A TEMPLATE FOR ARXIV STYLE [*]

Author1, Author2
Affiliation
Univ
City
`{Author1, Author2}email@email`

Author3
Affiliation
Univ
City
`email@email`

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

***K*eywords** First keyword · Second keyword · More

1 Introduction

In this article, we consider a continuous time stochastic differential equation defined as follows.

$$dx_t = f^*(x(s), u(s))dt + g^*(x(s))dW_t = F^*(x(s), u(s))\theta dt + F^*(x(s), u(s))\hat{\theta} dW_t \tag{1}$$

where the state vector $x_t \in \mathcal{X} \subset R^{d_x}$, $u(s)$, control input$u : [0, \infty) \to \mathbb{R}^{d_u}$, $f^*(\cdot, \cdot)$ and $f^*(\cdot, \cdot)$ are unkown functions we would learn from sampling over episode $n = 1 \cdots N$. Moreover, we consider the state feedback controllers denoted as a policy $\pi : \mathcal{X} \to \mathcal{U} \in R^{d_u}$, that is $\pi(x_t) = u_t$. In this work, we aim to identify the optimal policy with respect to the minimum of a given cost function $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}$. Specifically, our focus is to solve the following control problem over the policy space Π.

$$\pi^* \triangleq \arg\min_{\pi \in \Pi} C(\pi, f^*, g^*) = \arg\min_{\pi \in \Pi} E[\int_0^T c(x, \pi(x, t))\, dt] \tag{2}$$

$$\text{s.t.} \qquad \dot{x} = f^*(x(s), u(s))dt + g^*(x(s))dW_t, \quad x(0) = x_{t_0}. \tag{3}$$

In an ideally continuous time setting, one can observe the state x_t at each time slot $t \in [0, T]$. However, in some scenario, the sampling might be expensive, which promotes us to design the efficent sampling way to utilize the limited data. Therefore, we formally define a measurement selection strategy below.

Definition 1.1 (Measurement Selection Strategy)**.** A measurement selection strategy S is a sequence of sets $(S_n)_{n \geq 1}$, such that S_n contains m_n points at which measurements are taken, i.e., $S_n \subset [0, T], \quad |S_n| = m_n$.

During episode n, given a policy π_n and a Measurement Selection Strategy (MSS) S_n, we collect a dataset $D_n \sim (\pi_n, S_n)$. The dataset is defined as:

$$D_n \triangleq \{(z_n(t_{n,i}), \dot{y}_n(t_{n,i})) \mid t_{n,i} \in S_n, i \in \{1, \ldots, m_n\}\}$$

[*] *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

where

$$z_n(t_{n,i}) \triangleq (x_n(t_{n,i}), \pi_n(x_n(t_{n,i}))), \quad \dot{y}_n(t_{n,i}) \triangleq \dot{x}_n(t_{n,i}) + \epsilon_{n,i}.$$

Here, $x_n(t)$ and $\dot{x}_n(t)$ are the state and state derivative in episode n, and $\epsilon_{n,i}$ is i.i.d, σ-sub-Gaussian noise of the state derivative observations. Note, even though in practice only the state $x(t)$ might be observable, one can estimate its derivative $\dot{x}(t)$ (e.g., using finite differences, interpolation methods, etc. See Cullum (1971); Knowles and Wallace (1995); Chartrand (2011); Knowles and Renka (2014); Wagner et al. (2018); Treven et al. (2021)). We capture the noise in our measurements and/or estimation of $\dot{x}(t)$ with $\epsilon_{i,n}$.

In summary, at each episode n, we deploy a policy π_n for a horizon of T, observe the system according to a proposed MSS S_n, and learn the dynamics f^*. By deploying π_n instead of the optimal policy π^*, we incur a regret,

$$r_n(S) \triangleq C(\pi_n, f^*) - C(\pi^*, f^*).$$

Note that the policy π_n depends on the data $D_{1:n-1} = \bigcup_{i<n} D_i$ and hence implicitly on the MSS S.

1.1 Assumptions

Similar to those works on dicrete-time setting, here we make some assumption about continuity, mainly on dynamic, policies and costs.

Assumption 1.2 (Lipschitz continuity)**.** Given any norm $\|\cdot\|$, we assume that the system drifting f^*, diffusion g^* and cost c are L_f L_g and L_c-Lipschitz continuous, respectively, with respect to the induced metric. Moreover, we define Π to be the policy class of L_π-Lipschitz continuous policy functions and $\mathcal{F}$ a class of L_f Lipschitz continuous dynamics functions with respect to the induced metric.

In this article, we learn the model f^* and g^* from the collected data at each episode. For a given state-action pair $z = (x, u)$, our learned model predicts a mean estimate $\mu_n(z)$ and quantifies our epistemic uncertainty $\sigma_n(z)$ about the function f^*.

Assumption 1.3 (Well-calibration)**.** We assume that our learned model is an all-time well-calibrated statistical model of f^*. We further assume that the standard deviation functions $(\sigma_n(\cdot))_{n\geq 0}$ are L_σ-Lipschitz continuous.

Assumption 1.4. Let $\pi(x,t)$ denote the probability density function associated with the solution $x(t)$ of the stochastic differential equation 1. Assume that the second-order derivative of the logarithm of the probability density function $\log p(x,t)$ is bounded. Specifically, there exists a constant $C > 0$ such that

$$\left| \frac{\partial^2}{\partial x^2} \log p(x,t) \right| \leq L_\pi \quad \text{for all } x \in \mathbb{R} \text{ and for all } t \geq 0.$$

Assumption 1.5. There exists an unknown parameter vectors $\theta^* \in \mathbb{R}^d$ and $\Theta \in \mathbb{R}^{d^2}$ such that for any action (feature vector) $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$, function $f^*(x,u)$ and $g^*(x,u)$ are linear functiosn of $\mathbf{x}_t$:

$$f(x,u) = F(x,u)\theta$$

and

$$\mathrm{vec}(g(x,u)g(x,u)^\top) = \hat{F}(x,u)\Theta$$

where $\hat{F}(x,u) = F(x,u) \otimes F(x,u)$ and the function $F(x,u)$ satisfies the Lipschitz condition with constant L_F; that is, there exists a constant $L > 0$ such that for all x_1, x_2 in the domain of F,

$$\|F(x_1,u) - F(x_2,u)\|_2 \leq L_F \|x_1 - x_2\|_2$$

2 Proof

Lemma 2.1 (Gronwall's Inequality)**.** *Let $u(t)$ be a non-negative, continuous function on the interval $[a,b]$. Suppose that*

$$u(t) \leq K + \int_a^t \gamma(s) u(s)\, ds$$

for all $t \in [a,b]$, where K is a non-negative constant and $\gamma(s)$ is a non-negative, continuous function on $[a,b]$. Then,

$$u(t) \leq K \exp\left(\int_a^t \gamma(s)\, ds \right)$$

for all $t \in [a,b]$.

Lemma 2.2. *Suppose* $\forall t \in [0\ T], f(t) \in R^d$*, then their holds*

$$\int_0^T \|f(t)\|_2^2 dt \le d^2 \int_0^T \left\| \int_0^t f(s) ds \right\|_2^2 \tag{4}$$

and

$$\left\| \int_0^T f(t) f(t)^\top dt \right\|_2 \le d \int_0^T \left\| \int_0^t f(s) ds \right\|_2^2 \tag{5}$$

Proof. We start from the fact that

$$\int_0^T \|f(t)\|_2^2 dt = tr(\int_0^T f(t) f(t)^\top dt) \le d \left\| \int_0^T f(t) f(t)^\top dt \right\|_2 \tag{6}$$

Then we could construct

$$\left\| \int_0^T f(t) f(t)^\top dt \right\|_2 \le \left\| \int f(t) f(t)^\top dt + \left(\int_0^T f(t) dt \right) \left(\int f(t) dt \right)^\top \right\|_2 \tag{7}$$

$$= \left\| \int_0^T \int_0^t f(s) f(t)^\top ds dt + \int_0^T \int_0^t f(t) f(s)^\top ds dt \right\|_2 \tag{8}$$

$$\le 2 \left\| \int_0^T \int_0^t f(s) f(t)^\top ds dt \right\|_2 \tag{9}$$

$$\le \sqrt{\int_0^T \|f(t)\|_2^2 dt} \sqrt{\int_0^T \left\| \int_0^t f(s) ds \right\|_2^2} \tag{10}$$

Therefore, we could construct

$$\int_0^T \|f(t)\|_2^2 dt \le d^2 \int_0^T \left\| \int_0^t f(s) ds \right\|_2^2 \tag{11}$$

and

$$\left\| \int_0^T f(t) f(t)^\top dt \right\|_2 \le d \int_0^T \left\| \int_0^t f(s) ds \right\|_2^2 \tag{12}$$

