# Robustness Guarantees for Bayesian Inference with Gaussian Processes

Andrea Patane
Department of Computer Science, University of Oxford

Joint work with:
Luca Cardelli, Marta Kwiatkowska and Luca Laurenti

# Outline

- Motivations.
- Background: Bayesian Inference with Gaussian Processes.
- Problem Formulation: Probabilistic invariance.
- Methods:
    - Safe-approximation of invariance property.
    - Branch and Bound optimisation scheme for GPs.
- Case of Study: Empirical analysis of ReLU fully-connected Neural Networks via GP with ReLU kernel.

# Robustness for Bayesian Learning, Why?

- Bayesian methods are employed in safety critical applications, where uncertainty estimation is necessary (e.g. diagnosis, medicine intake, control systems…).

- Robustness guarantees are needed to prove the correctness of the model in a probabilistic fashion.

- Current methods either neglect uncertainty or are based on empirical approaches (e.g. variance thresholding)

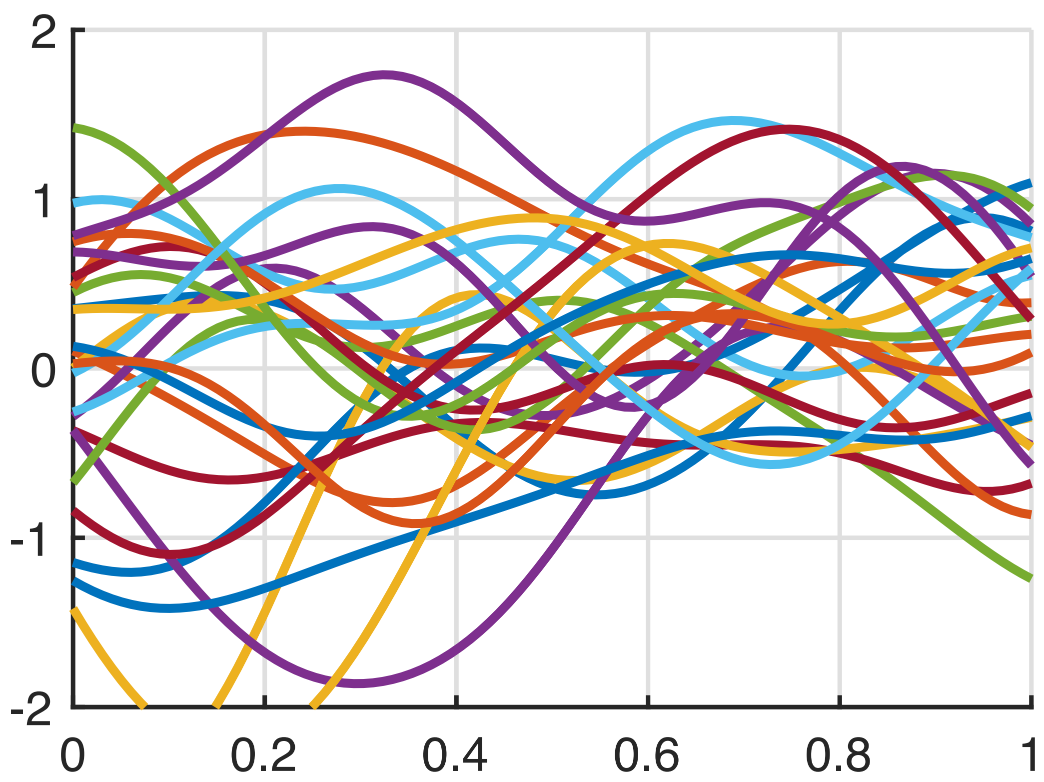
Problem: Provide probabilistic guarantees for GPs.

# Background

# Bayesian Inference with GPs (in Figures)

- Step 1: Definition of a GP prior distribution.

