

TU3A-1

Constrained Gaussian Process for Signal Integrity applications using Variational Inference

T. Nguyen[#], B. Shi[#], H. Ma^{*}, E. Li^{*},
A. Cangellaris[#], J. Schutt-Aine[#]

[#]University of Illinois Urbana-Champaign, USA

^{*}Zhejiang University, China

Outline

- Why surrogate modeling is needed for SPI problems
- Review: Linear regression – Bayesian point of view
- Gaussian Process
- Proposed method: bounded-output Gaussian Process via Variational Inference
- Example
- Conclusion

Surrogate modeling for SI

- Signal integrity is expensive:
 - Solving large scale EM models
 - Extremely long transient alike simulation
- Use surrogate model instead

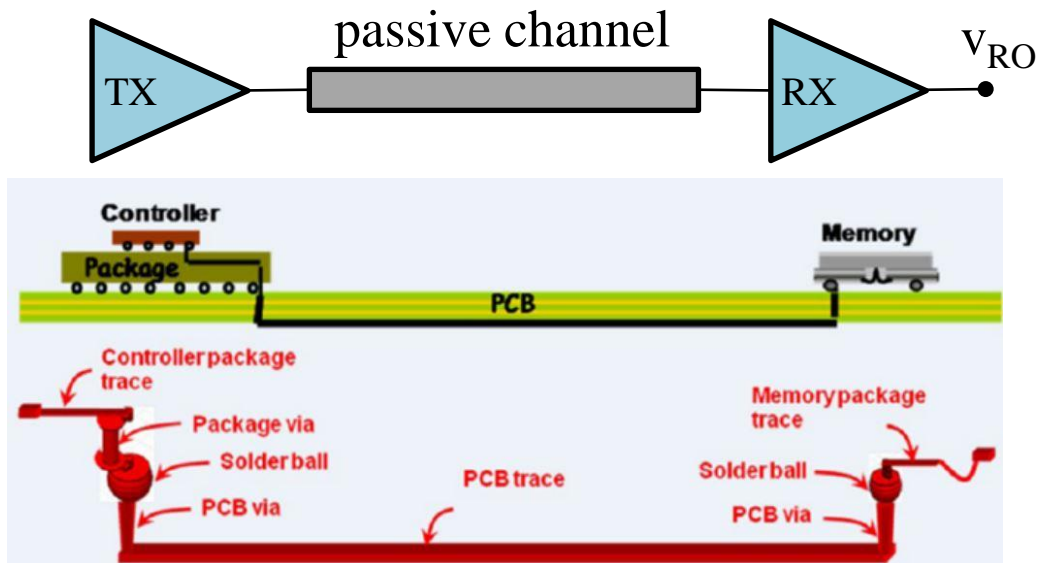
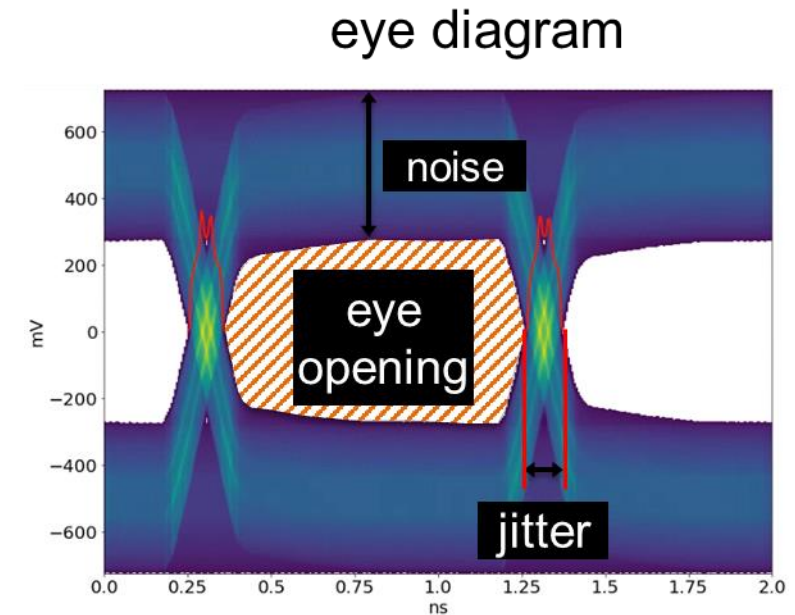


Image: ECE546 – UIUC



Surrogate modeling for SI

- Signal integrity is expensive:
 - Solving large scale EM models
 - Extremely long transient alike simulation
- Use surrogate model instead