□

$$\pi^*(x|u_n, t) \tag{13}$$

Theorem 2.3. *Consider two stochastic processes* $\hat{x}_n(s)$ *and* $x_n(s)$ *defined by the following dynamics:*

$$\dot{x}_n(s) = f^*(x_n(s), u_n(x_n(s))), ds + g^*(x_n(s), u_n(x_n(s))), dw_t, \tag{14}$$

$$\dot{\hat{x}}_n(s) = f_n(\hat{x}_n(s), u_n(\hat{x}_n(s))), ds + g_n(\hat{x}_n(s), u_n(\hat{x}_n(s))), dwt. \tag{15}$$

Let the probability density functions corresponding to these dynamics be denoted as $\pi^*(x_n(s), u_n(x_n(s)), t)$ *and* $\hat{\pi}_n(\hat{x}_n(s), u_n(\hat{x}_n(s)), t)$*, respectively. Then, the difference in the cost functions* $C(u_n, T, \pi^*)$ *and* $C(u_n, T, \hat{\pi}_n)$ *is bounded by:*

$$|C(u_n, T, \pi^*) - C(u_n, T, \hat{\pi}_n)| \le L_c e^{(2L_f + 1 + 4L_g)t} (\sigma_{n,t,1} + \sigma_{n,t,2}) \tag{16}$$

where $\sigma_{n,t,1}(\pi^*) = \int_0^t \int \left\| g^*(x_1, u_n(x_1)) g^*(x_1, u_n(x_1))^\top - g_n(x_1, u_n(x_1)) g_n(x_1, u_n(x_1))^\top \right\|_F \pi^*(x_1, u_n(x_1), s) dx ds$ *and* $\sigma_{n,t,2}(\pi^*) = \int_0^t \int \|f^*(x, u_n(x)) - f_n(x, u_n(x))\|_2^2 \pi^*(x, u_n(x), s) dx ds$.

Proof. We start by

$$|C(u_n, T, \pi^*) - C(u_n, T, \hat{\pi}_n)| = \left| \int_0^T \int c(x, u_n(x))(\pi^*(x|u_n, t) - \hat{\pi}_n(x, u_n(x), t))dxdt \right| \tag{17}$$

$$\leq L_c \int_0^T \int \int \|x_1 - x_2\|_2 \pi(x_1, x_2, u_n(), t)dx_1 dx_2 \tag{18}$$

where the inequality 18 is obtained by Lipschitz condition of the cost function c which is stated in Assumption 25.

Here, $\pi(x_1, x_2, u_n(), t)$ denotes the joint distribution of the two dynamics. Let us consider $\hat{x}_n(s)$ and $x_n(s)\|_2^2$ is generated by the following two dynamics separately.

$$\dot{x}_n(s) = f^*(x_n(s), u_n(x_n(s)))ds + g^*(x_n(s), u_n(\hat{x}_n(s)))dw_s \tag{19}$$

$$\dot{\hat{x}}_n(s) = f_n(\hat{x}_n(s), u_n(\hat{x}_n(s)))ds + g_n(\hat{x}_n(s), u_n(\hat{x}_n(s)))dw_s \tag{20}$$

Note that for function $h_n(s) = \|x_n(s) - \hat{x}_n(s)\|_2^2$ and $\hat{h}(x_1, x_2) = \|x_1^2 - x_2\|_2^2$, due to Ito lemma there holds

$$\dot{h}_n(s) = 2(x_n(s) - \hat{x}_n(s))^\top (f^*(x_n(s), u_n(x_n(s))) - f_n(\hat{x}_n(s), u_n(\hat{x}_n(s)))) \tag{21}$$

$$+ 2tr((g^*(x_n(s), u_n(x_n(s))) - g_n(\hat{x}_n(s), u_n(\hat{x}_n(s))))(g^*(x_n(s), u_n(x_n(s))) - g_n(\hat{x}_n(s), u_n(\hat{x}_n(s))))^\top) \tag{22}$$

$$+ 2(x_n(s) - \hat{x}_n(s))^\top (g^*(x_n(s), u_n(x_n(s))) - g_n(\hat{x}_n(s), u_n(\hat{x}_n(s))))dw_t \tag{23}$$

For simplification, we denote $\pi(x_1, x_2, u_n(), s)$ as the joint distribution of $x_n(s$ and $\hat{x}_n(s)$. Then we have

$$\int \dot{\hat{h}}(x_1, x_2)\pi(x_1, x_2, u_n(), s)dx = \int 2(x_1 - x_2)^\top (f^*(x_1, u_n(x_1)) - f_n(x_2, u_n(x_2)))\pi(x_1, x_2, u_n(), s)dx \tag{24}$$

$$+ 2 \int tr((g^*(x_1, u_n(x_1)) - g_n(x_2, u_n(x_2)))(g^*(x_1, u_n(x_1)) - g_n(x_2, u_n(x_2)))^\top)\pi(x_1, x_2, u_n(), s)dx \tag{25}$$

$$\leq 2\|x_1 - x_2\|_2(\|f^*(x_1, u_n(x_1)) - f_n(x_1, u_n(x_1))\|_2 + \|f_n(x_1, u_n(x_1)) - f_n(x_2, u_n(x_2))\|_2)\pi(x_1, x_2, u_n(), s)dx \tag{26}$$

$$+ 4 \int tr((g^*(x_1, u_n(x_1)) - g_n(x_1, u_n(x_1)))(g^*(x_1, u_n(x_1)) - g_n(x_1, u_n(x_1)))^\top)\pi(x_1, x_2, u_n(), s)dx \tag{27}$$

$$+ 4 \int tr((g_n(x_1, u_n(x_1)) - g_n(x_2, u_n(x_2)))(g_n(x_1, u_n(x_1)) - g_n(x_2, u_n(x_2)))^\top)\pi(x_1, x_2, u_n(), s)dx \tag{28}$$

$$\leq (2L_f + 1 + 4L_g) \int \|x_1 - x_2\|_2^2 \pi(x_1, x_2, u_n(), s)dx + \int \|f^*(x_1, u_n(x_1)) - f_n(x_1, u_n(x_1))\|_2^2 \pi(x_1, x_2, u_n(), s)dx \tag{29}$$

$$+ 4 \int tr((g^*(x_1, u_n(x_1)) - g_n(x_1, u_n(x_1)))(g^*(x_1, u_n(x_1)) - g_n(x_1, u_n(x_1)))^\top)\pi(x_1, x_2, u_n(), s)dx \tag{30}$$

$$\leq (2L_f + 1 + 4L_g) \int \|x_1 - x_2\|_2^2 \pi(x_1, x_2, u_n(), s)dx + \int \|f^*(x_1, u_n(x_1)) - f_n(x_1, u_n(x_1))\|_2^2 \pi^*(x_1, u_n(x_1), s)dx \tag{31}$$

$$+ 4 \int \|g^*(x_1, u_n(x_1))g^*(x_1, u_n(x_1))^\top - g_n(x_1, u_n(x_1))g_n(x_1, u_n(x_1))^\top\|_F \pi^*(x_1, u_n(x_1), s)dx \tag{32}$$

In the derivation provided, inequality (25) follows from the definition of the Wiener process. Inequality (28) is derived using the Cauchy–Schwarz inequality, and inequality (30) results from the Lipschitz conditions applied to the functions f_n and g_n (as stated in Assumption).

Then using Grönwall's inequality, we have

$$\int \hat{h}(x_1, x_2)\pi(x_1, x_2, u_n(), t)dx \leq e^{(2L_f+1+4L_g)t} \int_0^t \int \|f^*(x_1, u_n(x_1)) - f_n(x_1, u_n(x_1))\|_2^2 \pi^*(x_1, u_n(x_1), s)dxds \tag{33}$$

$$+ e^{(2L_f+1+4L_g)t} \int_0^t \int \int \|g^*(x_1, u_n(x_1))g^*(x_1, u_n(x_1))^\top - g_n(x_1, u_n(x_1))g_n(x_1, u_n(x_1))^\top\|_F \pi^*(x_1, u_n(x_1), s)dxds \tag{34}$$

$$ \tag{35} $$

Then in accordance with inequality 18, we could construct

$$|C(u_n,T,\pi^*) - C(u_n,T,\hat{\pi}_n)| \leq L_c e^{(2L_f+1+4L_g)t}\int_0^t \int \|f^*(x_1,u_n(x_1)) - f_n(x_1,u_n(x_1))\|_2^2 \pi^*(x_1,u_n(x_1),s)dxds \tag{36}$$

$$+ L_c e^{(2L_f+1+4L_g)t}\int_0^t \int\int \|g^*(x_1,u_n(x_1))g^*(x_1,u_n(x_1))^\top - g_n(x_1,u_n(x_1))g_n(x_1,u_n(x_1))^\top\|_F \pi^*(x_1,u_n(x_1),s)dxds \tag{37}$$

□

Theorem 2.4. *Consider two stochastic processes $\hat{x}_n(s)$ and $x_n(s)$ defined by the following dynamics:*

$$\dot{x}_n(s) = f^*(x_n(s),u_n(x_n(s)))ds + g^*(x_n(s),u_n(x_n(s)))dw_t, \tag{38}$$

$$\dot{\hat{x}}_n(s) = f_n(\hat{x}_n(s),u_n(\hat{x}_n(s)))ds + g_n(\hat{x}_n(s),u_n(\hat{x}_n(s)))dwt. \tag{39}$$