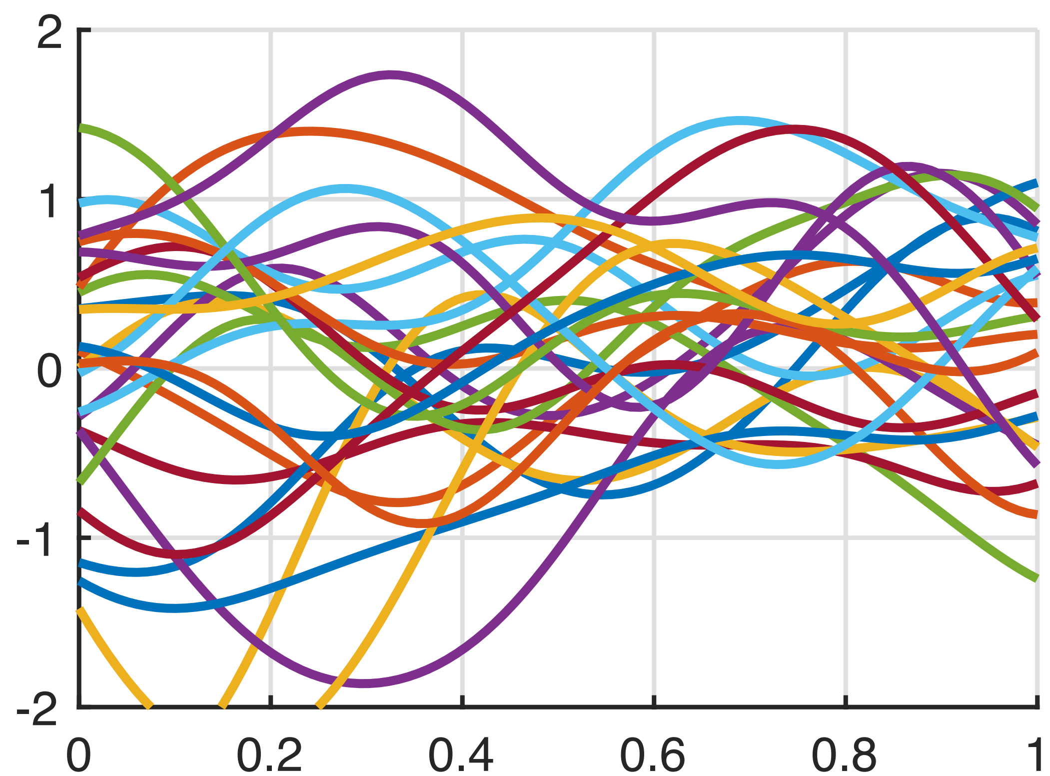
$$f \sim \mathcal{GP}$$



# Bayesian Inference with GPs (in Figures)

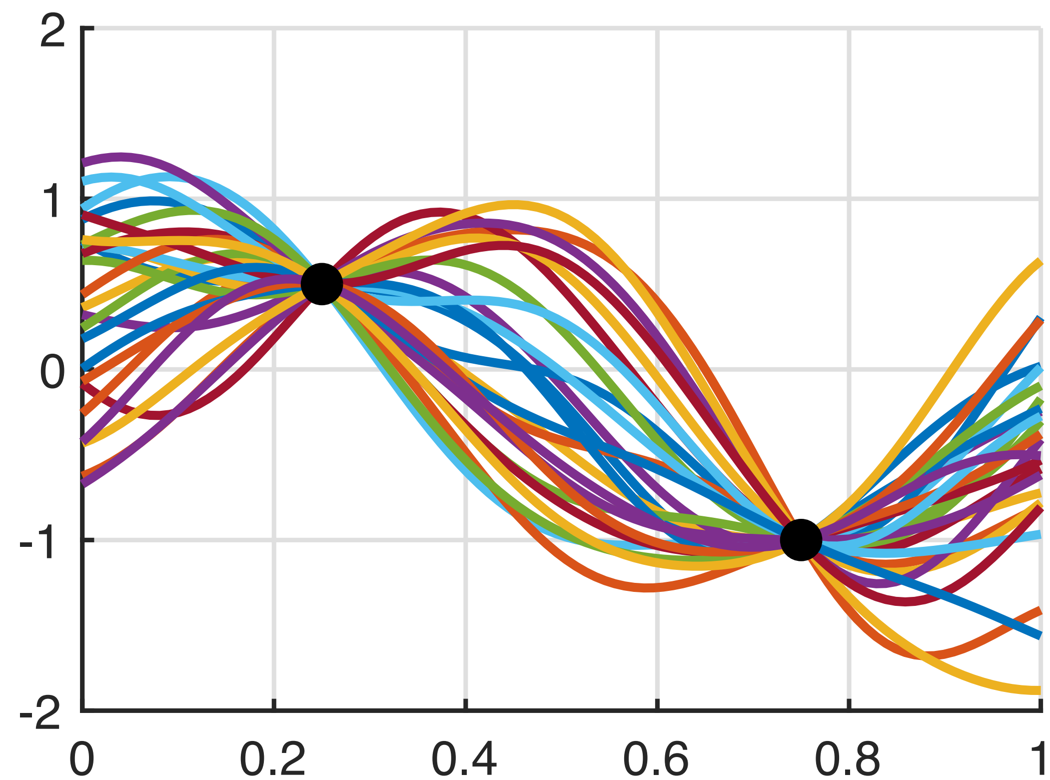
- Step 1: Definition of a GP prior distribution.

$$f \sim \mathcal{GP}$$




- Step 2: Conditioning on training data.

$$f \sim \mathcal{GP} \,|\, \mathbf{x}$$



# Bayesian Inference with GPs (in Formulas)

- Let $z$ be a GP with prior mean $\mu$ and variance $\Sigma$. Consider a training set $D = \{(x_i, y_i)\}_{i=1,\dots,N}$. The goal of Bayesian inference is to find:

$$\hat{z} = z \,|\, D$$

- For GPs this can be done analytically, obtaining a GP with posteriori mean and variance given by:

$$\hat{\mu}(x^*) = \mu(x^*) + \Sigma_{x^*,\mathcal{D}}\Sigma_{\mathcal{D},\mathcal{D}}^{-1}(\mathbf{y} - \mu_{\mathcal{D}})$$

$$\hat{\Sigma}_{x^*,x^*} = \Sigma_{x^*,x^*} - \Sigma_{x^*,\mathcal{D}}\Sigma_{\mathcal{D},\mathcal{D}}^{-1}\Sigma_{x^*,\mathcal{D}}^{T}$$

# Problem Formulation

# Probabilistic Invariance

- Probabilistic generalisation of problem associated with existence of local adversarial examples.

