}^{\bar\pi}(\bar{X}_{\tau+1}; \Xi) \right) \right] \\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau; \Xi) - Q_\tau^{\pi^\theta}(\bar{X}_\tau, \bar{U}_\tau; \Xi) + \sum_{t=\tau+1}^{T-1} \left(C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \right) \\
&= \sum_{t=\tau}^{T-1} \left(C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \right).
\end{aligned}
$$

This completes the induction.

Next, define $U_t := \pi_t^\theta(\bar{X}_t; I_t(\theta))$. Note that U_t is $\mathcal{F}_t(\theta)$-measurable, and $\bar{U}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable and therefore also $\mathcal{F}_t(\theta)$-measurable. Because we assume the matrices $M_{0:T-1}$ satisfy Condition 4.1,

$$
\mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) \mid I_t(\theta) \right] \le \mathbb{E}\left[Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \mid I_t(\theta) \right] - \mathrm{Tr}\{ M_t(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \}. \tag{28}
$$

Let $\tilde{U}_t := \mathbb{E}\left[U_t \mid I_t(\mathbf{0}) \right]$. We see that

$$
\begin{aligned}
& \mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0}) \right] \\
&= \mathbb{E}\left[(\tilde{U}_t - U_t)(\tilde{U}_t - U_t)^\top \mid I_t(\mathbf{0}) \right] + \mathbb{E}\left[(\tilde{U}_t - U_t)(\bar{U}_t - \tilde{U}_t)^\top \mid I_t(\mathbf{0}) \right] \\
&\quad + \mathbb{E}\left[(\bar{U}_t - \tilde{U}_t)(\tilde{U}_t - U_t)^\top \mid I_t(\mathbf{0}) \right] + \mathbb{E}\left[(\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top \mid I_t(\mathbf{0}) \right] \\
&= \mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0}) \right] + \mathbb{E}\left[\tilde{U}_t - U_t \mid I_t(\mathbf{0}) \right] (\bar{U}_t - \tilde{U}_t)^\top \\
&\quad + (\bar{U}_t - \tilde{U}_t)\mathbb{E}\left[\tilde{U}_t - U_t \mid I_t(\mathbf{0}) \right]^\top + (\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top \tag{29a} \\
&= \mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0}) \right] + (\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top, \tag{29b}
\end{aligned}
$$

where we use $(\bar{U}_t - \tilde{U}_t)$ is $\mathcal{F}_t(\mathbf{0})$-measurable in (29a); we use the definition of $\tilde{U}_t$ in (29b).

Applying the towering rule in (27) and substituting in (28) gives that

$$
\begin{aligned}
\mathbb{E}\left[C_0^{\pi^\theta}(x_0; \Xi) - C_0^{\bar\pi}(x_0; \Xi) \right] &= \sum_{t=0}^{T-1} \mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) \mid I_t(\theta) \right] - \mathbb{E}\left[Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \mid I_t(\theta) \right] \right] \\
&\le -\sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{Tr}\{ M_t(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \} \right], \\
&= -\sum_{t=0}^{T-1} \mathrm{Tr}\{ M_t\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \right] \}. \tag{30}
\end{aligned}
$$

If the stronger Condition 4.2 (a) holds, by (29), since $\bar{X}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable, we have

$$
\begin{aligned}
\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \right] &= \mathbb{E}\left[\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0}) \right] \right] \\
&\succeq \mathbb{E}\left[\mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0}) \right] \right] \succeq \Sigma_t. \tag{31}
\end{aligned}
$$

Then, we can apply (31) in (30) to obtain that

$$
\mathbb{E}\left[C_0^{\pi^\theta}(x_0; \Xi) - C_0^{\bar\pi}(x_0; \Xi) \right] \le -\sum_{t=0}^{T-1} \mathrm{Tr}\{ M_t \Sigma_t \}. \tag{32}
$$

Else, if the weaker Condition 4.2 (b) holds, by (29), since $\bar{X}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable, we have

$$\text{Tr}\{\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right]\} = \mathbb{E}\left[\text{Tr}\{\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right]\}\right]$$
$$\geq \mathbb{E}\left[\text{Tr}\{\mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0})\right]\}\right] \geq \sigma_t. \qquad (33)$$

Note that for any positive semi-definite matrices A, B, C such that $A \succeq C \succeq 0$, we have

$$\text{Tr}\{AB\} = \text{Tr}\{CB\} + \text{Tr}\{(A - C)B\} \geq \text{Tr}\{CB\}.$$

Since $M_t \succeq \mu_{\min}(M_t)I$, we can apply (33) in (30) to obtain that

$$\mathbb{E}\left[C_0^{\pi^\theta}(x_0; \Xi) - C_0^{\bar{\pi}}(x_0; \Xi)\right] \leq -\sum_{t=0}^{T-1} \text{Tr}\left\{\mu_{\min}(M_t)I \cdot \mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right]\right\}$$
$$\leq -\sum_{t=0}^{T-1} \mu_{\min}(M_t)\sigma_t.$$

C Proofs for LTV Dynamics with General Costs

In this section, we first provide a proof outline of Theorem 4.8 (Appendix C.1). Then, we discuss an example where the MPC in (7) is suboptimal (Appendix C.2). Lastly, we provide the proofs for the key technical results required by the proof of Theorem 4.8.

C.1 Proof Outline of Theorem 4.8

Assumption 4.5 makes two requirements about the well-conditioned cost functions, which are standard in the literature of online optimization and control [3, 4]. For the last requirement, we additionally require $\ell_A < 1$, which implies that the system is open-loop stable. Under Assumption 4.5, the expected cost-to-go function is a well-conditioned function, which is important for establishing Conditions 4.1 and 4.2 (b). We state this result formally in Lemma 4.6 in Section 4.1, which establishes uniform bounds for the strongly convexity/smoothness of the conditional expectation of cost-to-go functions: μ_t is uniformly bounded below by μ_x and ℓ_t is uniformly bounded above by $\frac{\ell_x}{1-\ell_A}$. We present a proof sketch of Lemma 4.6 and defer the formal proof to Appendix C.4.