Let the probability density functions corresponding to these dynamics be denoted as $\pi^*(x_n(s),u_n(x_n(s)),t)$ and $\hat{\pi}_n(\hat{x}_n(s),u_n(\hat{x}_n(s)),t)$, respectively. Then, the difference in the cost functions $C(u_n,T,\pi^*)$ and $C(u_n,T,\hat{\pi}_n)$ is bounded by:

$$|C(\pi^*,u_n,t) - C(\pi)_n,u_n,t)| \leq c_{\max}\sqrt{\frac{d^{1.5}T^2}{8\lambda_{\min}(g_n(x,u_n(x)))^2}(\sigma_{n,t,1}(\pi^*)+\sigma_{n,t,1}(\pi_n)) + T^3 d^{2.5}L_\pi(\sigma_{n,t,2}(\pi^*)+\sigma_{n,t,2}(\pi_n))} \tag{40}$$

Proof. Here, we would consider the difference of expectation of the cost function generated by estimated dynamic and real dynamic. For simplification of expression, we denote it as

$$c(u) - \hat{c}(u) = \int c(x,u_n(x))(\pi(x,u_n(x),t) - \hat{\pi}_n(x,u_n(x),t))dx \tag{41}$$

where $\pi(x,u_n(x),t)$ is the probability density function of the real dynamic and $\hat{\pi}_n(x,u_n(x),t)$ is the is the probability density function of the estimated dynamic. Then we have

$$c(u) - \hat{c}(u) \leq \int c(x,u_n(x))(\sqrt{\pi(x,u_n(x),t)} + \sqrt{\hat{\pi}(x,u_n(x),t)})(\pi(x,u_n(x),t)^{\frac{1}{2}} - \hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}})dx \tag{42}$$

Therefore, the main challenge here is to analyze $\sqrt{\int(\pi(x,u_n(x),t)^{\frac{1}{2}} - \hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}})^2 dx}$ Then we could construct

$$\int(\pi(x,u_n(x),t)^{\frac{1}{2}}dx - \hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}})^2 = \int(\pi(x,u_n(x),t) + \hat{\pi}_n(x,u_n(x),t) - 2\pi(x,u_n(x),t)^{\frac{1}{2}}\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}})dx \tag{43}$$

$$= \int \pi(x,u_n(x),t) + \hat{\pi}_n(x,u_n(x),t)dx - 2\int \pi(x,u_n(x),t)^{\frac{1}{2}}\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}dx \tag{44}$$

$$\leq 2 - \int \pi(x,u_n(x),t)^{\frac{1}{2}}\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}dx \tag{45}$$

To facilitate our analysis, we introudce $h(t) = -\int \pi(x,u_n(x),t)^{\frac{1}{2}}\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}dx$ and the operator A denoted as for any matrix $A \in R^{m\times n}$ $H(A) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}$ where a_{ij} is the i-th row and j-th column entry of A, Then we have

$$\dot{h}(t) = \int -\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\dot{\pi}^*(x,u_n(x),t) - \frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\dot{\hat{\pi}}(x,u_n(x),t)dx \tag{46}$$

$$= -\int \frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}(-\nabla(f^*(x,u_n(x))\pi^*(x|u_n,t)) + \Delta(g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)))dx \tag{47}$$

$$-\int \frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}(-\nabla(f_n(x,u_n(x))\hat{\pi}_n(x,u_n(x),t)) + \Delta(g_n(x,u_n(x))(x,u_n(x))g_n(x,u_n(x))^\top\hat{\pi}_n(x,u_n(x),t)))dx \tag{48}$$

$$= \underbrace{\int tr\left(\frac{\partial\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x} f^*(x,u_n(x))^\top\right)\pi^*(x|u_n,t)dx}_{\hat{\sigma}_{n,1}(t)} \tag{49}$$

$$-\underbrace{\int tr\left(\frac{\partial^2\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}(g^*(x,u_n(x))g^*(x,u_n(x))^\top\right)\pi^*(x|u_n,t))dx}_{\hat{\sigma}_{n,2}(t)} \tag{50}$$

$$+\underbrace{\int tr\left(\frac{\partial\left(\frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x} f_n(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)dx}_{\hat{\sigma}_{n,3}(t)} \tag{51}$$

$$-\underbrace{\int tr\left(\frac{\partial^2\left(\frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x^2}(g_n(x,u_n(x))g_n(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t))dx}_{\hat{\sigma}_{n,4}(t)} \tag{52}$$

$$\tag{53}$$

where the equality 46 is due to the definition of $h(t)$, the equality 48 is due to the Fokker–Planck equation. For the term $\hat{\sigma}_{n,1}(t)$ and $\hat{\sigma}_{n,3}(t)$, we have

$$\hat{\sigma}_{n,1}(t) = \int tr\left(\frac{\partial\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x} f^*(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}\frac{\pi^*(x|u_n,t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}dx \tag{54}$$

$$= \int tr\left(\frac{\partial\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x} f^*(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{55}$$

and

$$\hat{\sigma}_{n,3}(t) = \int tr\left(\frac{\partial\left(\frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x} f_n(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}} \tag{56}$$

$$= \int tr\left(\frac{\partial\left(\frac{\pi^*(x,u_n(x),t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x} f_n(x,u_n(x))^\top\right)\pi(x,u_n(x),t)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi(x,u_n(x),t)^{\frac{1}{2}}dx \tag{57}$$

where inequalities (55) and (57) follow from the property that $\nabla(\log(f(x))) = \frac{\nabla(f(x))}{f(x)}$, which is derived by applying the chain rule to the logarithm function.

Then we could construct

$$\hat{\sigma}_{n,1}(t)+\hat{\sigma}_{n,2}(t)=\int tr\left(\frac{\partial\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x}(f^*(x,u_n(x))-f_n(x,u_n(x)))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{58}$$

For the term $\hat{\sigma}_{n,2}(t)$ and $\hat{\sigma}_{n,4}(t)$, we construct

$$\hat{\sigma}_{n,2}(t)=-\int tr\left(\frac{\partial^2\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}(g^*(x,u_n(x))g^*(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}dx \tag{59}$$

$$=-\int tr\left(\left(\frac{\partial\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x}\right)\left(\frac{\partial\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x}\right)^\top(g^*(x,u_n(x))g^*(x,u_n(x))^\top\right) \tag{60}$$

$$\times\,\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{61}$$

$$-\int tr\left(\frac{\partial^2\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}(g^*(x,u_n(x))g^*(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{62}$$

and

$$\hat{\sigma}_{n,4}(t)=-\int tr\left(\frac{\partial^2\left(\frac{\pi^*(x|u_n,t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x^2}g^*(x,u_n(x))g^*(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}dx \tag{63}$$

$$=-\int tr\left(\left(\frac{\partial\log\left(\frac{\pi^*(x|u_n,t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x}\right)\left(\frac{\partial\log\left(\frac{\pi^*(x|u_n,t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)}{\partial x}\right)^\top g_n(x,u_n(x))g_n(x,u_n(x))^\top\right) \tag{64}$$

$$\times\,\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{65}$$

$$-\int tr\left(\frac{\partial^2\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}(g_n(x,u_n(x))g_n(x,u_n(x))^\top\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx \tag{66}$$

where the inequalities (62) and (66) follow from the relation $\frac{\partial^2\log(f(x))}{\partial x^2}=\frac{1}{f(x)}\frac{\partial^2(f(x))}{\partial x^2}-\left(\frac{\partial\log(f(x))}{\partial x}\right)\left(\frac{\partial\log(f(x))}{\partial x}\right)^\top$ and $\nabla(\log(f(x)))=\frac{\nabla(f(x))}{f(x)}$.

Then we have

$$\hat{\sigma}_{n,2}(t)+\hat{\sigma}_{n,4}(t)= \tag{67}$$

$$-\int tr\left(\left(\left(\frac{\partial\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x}\right)\left(\frac{\partial\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x}\right)^{\top}\left(g^*(x,u_n(x))g^*(x,u_n(x))^{\top}+g_n(x,u_n(x))g_n(x,u_n(x))^{\top}\right)\right)\tag{68}$$

$$\times\,\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx\tag{69}$$

$$-\int tr\left(\frac{\partial^2\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}g^*(x,u_n(x))g^*(x,u_n(x))^{\top}\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx\tag{70}$$

$$+\int tr\left(\frac{\partial^2\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}(g_n(x,u_n(x))g_n(x,u_n(x))^{\top}\right)\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}dx\tag{71}$$

where equalities 68 and 71 due to the fact that $\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)=-\partial\log\left(\frac{\pi^*(x|u_n,t)^{\frac{1}{2}}}{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}\right)$

