A Gain Randomization Framework Against Inference Attacks on Control Systems

Ehsan Nekouei¹, Mohammad Pirani¹, Chuanghong Weng¹, and Michael Antönie Van Wyk¹

¹Affiliation not available

May 21, 2024
A Gain Randomization Framework Against Inference Attacks on Control Systems

Ehsan Nekouei, Mohammad Pirani, Chuanghong Weng and Michael Antoine Van Wyk

Abstract—This paper develops a gain randomization framework against inference attacks on feedback control systems where an adversary with access to the states of the system attempts to infer the system model. To prevent the inference attack, the control gain of the system at each time step is randomly selected from a predefined set of control gains. We cast the gain selection problem as an optimal control problem where a gain selection policy at each time step selects a control gain according to a probability distribution such that (i) quadratic control cost is minimized and (ii) the uncertainty level of the adversary about selected control gain is maximized. In our formulation, the gain selection policy is allowed to depend on the entire history of the state measurements and the uncertainty level of the adversary about the control gain is captured by the Kullback-Leibler (KL) divergence between a uniform distribution and the posterior distribution of the feedback gains given the history of the system states. We first derive the backward Bellman optimality equation for the gain selection problem and study the structural properties of the optimal gain selection policy. Our results show that the optimal gain selection policy only depends on the current state of the system, rather than the entire history of the states, which renders the optimal gain selection problem to a non-linear Markov decision process. We next derive a policy gradient theorem for the gain selection problem which provides an expression for the gradient of the objective function of the gain selection problem with respect to the parameter of a stationary (time-invariant) policy. The policy gradient theorem allows us to develop a stochastic gradient descent algorithm for computing an optimal policy. We finally demonstrate the effectiveness of our results using a numerical example.

I. INTRODUCTION

A. Motivation

The role of humans is integrated into control systems in numerous industrial applications. Humans can engage with the control loop either as users or operators. In certain situations, however, these human roles can shift to that of an adversary, aiming to compromise the system, gather information, or cause a specific impact. To counteract such threats, three primary defense mechanisms are typically employed: prevention, detection, and mitigation. Among these, prevention takes precedence as the first line of defense to safeguard the system. One primary approach to preventing attacks in control systems is to enhance the ambiguity of system states and parameters. However, increasing ambiguity can compromise the system’s performance and optimality. In this paper, we explore a meticulously designed randomization of control gains. This approach aims to maximize the adversary’s uncertainty without compromising the control system’s performance.

B. Related Work

As control systems became more integrated with communication platforms, concerns about their security were raised. Security methods in control systems can be classified into three levels prevention algorithms, detection algorithms, and mitigation algorithms, see [1] and references therein. The first layer of defense is to prevent the attack from happening. In many cases, however, it is not possible to prevent all attacks (e.g., if attackers exploit subtle “zero-day” vulnerabilities [2], or rely on insider threats [3]). In those cases, the second layer comes into play which aims to detect and isolate the attack [4], [5]. Beside attack detection, the target system must be resilient enough to withstand the attacks or mitigate the impact of the attack [6]–[17].

Our focus in this paper is on a prevention mechanism and therefore belongs to the first defense layer. When estimation, control or actuation tasks in a NCS are performed by an untrusted party, sharing information might result in the leakage of private information. There is a rich literature pertaining attack prevention in control systems. Cryptography, network coding, model randomization, differential privacy, moving target method, and information-theoretic approaches are among well-known attack prevention mechanisms used for control systems in recent years [18]–[24].