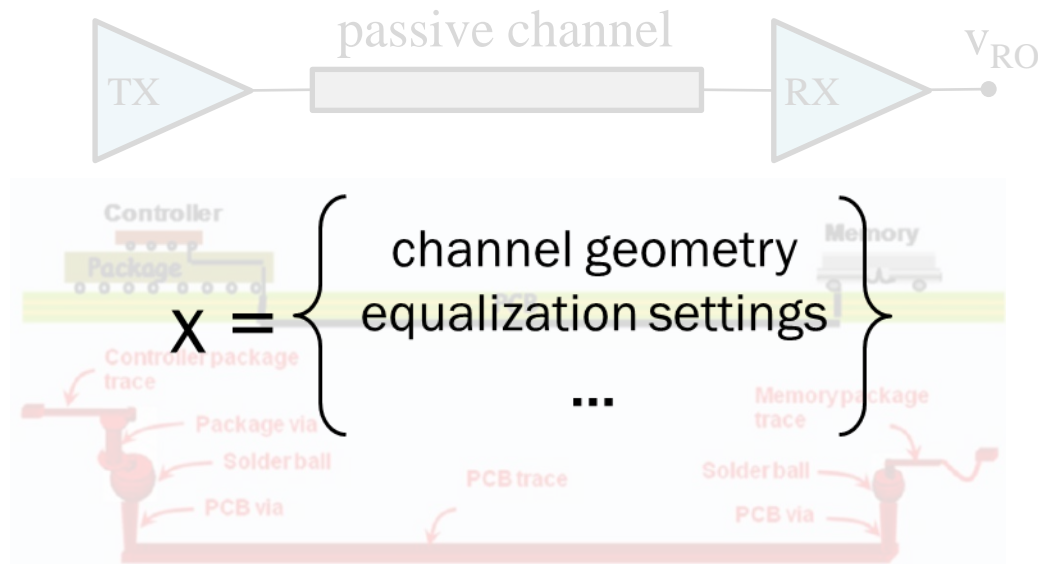

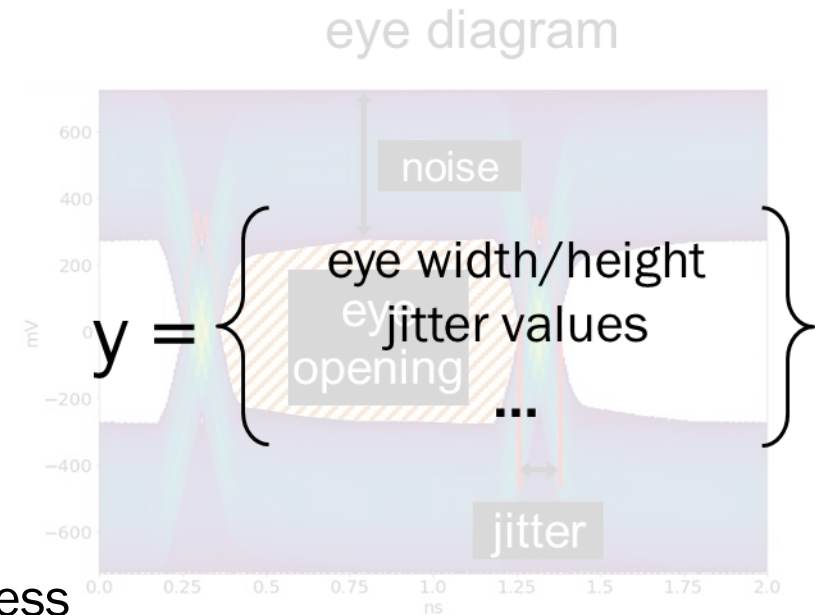
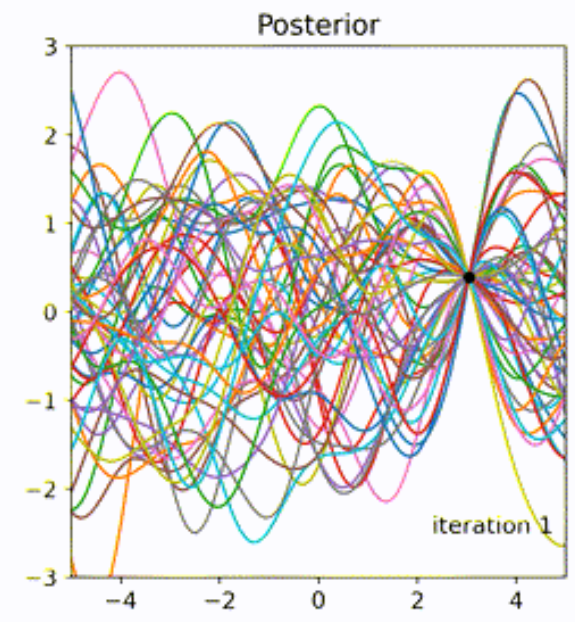