- Intuitively, we want to count the number of functions extracted from the GP for which deterministic invariance does not hold.

# Probabilistic Invariance

- Probabilistic generalisation of problem associated with existence of local adversarial examples.

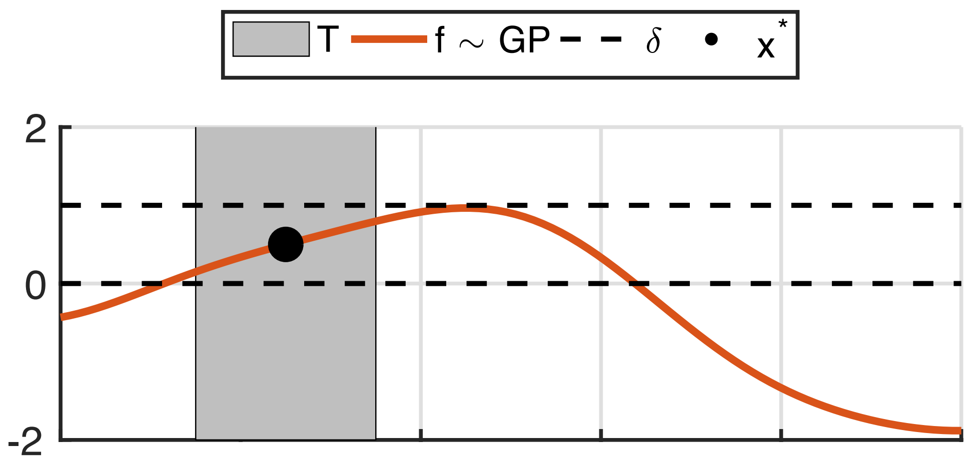
- Intuitively, we want to count the number of functions extracted from the GP for which deterministic invariance does not hold.

Consider $x^*$ and a neighbourhood $T$. Let $\delta$ be the adversarial threshold, then invariance probability is defined by:

$$\phi(x^*, T, \delta) = P(\exists x' \in T \; s.t. \, ||\hat{z}(x') - \hat{z}(x^*))|| > \delta)$$

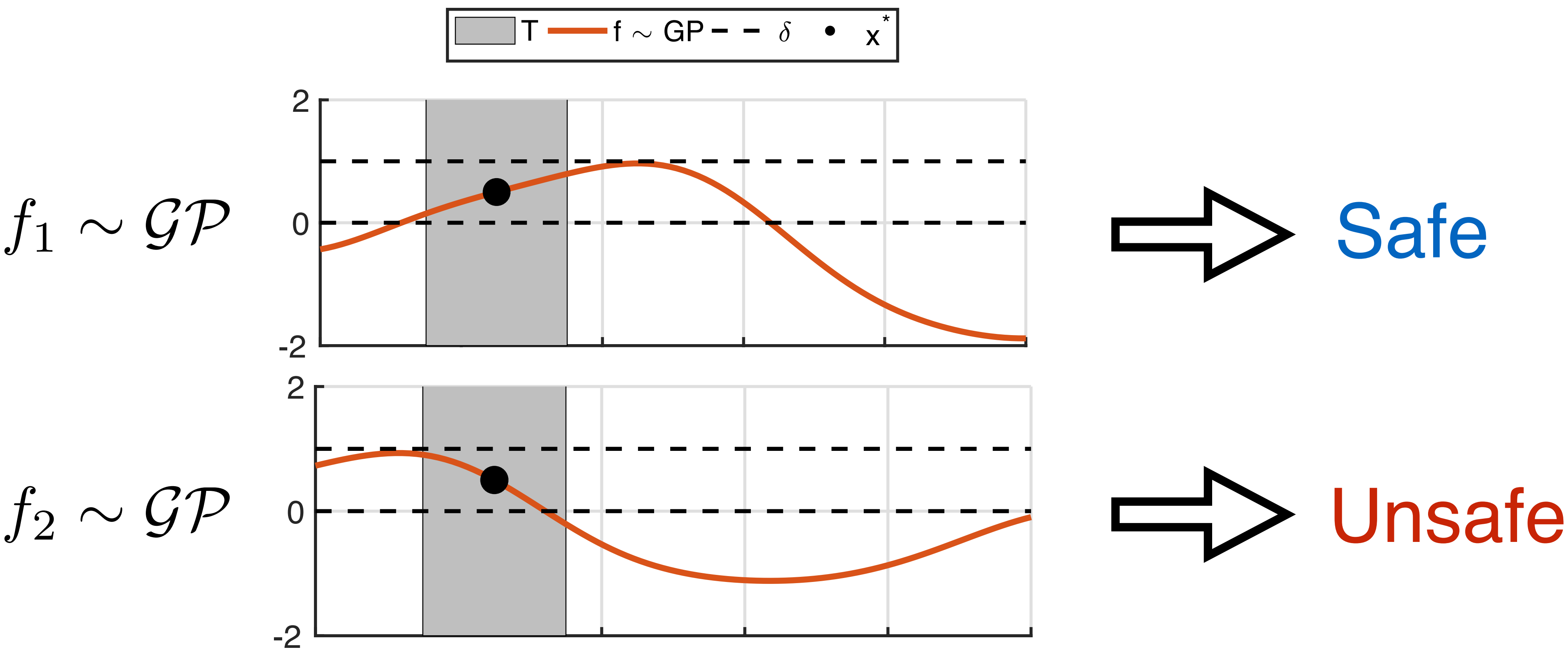
# Probabilistic Invariance (in Figures)