One practical prevention method is altering critical system parameters periodically to prevent attackers from inferring them. A recent approach, called moving target defense, achieves this by introducing random, time-varying parameters into the control system. This limits the attacker’s grasp of the model, making it difficult for them to devise covert attack strategies. Furthermore, the dynamic system nature discourages adaptive adversaries. One key parameter within the system, controlled by the system designer and, hence, can be easily altered, is the control gains. Gain randomization is a method aimed at randomly selecting control gains to prevent attackers from deducing the true set of gains. In references [25], [26], a gain randomization technique is introduced to maximize system ambiguity, quantified by the information accessible to potential attackers. The study of gain randomization strategies within the optimal control framework has not yet been explored, and this is the focus of the current paper.
C. Contributions

In this paper, we propose a security mechanism against inference attacks on control systems. In our setup, an adversary with access to history of the state measurements of a system attempts to infer the system’s model. As a security countermeasure against such an inference attack, we develop a gain randomization framework wherein a gain selection policy randomly selects a feedback gain at a predetermined rate (each time step) from a collection of predefined feedback gains. We formulate the gain selection policy as an optimal control problem where the objective is to minimize a quadratic control cost and maximize the uncertainty level of the adversary about the system model, captured by the Kullback-Leibler divergence between a uniform distribution and the posterior distribution of the feedback gains given the state measurements. We then study the structural properties of the optimal gain selection policy and develop a numerical algorithm for (computing-) executing an optimal policy. Our main contributions can be summarized as follows:

1) We derive the Bellman’s optimality principle for the optimal gain selection policy. Here, our results indicate that the optimal gain selection policy is only a function of the current state of the system, rather than the entire history of the state measurement. As a result, the gain selection policy becomes a non-linear Markov decision process. We also show that the optimal gain selection policy can be computed backward in time by solving a series of convex optimization problems.

2) We derive a policy gradient theorem which provides an expression for the gradient of the objective function of the gain randomization problem with respect to the parameters of a stationary policy.

3) Using the policy gradient theorem, we develop a stochastic gradient descent algorithm for computing an optimal gain selection policy.

The rest of this paper is organized as follows. The next section presents our system model and the gain selection problem. Section III is dedicated to the structural properties of the optimal gain selection policy and its numerical computation. Our numerical results are presented in Section IV, followed by the concluding remarks in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a linear stochastic system in which the system evolves according to

\[ X_{k+1} = AX_k + BU_k + W_k, \]

where \( X_k \) and \( W_k \) are the system’s state and the process noise at time-step \( k \), respectively. The process noise is modeled as a sequence of independent and identically distributed random vectors. The initial state of the system, \( i.e., X_1 \) is assumed to be independent of \( \{W_k\}_k \).

We consider an adversary interested in performing model-based attacks on the system, which requires the knowledge of the system model. Under a time-invariant control law, \( e.g., U_k = KX_k \), the adversary with access to the system states \( \{X_1, \ldots, X_T\} \) can infer the system model. The complete knowledge of the system model allows the adversary to launch powerful model-based attacks on the system. The framework proposed for countering such inference attacks against the system’s model, selects at a predetermined rate the control gain from a predefined set of gains denoted by \( \{K_1, \ldots, K_M\} \) in an apparently random manner. Fig. 1 provides a high level description of this framework. More specifically, the control gain at time \( k \) is selected based on the probability distribution \( \pi_k (\cdot | I_k) \) where \( I_k = \{X_1, \ldots, X_k\} \). Thus, the gain \( K_i \) is selected with probability \( \pi_k (i | I_k) = \Pr (L_k = K_i | I_k) \). Let \( L_k \) denote a random variable which denotes the index of the selected control gain at time \( k \), \( i.e., L_k = i \) if the control gain \( K_i \) is selected at time \( k \). Then, the control input of the system at time \( k \) can be written as \( U_k = K_{L_k} X_k \).