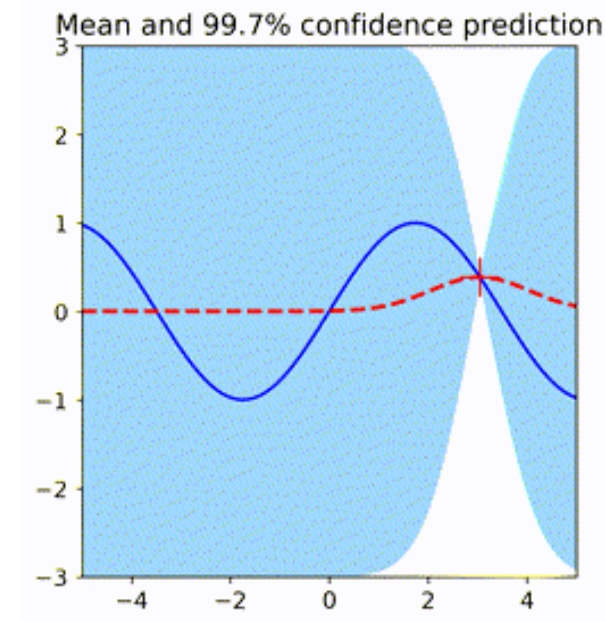
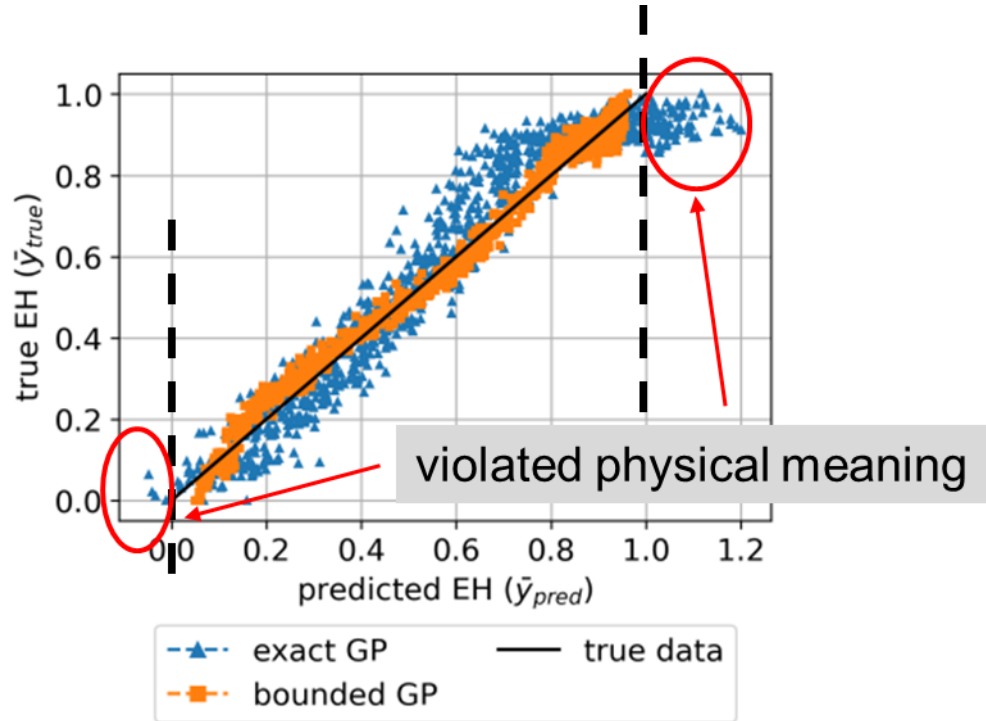


