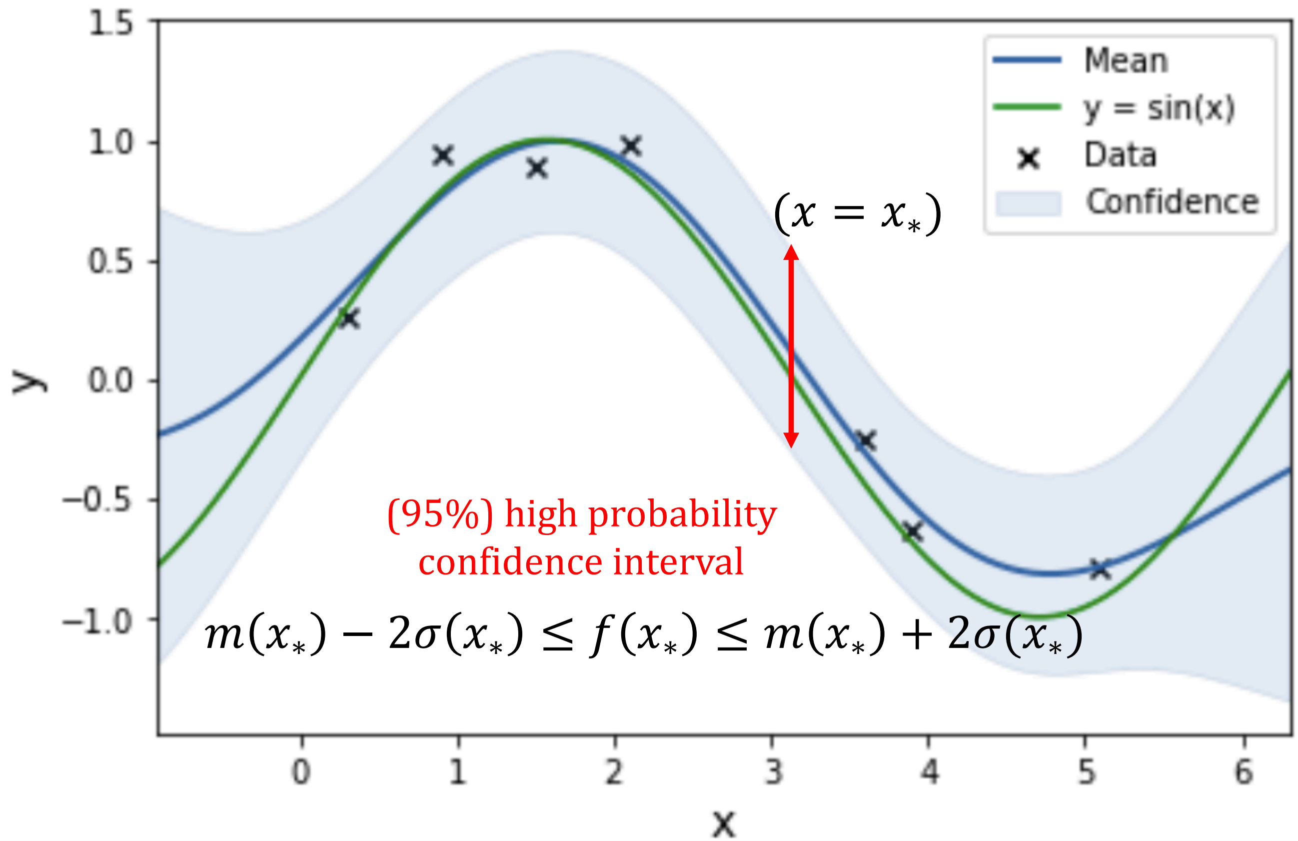
Weekly Report

M2 Ibuki Takeuchi

- GP regression of single output function $f(x)$

$$f(x_*) \sim \mathcal{N}(m(x_*), \sigma^2(x_*))$$



Weekly Report

M2 Ibuki Takeuchi

- In [1], they use GP regression of multi output function
  - Nominal control affine model

$$s_{t+1} = f(s_t) + g(s_t)a_t + d(s_t)$$

where f, g are known and d is uncertain → **GPR!**

In this case

$d(s)$ is $n - \mathit{dimensional}$ function : $\mathbb{R}^n \to \mathbb{R}^n$
How should we express the variance…?

- GP for each factor of $d_i(s)$: $\mathbb{R}^n \to \mathbb{R}$

$$\underbrace{\begin{pmatrix} m_1(s) \\ \vdots \\ m_n(s) \end{pmatrix}}_{\boldsymbol{m}(s)} - 2\underbrace{\begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_n(s) \end{pmatrix}}_{\boldsymbol{\sigma}(s)} \leq \begin{pmatrix} d_1(s) \\ \vdots \\ d_n(s) \end{pmatrix} \leq \begin{pmatrix} m_1(s) \\ \vdots \\ m_n(s) \end{pmatrix} + 2\begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_n(s) \end{pmatrix}$$

[1] : R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. " End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. " *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19),* 2019.

Weekly Report

M2 Ibuki Takeuchi

- What is proved ①
  - If there exists a solution to (*) for all $s \in C$(set of safe states) with $\varepsilon = 0$, the set C will be forward invariant w.p. $1-\delta$.

  - If there exists a state $s \in C$ such that (*) has solution with $\varepsilon = \varepsilon_{max} > 0$. If for all $s \in C$, the solution to (*) satisfies $\varepsilon < \varepsilon_{max}$, then set C_ε(larger set to C) will be forward invariant w.p. $1-\delta$.

(*)

$$
\begin{aligned}
(a_t, \varepsilon) = \underset{a_t, \varepsilon}{\operatorname{argmin}} \ & \|a_t\|_2 + K_\varepsilon \varepsilon \\
\text{s.t.} \quad & p^\top f(s_t) + p^\top g(s_t)\left(u_{\theta_k}^{RL}(s) + a_t\right) + p^\top \mu_d(s_t) \\
& - k_\delta |p|^\top \sigma_d(s_t) + q \geq (1-\eta)h(s_t) - \varepsilon \\
& a_{low}^i \leq u_{\theta_k}^{RL(i)}(s) + a_t \leq a_{high}^i \text{ for } i = 1, \ldots, M
\end{aligned}
$$

Weekly Report

M2 Ibuki Takeuchi

- What is proved ②
  - If we use TRPO for RL, the algorithm achieve performance guarantee

$$J(\pi_k^{prop}) \geq J(\pi_{k-1}) - \frac{2\lambda\gamma}{(1-\gamma)^2}\delta_\pi$$

$$J(\pi) = \mathbb{E}_{a\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r(s_t)]$$