Then we could construct

$$\hat{\sigma}_{n,1}(t)+\hat{\sigma}_{n,2}(t)+\hat{\sigma}_{n,3}(t)+\hat{\sigma}_{n,4}(t)\le\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}\pi^*(x|u_n,t)^{\frac{1}{2}}\sqrt{d}\times\tag{72}$$

$$(\frac{1}{4}\left(f^*(x,u_n(x))-f_n(x,u_n(x))\right)^{\top}\left(g^*(x,u_n(x))g^*(x,u_n(x))^{\top}+g_n(x,u_n(x))g_n(x,u_n(x))^{\top}\right)\left(f^*(x,u_n(x))-f_n(x,u_n(x))\right)\tag{73}$$

$$+\left\|\frac{\partial^2\log\left(\frac{\hat{\pi}_n(x,u_n(x),t)^{\frac{1}{2}}}{\pi^*(x|u_n,t)^{\frac{1}{2}}}\right)}{\partial x^2}\right\|_F\left\|g^*(x,u_n(x))g^*(x,u_n(x))^{\top}-g_n(x,u_n(x))g_n(x,u_n(x))^{\top}\right\|_F)\tag{74}$$

which is obtained by Cauchy-schwarz inequality. Then using Lemma 2.2, we could construct

$$h(t)\le\frac{d^{1.5}}{8\lambda_{\min}(g_n(x,u_n(x)))^2}\left\|\int_0^t f^*(x,u_n(x))-f_n(x,u_n(x)(\pi^*(x|u_n,t)+\pi_n(x|u_n,t))\right\|_2^2\tag{75}$$

$$+td^{2.5}L_\pi\left\|\int_0^t g^*(x,u_n(x))g^*(x,u_n(x))^{\top}-g_n(x,u_n(x))g_n(x,u_n(x))^{\top}(\pi^*(x|u_n,t)+\pi_n(x|u_n,t))\right\|_F\tag{76}$$

Then we could construct

$$\int_0^T h(t)\le\frac{d^{1.5}T}{8\lambda_{\min}(g_n(x,u_n(x)))^2}\left\|\int_0^T f^*(x,u_n(x))-f_n(x,u_n(x)(\pi^*(x|u_n,t)+\pi_n(x|u_n,t))\right\|_2^2\tag{77}$$

$$+T^2d^{2.5}L_\pi\left\|\int_0^0 g^*(x,u_n(x))g^*(\right.\tag{78}$$

Then we could construct

$$|C(\pi^*,u_n,t)-C(\pi)_n,u_n,t)|\le c_{\max}\sqrt{T(\hat{\sigma}_{n,1}(t)+\hat{\sigma}_{n,2}(t))}\tag{79}$$

where $\hat{\sigma}_{n,1}(t)=\frac{d^{1.5}T}{8\lambda_{\min}(g_n(x,u_n(x)))^2}\left\|\int_0^T f^*(x,u_n(x))-f_n(x,u_n(x)(\pi^*(x|u_n,t)+\pi_n(x|u_n,t))\right\|_2^2$ and $\hat{\sigma}_{n,2}(t)=T^2d^{2.5}L_\pi\left\|\int_0^0 g^*(x,u_n(x))g^*(x,u_n(x))^{\top}-g_n(x,u_n(x))g_n(x,u_n(x))^{\top}(\pi^*(x|u_n,t)+\pi_n(x|u_n,t))\right\|_F$. □

Then based on the proposed theorems, we could establish the regret of our proposed algorithms. We first start from the bound of the noise $\int g(x)dw_t$ at two sampling point.

$$\|x_n(t)-x(0)\|_2^2\le 2\|\int f^*(x_n(t),u_n(x_n(t)))dt\|_2^2+2\int\|g^*(x_n(t),u_n(x_n(t)))\|_2^2\tag{80}$$

$$\leq 2(L_f+L_g)\int_0^t \|x_n(s)-x(0)\|_2^2 ds + (2t^2\|f^*(x(0),u_n(x(0)))\|_2^2 + 2t\|g^*(x(0),u_n(x(0)))\|_2^2) \tag{81}$$

$$\leq e^{2(L_f+L_g)t}(2t^2\|f^*(x(0),u_n(x(0)))\|_2^2 + 2t\|g^*(x(0),u_n(x(0)))\|_2^2) \tag{82}$$

Then we have

$$\|x_n(t)-x(0)\|_2 \leq e^{2(L_f+L_g)t}(\sqrt{2}t\|f^*(x(0),u_n(x(0)))\|_2 + \sqrt{2t}\|g^*(x(0),u_n(x(0)))\|_2) \tag{83}$$

Similarly, for $t_1 \geq t_2$, we have Then we have

$$\|x_n(t_1)-x(t_2)\|_2 \leq e^{2(L_f+L_g)(t_1-t_2)}(\sqrt{2}(t_1-t_2)\|f^*(x(t_2),u_n(x(t_2)))\|_2 + \sqrt{2t}\|g^*(x(t_2),u_n(x(t_2)))\|_2) \tag{84}$$

$$\leq e^{2(L_f+L_g)(t_2+(t_1-t_2))}(\sqrt{2}(t_1-t_2)t_2\|f^*(x(0),u_n(x(0)))\|_2 + \sqrt{2(t_1-t_2)t_2}\|g^*(x(0),u_n(x(0)))\|_2) \tag{85}$$

Then by choosing suitable sampling point, we could bound $\|x_n(t_1)-x(t_2)\|_2$ with a constant. For simplification, we denote it as υ.

Lemma 2.5. Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of matrices in $\mathbb{R}^{d_1\times d_2}$, V a $d_1 \times d_1$ positive definite matrix and define

$$V_t = V + \sum_{s=1}^t X_s X_s^\top.$$

Then, we have that

$$\sum_{i=1}^t \|X_i V_t^{-1} X_i^\top\|_2 \leq d$$

Proof. First, we could construct

$$\sum_{i=1}^t \|X_i^\top V_t^{-1} X_i\|_2 \leq \sum_{i=1}^t \|V_t^{-\frac{1}{2}} X_i\|_F^2 \tag{86}$$

Note that

$$\sum_{i=1}^t \|V_t^{-\frac{1}{2}} X_i\|_F^2 = \sum_{i=1}^t tr(X_i^\top V_t^{-1} X_i) = d \tag{87}$$

□

Proof. By definition, we could construct

$$\mathrm{Var}(V_{k-1}^{-\frac{1}{2}}(X_i H_i)) = (X_i\Sigma_i)^\top V_k^{-1} X_i \Sigma_i) \preceq \|X_i^\top V_k^{-1} X_i\|_2 \Sigma_i^\top \Sigma_i \tag{88}$$

and

$$\|V_{k-1}^{-\frac{1}{2}}(X_i H_i)]\|_2 \leq \|V_k^{-\frac{1}{2}} X_i\|_2 \eta \tag{89}$$

Then utilizing the matrix-type freedman inequality, with possibility $1-\delta$, there holds

$$\|V_{k-1}^{-\frac{1}{2}}(\sum_{i=1}^k X_i H_i))\|_2 \leq \frac{2\max_i \|V_k^{-\frac{1}{2}} X_i\|_2 \log\left(\frac{d}{\delta}\right)\eta}{3} + \sqrt{2\log\left(\frac{d}{\delta}\right)\max_i \|X_i^\top V_k^{-1} X_i\|_2 \|\sum_{i=1}^k \Sigma_i^\top \Sigma_i\|_2} \tag{90}$$

□

Theorem 2.6. *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of matrices in $\mathbb{R}^{d_1\times d_2}$, V a $d_1 \times d_1$ positive definite matrix and define*

$$V_t = V + \sum_{s=1}^t X_s X_s^\top.$$

Let $\{Y_t\}_{t=1}^{\infty}$ be a sequence of matrices in $\mathbb{R}^{d_1^2\times d_2^2}$, V a $d_1^2 \times d_1^2$ positive definite matrix and define

$$\hat{V}_t = \hat{V} + \sum_{s=1}^t Y_s Y_s^\top.$$

there hold

$$\|\sum_{s=1}^k \hat{V}_{k-1}^{-\frac{1}{2}} Y_s^\top 2(X_s H_s)^\top V_{k-1}^{-1} \sum_{i=1}^k X_i H_i\|_2 \tag{91}$$

Proof.

$$\|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i\|_2 \leq \|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i1_{i=s}\|_2 \tag{92}$$

$$+\|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i1_{i\neq s}\|_2 \tag{93}$$

$$\leq\sqrt{\sum_{i=1}^{k}\|\hat{V}_{k-1}^{-\frac{1}{2}}Y_s\|_2^2}\sqrt{\sum_{i=1}^{k}\|V_{k-1}^{-\frac{1}{2}}X_s\|_2^2\eta^2} \tag{94}$$

$$+\|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i1_{i\neq s}\|_2 \tag{95}$$

$$\leq d^{\frac{3}{2}}\eta^2+\|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i1_{i\neq s}\|_2 \tag{96}$$

Then we have

$$\mathrm{Var}(\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}\sum_{i=1}^{k}X_iH_i1_{i\neq s})\preceq 2\|\hat{V}_{k-1}^{-\frac{1}{2}}Y_s\|_2^2\|V_{k-1}^{-\frac{1}{2}}X_s\|_2^2\|V_{k-1}^{-\frac{1}{2}}X_i\|_2^2\Sigma_i^{\top}\Sigma_i\Sigma_s^{\top}\Sigma_s \tag{97}$$

and

$$\|\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}X_iH_i1_{i\neq s}\|_2\leq 2\|\hat{V}_{k-1}^{-\frac{1}{2}}Y_s\|_2\|V_{k-1}^{-\frac{1}{2}}X_s\|_2\|V_{k-1}^{-\frac{1}{2}}X_i\|_2\eta^2 \tag{98}$$

Then we could construct

$$\|\sum_{s=1}^{k}\hat{V}_{k-1}^{-\frac{1}{2}}Y_s^{\top}2(X_sH_s)^{\top}V_{k-1}^{-1}X_iH_i\|_2\leq\frac{2\max_s\|\hat{V}_k^{-\frac{1}{2}}Y_s\|_2\max_i\|V_k^{-\frac{1}{2}}X_i\|_2^2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2 \tag{99}$$

$$+\sqrt{2\log\left(\frac{d}{\delta}\right)\max_s\|\hat{V}_k^{-\frac{1}{2}}Y_s\|_2^2\max_i\|V_k^{-\frac{1}{2}}X_i\|_2^4\|\sum_{i=1}^{k}\Sigma_i^{\top}\Sigma_i\|_2^2} \tag{100}$$

□

Facilitating the technique we have now, we could construct the theoretical analysis of our proposed algorithms.