Starting from time step T, we know the cost-to-go $C_T^{\pi^\theta}(x; \Xi)$ equals to the terminal cost $h_t^x(x)$. It satisfies the strong convexity/smoothness directly by Assumption 4.5. We repeat the following induction iterations: Given $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x; \Xi) \mid I_{t+1}(\theta)\right]$ at time $t+1$, we define an auxiliary function that adds in the disturbance residual $W_t - W_{t|t}^\theta$ and condition on the history at time t:

$$\bar{C}_{t+1}^{\pi^\theta}(x; \iota_t(\theta)) := \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]. \qquad (34)$$

It can be expressed as $\mathbb{E}\left[\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta; \Xi) \mid I_{t+1}(\theta)\right] \Big| I_t(\theta) = \iota_t(\theta)\right]$ by the tower rule. Thus, we know function $\bar{C}_{t+1}^{\pi^\theta}$ is strongly convex and smooth in x because these properties are preserved after taking the expectation. Then, we can obtain the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right] = h_t^x(x) + \min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + w_{t|t}^\theta; \iota_t(\theta))\right)$. We use an existing tool called *infimal convolution* to study the optimal value of the this optimization problem as a function of x. Specifically, define an operator $\Box_B$:[4]

$$(f\Box_B\omega)(x) := \min_{u \in \mathbb{R}^m} \{f(u) + \omega(x - Bu)\} \text{ for } f: \mathbb{R}^m \to \mathbb{R} \text{ and } \omega: \mathbb{R}^n \to \mathbb{R}. \qquad (35)$$

One can show that if f and ω are well-conditioned functions, then $(f\Box_B\omega)$ is also well-conditioned (see Appendix C.6 for the formal statement and proof). We can use this result to show the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right] = h_t^x(x) + (h_t^u\Box_{(-B_t)}\bar{C}_{t+1}^{\pi^\theta})(A_t x + w_{t|t}^\theta; \iota_t(\theta))$, is also well-conditioned in x at time step t, which completes the induction.

---

[4]If ω takes an additional parameter w, we denote $(f\Box_B\omega)(x; w) := \min_{u \in \mathbb{R}^m} \{f(u) + \omega(x - Bu; w)\}$

For the second condition on the covariance of π_t^θ's actions, we note that $\lambda_t(\theta)$ in Assumption 4.7 should be positive as long as $V_t(\theta)$ has some weak correlation with W_t. Under Assumption 4.7, we can express the optimal policy as

$$\pi_t^\theta(x; I_t(\theta)) := \arg\min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta) \right). \tag{36}$$

While the original definition of $\bar{C}_{t+1}^{\pi^\theta}$ in (34) requires the history $\iota_t(\theta)$ as an input, it no longer depends on the history under Assumption 4.7. We defer the proof to Appendix C.5.

We can express $\pi_t^\theta(x; I_t(\theta))$ as the solution to $(h_t^u \square_{(-B_t)} \bar{C}_{t+1}^{\pi^\theta})(A_t x + W_{t|t})$. For some distributions including Gaussian, the covariance in the input of an infimal convolution will be passed through to its optimal solution. Specifically, let $u_{(f \square_B \omega)}(x)$ denote the solution to the optimization problem (35). When ω and f are well-conditioned, we can derive a lower bound on the trace of the covariance $\text{Tr}\{\mathbf{Cov}\left[u_{(f \square_B \omega)}(X)\right]\}$ that depends on the covariance of X. Due to space limit, we defer the formal statement of this result and its proof to Lemma C.2 in Appendix C.6. Using this property and the observation that $\pi_t^\theta(x; I_t(\theta))$ can be expressed as $u_{(h_t^u \square_{-B_t} \bar{C}_{t+1}^{\pi^\theta})}(A_t x + W_{t|t}^\theta)$, we can directly verify that Condition 4.2 (b) holds with

$$\text{Tr}\{\mathbf{Cov}\left[\pi_t^\theta(x; I_t(\theta)) \mid \mathcal{F}_t(0)\right]\} \geq \sigma_t := \frac{n \lambda_t(\theta) \mu_{t+1}^2 \cdot \mu_B}{2(\ell_u + \ell_{t+1}\sqrt{\ell_B})^2}. \tag{37}$$

Since Lemma 4.6 and (37) imply that Conditions 4.1 and 4.2 (b) hold with $M_t = \mu_t I$ and σ_t respectively, we can apply Theorem 4.3 to obtain the prediction power lower bound in Theorem 4.8.

C.2 Example: MPC can be suboptimal

We first highlight the challenge by showing that MPC can be suboptimal, i.e., only planning and optimizing based on the current information might be suboptimal when the cost functions are not quadratic.

Consider a 2-step optimal control problem (1-dimension):

$$X_1 = X_0 + U_0, \text{ and } X_2 = X_1 + U_1 + W_1.$$

The cost functions are given by

$$h_0(x, u) = x^2 + u^2, \; h_1(x, u) = x^2 + u^2, \text{ and } h_2(x) = \begin{cases} x^2, & \text{if } x \leq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Suppose W_1 is a random variable that satisfies $\mathbb{P}(W_1 = 1) = p$ and $\mathbb{P}(W_1 = 0) = 1 - p$, where $0 < p < 1$. At time 0, we don't have any knowledge about W_1 (i.e., W_1 is independent with $I_0(\theta)$). However, at time 1, we can predict W_1 exactly, which means $\sigma(W_1) \subseteq \mathcal{F}_1(\theta)$.

Suppose the system starts at $x_0 = 0$. At time step 0, MPC (7) solves the optimization

$$\min_{u_0, u_1} \mathbb{E}\left[h_0(X_0, u_0) + h_1(X_1, u_1) + h_2(X_2) \mid I_0(\theta)\right]$$
$$\text{s.t. } X_0 = 0, \; X_1 = X_0 + u_0, \; X_2 = X_1 + u_1 + W_1. \tag{38}$$

Since $I_0(\theta)$ is independent with W_1, the optimization problem can be expressed equivalently as

$$\min_{u_0, u_1} u_0^2 + (u_0^2 + u_1^2) + \mathbb{E}\left[h_2(u_0 + u_1 + W_1)\right]$$
$$= \min_{u_0, u_1} 2u_0^2 + u_1^2 + 1, \text{ s.t. } u_0 + u_1 = -1.$$

The equation holds because the planned trajectory must avoid the huge cost at time step 2. Solving this gives $u_0 = -\frac{1}{3}$. Thus, implementing MPC incurs a total cost that is at least $2u_0^2 = \frac{2}{9}$. In contrast, if one just pick $u_0 = 0$, the agent can pick u_1 based on the prediction revealed at time step 2:

$$u_1 = \begin{cases} 0 & \text{if } W_1 = 0, \\ -1 & \text{otherwise.} \end{cases}$$

In this case, the expected cost incurred is p. Thus, we can claim that MPC is not the optimal policy when $p < \frac{2}{9}$. The underlying reason that MPC is suboptimal is because it does not consider what

information may be available when we make the decision in the future. In this specific example, since W_1 is revealed at time 1, we don't need to verify about the small probability event that leads to a huge loss.