Since the parameters \( A, B \) depend on the physical properties of the system, we assume that they are known by the adversary. We further assume that the adversary has access to the gain collection \( \{K_1, \ldots, K_M\} \). If the adversary has access to \( L_k \), \( i.e., \) the control gain at time \( k \), it will have complete knowledge of the system model. To measure the inference ability of the adversary at time \( k \), we use the posterior probability of \( L_k \) when the adversary has access to \( X_{1:k+1} = \{X_1, \ldots, X_{k+1}\} \), \( i.e., \) \( \Pr (L_k = i | X_{1:k+1}) \). Note a large value of \( \Pr (L_k = i | X_{1:k+1}) \) indicates that the adversary has a large confidence that the feedback gain \( K_i \) is used at time \( k \), whereas a small value of \( \Pr (L_k = i | X_{1:k+1}) \) indicates a low confidence level.

The optimal gain selection policies are obtained by solving the optimization problem (2) where \( R \) and \( Q \) are positive semi-definite matrices, \( \lambda \) is a positive constant, and

\[ D[\text{Uni}_M (j) \| \Pr (L_k = j | X_{1:k+1})] \] is the Kullback-Leibler divergence between a uniform distribution on \( \{K_1, \ldots, K_M\} \) and the posterior distribution \( L_k \Pr (L_k | I_k) \). Therefore, the term \( D[\text{Uni}_M (j) \| \Pr (L_k = j | X_{1:k+1})] \) penalizes the deviation of the posterior distribution \( \Pr (L_k = j | X_{1:k+1}) \) from the uniform distribution. Note that the adversary has maximum uncertainty about \( L_k \) when \( \Pr (L_k = j | X_{1:k+1}) \) is uniformly distributed. Therefore, the optimization problem (2) finds the optimal gain selection policy such that the control cost is minimized while ensuring the cannot reliably infer the control gains.

Fig. 1. The schematic representation of the proposed gain selection scheme.

III. STRUCTURE AND COMPUTATION OF THE OPTIMAL POLICY

In this subsection, we first derive the optimality equations for the gain selection problem, which is used to study the
structural properties of the optimal gain selection policy. We then derive a numerical algorithm for computing an optimal policy.

**Theorem 1:** Let \( \pi^*_k (\cdot | \cdot) \) denote the optimal gain selection policy at time \( k \). Then, the following statements are true.

1. Without loss of optimality, the search space of the optimization problem (2) can be reduced to the collection of the policies of the form \( \{ \pi_k (\cdot | X_k) \}_{k=1}^T \), in which the gain selection policy at time \( k \) only depends on \( X_k \).
2. The optimal gain selection policy \( \pi^*_k (\cdot | \cdot) \) is only a function of \( X_k \).
3. The optimal gain selection policy at time \( k \) can be computed by solving the optimization problem (3) where \( p (x_{k+1} | x_k, L_k = j) \) is the conditional density of \( X_{k+1} \) given \( X_k = x_k, L_k = j \).
4. The optimization problem (2) is a non-linear Markov decision process (MDP) with \( V^*_k (X_k) \) as the optimal value function associated with \( X_k \) and \( V^*_{k+1} (X_{k+1}) = X^T_{k+1} Q X_{k+1} \).
5. The optimization problem in (3) is convex.

**Proof:** See Appendix A.

According to Theorem 1, the optimal gain selection problem is a feedback control problem in the form of an MDP. This is due to the fact that the choice of the feedback gain at time \( k \) affects the evolution of the system state, and consequently the next state \( (X_{k+1}) \) will depend on the gain selection policy at time \( k \). Thus, the impact of the gain selection policy at time \( k \) (\( \pi_k (\cdot | X_k) \)) on the future states must be taken into account in the design of optimal gain selection policy. Different from standard MDPs, the optimal gain selection problem (2) is a non-linear MDP, where the objective function is a non-linear function of the policy due to the KL-divergence term in the objective function. As a result, the optimal gain selection policy will be a randomized function of the current state of the system. However, in a standard MDP, the objective function will be linear in the policy and as a result the optimal policy will be deterministic.