Image: ECE546 – UIUC

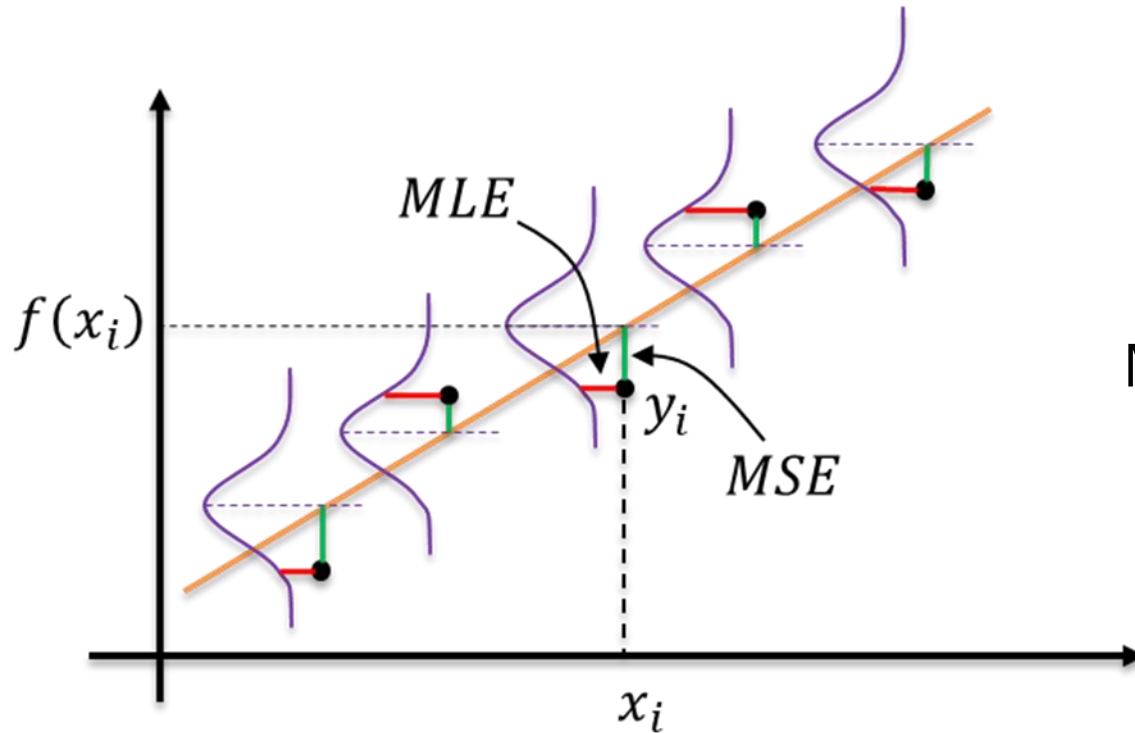
f

 Neural nets
 SVR
 PC
 Gaussian Process



- GP is a set of points which sampled from a multi-dim Gaussian
 - GP: $f: D \mapsto \mathbb{R}$
- The need for constrained GP: outputs are bounded by physical meanings