$f_1 \sim \mathcal{GP}$

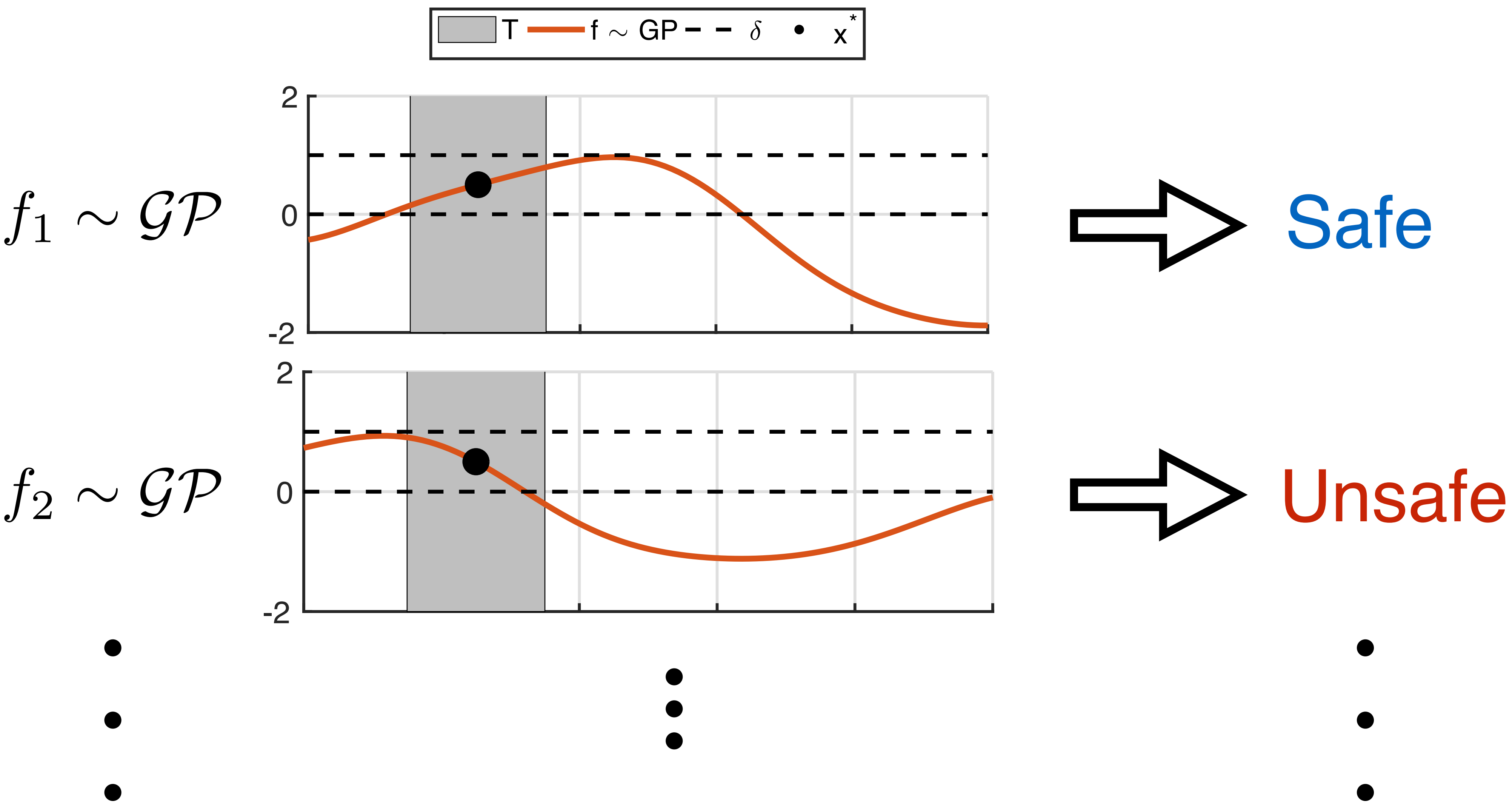



Safe