The feasible set of the optimization problem above is the set of all conditional probabilities of the form \( \{ \pi_k (\cdot | X_k) \}_{k=1}^T \) which depend on history of the states. However, based on Theorem 1, it is optimal to reduce the search space for the optimal policy to the policies which only depend on the current state, i.e., \( \{ \pi_k (\cdot | X_k) \}_{k=1}^T \), which is substantially smaller than \( \{ \pi_k (\cdot | X_1, \ldots, X_k) \}_{k=1}^T \). This would significantly reduce the computational complexity of the search for the optimal policy. It also allows the design of stationary (time-invariant) optimal policies, which is impossible when the optimal policy depends on the entire history of state trajectory. Finally, the optimal value function at time \( k \) only depends on \( X_k \) rather than \( X_1, \ldots, X_k \), which is especially helpful when the value-based methods are used to compute an optimal policy.

### A. Policy Gradient Theorem

In this subsection, we develop a policy gradient algorithm for finding the optimal gain selection policy. To this end, let \( \pi_\theta (\cdot | \cdot) \) denote a policy parameterized by \( \theta \). Thus, at time \( k \), a feedback gain is randomly selected according to the conditional probability distribution \( \pi_\theta (\cdot | X_k) \) which depends on the state at time \( k \). We first derive an expression for the gradient of the objective function of the optimization problem (2) with respect to \( \theta \). Using the expression of the gradient, we develop the a policy gradient algorithm for solving (2).

**Theorem 2:** Consider a randomization policy of the form \( \pi_\theta (\cdot | \cdot) \), which is parameterized by \( \theta \). Let \( J_\theta \) denote the value of the objective function of the optimization problem (2) under...
\[ \pi_{\theta}(\cdot | \cdot). \]

Then, the gradient of \( J_\theta \) with respect to \( \theta \) can be written as (4) where \( \Phi_\theta(X_{1:T+1}) \) is give by
\[
\Phi_\theta(X_{1:T+1}) = \frac{1}{M} \sum_{k,j} \log \sum_i \frac{p(X_{k+1} | X_k, L_k = i) \pi_{\theta}(i | X_k)}{p(X_{k+1} | X_k, L_k = j) \pi_{\theta}(j | X_k)}.
\]

**Proof:** See Appendix B.

According to Theorem 2, the gradient of the objective function with respect to \( \theta \) can be expressed as the sum of two terms. The first term in (4) follows from the standard Policy Gradient Theorem of MDPs, where as the second term in (4) is due to the non-linear term introduced in the objective function by the KL-divergence. Note that the second term vanishes when \( \lambda \) is equal to zero and the optimization problem (2) reduces to a standard MDP.

### B. Stochastic Gradient Descent Algorithm

In this subsection, we use Theorem 2 to develop a stochastic gradient algorithm for computing an optimal policy. Let \( s \) denote the iteration index of the algorithm. Also, let \( \pi_{\theta}(\cdot | \cdot) \) and \( \theta \) denote the gain selection policy and the policy parameter at iteration \( s \), respectively. To update the policy parameter at iteration \( s \), we first use the current policy to generate a sequence of the system state and the selected gains over the horizon \( 1, \ldots, T + 1 \), which are denoted as \( \{X_1, X_2^*, \ldots, X_{T+1}^*\} \) and \( \{L_1^*, L_2^*, \ldots, L_T^*\} \). We then update the policy parameter according to (5) where \( \mu_s \) is the steps-size at time-step \( s \). Note that the expectation in (4) is costly to compute numerically. Therefore, in (5), we approximate the expectation using one sample path of system state and the selected gains over the horizon \( 1, \ldots, T + 1 \).