- Review: Linear regression in Bayes' view



$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

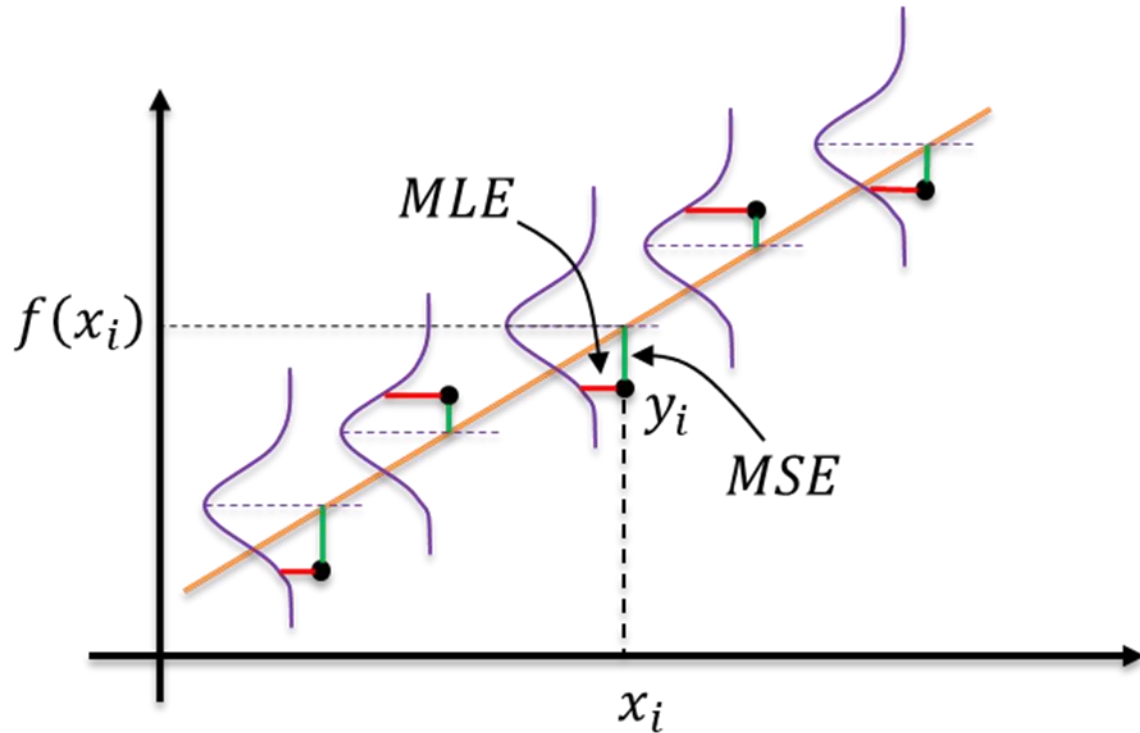
$$y = f(\mathbf{x}) + \epsilon_d$$

$$f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} = \sum_{k=1}^d x_k \theta_k$$

$$\text{MLE: } \underset{\boldsymbol{\theta}}{\operatorname{argmax}} p(\mathcal{D} | \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^N p(y^{(i)} | \mathbf{x}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \leftarrow \text{Point estimate}$$

- Review: Linear regression in Bayes' view



$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y = f(\mathbf{x}) + \epsilon$$

$$f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} = \sum_{k=1}^d x_k \theta_k$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(0, \frac{1}{d} \Sigma_{\boldsymbol{\theta}}\right) \leftarrow \text{Place a prior on } \boldsymbol{\theta}$$

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \leftarrow \text{Interval estimate}$$

- Sample from the posterior for prediction

$$p(y_* | \mathbf{x}_*, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}$$

- Relationship with Gaussian Process:
 - Use nonlinear mapping (feature map): φ
 - Posterior involves the term

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{d} \varphi(\mathbf{x})^T \Sigma_{\theta} \varphi(\mathbf{x}')$$

Kernel function

for any pair of input \mathbf{x} and \mathbf{x}'

- Mercer's theorem: **choose k** instead of φ

$$y = f(\mathbf{x}) + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$f(\mathbf{x}) = \varphi(\mathbf{x})^T \boldsymbol{\theta}$$

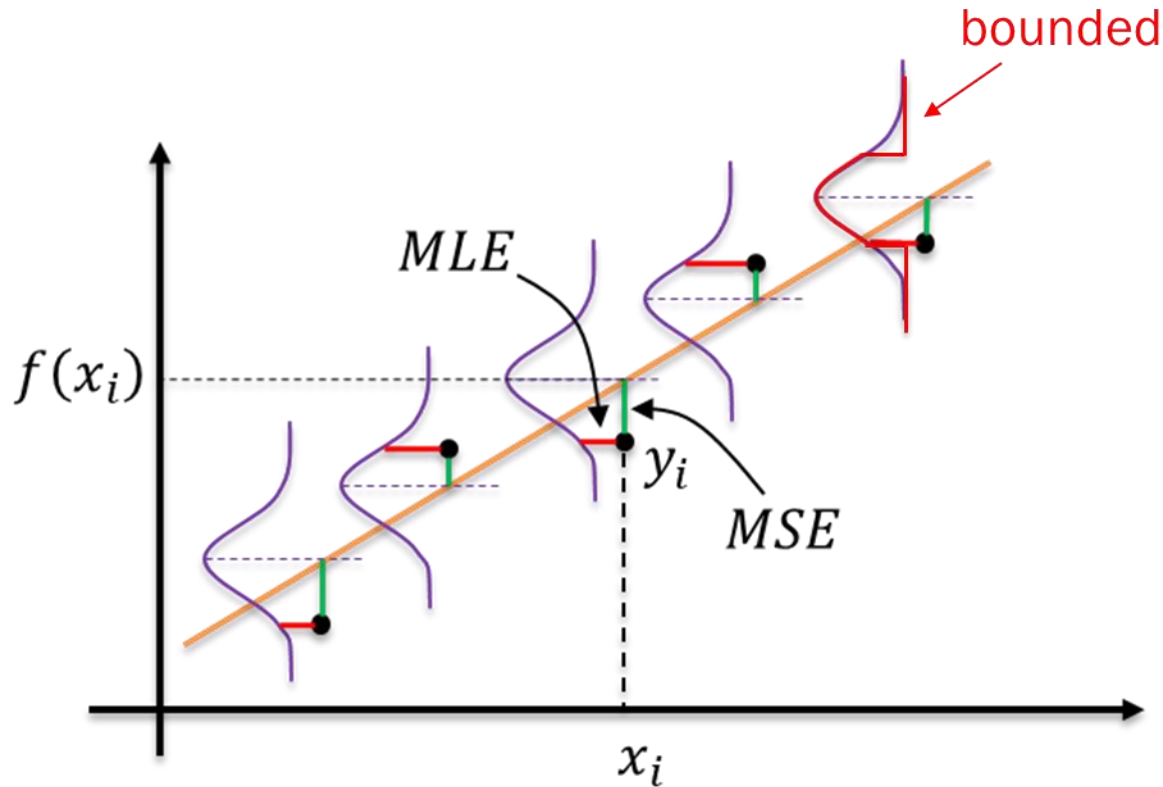
$$\boldsymbol{\theta} \sim \mathcal{N}\left(0, \frac{1}{d} \Sigma_{\theta}\right) \quad \leftarrow \text{Place a prior on } \boldsymbol{\theta}$$