We dive deeper into the reason why MPC (7) is optimal in the LQR setting (Section 3). Note that the expected optimal cost-to-go function at time step 1 is

$$\mathbb{E}\left[C_1^{\pi^\theta}(x;\Xi) \,\Big|\, I_1(\theta)\right] = \min_{u_1} \mathbb{E}\left[h_1(x,u_1) + h_2(X_2) \mid I_1(\theta)\right], \text{ s.t. } X_2 = x + u_1 + W_1. \quad (39)$$

Here, u_1 is $\mathcal{F}_1(\theta)$-measurable. And the true optimal policy at time 0 is decided by solving

$$\min_{u_0} h_0(x,u_0) + \mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right], \text{ s.t. } X_1 = x + u_0.$$

In general, we cannot use

$$\min_{u_1} \mathbb{E}\left[h_1(X_1,u_1) + h_2(X_2) \mid I_0(\theta)\right], \text{ s.t. } X_2 = X_1 + u_1 + W_1, \quad (40)$$

to replace $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \,\Big|\, I_0(\theta)\right]$ like what MPC does in (38) because here u_1 is $\mathcal{F}_0(\theta)$-measurable in (40). Recall that u_1 is $\mathcal{F}_1(\theta)$-measurable in (39) and $\mathcal{F}_0(\theta)$ is a subset of $\mathcal{F}_1(\theta)$. However, in the LQR setting, as the closed-form expression (16), the part of $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right]$ that depends on X_1 will not change even if $\mathcal{F}_1(\theta)$ changes. Thus, we can assume $\mathcal{F}_1(\theta) = \mathcal{F}_0(\theta)$ without affecting the optimal action at time 0. Therefore, MPC's replacement of $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right]$ with (40) is valid in the LQR setting.

C.3 Infimal Convolution Properties

The first result states that the variant of infimal convolution preserves the strong convexity/smoothness of the input functions. The proof can be found in Appendix C.6.

Lemma C.1. *Consider a variant of infimal convolution defined as*

$$\left(f\Box_B\omega\right)(x) = \min_u \left\{f(u) + \omega(x - Bu)\right\}, \quad (41)$$

where $f : \mathbb{R}^m \to \mathbb{R}$, $\omega : \mathbb{R}^n \to \mathbb{R}$, and $B \in \mathbb{R}^{n\times m}$ is a matrix. Suppose that f is a μ_f-strongly convex function, and ω is a μ_ω-strongly convex and ℓ_ω-smooth function. Then, $f\Box_B\omega$ is a $\left(\frac{\mu_\omega\mu_f}{\mu_f+\|B\|^2\mu_\omega}\right)$-strongly convex and ℓ_ω-smooth function. We also have $\nabla(f\Box_B\omega)(x) = \nabla\omega(x - Bu(x))$.

The second result is about the optimal solution of the variant of infimal convolution. It states that for some distributions, the covariance on the input will induce a variance on the optimal solution. We state it in Lemma C.2 and defer the proof to Appendix C.7.

Lemma C.2. *Let $u_{(f\Box_B\omega)}(x)$ denote the solution to the optimization problem (35). Suppose function f is μ_f-strongly convex. Function ω is μ_ω-strongly convex and ℓ_ω-smooth. Suppose X is a random vector with bounded mean and $\mathbf{Cov}\,[X] = \Sigma \succeq \sigma_0 I$. Further, there exists a constant $C > 0$ such that for any positive integer N, X can be decomposed as $X = \sum_{i=1}^N X_i$ for i.i.d. random vectors X_i that satisfies $\mathbb{E}\left[\|X_i\|^4\right] \leq C \cdot N^{-2}$. Then,*

$$\mathrm{Tr}\big\{\mathbf{Cov}\left[u_{(f\Box_B\omega)}(X)\right]\big\} \geq \frac{n\sigma_0\mu_\omega^2 \cdot \sigma_{min}(B)^2}{2(\ell_f + \ell_\omega\|B\|)^2}.$$

As a remark, an example of X that satisfies the assumptions is the normal distribution $X \sim N(0,\Sigma)$. We have $X_i \sim N(0,\Sigma/N)$, thus $\mathbb{E}\left[\|X_i\|^4\right] \leq 3\,\mathrm{Tr}\{\Sigma\}N^{-2}$.

The next result (Lemma C.3) considers the case when there is an additional input w to function ω in the infimal convolution. When this additional parameter causes a covariance on the gradient $\nabla_1\omega(x,W)$, the optimal solution of the infimal convolution will also have a nonzero variance.

Lemma C.3. *Suppose that $\omega(x, w)$ satisfies that $\omega(\cdot, w)$ is an ℓ_ω-smooth convex function for all w. For a random variable W, suppose that the following inequality holds for arbitrary fixed vector $x \in \mathbb{R}^n$,*

$$\mathbf{Cov}\left[\nabla_1 \omega(x, W)\right] \succeq \sigma_0 I.$$

Suppose that $f : \mathbb{R}^m \to \mathbb{R}$ is a μ_f-strongly convex and ℓ_f-smooth function ($m \leq n$). Let B be a matrix in $\mathbb{R}^{n \times m}$. Then, the optimal solution of the infimal convolution

$$u_{(f \square_B \omega)}(x, w) := \arg\min_u \left(f(u) + \omega(x - Bu, w)\right)$$

satisfies that

$$\mathrm{Tr}\big\{\mathbf{Cov}\left[u_{(f\square_B\omega)}(x, W)\right]\big\} \geq \frac{n\sigma_0 \cdot \sigma_{min}(B)^2}{2(\ell_f + \ell_\omega \|B\|)^2}.$$

holds for arbitrary fixed vector x, where $\sigma_{min}(B)$ denotes the minimum singular value of B.

Lemma C.3 is useful for showing Lemma C.2. We defer its proof to Appendix C.8.

C.4 Proof of Lemma 4.6

We use induction to show that $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$ is a μ_t-strongly convex and ℓ_t-smooth function for any $\iota_t(\theta)$, where the coefficients μ_t and ℓ_t are defined recursively in (14). To simplify the notation, we will omit "$I_t(\theta) =$" in the conditional expectations throughout this proof when conditioning on a realization of the history $\iota_t(\theta)$.

Note that the statement holds for $t = T$, because $\mathbb{E}\left[C_T^{\pi^\theta}(x; \Xi) \mid \iota_T(\theta)\right] = h_T^x(x)$ and the terminal cost h_T^x is μ_x-strongly convex and ℓ_x-smooth.

Suppose the statement holds for $t + 1$. We see that

$$\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid \iota_t(\theta)\right] = h_t^x(x) + \min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right).$$

By the induction assumption, we know that $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(\cdot; \Xi) \mid \iota_{t+1}(\theta)\right]$ is a μ_{t+1}-strongly convex and ℓ_{t+1}-smooth function for any $\iota_{t+1}(\theta)$. Thus, $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(\cdot + W_t; \Xi) \mid \iota_t(\theta)\right]$ is also a μ_{t+1}-strongly convex and ℓ_{t+1}-smooth function. Therefore,

$$\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right)$$

is a $\frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}$-strongly convex and ℓ_{t+1}-smooth function of x by Lemma C.1. By changing the variable from x to $A_t x$, we see that

$$\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right)$$

is a $\mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}$-strongly convex and $\ell_A \cdot \ell_{t+1}$-smooth function by Assumption 4.5. Since h_t^x is a μ_x-strongly convex and ℓ_x-smooth function, we see that $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid \iota_t(\theta)\right]$ is also a μ_t-strongly convex and ℓ_t-smooth function because

$$\mu_t = \mu_x + \mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}, \text{ and } \ell_t = \ell_x + \ell_A \cdot \ell_{t+1}.$$

C.5 Proof of Theorem 4.8

Note that the optimal action at time step t is determined by

$$\pi_t^\theta(x; I_t(\theta)) := \arg\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid I_t(\theta)\right]\right). \tag{42}$$

This can be further simplified to

$$\pi_t^\theta(x; I_t(\theta)) := \arg\min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta) \right).$$

The additional input $I_t(\theta)$ is not required for $\bar{C}_{t+1}^{\pi^\theta}$ because the function $\bar{C}_{t+1}^{\pi^\theta}(x; \iota_t(\theta))$ does not change with the history $\iota_t(\theta)$ under Assumption 4.7. The reason is that $W_t - W_{t|t}^\theta$ and all future predictions and disturbances $W_{t+1:T-1}, V_{t+1:T-1}^\theta$ are independent with the history $I_t(\theta)$.