### IV. Numerical Results

In this section, we numerically study the performance of the proposed gain selection problem. To this end, we consider a linear system where the parameters \( A, B \) are given by
\[
A = \begin{bmatrix} 1.98 & -0.98 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0.0625 \\ 0 \end{bmatrix},
\]

respectively. We set \( R = 1 \) and \( Q = 0.0414 \) where \( I_4 \) is a 4-by-4 matrix with all entries equal to 1. The gain selection policy randomly selects one of the following feedback gains
\[
K_1 = [3.27, -2.9], \\
K_2 = [2.95, -2.9], \\
K_3 = [3.27, -3.2],
\]

where \( K_2 \) is the optimal gain minimizing the control performance in (2) when \( \lambda \) is zero, and \( K_1, K_3 \) are perturbed versions of \( K_2 \). In our simulations, the process noise \( W_k \) is modeled as a sequence of independent and identically random vectors with zero mean and covariance matrix \( \begin{bmatrix} 0.05 & 0 \\ 0 & 0.01 \end{bmatrix} \).

Fig. 2 shows the sample trajectories of the selected gain index and the belief of adversary about each control gain when \( \lambda \) is equal to 0.5. Note that when \( \lambda \) is small the optimization problem (2) assigns more weight to the control performance rather than the ambiguity of the adversary about the selected control gain. As a result, the gain selection policy selects \( K_2 \) with a high probability, and as a result, the index of selected gain is often equal to 2, as shown in Fig. 2-(a). When the gain selection policy chooses \( K_2 \) in a relatively long period, the belief of the adversary about \( K_2 \) becomes close to 1, as shown in Fig. 2-(b). As a result, the gain selection policy randomly selects a different gain in order to increase the uncertainty of the adversary about the system model.

Fig. 3 illustrates the sample trajectories of the gain index and the belief of adversary under the optimal policy when \( \lambda \) is equal to 5.5. As \( \lambda \) becomes large, the gain selection problem assigns more weight to the uncertainty level of the adversary. As a result, the gain selection policy becomes more random compared with \( \lambda = 0.5 \), as shown in Fig. 3-(a). According to
allowed us to develop a stochastic gradient descent algorithm policy. We also derived a policy gradient theorem, which problem and studied the structural properties of the optimal formulated the gain selection problem as an optimal control against inference attacks on feedback control systems. We 

Fig. 4 shows the sample trajectories of the gain index and the belief of adversary when $\lambda = 18.5$. According to this figure, the gain selection policy is more random compared with $\lambda = 0.5$ and $\lambda = 5.5$. Thus, the feedback is changed more often. Moreover, the adversary’s belief about the system model is very low in this case due to the higher weight on the uncertainty level of the adversary.

V. CONCLUSIONS

In this paper, we developed a gain randomization framework against inference attacks on feedback control systems. We formulated the gain selection problem as an optimal control problem and studied the structural properties of the optimal policy. We also derived a policy gradient theorem, which allowed us to develop a stochastic gradient descent algorithm for computing an optimal policy.

REFERENCES


APPENDIX A

PROOF OF THEOREM 1

The value function at time $T + 1$ is given by

$$V_{T+1}^*(X_{T+1}) = X_{T+1}^TQX_{T+1}.$$ 

Thus, the optimal value function at time $T + 1$ only depends on $X_{T+1}$. Next, we show by induction that if the optimal value function at time $k + 1$ only depends on $X_{k+1}$, then the value optimal value function at time $k$ only depends on $X_k$. Note
that the optimal value function at time $k$ can be written as (6). Since $X_k$ is in $I_k$, we have
\[
    E \left[ X^T_k Q X_k \mid I_k \right] = X^T_k Q X_k.
\] (7)
Moreover, $E \left[ U^T_k R U_k \mid I_k \right]$ can be written as
\[
    \text{(a)} \ E \left[ U^T_k R U_k \mid I_k \right] = X^T_k R X_k + \lambda U \mathcal{D} \left[ \text{Unif} \left( M \right) \mid \text{Pr} \left( L_k \mid X_{1:k+1} \right) \right] + V^*_{k+1} \left( X_{k+1} \mid I_k \right)
\] (8)
where (a) follows from the definition of $U_k$, (b) follows from the fact that $X_k$ is in $I_k$, (c) and (d) follow from the definition of the conditional expectation and the definition of the policy, respectively.