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad \leftarrow \text{Interval estimate}$$

- Sample from the posterior for prediction

$$p(y_* | \mathbf{x}_*, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \quad \leftarrow \text{A Gaussian}$$

- Constrained GP



$$y = f(\mathbf{x}) + \epsilon$$

~~$\epsilon \sim \mathcal{N}(0, \sigma^2)$~~
Choose a bounded distribution

$$f(\mathbf{x}) = \varphi(\mathbf{x})^T \boldsymbol{\theta}$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(0, \frac{1}{d} \Sigma_{\boldsymbol{\theta}}\right)$$

← Place a prior on $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

← Intractable

- Sample from the posterior for prediction

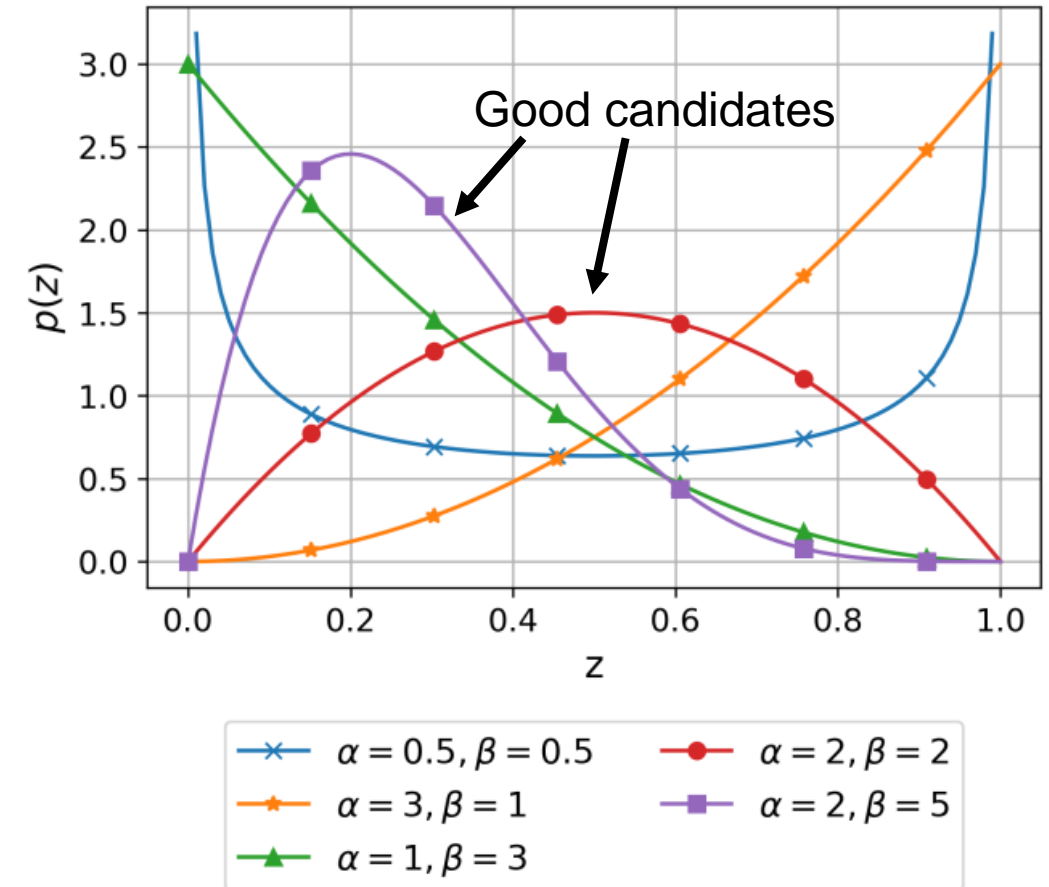
$$p(y_* | \mathbf{x}_*, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}$$