By (36), we see that

$$\pi_t^\theta(x; I_t(\theta)) = u_{(h_t^u \square_{-B_t} \bar{C}_{t+1}^{\pi^\theta})}(A_t x + W_{t|t}^\theta).$$

Under Assumption 4.7, we see that

$$\mathbf{Cov}\left[W_{t|t}^\theta\right] = \mathbf{Cov}\left[W_t\right] - \mathbf{Cov}\left[W_t \mid V_t(\theta)\right] \succeq \lambda_t(\theta) I$$

and $W_{t|t}^\theta$ is Gaussian. Therefore, we can apply Lemma C.2 to obtain that

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[\pi_t^\theta(x; I_t(\theta)) \mid \mathcal{F}_t(0)\right]\right\} \geq \sigma_t := \frac{n\lambda_t(\theta)\mu_{t+1}^2 \cdot \mu_B}{2(\ell_u + \ell_{t+1}\sqrt{\ell_B})^2}.$$

Thus, Condition 4.2 (b) holds with σ_t.

On the other hand, Condition 4.1 holds with $M_t = \mu_t I$ by Lemma 4.6. Therefore, by Theorem 4.3, we obtain that $P(\theta) \geq \sum_{t=0}^{T-1} \mu_u \sigma_t$.

C.6 Proof of Lemma C.1

By the definition of conjugate, we see that

$$
\begin{align}
(f\square_B\omega)^*(y) &= \max_x \left\{\langle y, x\rangle - \min_u \{f(u) + \omega(x - Bu)\}\right\} \tag{43a}\\
&= \max_x \max_u \left\{\langle y, x\rangle - f(u) - \omega(x - Bu)\right\}\\
&= \max_x \max_u \left\{\langle y, x - Bu\rangle + \langle y, Bu\rangle - f(u) - \omega(x - Bu)\right\}\\
&= \max_u \max_x \left\{(\langle y, x - Bu\rangle - \omega(x - Bu)) + (\langle B^\top y, u\rangle - f(u))\right\} \tag{43b}\\
&= \max_u \left\{\max_x \{\langle y, x - Bu\rangle - \omega(x - Bu)\} + \langle B^\top y, u\rangle - f(u)\right\}\\
&= \max_u \left\{\omega^*(y) + \langle B^\top y, u\rangle - f(u)\right\} \tag{43c}\\
&= \omega^*(y) + f^*(B^\top y), \tag{43d}
\end{align}
$$

where we use the definition of $f\square_B\omega$ in (43a); we change the order of taking the maximum and use $\langle y, Bu\rangle = \langle B^\top y, u\rangle$ in (43b); we use the definition of ω^* in (43c); we use the definition of f^* in (43d).

Since $f\square_B\omega$ is convex, by Theorem 4.8 in [26], we know that

$$(f\square_B\omega)(y) = \left(\omega^*(y) + f^*(B^\top y)\right)^*. \tag{44}$$

Since ω is a μ_ω-strongly convex and ℓ_ω-smooth function, we know ω^* is an $\frac{1}{\ell_\omega}$-strongly convex and $\frac{1}{\mu_\omega}$-smooth function by the conjugate correspondence theorem [26]. Similarly, we know that f^* is a $\frac{1}{\mu_f}$-smooth convex function. Thus, we know that $\omega^*(y) + f^*(B^\top y)$ is an $\frac{1}{\ell_\omega}$-strongly convex and $\left(\frac{1}{\mu_\omega} + \frac{\|B\|^2}{\mu_f}\right)$-smooth function. Therefore, by the conjugate correspondence theorem, we know that $f\square_B\omega$ is a $\left(\frac{\mu_\omega\mu_f}{\mu_f + \|B\|^2\mu_\omega}\right)$-strongly convex and ℓ_ω-smooth function.

Now, we show that

$$\nabla(f\square_B\omega)(x) = \nabla\omega(x - Bu(x)). \tag{45}$$

Following a similar approach with the proof of Theorem 5.30 in [26], we define $z = \nabla\omega(x - Bu(x))$. Define function $\phi(\xi) := (f\Box_B\omega)(x + \xi) - (f\Box_B\omega)(x) - \langle\xi, z\rangle$. We see that

$$\phi(\xi) = (f\Box_B\omega)(x + \xi) - (f\Box_B\omega)(x) - \langle\xi, z\rangle$$

$$\leq \omega(x + \xi - Bu(x)) - \omega(x - Bu(x)) - \langle\xi, z\rangle \tag{46a}$$

$$\leq \langle\xi, \nabla\omega(x + \xi - Bu(x))\rangle - \langle\xi, z\rangle \tag{46b}$$

$$= \langle\xi, \nabla\omega(x + \xi - Bu(x)) - \nabla\omega(x - Bu(x))\rangle$$

$$\leq \|\xi\| \cdot \|\nabla\omega(x + \xi - Bu(x)) - \nabla\omega(x - Bu(x))\| \tag{46c}$$

$$\leq \ell_\omega\|\xi\|^2, \tag{46d}$$

where in (46a), we use

$$(f\Box_B\omega)(x + \xi) \leq f(u(x)) + \omega(x + \xi - Bu(x)), \text{ and}$$
$$(f\Box_B\omega)(x) = f(u(x)) + \omega(x - Bu(x));$$

we use the convexity of ω in (46b); we use the Cauchy-Schwarz inequality in (46c); we use the assumption that ω is ℓ_ω-smooth in (46d).

Since $(f\Box_B\omega)$ is a convex function, ϕ is also convex, thus we see that

$$\phi(\xi) \geq 2\phi(0) - \phi(-\xi) = -\phi(-\xi) \geq -\ell_\omega\|\xi\|^2.$$

Combining this with (46), we conclude that $\lim_{\|\xi\|\to 0} |\phi(\xi)|/\|\xi\| = 0$. Thus, (45) holds.