Theorem 2.7. *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of matrices in $\mathbb{R}^{d_1\times d_2}$, V a $d_1\times d_1$ positive definite matrix and define*

$$V_t=V+\sum_{s=1}^{t}X_sX_s^{\top}.$$

and vector η_s follows $\eta_s=\int_{\hat{t}_{s-1}}^{\hat{t}_s}Y_t\vartheta^*dw_t$, then there holds

Proof. We start by defining $\vartheta=V_t^{-1}\sum_{s=1}^{t}X_s\eta_s$ and $\hat{X}_\varrho=\sum_{s=1}^{t}X_s1_{\hat{t}_{s-1}\leq\varrho\leq\hat{t}_s}$.

Then we could construct

$$\|\vartheta\vartheta^{\top}\|_2=\|V_t^{-1}\int_0^{\hat{t}_t}\hat{X}_s\sigma(s)\sigma(s)^{\top}\hat{X}_s^{\top}dsV_t^{-1}\|_2 \tag{101}$$

$$\leq\int_0^t\|V_t^{-\frac{1}{2}}\hat{X}_s\|_2^2\|V_t^{-\frac{1}{2}}Y_s\|_2^2ds \tag{102}$$

$$\leq\max_{s\in[t]}\int_{\hat{t}_{s-1}}^{\hat{t}_s}\|V_t^{-\frac{1}{2}}Y_i\|_2^2di\sum_{s=1}^{t}\|V_t^{-\frac{1}{2}}\hat{X}_s\|_2^2 \tag{103}$$

$$\leq d+1 \tag{104}$$

Then the theorem is proved. □

Theorem 2.8. *The cost function C follows*

$$\|C(u_n,T,\pi^*)-C(u_n,T,\hat{\pi}_n)\| \leq L_c, e^{KT}\int_0^T\int |f^*(x_1,u_n(x_1))-f_n(x_1,u_n(x_1))|_2^2, \hat{\pi}_n(x_1,s), dx, ds \tag{105}$$

$$+4L_c, e^{KT}\int_0^T\int \|G^*(x,u_n(x_1))-G_n(x_1,u_n(x_1))\|_F, \hat{\pi}_n(x_1,s), dx, ds, \tag{106}$$

where $K=2L_f+1+4L_g+(d+1)L_F+(d^2+1)L_F$.

Proof. We start with the inequality:

$$\sigma_{n,1}(t) \leq L_F, \mathbb{E}\left[\|_2 x_1 - x_2\|_2\right]\|\theta_n-\theta^*\|_2, \tag{107}$$

where $\sigma n, 1(t)$ is a non-negative error term at time t, L_F is the Lipschitz constant of the function F, $\mathbb{E}$ denotes the expectation operator, $|\cdot|_2$ denotes the Euclidean norm, x_1 and x_2 are state variables, and θ_n and θ^* are the estimated and true parameters, respectively. By applying Theorem 2.7, which provides a bound on the parameter estimation error $\|\theta_n-\theta^*\|_2$, we can further bound $\sigma n, 1(t)$:

$$\sigma_{n,1}(t) \leq (d+1)L_F\int \|x_1-x_2\|_2, \pi(x_1,x_2,u_n,s), dx + \int \|f^*(x_1,u_n(x_1))-f_n(x_1,u_n(x_1))\|_2^2, \hat{\pi}_n(x_1,s), dx, \tag{108}$$

where d is the dimension of the state space, $\pi(x_1,x_2,u_n,s)$ is the joint probability density function of x_1 and x_2 under the control policy u_n at time s, f^* and f_n are the true and estimated system dynamics, respectively, $\hat{\pi}_n(x_1,s)$ is the marginal probability density function of x_1 under u_n at time s, and $u_n(x_1)$ is the control input at state x_1. Similarly, we can bound another error term $\sigma_{n,2}(t)$:

$$\sigma_{n,2}(t) \leq 2(d^2+1)L_F, \mathbb{E}\left[\|x_1-x_2\|_2\right] + \int \|G^*(x_1,u_n(x_1))-G_n(x_1,u_n(x_1))\|_F, \hat{\pi}_n(x_1,s), dx, \tag{109}$$

where $|\cdot|_F$ denotes the Frobenius norm, $G^*(x_1,u_n(x_1)) = g^*(x_1,u_n(x_1))g^\top(x_1,u_n(x_1))$, $G_n(x_1,u_n(x_1)) = g_n(x_1,u_n(x_1))g_n^\top(x_1,u_n(x_1))$, and g^* and g_n are the true and estimated diffusion terms, respectively. Returning to inequality (32), we have:

$$\int \dot{h}(x_1,x_2), \pi(x_1,x_2,u_n,s), dx \leq \left(2L_f+1+4L_g+(d+1)L_F+(d^2+1)L_F\right)\int \|x_1-x_2\|_2^2, \pi(x_1,x_2,u_n,s), dx \tag{110}$$

$$+\int |f^*(x_1,u_n(x_1))-f_n(x_1,u_n(x_1))|_2^2, \hat{\pi}_n(x_1,s), dx \tag{111}$$

$$+4\int \|G^*(x_1,u_n(x_1))-G_n(x_1,u_n(x_1))\|_F, \hat{\pi}_n(x_1,s), dx, \tag{112}$$

where L_f and L_g are the Lipschitz constants of the functions f and g, respectively. Applying Grönwall's inequality, we obtain:

$$\int h(x_1,x_2), \pi(x_1,x_2,u_n,t), dx \leq e^{Kt}\int_0^t\int |f^*(x_1,u_n(x_1))-f_n(x_1,u_n(x_1))|_2^2, \hat{\pi}_n(x_1,s), dx, ds \tag{113}$$

$$+4e^{Kt}\int_0^t\int \|G^*(x_1,u_n(x_1))-G_n(x_1,u_n(x_1))\|_F, \hat{\pi}_n(x_1,s), dx, ds, \tag{114}$$

where $K=2L_f+1+4L_g+(d+1)L_F+(d^2+1)L_F$. According to inequality (18), the difference in cost functions can be bounded as:

$$\|C(u_n,T,\pi_n)-C(u_n,T,\hat{\pi}_n)\| \leq L_c, e^{KT}\int_0^T\int |f^*(x_1,u_n(x_1))-f_n(x_1,u_n(x_1))|_2^2, \hat{\pi}_n(x_1,s), dx, ds \tag{115}$$

$$+4L_c, e^{KT}\int_0^T\int \|G^*(x_1,u_n(x_1))-G_n(x_1,u_n(x_1))\|_F, \hat{\pi}_n(x_1,s), dx, ds, \tag{116}$$

where $C(u_n,T,\pi_n)$ and $C(u_n,T,\hat{\pi}_n)$ are the costs associated with the control policy u_n under the distributions π_n and $\hat{\pi}_n$, respectively, L_c is the Lipschitz constant of the cost function, and T is the time horizon. □

Theorem 2.9. *At the layer l, he cost function C fulfills*

$$|C(\pi, f^*, g^*) - C(\pi, f_n, g_n)| \leq e^{KT} \frac{\gamma^{-2l}}{\sqrt{(1-2d\gamma^{-2l})}} \left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3} + \sqrt{2\log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g_n(x_n(t), u_n(x_n(t))) g_n(x, u_n(x_n(t)))^\top \tag{117}$$

$$+ \gamma^{-2l} \sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3} + d^{\frac{3}{2}}\eta^2} \tag{118}$$