← Intractable

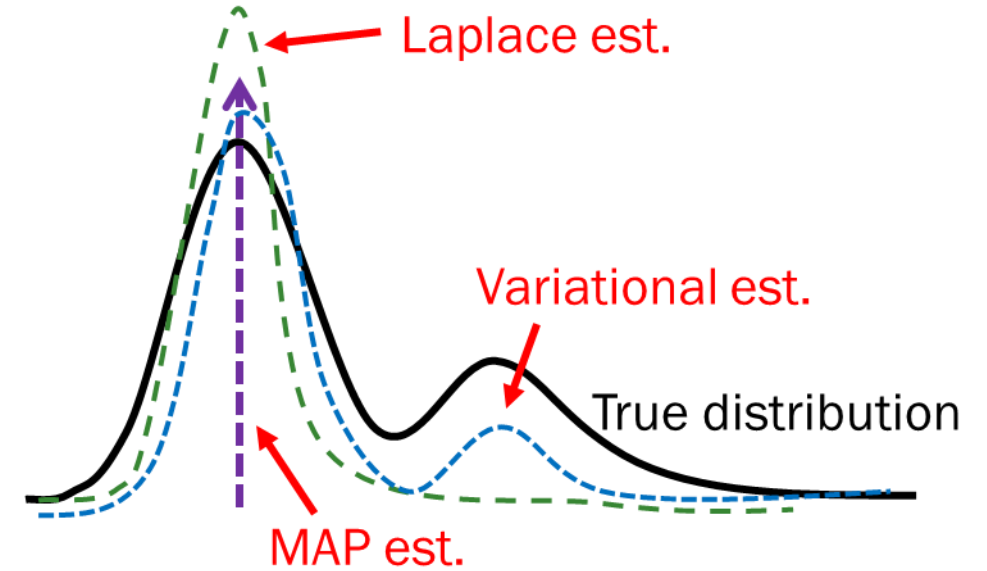
- Constrained GP:

- Choosing a bounded likelihood:
Beta distribution

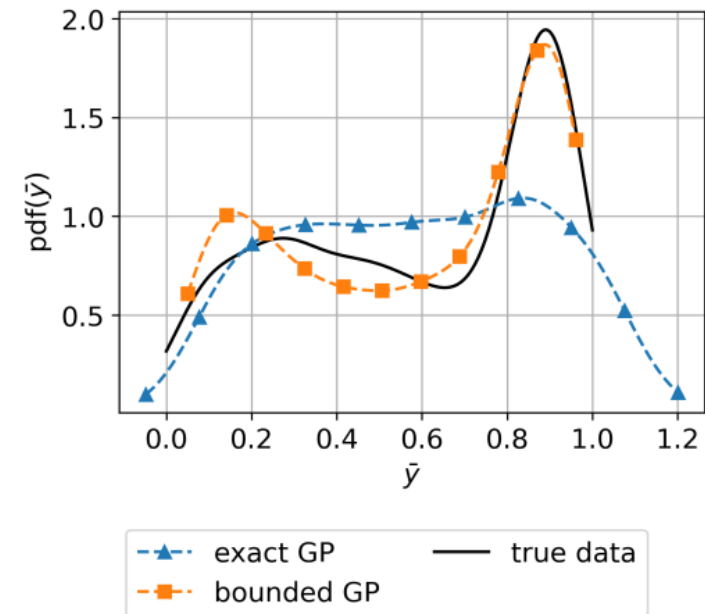
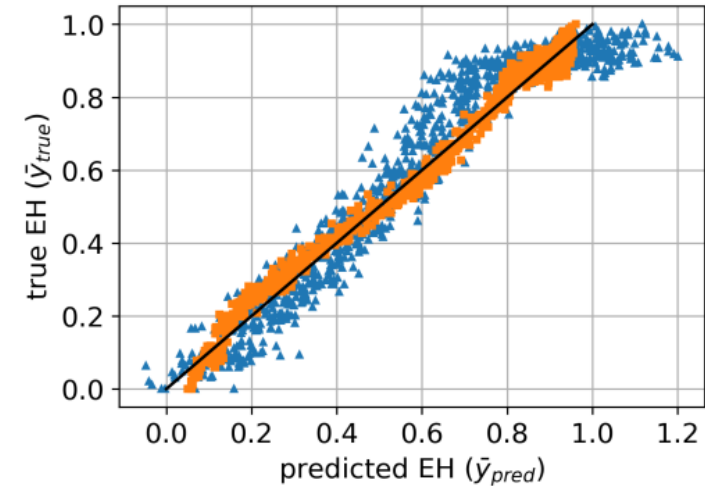
$$p(z|\alpha, \beta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1} (1-z)^{\beta-1}$$