C.7 Proof of Lemma C.2

By Theorem D.1, we see that

$$\mathbf{Cov}\left[\nabla\omega(X)\right] \geq \sigma_0\mu_\omega^2.$$

Then, we apply Lemma C.3 with the second function input to the infimal convolution as $\tilde{\omega}(x, w) := \omega(x + w)$. In the context of Lemma C.3, we set $W = X$, so the assumption about the covariance of the gradient holds with

$$\mathbf{Cov}\left[\nabla_1\tilde{\omega}(x, W)\right] \succeq \sigma_0\mu_\omega^2.$$

Note that for any fixed w, $\tilde{\omega}(\cdot, w)$ is μ_ω-strongly convex. Therefore, we obtain that

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f\Box_B\omega)}(X)\right]\right\} = \mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f\Box_B\tilde{\omega})}(0, W)\right]\right\} \geq \frac{n\sigma_0\mu_\omega^2 \cdot \sigma_{\min}(B)^2}{2(\ell_f + \ell_\omega\|B\|)^2}$$

C.8 Proof of Lemma C.3

Because function c is ℓ_c-smooth, we have

$$\|\nabla c(u(x, w)) - \nabla c(u(x, w'))\| \leq \ell_c\|u(x, w) - u(x, w')\| \tag{47}$$

Because function f is ℓ_f-smooth, we have

$$\left\|B^\top\nabla_1 f(x - Bu(x, w), w) - B^\top\nabla_1 f(x - Bu(x, w'), w')\right\|$$

$$\geq \left\|B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w) - B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w')\right\|$$

$$- \left\|B^\top\nabla_1 f(x - B \cdot u(x, w), w) - B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w)\right\|$$

$$- \left\|B^\top\nabla_1 f(x - B \cdot u(x, w'), w') - B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w')\right\| \tag{48a}$$

$$\geq \left\|B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w) - B^\top\nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x, W)\right], w')\right\|$$

$$- \ell_f\|B\| \cdot \left(\|u(x, w) - \mathbb{E}_W\left[u(x, W)\right]\| + \|u(x, w') - \mathbb{E}_W\left[u(x, W)\right]\|\right), \tag{48b}$$

where we use the triangle inequality in (48a); we use the smoothness of f in (48b).

Note that by the first-order optimality condition, we have

$$\nabla c(u(x, w)) - B^\top\nabla_1 f(x - B \cdot u(x, w), w) = 0.$$

Therefore, for any w, w', we have that

$$\nabla c(u(x,w)) - \nabla c(u(x,w')) = B^\top \nabla_1 f(x - B \cdot u(x,w), w) - B^\top \nabla_1 f(x - B \cdot u(x,w'), w'). \tag{49}$$

By combining (49) with (47) and (48), we obtain that

$$\ell_c \|u(x,w) - u(x,w')\|$$
$$+ \ell_f \cdot \|B\| \cdot (\|u(x,w) - \mathbb{E}_W[u(x,W)]\| + \|u(x,w') - \mathbb{E}_W[u(x,W)]\|)$$
$$\geq \|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], w) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], w')\|$$

holds for arbitrary w and w'. Let W' be a random vector independent of W and have the same distribution. By replacing w/w' with W/W' respectively, we see

$$\ell_c \|u(x,W) - u(x,W')\|$$
$$+ \ell_f \cdot \|B\| \cdot (\|u(x,W) - \mathbb{E}_W[u(x,W)]\| + \|u(x,W') - \mathbb{E}_W[u(x,W)]\|)$$
$$\geq \|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W')\|,$$

which implies

$$(\ell_c + \ell_f \|B\|) \left(\|u(x,W) - \mathbb{E}_W[u(x,W)]\| + \|u(x,W') - \mathbb{E}_W[u(x,W)]\|\right)$$
$$\geq \|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W')\| \tag{50}$$

by the triangle inequality. Taking the square of both sides of (50) and applying the AM-GM inequality gives that

$$2(\ell_c + \ell_f\|B\|)^2\|u(x,W) - \mathbb{E}_W[u(x,W)]\|^2 + 2(\ell_c + \ell_f\|B\|)^2\|u(x,W') - \mathbb{E}_W[u(x,W)]\|^2$$
$$\geq \|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W')\|^2. \tag{51}$$

Let $Y := \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W) - \nabla_1 f(x - B \cdot \mathbb{E}_W[u(x,W)], W')$. Note that the right-hand side of (51) can be expressed as $\|B^\top Y\|^2 = \text{Tr}\{B^\top(YY^\top)B\}$. By taking the expectations of both sides, we obtain that

$$4(\ell_c + \ell_f\|B\|)^2 \text{Tr}\{\mathbf{Cov}[u(x,W)]\} \geq 2\,\text{Tr}\{B^\top \mathbf{Cov}[\nabla_1 f(x - B\mathbb{E}_W[u(x,W)], W)]B\}$$
$$\geq 2n\sigma_0 \sigma_{\min}(B)^2.$$

In the last inequality, we use the property that the trace of a positive semi-definite matrix equals the sum of its eigenvalues. Thus, it is greater than or equal to n times the smallest eigenvalue $\sigma_0 \sigma_{\min}(B)^2$. Rearranging the terms finishes the proof.

D Useful Technical Results

In this section, we state a useful result about what functions can pass the covariance of its input to the output in Theorem D.1, which is used to show Lemma C.2. We defer the proof to Appendix D.1.

Theorem D.1. *Suppose that a function $g : \mathbb{R}^d \to \mathbb{R}^d$ satisfies*

$$\langle g(x) - g(x'), x - x' \rangle \geq \gamma \|x - x'\|^2, \text{ and } \|g(x) - g(x')\| \leq L\|x - x'\|, \ \forall\, x, x' \in \mathbb{R}^d. \tag{52}$$

Additionally, there exists a positive constant ℓ such that

$$-\ell I \preceq \nabla^2 g_i(x) \preceq \ell I, \ \forall\, x \in \mathbb{R}^d, i \in [d]. \tag{53}$$

Suppose X is a random vector that satisfies $|\mathbb{E}[X]| < \infty$ and $\mathbf{Cov}[X] = \Sigma \succeq \mu I$. Further, there exists a constant $C > 0$ such that for any positive integer N, X can be decomposed as $X = \sum_{i=1}^N X_i$ for i.i.d. random vectors X_i that satisfy $\mathbb{E}\left[\|X_i\|^4\right] \leq C \cdot N^{-2}$. Then, we have

$$\mathbf{Cov}[g(X)] \succeq \mu\gamma^2 I.$$

As a remark, the gradient of a well-conditioned function satisfies the conditions in (52).

D.1 Proof of Theorem D.1

Without any loss of generality, we assume $\mathbb{E}[X] = 0$ because we can view $g(\mathbb{E}[X] + \cdot)$ as the function and subtract the mean from the random variables. The assumptions about g and X in Theorem D.1 still hold.