Using the definition of the KL divergence, we have (9). Using the Bayes’ rule, $\text{Pr} \left( X_k = j \mid X_{1:k+1} \right)$ can be expanded as (10) where (a) and (b) follow from the following Markov chain $X_{1:k-1} \rightarrow (X_k, L_k) \rightarrow X_{k+1}$, and (c) follows from the definition of the policy. Combining (10) and (9), we have (11) where (a) follows from the Bayes’ rule and the Markov chain $X_{1:k-1} \rightarrow (X_k, L_k) \rightarrow X_{k+1}$.

Finally, $E \left[ V^*_{k+1} \left( X_{k+1} \mid I_k \right) \right]$ can be written as (12). Combining (6)-(12), we have (13). Note that the objective function of the optimization above only depends on $X_k$, the gain selection policy and system parameters. Then, the optimal gain selection policy $\pi^*_k (\cdot \mid I_k)$ is only a function of $X_k$, and it does not depend on the entire history of the state measurements. Moreover, the optimal value function at time $k$ only depends on $X_k$.

To show that (2) is a non-linear MDP, note that the dynamic of the system in (7) is Markovian. Using (9), the per stage cost of the (2) at time $k < T$ can be written as
\[
    V^*_k (I_k) = \min_{\pi_k (\cdot \mid I_k)} \ E \left[ U^T_k R U_k + X^T_k Q X_k + \lambda \mathcal{D} \left[ \text{Unif} \left( M \right) \mid \text{Pr} \left( L_k \mid X_{1:k+1} \right) \right] + V^*_{k+1} \left( X_{k+1} \mid I_k \right) \right]
\] (6)
Finally, $E \left[ D \left[ \text{Unif} \left( M \right) \mid \text{Pr} \left( L_k = j \mid X_{1:k+1} \right) \right] \right]$ is convex in the policy due to the negative sign outside the integral in (11).

**Appendix B**

**Proof of Theorem 2**

Note that the objective function of the optimization problem (2) can be written as $O^\theta_1 + O^\theta_2$ where
\[
    O^\theta_1 = E \left[ X^T_{1:T+1} Q X_{1:T+1} + \sum_{k=1}^T x^T_k Q x_k + K^T_k x_k \text{Pr} \left( L_k = l \mid X_{1:k+1} \right) \right],
\]
\[
    O^\theta_2 = E \sum_{k=1}^T D \left[ \text{Unif} \left( M \right) \mid \text{Pr} \left( L_k \mid X_{1:k+1} \right) \right] \right]
\]
We first derive an expression for $\nabla \phi O^\theta_2$ in (15) where $r \left( x_{1:T+1}, l_{1:T} \right) = x^T_{1:T+1} Q x_{1:T+1} + \sum_{k=1}^T x^T_k Q x_k + K^T_k x_k$, and $l_{1:T}$ is a realization of $L_{1:T} = \{L_1, \ldots, L_T\}$. Moreover, $\nabla \phi O^\theta_2$ can be written as (15) where $\phi (x_{1:T+1})$ is given by
\[
    \phi (x_{1:T+1}) = \frac{1}{M} \sum_{k=1}^M \log \sum_{j=1}^M p (X_{k+1} | x_k, L_k = j) \pi_k (j | X_k)
\]
Note that $\nabla \phi \phi (x_{1:T+1})$ can be expanded as (16). We also have (17) Thus, $\nabla \phi O^\theta_2$ can be written as (18).
\[
D[\text{Uni}_M(j) \| \text{Pr}(L_k = j \mid X_{1:k+1})] = \frac{1}{M} \sum_j \log \left( \frac{1/M}{\text{Pr}(L_k = j \mid X_{1:k+1})} \right) \\
= -\log(M) - \frac{1}{M} \sum_j \log(\text{Pr}(L_k = j \mid X_{1:k+1})) 
\]