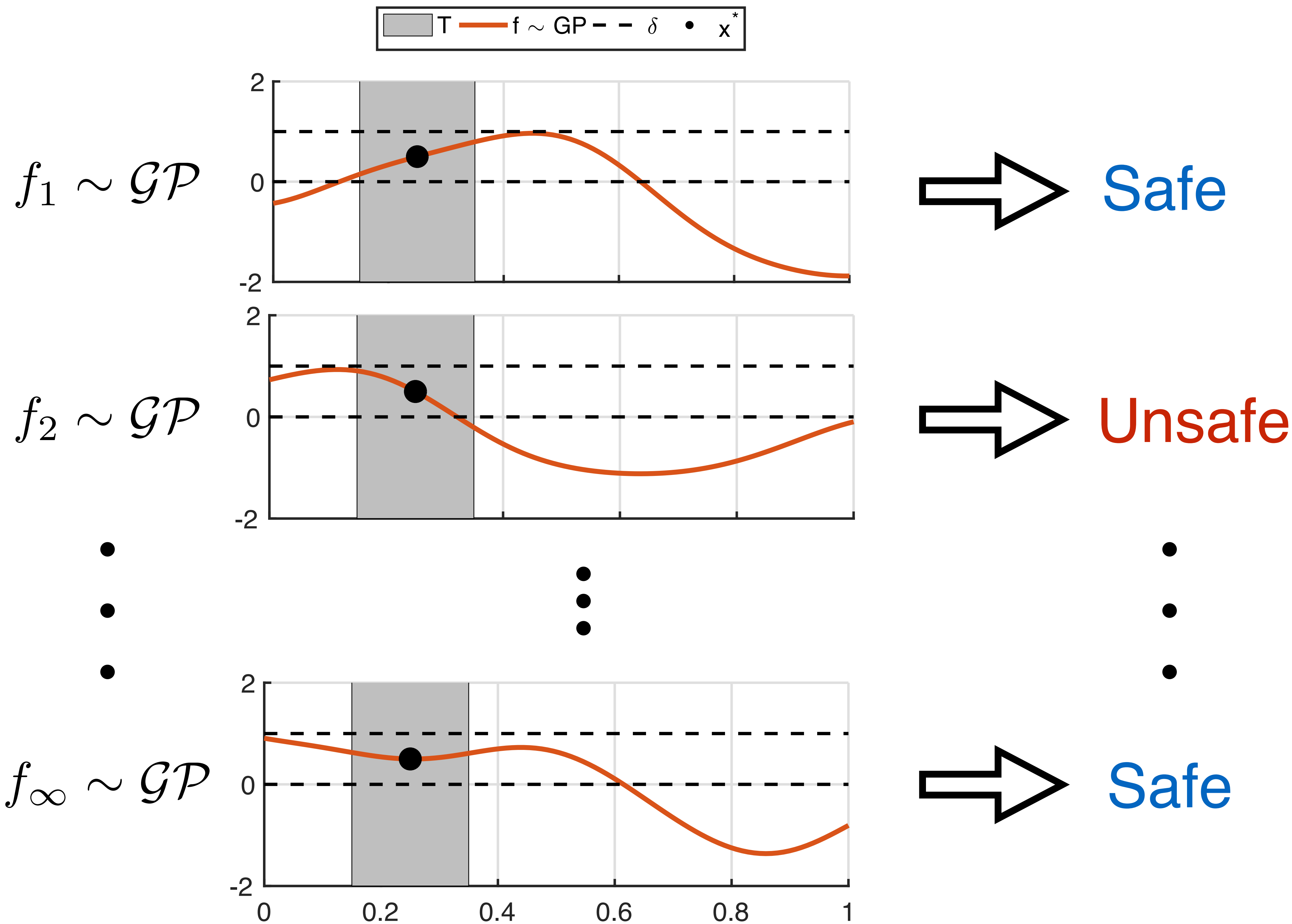
# Probabilistic Invariance (in Figures)