For any $i \in [d]$ and $\epsilon \in \mathbb{R}^d$, we have the Taylor series expansion Lagrangian form (see Chapter 3.2 of [27])

$$g_i(x + \epsilon) = g_i(x) + \nabla g_i(x)^\top \epsilon + \frac{1}{2}\epsilon^\top \nabla^2 g_i(\bar{x}^{(i)})\epsilon, \tag{54}$$

where $\bar{x}^{(i)}$ is a point on the line segment between x and $x + \epsilon$. For notational convenience, let

$$\nabla g(x) := \begin{bmatrix} \nabla g_1(x)^\top \\ \vdots \\ \nabla g_d(x)^\top \end{bmatrix} \in \mathbb{R}^{d \times d}, \text{ and } v_1(x, \epsilon) := \begin{bmatrix} \epsilon^\top \nabla^2 g_1(\bar{x}^{(1)})\epsilon \\ \vdots \\ \epsilon^\top \nabla^2 g_d(\bar{x}^{(d)})\epsilon \end{bmatrix} \in \mathbb{R}^d.$$

With the above notation, Eq. (54) can be equivalently written as

$$g(x + \epsilon) - g(x) = \nabla g(x) \cdot \epsilon + \frac{1}{2}v_1(x, \epsilon). \tag{55}$$

From Eq. (53), we know that $|v(x, \epsilon)_i| \le \ell\|\epsilon\|^2$, which implies

$$\|v_1(x, \epsilon)\| \le \ell\sqrt{d}\|\epsilon\|^2. \tag{56}$$

In addition, by Eq. (52), we see that

$$\langle g(x + \epsilon) - g(x), \epsilon \rangle \ge \gamma\|\epsilon\|^2.$$

Substituting Eq. (55) into the above equation and rearranging the terms, we obtain

$$\epsilon^\top \cdot \nabla g(x) \cdot \epsilon \ge \gamma\|\epsilon\|^2 - \epsilon^\top \cdot v_1(x, \epsilon),$$

which is equivalent to

$$\epsilon^\top \cdot \frac{\nabla g(x) + \nabla g(x)^\top}{2} \cdot \epsilon \ge \gamma\|\epsilon\|^2 - \epsilon^\top \cdot v_1(x, \epsilon).$$

Observe that the term subtracted from the right-hand side satisfies $\left|\epsilon^\top \cdot v_1(x, \epsilon)\right| \le \ell\sqrt{d}\|\epsilon\|^3$, which follows from Cauchy–Schwarz inequality and Eq. (56). Therefore, since the previous inequality holds for any $\epsilon \in \mathbb{R}^d$, taking $\epsilon \to 0$ gives that

$$\frac{\nabla g(x) + \nabla g(x)^\top}{2} \succeq \gamma I. \tag{57}$$

Before we proceed, we first state and prove a lemma that can convert the summation in Eq. (57) into a product form.

Lemma D.2. *Let $M \in \mathbb{R}^{d \times d}$ be a real-valued matrix satisfying $M + M^\top \succeq 2\gamma I$. Then, for any positive definite matrix $\Sigma \succeq \mu I$, we have $M\Sigma M^\top \succeq \mu\gamma^2 I$.*

Proof of Lemma D.2. Since $M + M^\top \succeq 2\gamma I$, we have for any $x \in \mathbb{R}^d$ that

$$2\gamma\|x\|^2 \le 2x^\top M^\top x = 2x^\top \Sigma^{-1/2}\Sigma^{1/2}M^\top x \le 2\|\Sigma^{-1/2}x\|\|\Sigma^{1/2}M^\top x\|$$
$$\le 2\mu^{-1/2}\|x\|\|\Sigma^{1/2}M^\top x\|,$$

where the last inequality follows from $\Sigma \succeq \mu I \Rightarrow \|\Sigma^{-1/2}x\| = \sqrt{x^\top \Sigma^{-1}x} \le \mu^{-1/2}\|x\|$. Rearranging terms, we obtain

$$\gamma\mu^{1/2}\|x\| \le \|\Sigma^{1/2}M^\top x\|.$$

Squaring both sides concludes the proof. $\square$

Next, we state and prove a lemma about the lower bound of the covariance induced by an additive random noise on the input that is useful when the noise is sufficiently small.

Lemma D.3. *Let ε be a mean-zero random vector in $\mathbb{R}^d$ that satisfies $\underline{\delta} I \preceq \mathbf{Cov}\left[\varepsilon\right]$ and $\mathbb{E}\left[\|\varepsilon\|^4\right] \leq \overline{\gamma}$. Let g be a function that satisfies (52) and (53). Then, for arbitrary fixed real vector $x \in \mathbb{R}^d$, we have*

$$\mathbf{Cov}\left[g(x+\varepsilon)\right] \succeq \left(\gamma^2 \underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right) I.$$

Proof of Lemma D.3. We first derive bounds on the i th moment of $\|\varepsilon\|$ ($i = 1, 2, 3$). By Jensen's inequality, we have

$$\mathbb{E}\left[\|\varepsilon\|^2\right] = \mathbb{E}\left[\left(\|\varepsilon\|^4\right)^{\frac{1}{2}}\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^4\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{1}{2}}. \tag{58}$$

Using Jensen's inequality again, we obtain that

$$\mathbb{E}\left[\|\varepsilon\|\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^2\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{1}{4}}. \tag{59}$$

Lastly, by the Cauchy-Schwartz inequality, we see that

$$\mathbb{E}\left[\|\varepsilon\|^3\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^4\right] \cdot \mathbb{E}\left[\|\varepsilon\|^2\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{3}{4}}. \tag{60}$$

Note that by (55), we have

$$\mathbf{Cov}\left[g(x+\varepsilon)\right] = \mathbf{Cov}\left[g(x+\varepsilon) - g(x)\right] = \mathbf{Cov}\left[\nabla g(x) \cdot \varepsilon + \frac{1}{2}v_1(x,\varepsilon)\right]. \tag{61}$$

Since $\mathbb{E}\left[\varepsilon\right] = 0$, we can further decompose (61) as

$$\mathbf{Cov}\left[\nabla g(x) \cdot \varepsilon + \frac{1}{2}v_1(x,\varepsilon)\right]$$

$$= \mathbb{E}\left[\left(\nabla g(x) \cdot \varepsilon + \frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)\left(\nabla g(x) \cdot \varepsilon + \frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]$$

$$= \nabla g(x) \cdot \mathbf{Cov}\left[\varepsilon\right] \cdot \nabla g(x)^\top + \nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]$$

$$+ \mathbb{E}\left[\left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right) \cdot \varepsilon^\top\right] \cdot \nabla g(x)^\top + \frac{1}{4}\mathbf{Cov}\left[v_1(x,\varepsilon)\right]. \tag{62}$$

By Lemma D.2 and (57), we know the first term in (62) is lower bounded by

$$\nabla g(x) \cdot \mathbf{Cov}\left[\varepsilon\right] \cdot \nabla g(x)^\top \succeq \gamma^2 \underline{\delta} I. \tag{63}$$