Proof. First, using the theorem 2.3 and 2.4, we have

$$|C(\pi, f^*, g^*) - C(\pi, f_n, g_n)| \leq L_c e^{KT} \times \left\| \int_0^T \int (f(x, u_n(x)) - \hat{f}(x, u_n(x)))((f(x, u_n(x)) - \hat{f}(x, u_n(x))))^\top \pi(x, u_n(x), t) dx dt \right\|_2 \tag{119}$$

$$+ e^{KT} \sqrt{\| \int E_t[(g^*(x, u_n(x)) g^*(x, u_n(x))^\top - g_n(x, u_n(x)) g_n(x, u_n(x))^\top)] dt \|_2} \tag{120}$$

$$= e_f \left\| \int_0^T \int (f(x, u_n(x)) - \hat{f}(x, u_n(x)))((f(x, u_n(x)) - \hat{f}(x, u_n(x))))^\top \pi(x, u_n(x), t) dx dt \right\|_2 \tag{121}$$

$$+ e_g \sqrt{\| \int E_t[(g^*(x, u_n(x)) g^*(x, u_n(x))^\top - g_n(x, u_n(x)) g_n(x, u_n(x))^\top)] dt \|_2} \tag{122}$$

Here, we denote the sampling at the epsiode n is $x_n(t)$. At the layer l, with possibility $1 - 2\delta$, there holds

$$\sigma_{n,t,1}(\pi^*) \leq \| \int_0^T (f(x_n(t), u) - \hat{f}(x_n(t), u))((f(x_n(t), u) - \hat{f}(x_n(t), u)))^\top dt \|_F^{\frac{1}{2}} \tag{123}$$

$$\leq (\gamma^{-l}\tau) \| \Gamma_n^{\frac{1}{2}} (\theta_n - \theta^*) \|_2 \tag{124}$$

$$\leq (\gamma^{-2l}\tau) \left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3} + \sqrt{2\log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g^*(x_n(s), u_n(x_n(s)), s) g^*(x_n(s), u_n(x_n(s)), s)^\top} \right) \tag{125}$$

and

$$\sigma_{n,t,2}(\pi^*) \leq \| \int_0^T \int (g^*(x_n(t), u_n(x_n(t))) g^*(x_n(t), u_n(x_n(t))) \top - g_n(x_n(t), u_n(x_n(t))) g_n(x_n(t), u_n(x_n(t))) \|_F^{\frac{1}{2}} \tag{126}$$

$$\leq \gamma^{-\frac{l}{2}} \tau \| \hat{\Gamma}_n^{\frac{1}{2}} (\Theta_n - \Theta^*) \|_2^{\frac{1}{2}} \tag{127}$$

$$\leq \gamma^{-2l} \left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3} + \sqrt{2\log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g^*(x_n(s), u_n(x_n(s)), s) g^*(x_n(s), u_n(x_n(s)), s)^\top} + d^{\frac{3}{2}}\eta \right) \tag{128}$$

Hower, since we have no access to the variance information, we would consider

$$\| \sum_{i=1}^{n} \int_0^T \hat{F}_i(x_i(t), u_i(x_i(t))) \theta_n - \text{vec}(\sum_{n=1}^{N} \int_0^T g^*(x, u_n(x)) g^*(x, u_n(x))^\top \pi^*(x|u_n, t) dt) \|_2 \tag{129}$$

$$\leq \gamma^{-4l}(\tau n)\| \sum_{n=1}^{N} \int_0^T g^*(x, u_n(x)) g^*(x, u_n(x))^\top \pi^*(x|u_n, t) dt \|_2 \sqrt{2 \log\left(\frac{d}{\delta}\right)} + \gamma^{-4l}(\tau n)(\frac{2 \log\left(\frac{d}{\delta}\right) \eta^2}{3} + d^{\frac{3}{2}} \eta^2) \tag{130}$$

Then we have

$$\| \sum_{i=1}^{n} \int_0^T \hat{F}_i(x_i(t), u_i(x_i(t))) \theta_n \|_2 \leq (1 + 2d\gamma^{-2l}) \times \tag{131}$$

$$(\| \sum_{n=1}^{N} \int_0^T g^*(x, u_n(x)) g^*(x, u_n(x))^\top \pi^*(x|u_n, t) dt \|_2 + \frac{2 \log\left(\frac{d}{\delta}\right) \eta^2}{3} + d^{\frac{3}{2}} \eta^2) \tag{132}$$

and

$$\| \sum_{i=1}^{n} \int_0^T \hat{F}_i(x_i(t), u_i(x_i(t))) \theta_n \|_2 \geq (1 - 2d\gamma^{-2l}) \times \tag{133}$$

$$(\| \sum_{n=1}^{N} \int_0^T g^*(x, u_n(x)) g^*(x, u_n(x))^\top \pi^*(x|u_n, t) dt \|_2 + \frac{2 \log\left(\frac{d}{\delta}\right) \eta^2}{3} + d^{\frac{3}{2}} \eta^2) \tag{134}$$

Then we could construct

$$|C(\pi, f^*, g^*) - C(\pi, f_n, g_n)| \leq e^{KT} \frac{\gamma^{-2l}}{\sqrt{(1 - 2d\gamma^{-2l})}} \left(\frac{2 \log\left(\frac{d}{\delta}\right) \eta}{3} + \sqrt{2 \log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g_n(x_n(t), u_n(x_n(t))) g_n(x, u_n(x_n(t)))^\top \tag{135}$$

$$+ \gamma^{-2l} \sqrt{\frac{2 \log\left(\frac{d}{\delta}\right) \eta^2}{3} + d^{\frac{3}{2}} \eta^2} \tag{136}$$

□

Theorem 2.10. *R_T follows*

Proof. We start from the last layer, i.e. $\max(\|\Gamma_n^{-\frac{1}{2}} \int_{t_{i-1}}^{t_i} F(x_n(s), u_n(x_n(s)))\|_2, \|\Gamma_n^{-\frac{1}{2}}, \int_{t_{i-1}}^{t_i} \hat{F}(x_n(s), u_n(x_n(s))) ds\|_2 \leq \alpha$. For simplification, we denote it as condition $E_{\alpha,n,i}$. Similarly, we denote $\max_n \max_i \max(\|\Gamma_n^{-\frac{1}{2}} \int_{t_{i-1}}^{t_i} F(x_n(s), u_n(x_n(s)))\|_2, \|\Gamma_n^{-\frac{1}{2}}, \int_{t_{i-1}}^{t_i} \hat{F}(x_n(s), u_n(x_n(s))) ds\|_2 \leq \gamma^{-l}$ as condition $E_{l,n,i}$. Given $\alpha \leq \max(\frac{1}{N\tau}, \frac{1}{2\sqrt{d}})$ we could construct

$$\sum_{i=1}^{\tau} \sum_{n=1}^{N} |C(\pi^*, f^*, g^*) - C(\hat{\pi}_n, f_n, g_n)| 1_{E_{\alpha,n,i}} \leq e^{KT} 2d\sqrt{T}(\tau + \sqrt{\tau}) \times \tag{137}$$

$$\frac{\alpha^2}{\sqrt{(1 - 2d\alpha^2)}} \left(\frac{2 \log\left(\frac{d}{\delta}\right) \eta}{3} + \sqrt{2 \log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g_n(x_n(t), u_n(x_n(t))) g_n(x, u_n(x_n(t)))^\top dt \|_2} \right) \tag{138}$$

$$+ \frac{\alpha^2}{\sqrt{(1 - 2d\alpha^2)}} (2 \log\left(\frac{d}{\delta}\right))^{\frac{1}{4}} \sqrt{\| \sum_{n=1}^{N} \int_0^T g^*(x, u_n(x)) g^*(x, u_n(x))^\top \pi^*(x|u_n, t) dt \|_2} \tag{139}$$

$$+ \alpha^2 \sqrt{\frac{2 \log\left(\frac{d}{\delta}\right) \eta^2}{3} + d^{\frac{3}{2}} \eta^2} \tag{140}$$

$$\leq 2e^{KT}(\tau + \sqrt{\tau}) \times \tag{141}$$

$$\left(\frac{2 \log\left(\frac{d}{\delta}\right) \eta}{3} + \sqrt{2 \log\left(\frac{d}{\delta}\right) \| \sum_{n=1}^{N} \int_0^T g_n(x_n(t), u_n(x_n(t))) g_n(x, u_n(x_n(t)))^\top dt \|_2} \right) \tag{142}$$

$$+ (2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top \pi^*(x|u_n,t)dt\|_2} \tag{143}$$

$$+\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2} \tag{144}$$

$$\tag{145}$$

Then for layer l, there holds

$$\sum_{n=1}^{N}|C(f^*,g^*)-C(f_n,g_n)|1_{E_{l,n,i}} \leq e^{KT}2d\sqrt{T}(\tau+\sqrt{\tau})\times \tag{146}$$

$$\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{147}$$

$$+\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}(2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{148}$$

$$+2^{-2l}\tau N\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2} \tag{149}$$

$$\leq 2e^{KT}2d^2\sqrt{T}(\tau+\sqrt{\tau})\times \tag{150}$$

$$\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{151}$$

$$+ (2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}d\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{152}$$

$$+\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2 d} \tag{153}$$

$$\tag{154}$$

Then

$$R_T=\sum_{n=1}^{N}C(u_n(x),f^*,g^*)-C(u^*(x),f^*,g^*) \tag{155}$$

$$\leq R_T=\sum_{n=1}^{N}C(u_n(x),f^*,g^*)-C(u_n(x),f_n,g_n)+C(u^*(x),f_n,g_n)-C(u^*(x),f^*,g^*) \tag{156}$$