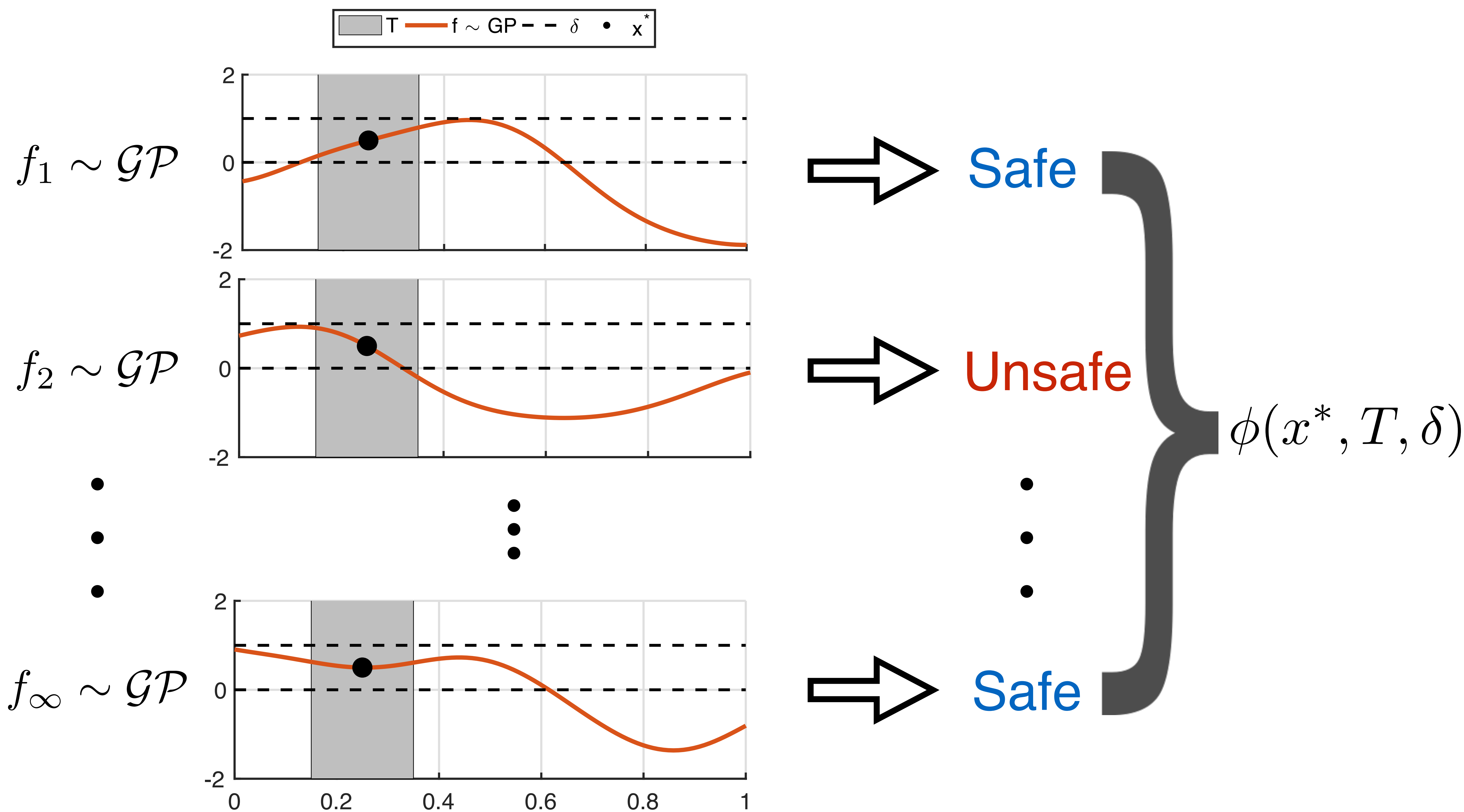
# Probabilistic Invariance (in Figures)



# Probabilistic Invariance (in Figures)



# Probabilistic Invariance (in Figures)



# Methods

# Upper-bound on Invariance

- Computation of $\phi(x^*, T, \delta)$ is far from trivial as it involves solutions of uncountably many optimisation problems. Instead, we compute a safe approximation.

# Upper-bound on Invariance

- Computation of $\phi(x^*, T, \delta)$ is far from trivial as it involves solutions of uncountably many optimisation problems. Instead, we compute a safe approximation.

Theorem 1: For every output dimension $i$ let:

$$\eta_i = \frac{\delta \, - \, sup_{x \in T} |\mu^o(x^*, x)|_1}{n} - 12 \int_0^{\frac{1}{2} sup_{x_1, x_2 \in T} d_{x^*}^{(i)}(x_1, x_2)} \sqrt{ln\left( (\frac{\sqrt{m} K_{x^*}^{(i)} D}{z} + 1)^m \right)} dz$$

Then:

$$\phi(x^*, T, \delta | \mathcal{D}) \leq \hat{\phi}(x^*, T, \delta | \mathcal{D}) := 2 \sum_{i=1}^{n} e^{-\frac{\bar{\eta}_i^2}{2\xi^{(i)}}}$$

# Upper-bound on Invariance

Theorem 1: For every output dimension $i$ let:

Maximum mean difference

$$\eta_i = \frac{\delta - \boxed{sup_{x \in T} |\mu^o(x^*, x)|_1}}{n} -$$

$$12 \int_0^{\frac{1}{2} sup_{x_1, x_2 \in T} d_{x^*}^{(i)}(x_1, x_2)} \sqrt{ln \left( (\frac{\sqrt{m} K_{x^*}^{(i)} D}{z} + 1)^m \right)} dz$$

Then:

$$\phi(x^*, T, \delta | \mathcal{D}) \leq \hat{\phi}(x^*, T, \delta | \mathcal{D}) := 2 \sum_{i=1}^{n} e^{-\frac{\bar{\eta}_i^2}{2\xi^{(i)}}}$$

# Upper-bound on Invariance

Theorem 1: For every output dimension $i$ let:

Maximum mean difference

Lipschitz constant on Variance

$$\eta_i = \frac{\delta - \boxed{sup_{x \in T} |\mu^o(x^*, x)|_1}}{n} - 12 \int_0^{\frac{1}{2} sup_{x_1, x_2 \in T} d_{x^*}^{(i)}(x_1, x_2)} \sqrt{ln\left( \left(\frac{\sqrt{m} \boxed{K_{x^*}^{(i)}} D}{z} + 1\right)^m \right)} dz$$

Then:

$$\phi(x^*, T, \delta | \mathcal{D}) \leq \hat{\phi}(x^*, T, \delta | \mathcal{D}) := 2 \sum_{i=1}^{n} e^{-\frac{\bar{\eta}_i^2}{2\xi^{(i)}}}$$

# Upper-bound on Invariance

Theorem 1: For every output dimension $i$ let:

Maximum mean difference

Lipschitz constant on Variance

$$\eta_i = \frac{\delta - \boxed{sup_{x \in T} |\mu^o(x^*, x)|_1}}{n} - 12 \int_0^{\frac{1}{2} sup_{x_1, x_2 \in T} d_{x^*}^{(i)}(x_1, x_2)} \sqrt{ln\left( \left( \frac{\sqrt{m} \boxed{K_{x^*}^{(i)}} D}{z} + 1 \right)^m \right)} dz$$

Then:

$$\phi(x^*, T, \delta | \mathcal{D}) \leq \hat{\phi}(x^*, T, \delta | \mathcal{D}) := 2 \sum_{i=1}^{n} e^{-\frac{\bar{\eta}_i^2}{\boxed{2\xi^{(i)}}}}$$

Maximum correlation

# Proof Sketch

- We want to upper-bound:

$$\phi(x^*, T, \delta | \mathcal{D}) = P\left(\sup_{x \in T} ||z(x) - z(x^*)|| > \delta\right)$$

- Since $z^o(x^*, x) = z(x^*) - z(x)$ is still a GP we can employ the Borell-TIS inequality, which upper-bounds the supremum:

$$P\left(sup_{x \in T} ||z^o(x^*, x)|| > \delta\right) \leq e^{\frac{\left(\delta - E[\sup_{x \in T} z^o(x^*, x)]\right)^2}{2\sigma_T^2}}$$

- Finally, $E[\sup_{x \in T} z^o(x^*, x)]$ can be over-approximated using the Dudley entropy integral.

# Constant Computation

- The upper-bound computation requires computation of different constants e.g.:

$$\sup_{x \in T} \mu(x^*) - \mu(x) = \mu(x^*) - \inf_{x \in T} \mu(x) = \mu(x^*) - \inf_{x \in T} \Sigma_{x,\mathcal{D}} \Sigma^{-1}_{\mathcal{D},\mathcal{D}} \mathbf{y}$$

- We define two functions $\varphi$ and $\psi$ that decompose the GP variance as: $\Sigma_{x,x_i} = \psi\left(\varphi\left(x, x_i\right)\right)$.
- Using interval analysis on $\varphi$ and optimising $\psi$ we can compute lower and upper bounds on each $\Sigma_{x,x_i}$
- Thanks to linearity, we propagate these to get bounds on the sup; and refine via Branch and Bound.