\begin{align*}
\text{Pr}(L_k = j \mid X_{1:k+1}) &= p(X_{k+1} \mid X_k, L_k = j) \cdot \text{Pr}(L_k = j \mid X_{1:k}) \cdot p(X_{1:k}) \\
&= \frac{p(X_{k+1} \mid X_k, L_k = j) \cdot \text{Pr}(L_k = j \mid X_{1:k}) \cdot p(X_{1:k})}{\sum_i p(X_{k+1} \mid X_k, L_k = i) \cdot \text{Pr}(L_k = i \mid X_{1:k}) \cdot p(X_{1:k})} \\
&= \frac{p(X_{k+1} \mid X_k, L_k = j) \cdot \pi_k(j \mid I_k)}{\sum_i p(X_{k+1} \mid X_k, L_k = i) \cdot \pi_k(i \mid I_k)} 
\end{align*}

\begin{align*}
E[D[\text{Uni}_M(j) \| \text{Pr}(L_k = j \mid X_{1:k+1})] \mid I_k] \\
= -\log(M) - \frac{1}{M} \sum_j E \left[ \log \frac{p(X_{k+1} \mid X_k, L_k = j) \cdot \pi_k(j \mid I_k)}{\sum_i p(X_{k+1} \mid X_k, L_k = i) \cdot \pi_k(i \mid I_k)} \right] \\
= -\log(M) - \frac{1}{M} \sum_j \int \log \frac{p(x_{k+1} \mid X_k, L_k = j) \cdot \pi_k(j \mid I_k)}{\sum_i p(x_{k+1} \mid X_k, L_k = i) \cdot \pi_k(i \mid I_k)} p(x_{k+1} \mid I_k) \, dx_{k+1} \\
\equiv -\log(M) - \frac{1}{M} \sum_{j,l} \int \log \frac{p(x_{k+1} \mid X_k, L_k = j) \cdot \pi_k(j \mid I_k)}{\sum_i p(x_{k+1} \mid X_k, L_k = i) \cdot \pi_k(i \mid I_k)} p(x_{k+1} \mid X_k, L_k = l) \cdot \pi_k(l \mid I_k) \, dx_{k+1} 
\end{align*}

\begin{align*}
E[V_{k+1}^* (X_{k+1}) \mid I_k] &= \int V_{k+1}^* (x_{k+1}) \cdot p(x_{k+1} \mid I_k) \, dx_{k+1} \\
&= \sum_l \int V_{k+1}^* (x_{k+1}) \cdot p(x_{k+1} \mid X_k, L_k = l) \cdot \pi_k(l \mid I_k) \, dx_{k+1} 
\end{align*}

\begin{align*}
V_k^* (I_k) &= \min_{\pi_k(\cdot \mid I_k)} X_k^T Q X_k + \sum_l X_k^T K_i^T R K_i X_k \pi_k(l \mid I_k) + \log \frac{1}{M} \\
&+ \frac{1}{M} \sum_{j,l} \int V_{k+1}^* (x_{k+1}) - \lambda \log \frac{p(x_{k+1} \mid X_k, L_k = j) \cdot \pi_k(j \mid I_k)}{\sum_i p(x_{k+1} \mid X_k, L_k = i) \cdot \pi_k(i \mid I_k)} \cdot p(x_{k+1} \mid X_k, L_k = l) \cdot \pi_k(l \mid I_k) \, dx_{k+1} 
\end{align*}