$$\leq O(e^{KT}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2}(\log\left(\frac{d}{\delta}\right))^{\frac{1}{2}}) \tag{157}$$

$$+O(d^{2.5})+\sum_{n=1}^{N}C(u_n(x),f_n,g_n)-C(u^*(x),f_n,g_n) \tag{158}$$

We define $\max(\|\Gamma_n^{-\frac{1}{2}}\int_{t_{i-1}}^{t_i}F(x_n(s),u^*(x_n(s)))\|_2, \|\Gamma_n^{-\frac{1}{2}},\int_{t_{i-1}}^{t_i}\hat{F}(x_n(s),u_n(x_n(s)))ds\|_2) \leq \alpha$ as $\mathcal{E}_{\alpha,n,i}$ and $\max(\|\Gamma_n^{-\frac{1}{2}}\int_{t_{i-1}}^{t_i}F(x_n(s),u^*(x_n(s)))\|_2, \|\Gamma_n^{-\frac{1}{2}},\int_{t_{i-1}}^{t_i}\hat{F}(x_n(s),u_n(x_n(s)))ds\|_2) \leq \gamma^{-l}$ as $\hat{\mathcal{E}}_{l,n,i}$. Then we have

$$\sum_{i=1}^{\tau}\sum_{n=1}^{N}C(u_n(x),f_n,g_n)-C(u^*(x),f_n,g_n)1_{\mathcal{E}_{\alpha,n,i}} \leq e^{KT}2d\sqrt{T}(\tau+\sqrt{\tau})\times \tag{159}$$

$$\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{160}$$

$$+\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}(2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{161}$$

$$+2^{-2l}\tau N\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2} \tag{162}$$

$$\leq 2e^{KT}2d^2\sqrt{T}(\tau+\sqrt{\tau})\times \tag{163}$$

$$\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{164}$$

$$+(2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}d\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{165}$$

$$+\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2 d} \tag{166}$$

$$\tag{167}$$

$$\sum_{i=1}^{\tau}\sum_{n=1}^{N}C(u_n(x),f_n,g_n)-C(u^*(x),f_n,g_n)1_{\mathcal{E}_{\alpha,n,i}}\leq e^{KT}2d\sqrt{T}(\tau+\sqrt{\tau})\times \tag{168}$$

$$\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{169}$$

$$+\frac{2^{-2l}\tau N}{\sqrt{(1-2d2^{-2l})}}(2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{170}$$

$$+2^{-2l}\tau N\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2} \tag{171}$$

$$\leq 2e^{KT}2d^2\sqrt{T}(\tau+\sqrt{\tau})\times \tag{172}$$

$$\left(\frac{2\log\left(\frac{d}{\delta}\right)\eta}{3}+\sqrt{2\log\left(\frac{d}{\delta}\right)\|\sum_{n=1}^{N}\int_0^T g_n(x_n(t),u_n(x_n(t)))g_n(x,u_n(x_n(t)))^\top dt\|_2}\right) \tag{173}$$

$$+(2\log\left(\frac{d}{\delta}\right))^{\frac{1}{4}}d\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2} \tag{174}$$

$$+\sqrt{\frac{2\log\left(\frac{d}{\delta}\right)\eta^2}{3}+d^{\frac{3}{2}}\eta^2 d} \tag{175}$$

$$\tag{176}$$

Then the regret

$$R_N = O(e^{KT}\sqrt{\|\sum_{n=1}^{N}\int_0^T g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)dt\|_2}(\log\left(\frac{d}{\delta}\right))^{\frac{1}{2}}) \tag{177}$$

$$+O(d^{2.5}) \tag{178}$$

□

Algorithm 1 SupLin + Adaptive Variance-aware Exploration (SAVE)

Require: horizon T, action set $\mathcal{A}$, all $P(Pa_Y|a)$. $\alpha > 0$, and the upper bound on the ℓ_2-norm of a in $\mathcal{D}_k (k \geq 1)$, i.e., A.

Initialize $L \leftarrow \lceil \log_2(1/\alpha) \rceil$.

Initialize the estimators for all layers: $\hat{\Sigma}_{1,\ell} \leftarrow \gamma^{-2\ell} \cdot \mathbf{I},\ \hat{b}_{1,\ell} \leftarrow 0,\ \hat{\theta}_{1,\ell} \leftarrow 0,\ \Psi_{k,\ell} \leftarrow \varnothing, \hat{\beta}_{1,\ell} \leftarrow \gamma^{-\ell+1}$ for all $\ell \in [L]$.

for $k = 1, \ldots, K$ **do**

Construct function space $\mathcal{D}_k = \{(\hat{f}_n(x,t), \hat{g}_n(x, u_n(x))) | \forall x \in R^n$ and $u(x) \in R^m, \|\hat{f}_n(x,t) - F(x, u_n(x))\theta_n\|_2 \leq \sigma_f(n,x)\ \|\vec{(}\hat{g}_n(x,u_n(x))\hat{G}_n(x,u_n(x))^\top) - \hat{F}(x,u_n(x))\theta_n\|_2 \leq \sigma_g(n,x)\ \}$

Let $\mathcal{A}_{k,1} \leftarrow \mathcal{D}_k,\ \ell \leftarrow 1$.

while the optimal solution of pairs $(\hat{f}_n(x_{k,i},t), \hat{g}_n(x, u_n(x)))$ not found **do**

if $\forall (x, u_n(x)), \|\hat{F}(x, u_n(x))\|_{\hat{\Gamma}^{-1}_{k,\ell}} \leq \alpha$ for all $(\hat{f}_n(x_{k,i},t), \hat{g}_n(x, u_n(x))) \in \mathcal{A}_{k,\ell}$ **then**

Choose $(\hat{f}_n(x_{k,i},t), \hat{g}_n(x, u_n(x))) \leftarrow \arg\min_{f,g} \mathrm{LCB}_C(u, f, g)$.

Choose $u \leftarrow \arg\min_u \mathrm{LCB}_C(u, \hat{f}_n, \hat{g}_n)$.

Compute the weight: $w_{k,i} \leftarrow T\gamma^i$.

Compute the weight: $\hat{w}_{k,i} \leftarrow T\gamma^i$.

for $i = 1, \cdots, \tau$ **do**

$t_{k,i} = \arg\max_{t \in [\sum_{s=1}^{i-1} \Delta_{k,i} \sum_{s=1}^{i} \Delta_{k,i}]} \max(\|\Gamma_n^{-1} \int_{t_{k,i-1}}^t F(x,u)\pi_n(x|u_n,t)dt\|_2, \|\hat{\Gamma}_n^{-1} \int_{t_{k,i-1}}^t \hat{F}(x,u)\pi_n(x|u_n,t)dt\|_2$

Observe the state at time $t_{k,i}$

end for

Keep the same index sets at all layers: $\Psi_{k+1,\ell'} \leftarrow \Psi_{k,\ell'}$ for all $\ell' \in [L]$.

else if $\|\hat{F}(x, u_n(x))\|_{\hat{\Gamma}^{-1}_{k,\ell}} \leq \gamma^{-\ell}$ for all $a \in \mathcal{A}_{k,\ell}$ **then**

$\mathcal{A}_{k,\ell+1} \leftarrow \{(\hat{f}_n(x,t), \hat{g}_n(x, u_n(x))) \in \mathcal{A}_{k,\ell} \mid \langle a, \hat{\theta}_{k,\ell} \rangle \geq \max_{(\hat{f}_n, \hat{g}_n) \in \mathcal{A}_{k,\ell}} C(\hat{f}_n, \hat{g}_n, u(\hat{f}_n, \hat{g}_n)) - 3 \times 118.$

else

Choose $(f_n g_n)$ such that $\|F(x, u_n(x))\|_{\hat{\Gamma}^{-1}_{k,\ell}}\| \geq \gamma^{-\ell}$

for $i = 1, \cdots, \tau$ **do**

$t_{k,i} = \arg\max_{t \in [\sum_{s=1}^{i-1} \Delta_{k,i} \sum_{s=1}^{i} \Delta_{k,i}]} \max(\|\Gamma_n^{-1} \int_{t_{k,i-1}}^t F(x,u)\pi_n(x|u_n,t)dt\|_2, \|\hat{\Gamma}_n^{-1} \int_{t_{k,i-1}}^t \hat{F}(x,u)\pi_n(x|u_n,t)dt\|_2$

Compute the weight: $w_{k,i} \leftarrow \gamma^{-\ell} / \|\hat{F}(x, u_n(x))\|_{\hat{\Gamma}^{-1}_{k,\ell}} (t_{k,i} - t_{k,i-1})^{0.5}$.