Define the residual term as the sum of the last 3 terms in (62):

$$R := \nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]$$

$$+ \mathbb{E}\left[\left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right) \cdot \varepsilon^\top\right] \cdot \nabla g(x)^\top + \frac{1}{4}\mathbf{Cov}\left[v_1(x,\varepsilon)\right]. \tag{64}$$

To show Lemma D.3, we only need to show

$$\|R\| \leq 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} + \ell^2 d\overline{\gamma}. \tag{65}$$

To see this, note that

$$\left\|\nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]\right\|$$

$$\leq \|\nabla g(x)\| \cdot \left\|\mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2}v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]\right\| \tag{66a}$$

37

$$\leq \frac{L}{2}\left(\left\|\mathbb{E}\left[\varepsilon\cdot v_1(x,\varepsilon)^\top\right]\right\| + \left\|\mathbb{E}\left[\varepsilon\right]\cdot\mathbb{E}\left[v_1(x,\varepsilon)\right]^\top\right\|\right) \tag{66b}$$

$$\leq \frac{L}{2}\left(\mathbb{E}\left[\left\|\varepsilon\cdot v_1(x,\varepsilon)^\top\right\|\right] + \left\|\mathbb{E}\left[\varepsilon\right]\right\|\cdot\left\|\mathbb{E}\left[v_1(x,\varepsilon)\right]\right\|\right) \tag{66c}$$

$$\leq \frac{L}{2}\left(\mathbb{E}\left[\|\varepsilon\|\cdot\|v_1(x,\varepsilon)\|\right] + \mathbb{E}\left[\|\varepsilon\|\right]\cdot\mathbb{E}\left[\|v_1(x,\varepsilon)\|\right]\right) \tag{66d}$$

$$\leq \frac{L\ell\sqrt{d}}{2}\cdot\left(\mathbb{E}\left[\|\varepsilon\|^3\right] + \mathbb{E}\left[\|\varepsilon\|\right]\cdot\mathbb{E}\left[\|\varepsilon\|^2\right]\right) \tag{66e}$$

$$\leq L\ell d^2\cdot\overline{\gamma}^{\frac{3}{4}}, \tag{66f}$$

where we use the definition of the induced matrix norm in (66a); we use (52) and the triangle inequality in (66b); we use the Jensen's inequality and the definition of the induced matrix norm in (66c) and (66d); we use (56) in (66e); we use the bounds on the moments of $\|\varepsilon\|$ (58), (59), and (60) in (66f).

On the other hand, we know that $\mathbf{Cov}\left[v_1(x,\varepsilon)\right]$ is a positive semi-definite matrix that satisfies

$$\mathbf{Cov}\left[v_1(x,\varepsilon)\right] = \mathbb{E}\left[v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right] - \mathbb{E}\left[v_1(x,\varepsilon)\right]\cdot\mathbb{E}\left[v_1(x,\varepsilon)\right]^\top \preceq \mathbb{E}\left[v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right].$$

Thus, its induced matrix norm can be upper bounded by

$$\left\|\mathbf{Cov}\left[v_1(x,\varepsilon)\right]\right\| \leq \left\|\mathbb{E}\left[v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right]\right\| \leq \mathbb{E}\left[\left\|v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right\|\right] \leq \mathbb{E}\left[\|v_1(x,\varepsilon)\|^2\right].$$

Using the bound of $\|v_1(x,\varepsilon)\|$ in (56) and the 4 th moment bound of $\|\varepsilon\|$, we obtain that

$$\left\|\mathbf{Cov}\left[v_1(x,\varepsilon)\right]\right\| \leq \ell^2 d\mathbb{E}\left[\|\varepsilon\|^4\right] \leq \ell^2 d\overline{\gamma}. \tag{67}$$

Note that the norm of R (Equation (65)) can be upper bounded by the sum of the norms of the 3 separate terms. Thus, by combining the (66) and (67), we see that (65) holds. $\qquad\square$

Lastly, we consider the case when the input of g can be expressed as the sum of a sequence of mutual independent random vectors.

Lemma D.4. *Let $\{X_i\}_{1\leq i\leq N}$ be a sequence of mean-zero random vectors in $\mathbb{R}^d$ that are mutually independent and satisfies $\underline{\delta}I \preceq \mathbf{Cov}\left[X_i\right]$ and $\mathbb{E}\left[\|X_i\|^4\right] \leq \overline{\gamma}$. Let g be a function that satisfies (52) and (53). Then, for any positive integer N, we have*

$$\mathbf{Cov}\left[g\left(\sum_{i=1}^{N}X_i\right)\right] \succeq N\left(\gamma^2\underline{\delta} - 2L\ell d^2\cdot\overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right)I. \tag{68}$$

Proof of Lemma D.4. We use an induction on N to show that (68) holds.

When $N = 1$, (68) holds by setting $x = 0$ and $\varepsilon = X_1$ in Lemma D.3.

Suppose (68) holds for $N - 1$. Then, for N, by the law of total variance, we see that

$$\mathbf{Cov}\left[g\left(\sum_{i=1}^{N}X_i\right)\right] = \mathbf{Cov}\left[\mathbb{E}\left[g\left(\sum_{i=1}^{N}X_i\right)\middle|\sum_{i=1}^{N-1}X_i\right]\right] + \mathbb{E}\left[\mathbf{Cov}\left[g\left(\sum_{i=1}^{N}X_i\right)\middle|\sum_{i=1}^{N-1}X_i\right]\right]. \tag{69}$$

For the first term in (69), we define a new function

$$\bar{g}(x) := \mathbb{E}\left[g(x + X_N)\right].$$

Since the random variables $\{X_i\}_{1\leq i\leq N}$ are mutually independent, we observe that

$$\mathbb{E}\left[g\left(\sum_{i=1}^{N}X_i\right)\middle|\sum_{i=1}^{N-1}X_i\right] = \bar{g}\left[\sum_{i=1}^{N-1}X_i\right].$$

One can verify that if g satisfies the conditions in (52) and (53), then $\bar{g}$ also satisfies the same conditions as g because

$$\|\bar{g}(x) - \bar{g}(x')\| = \|\mathbb{E}\left[g(x + X_N) - g(x' + X_N)\right]\| \leq \mathbb{E}\left[\|g(x + X_N) - g(x' + X_N)\|\right] \leq L\|x - x'\|.$$

On the other hand, we have

$$\langle \bar{g}(x) - \bar{g}(x'), x - x' \rangle = \langle \mathbb{E}\left[g(x + X_N) - g(x' + X_N)\right], x - x' \rangle$$
$$= \mathbb{E}\left[\langle g(x + X_N) - g(x' + X_N), x - x' \rangle\right] \geq \gamma \|x - x'\|^2.$$