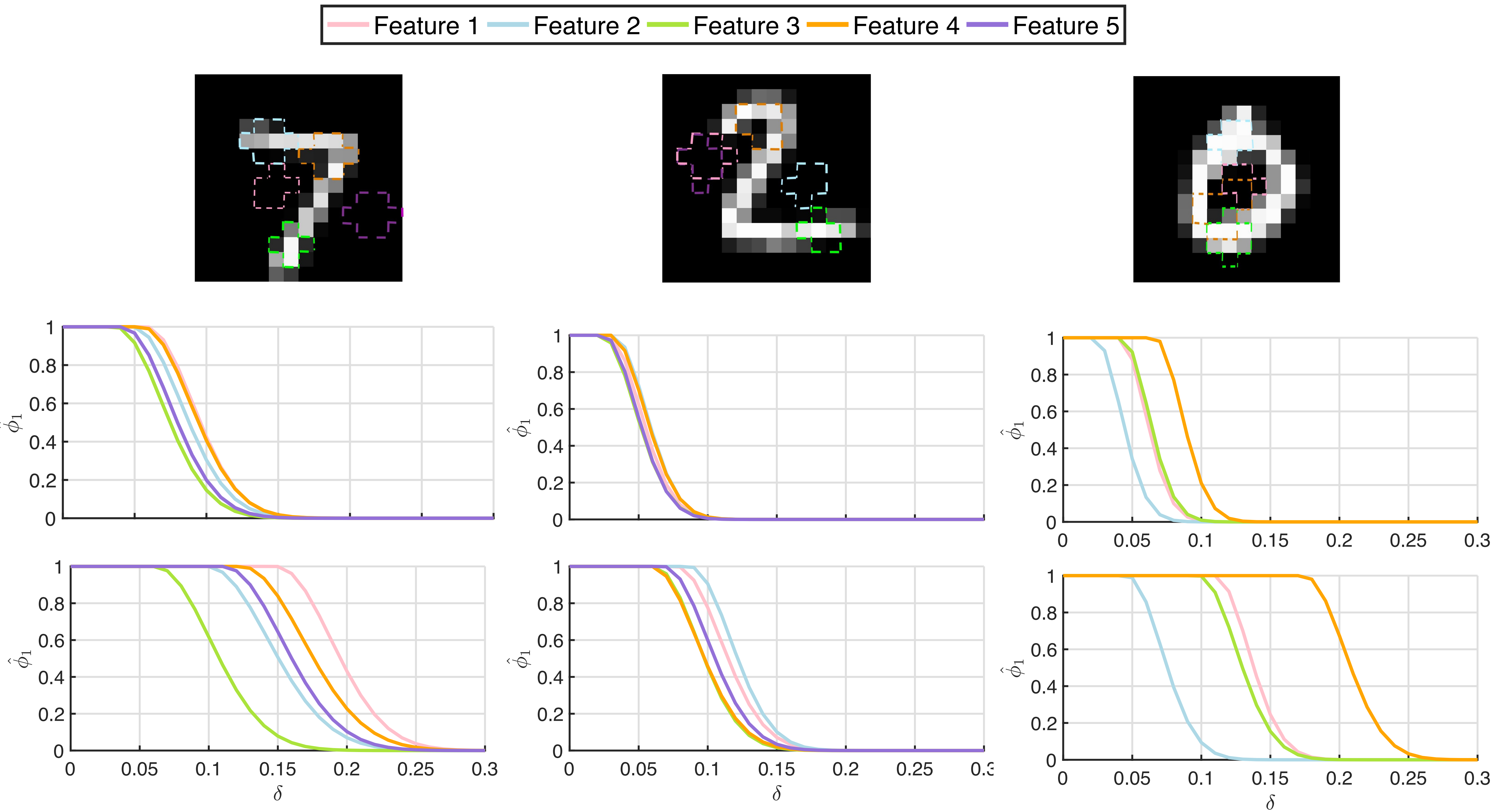
# Case of Study

# GPs and Neural Networks: Experimental Settings

- Bayesian fully-connected neural networks converge in distribution to specific GPs, as the number of neurons approaches infinity*.
- We can employ the method we developed to perform empirical analysis of fully connected NNs.
- We focus on ReLU NNs applied to the MNIST dataset.
- For scalability, we provide feature-level analysis using SIFT.

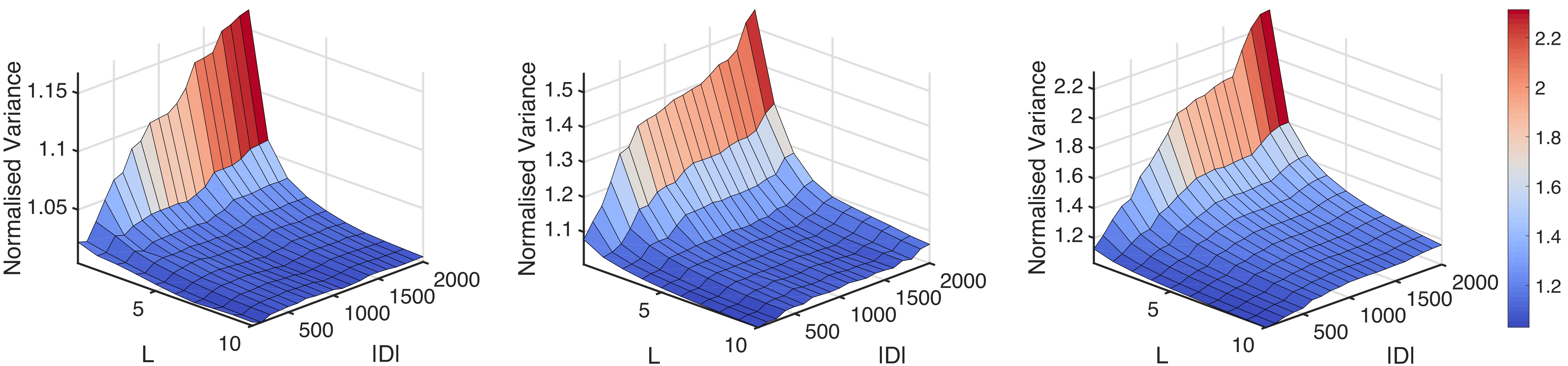
*Neal, R. M. Bayesian learning for neural networks. Springer, 2012.

# Parametric Analysis on Adversarial Thresholds



# Parametric Analysis on Variance

Analysis of how variance changes in *T* depending on number of training samples and layers.



# Conclusions

- We developed a formal approach for invariance analysis of Bayesian inference with Gaussian Processes.
- Developed an algorithmic approach for computation of upper-bound on invariance probability.
- We relied on the relationship between Bayesian NNs and GPs, to analyse NN behaviour at infinity width limit.
- Provided experimental results on MNIST.