Compute the weight: $\hat{w}_{k,i} \leftarrow \gamma^{-\ell} / \|\hat{F}(x, u_n(x))\|_{\hat{\Gamma}^{-1}_{k,\ell}} (t_{k,i} - t_{k,i-1})$.

end for

Update the index sets: $\Psi_{k+1,\ell} \leftarrow \Psi_{k,\ell} \cup \{k\}$ and $\Psi_{k+1,\ell'} \leftarrow \Psi_{k,\ell'}$ for $\ell' \in [L] \setminus \{\ell\}$.

end if

$\ell \leftarrow \ell + 1$.

end while

for $\alpha \in [L]$ such that $\Psi_{k+1,\alpha} \neq \Psi_{k,\alpha}$ **do**

Update the estimators as follows:

$$\Gamma_{k+1,\ell} \leftarrow \Gamma_{k,\ell} + w_{k,i}^2 F(x_{k,i}, u) F(x_{k,i}, u)^\top,\ b_{k+1,\ell} \leftarrow b_{k,\ell} + w_{k,i}^2 F(x_{k,i}, u)(x_{t,i} - x_{t,i-1}),\ \theta_{k+1,\ell} \leftarrow \Gamma^{-1}_{k+1,\ell} b_{k+1,\ell}.$$

$$\hat{\Gamma}_{k+1,\ell} \leftarrow \hat{\Gamma}_{k,\ell} + \hat{w}_{k,i}^2 \hat{F}(x_{k,i}, u) \hat{F}(x_{k,i}, u)^\top,\ b_{k+1,\ell} \leftarrow \hat{b}_{k,\ell} + \hat{w}_{k,i}^2 \hat{F}(x_{k,i}, u) \mathrm{vec}((x_{t,i} - x_{t,i-1} - F(x_{k,i}, u)\theta_n)$$
$$(x_{t,i} - x_{t,i-1} - F(x_{k,i}, u)\theta_n)^\top),\ \Theta_{k+1,\ell} \leftarrow \hat{\Gamma}^{-1}_{k+1,\ell} \hat{b}_{k+1,\ell}.$$

end for

Compute the adaptive confidence radius $\hat{\beta}_{k+1,\ell}$ for the next round according to (2.3).

for $\ell \in [L]$ **do**

$\Psi_{k+1,\ell} = \Psi_{k,\ell}$, let $\hat{\Sigma}_{k+1,\ell} \leftarrow \hat{\Sigma}_{k,\ell}, \hat{b}_{k+1,\ell} \leftarrow \hat{b}_{k,\ell},\ \hat{\theta}_{k+1,\ell} \leftarrow \hat{\theta}_{k,\ell},\ \hat{\beta}_{k+1,\ell} \leftarrow \hat{\beta}_{k,\ell}$.

end for

end for

Theorem 2.11.

Proof. We consider

$$h(t) = \int \max(\|f_n(x,u_n(x)) - f^*(x,u_n(x))\|_2^2, \|g_n(x,u_n(x))g_n(x,u_n(x))^\top - g^*(x,u_n(x))g^*(x,u_n(x))^\top\|_F)(\pi^*(x|u_n,t)) - \pi_n(x \tag{179}$$

Then we have

$$\begin{aligned}\dot{h}(t) =& \max(\|f_n(x,u_n(x)) - f^*(x,u_n(x))\|_2^2, \|g_n(x,u_n(x))g_n(x,u_n(x))^\top - g^*(x,u_n(x))g^*(x,u_n(x))^\top\|_F)\times && (180)\\ & \int(\nabla(f^*(x,u_n(x))\pi^*(x|u_n,t)) - \nabla(f_n(x,u_n(x))\pi_n(x|u_n,t)) && (181)\\ & + \Delta(g^*(x,u_n(x))g^*(x,u_n(x))^\top\pi^*(x|u_n,t)) - \Delta(g_n(x,u_n(x))g_n(x,u_n(x))^\top\pi_n(x|u_n,t)))dx && (182)\\ =& tr((\nabla h(x))^\top(f^*(x,u_n(x)) - f_n(x,u_n(x)))\pi_n(x|u_n,t)) && (183)\\ & + tr((\nabla h(x))^\top(f^*(x,u_n(x))(\pi_n(x|u_n,t) - \pi^*(x|u_n,t)))) && (184)\\ & + tr(\text{vec}(\Delta(h(t)))h(t)\frac{g^*(x,u_n(x))g^*(x,u_n(x))^\top - g_n(x,u_n(x))g_n(x,u_n(x))^\top\pi_n(x|u_n,t)}{h(t)}) && (185)\\ & + tr(\text{vec}(\Delta(h(t)))g^*(x,u_n(x))g^*(x,u_n(x))^\top(\pi_n(x|u_n,t) - \pi^*(x|u_n,t))) && (186)\\ & && (187)\end{aligned}$$

Note that

$$185 = -tr\left(\text{vec}(\nabla(h(t))\nabla(h(t))^\top)\frac{g^*(x,u_n(x))g^*(x,u_n(x))^\top + g_n(x,u_n(x))g_n(x,u_n(x))^\top\pi_n(x|u_n,t)}{h(t)}\right) \tag{188}$$

$$+tr\left(\text{vec}(\Delta(h(t)^2))\frac{g^*(x,u_n(x))g^*(x,u_n(x))^\top - g_n(x,u_n(x))g_n(x,u_n(x))^\top\pi_n(x|u_n,t)}{h(t)}\right) \tag{189}$$

Then we have □

Lemma 2.12. *The cost function C follows*

$$C^*(\pi^*, f_n, g_n) \geq C_n(\pi_n, f_n, g_n) - 2|C^*(\pi_n, f_n, g_n) - C^*(\pi^*, f_n, g_n)| - |C_n(\pi_n, f_n, g_n) - C_n(\pi^*, f_n, g_n)| \tag{190}$$

Proof. First, we could construct

$$\begin{aligned}C^*(\pi_n, f_n, g_n) - C_n(\pi_n, f_n, g_n) \geq & C^*(\pi^*, f_n, g_n) - C_n(\pi^*, f_n, g_n) - |C^*(\pi_n, f_n, g_n) - C^*(\pi^*, f_n, g_n)| && (191)\\ & - |C_n(\pi_n, f_n, g_n) - C_n(\pi^*, f_n, g_n)| \geq -|C^*(\pi_n, f_n, g_n) - C^*(\pi^*, f_n, g_n)| && (192)\\ & - |C_n(\pi_n, f_n, g_n) - C_n(\pi^*, f_n, g_n)| && (193)\end{aligned}$$

Then we have

$$C^*(\pi^*, f_n, g_n) - C_n(\pi_n, f_n, g_n) \geq -2|C^*(\pi_n, f_n, g_n) - C^*(\pi^*, f_n, g_n)| - |C_n(\pi_n, f_n, g_n) - C_n(\pi^*, f_n, g_n)| \tag{194}$$

Then the theorem is proved. □

Lemma 2.13.

Proof. We consider the cost function $L(\theta)$ is defined as

$$L(\theta) = \frac{1}{2}\|f(\theta) - y\|_2^2 + \frac{m\lambda}{2}\|\theta - \theta^{(0)}\|_2^2 \tag{195}$$

For ease of expression, we denote $g(\theta) = ((g(x_1, a_1, \theta), \ldots, g(x_t, a_t; \theta)))^\top$ $f(\theta) = (f(x_1, a_1; \theta), \ldots, f(x_t, a_t; \theta))^\top$
Then we could construct

$$\begin{aligned}
\dot{L}(\theta) =& -(f(\theta) - y + m\lambda(\theta - \theta^{(0)}))^\top (g(\theta) + \lambda m I)(f(\theta) - y + m\lambda(\theta - \theta^{(0)})) && (196)\\
\leq& -\|f(\hat{\theta}) - y + 2m\lambda(\hat{\theta} - \theta^{(0)})\|_2^2 - (m^2\lambda^2)\|\theta - \theta^{(0)}\|_2^2 + m\lambda\|\theta - \theta^{(0)}\|_2 \|f(\hat{\theta}) - y + 2m\lambda(\hat{\theta} - \theta^{(0)})\|_2 && (197)\\
\leq& -\|f(\hat{\theta}) - y + 2m\lambda(\hat{\theta} - \theta^{(0)})\|_2^2 + m^2\lambda^2\|\theta - \theta(0)\|_2^2 + m\lambda\| && (198)
\end{aligned}$$

Then we have

$$\dot{L}(\theta) - \dot{L}(\hat{\theta}) = -L(\theta) + L(\hat{\theta}) + (f(\hat{\theta}) - y + m\lambda(\hat{\theta} - \theta^{(0)}))^\top (g(\hat{\theta}) - g(\theta))(f(\hat{\theta}) - y + m\lambda(\hat{\theta} - \theta^{(0)})) \quad (199)$$

Then there hold

$$L(\theta) - L(\hat{\theta}) \leq \|(g(\theta) + \lambda m I)^{-1}\|_2 \|g(\hat{\theta}) - g(\theta)\|_2 L(\hat{\theta})(1 - e^{-\frac{1}{\|(g(\theta)+\lambda m I)^{-1}\|_2} t}) \quad (200)$$

We also have

$$\dot{L}(\theta) - \dot{L}(\hat{\theta}) \geq -\|(g(\theta) + \lambda m I)\|_2 (L(\theta) - L(\hat{\theta})) - \|g(\hat{\theta}) - g(\theta)\|_2 L(\hat{\theta}) \quad (201)$$

Then we could construct $\square$

References