\begin{align*}
\nabla_\theta Q_1^\theta &= \nabla_\theta \sum_{l_1:T} \int r(x_{1:T+1}, l_{1:T}) \cdot p_1(x_1) \prod_{k=1}^T p(x_{k+1} \mid x_k, L_k = l_k) \cdot \pi_\theta(l_k \mid x_k) \, dx_{1:T+1} \\
&= \sum_{l_1:T} \int r(x_{1:T+1}, l_{1:T}) \left( \nabla_\theta \log \prod_{k=1}^T \pi_\theta(l_k \mid x_k) \right) \cdot p_1(x_1) \prod_{k=1}^T p(x_{k+1} \mid x_k, L_k = l_k) \cdot \pi_\theta(l_k \mid x_k) \, dx_{1:T+1} \\
&= \sum_{l_1:T} \int r(x_{1:T+1}, l_{1:T}) \left( \sum_{k=1}^T \nabla_\theta \log \pi_\theta(l_k \mid x_k) \right) \cdot p_1(x_1) \prod_{k=1}^T p(x_{k+1} \mid x_k, L_k = l_k) \cdot \pi_\theta(l_k \mid x_k) \, dx_{1:T+1} \\
&= \mathbb{E} \left[ \left( X_{T+1}^T Q X_{T+1} + \sum_{k=1}^T X_k^T Q X_k + U_k^T R U_k \right) \sum_{k=1}^T \nabla_\theta \log \pi_\theta(l_k \mid x_k) \right] 
\end{align*}
\[ \nabla_\theta O_2^\theta = \nabla_\theta \sum_{l_{1:T}} \Phi_\theta (x_{1:T+1}) p_1(x_1) \prod_{k=1}^T p(x_{k+1} | x_k, L_k = l_k) \pi_\theta (l_k | x_k) dx_{1:T+1} \]

\[ = \sum_{l_{1:T}} \int \nabla_\theta \Phi_\theta (x_{1:T+1}) p_1(x_1) \prod_{k=1}^T p(x_{k+1} | x_k, L_k = l_k) \pi_\theta (l_k | x_k) dx_{1:T+1} \]

\[ + \sum_{l_{1:T}} \Phi_\theta (x_{1:T+1}) \left( \nabla_\theta \log \prod_{k=1}^T \pi_\theta (l_k | x_k) \right) p_1(x_1) \prod_{k=1}^T p(x_{k+1} | x_k, L_k = l_k) \pi_\theta (l_k | x_k) dx_{1:T+1} \]

\[ = E \left[ \nabla_\theta \Phi_\theta (X_{1:T+1}) + \Phi_\theta (X_{1:T+1}) \left( \nabla_\theta \log \prod_{k=1}^T \pi_\theta (L_k | X_k) \right) \right] \quad (15) \]

\[ \nabla_\theta \Phi_\theta (x_{1:T+1}) = \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^M \sum_{k=1}^T \frac{p(X_{k+1} | X_k, L_k = i) \nabla_\theta \pi_\theta (i | X_k)}{\sum_i p(X_{k+1} | X_k, L_k = i) \pi_\theta (i | X_k)} \frac{\nabla_\theta \pi_\theta (j | X_k)}{\pi_\theta (j | X_k)} \]

\[ = \sum_{k=1}^T \sum_{i=1}^M \frac{p(X_{k+1} | X_k, L_k = i) \nabla_\theta \pi_\theta (i | X_k)}{\sum_i p(X_{k+1} | X_k, L_k = i) \pi_\theta (i | X_k)} \sum_{j=1}^M \frac{\nabla_\theta \pi_\theta (j | X_k)}{\pi_\theta (j | X_k)} \]

\[ = \nabla_\theta \int \sum_{i} p(x_{k+1} | x_k, L_k = i) \pi_\theta (i | x_k) p(x_k) dx_{k+1} dx_k \]

\[ = \nabla_\theta \int p_\theta (x_{k+1}, x_k) dx_{k+1} dx_k \]

\[ = 0 \quad (16) \]

\[ \nabla_\theta O_2^\theta = E \left[ \Phi_\theta (X_{1:T+1}) \sum_{k=1}^T \nabla_\theta \log \pi_\theta (L_k | X_k) \right] - \frac{1}{M} \sum_{k=1}^T \sum_{j=1}^M \nabla_\theta \log \pi_\theta (j | X_k) \quad (17) \]