For the Hessian upper/lower bounds, because $\nabla^2 \bar{g}_i(x) = \nabla^2 \mathbb{E}\left[g_i(x + X_N)\right] = \mathbb{E}\left[\nabla^2 g_i(x + X_N)\right]$,

$$-\ell I \preceq \bar{g}_i(x) \preceq \ell I.$$

Therefore, by the induction assumption, we see that

$$\mathbf{Cov}\left[\mathbb{E}\left[g\left(\sum_{i=1}^{N} X_i\right) \middle| \sum_{i=1}^{N-1} X_i\right]\right] = \mathbf{Cov}\left[\bar{g}\left[\sum_{i=1}^{N-1} X_i\right]\right]$$
$$\succeq (N-1)\left(\gamma^2 \underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right) I. \qquad (70)$$

For the second term in (69), we note that for any realization x of $\sum_{i=1}^{N-1} X_i$, we have

$$\mathbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right) \middle| \sum_{i=1}^{N-1} X_i = x\right] = \mathbf{Cov}\left[g(x + X_N) \middle| \sum_{i=1}^{N-1} X_i = x\right] = \mathbf{Cov}\left[g(x + X_N)\right]$$
$$\succeq \left(\gamma^2 \underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right) I,$$

where the conditioning can be removed in the second step because the random variables $\{X_i\}_{1 \leq i \leq N}$ are mutually independent, so $g(x + X_N)$ is independent with $\sum_{i=1}^{N-1} X_i$; and we use Lemma D.3 in the last inequality. Therefore, we obtain that

$$\mathbb{E}\left[\mathbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right) \middle| \sum_{i=1}^{N-1} X_i\right]\right] \succeq \left(\gamma^2 \underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right) I. \qquad (71)$$

Substituting (70) and (71) into (69) shows that (68) still holds for N. Thus, we have proved Lemma D.4 by induction. $\qquad \square$

Now we come back to the proof of Theorem D.1. By the assumption, we know the distribution of X is identical to the distribution of $\sum_{i=1}^{N} X_i$, where X_i are i.i.d. random vectors that satisfy $\mathbb{E}\left[\|X_i\|^4\right] \leq C \cdot N^{-2}$. Thus, we have

$$\mathbf{Cov}\left[g(X)\right] = \mathbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right)\right].$$

Note that each X_i satisfies $\mathbf{Cov}\left[X_i\right] = \frac{1}{N}\mathbf{Cov}\left[X\right] \succeq \frac{\mu}{N} I$. Applying Lemma D.4 gives that

$$\mathbf{Cov}\left[g(X)\right] \succeq \left(\mu\gamma^2 - \frac{C^{3/4}}{\sqrt{N}} \cdot 2L\ell d^2 - \frac{C}{N} \cdot \ell^2 d\right) \cdot I.$$

By letting N tend to infinity in the above inequality, we finish the proof of Theorem D.1.

E   Roadmap to Multi-step Prediction under Well-Conditioned Costs

A limitation of Assumption 4.7 in Section 4.1 is that it only allows the prediction $V_t(\theta)$ to depend on the disturbance W_t at time step t. A natural question is whether we can relax the assumption by allowing $V_t(\theta)$ to depend on all future disturbances $W_{t:(T-1)}$. In this section, we present a roadmap towards this generalization and discuss about the potential challenges.

First, we show that the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]$ can be expressed as a function that only depends on the conditional expectations $W_{\tau|t}^\theta$ for all $\tau \geq t$, i.e., there exists a function $\tilde{C}_t^{\pi^\theta}$ that satisfies

$$\tilde{C}_t^{\pi^\theta}(x; W_{t:(T-1)|t}^\theta) = \mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]. \qquad (72)$$

We show (72) by induction on $t = T, T-1, \ldots, 0$. Note that the statement holds for T. Suppose it holds for $t+1$, by (34), we have

$$\bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta)) = \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta; \Xi) \mid I_t(\theta)\right]$$

$$= \mathbb{E}\left[\tilde{C}_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta; W_{(t+1):(T-1)|t+1}^\theta) \,\Big|\, I_t(\theta)\right],$$

where we use the induction assumption in the last equation. Define the random variables $\varepsilon_{t|t}^\theta := W_t - W_{t|t}^\theta$ and $\varepsilon_{\tau|t}^\theta := W_{\tau|(t+1)}^\theta - W_{\tau|t}^\theta$. Using the properties of joint Gaussian distribution, we know that $\varepsilon_{t:(T-1)|t}^\theta$ are independent with $I_t(\theta)$. Therefore,

$$\bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta)) = \mathbb{E}\left[\tilde{C}_{t+1}^{\pi^\theta}(x + \varepsilon_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta + \varepsilon_{(t+1):(T-1)|t}^\theta) \,\Big|\, I_t(\theta)\right]$$

$$= \mathbb{E}_{\varepsilon_{t:(T-1)|t}^\theta}\left[\tilde{C}_{t+1}^{\pi^\theta}(x + \varepsilon_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta + \varepsilon_{(t+1):(T-1)|t}^\theta)\right].$$

Thus, $\bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta))$ can be expressed as a function of x and $W_{(t+1):(T-1)|t}^\theta$, and we denote it as

$$\tilde{\bar{C}}_{t+1}(x; W_{(t+1):(T-1)|t}^\theta) := \bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta)). \tag{73}$$

Therefore, we obtain that

$$\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right] = h_t^x(x) + (h_t^u \square_{(-B_t)} \bar{C}_{t+1}^{\pi^\theta})(A_t x + W_{t|t}^\theta; I_t(\theta))$$

$$= h_t^x(x) + (h_t^u \square_{(-B_t)} \tilde{\bar{C}}_{t+1}^{\pi^\theta})(A_t x + W_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta).$$

Therefore, $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]$ can also be expressed in the form $\tilde{C}_t^{\pi^\theta}(x; W_{t:(T-1)|t}^\theta)$. Thus, we have shown (72) by induction, with (73) as an intermediate result.

Note that the optimal policy is given by

$$\pi_t^\theta(x; I_t(\theta)) := \arg\min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta; I_t(\theta))\right)$$

$$= \arg\min_u \left(h_t^u(u) + \tilde{\bar{C}}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta)\right)$$

$$= u_{(h_t^u \square_{-B_t} \tilde{\bar{C}}_{t+1}^{\pi^\theta})}(A_t x + W_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta).$$

Therefore, by Lemma C.3, we need to establish a covariance lower bound of the gradient

$$\nabla_x \tilde{\bar{C}}_{t+1}^{\pi^\theta}(x + W_{t|t}^\theta; W_{(t+1):(T-1)|t}^\theta)$$

in order to derive a lower bound for the trace of the covariance matrix of $\pi_t^\theta(x; I_t(\theta))$. While this is relatively straightforward when we only have $W_{t|t}^\theta$ because it is added directly with x, it is much more challenging to also consider the covariance caused by $W_{(t+1):(T-1)|t}^\theta$. This is because they affect $\tilde{\bar{C}}_{t+1}^{\pi^\theta}$ through multiple steps of infimal convolutions. Nevertheless, we feel the approach that we describe here is promising if we can derive more properties that are preserved through the infimal convolution operators. We leave this direction as future work.