- Constrained GP:
 2. Approximate the posterior:
 - Laplace appr. works well for unimodal distribution only.
 - Variational approach is more powerful
- Variational inference:
 - Use a variational distribution q to approximate the true distribution
 - Need to optimize some metrics: Kullback–Leibler (KL) divergence
 - Minimizing KL divergence is equivalent of maximizing evidence lower bound (ELBO)



- A D2D channel:
 - Inputs: channel geometry, RX EQ
 - Outputs: eye width, eye height
 - SGD with Adam optimizer was used to train.
 - RBF kernel, 50 training samples (uniformly sampled)
- Result:
 - Bounded GP does not violate physical meanings
 - Bounded GP has lower errors, converge sooner
 - Training time is only slightly longer (for the same number of epochs)



Conclusion

- GP was shown to outperform other surrogate models in certain tasks in SPI/RF microwave: fewer samples, lower errors.
- Modified GP can be done to enforce physical constraints in exchange for the analytical solution. In which case, an approximate GP implementation is needed.
- Performance of multi-output bounded GP is up next.

- Marginal likelihood (or the evidence)

$$\begin{aligned}
 p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^N p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}) \\
 &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} \left[y^{(i)} - [\mathbf{x}^{(i)}]^T \boldsymbol{\theta}\right]^2\right\} \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^N \left[y^{(i)} - [\mathbf{x}^{(i)}]^T \boldsymbol{\theta}\right]^2\right\} \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\frac{1}{2\sigma^2} \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}_{\ell(\boldsymbol{\theta})}\right\}
 \end{aligned}$$

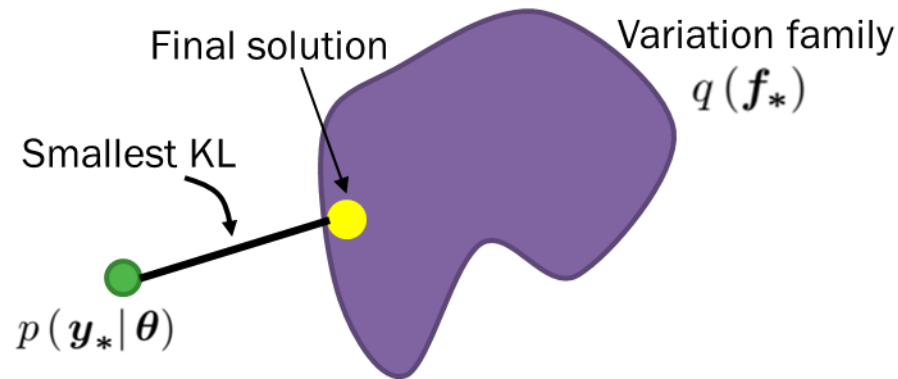
$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = 0$$

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

• ELBO derivation

$$\begin{aligned} \log p(\mathbf{y}_* | \boldsymbol{\theta}) &= \log \int p(\mathbf{y}_*, \mathbf{f}_* | \boldsymbol{\theta}) d\mathbf{f}_* \\ \text{Log evidence} &\geq \underbrace{\int q(\mathbf{f}_*) \log \frac{p(\mathbf{y}_*, \mathbf{f}_* | \boldsymbol{\theta})}{q(\mathbf{f}_*)} d\mathbf{f}_*}_{\mathcal{L}(\boldsymbol{\theta})} \end{aligned}$$

Evidence Lower Bound (ELBO)



$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \int q(\mathbf{f}_*) \log \frac{p(\mathbf{y}_*, \mathbf{f}_* | \boldsymbol{\theta})}{q(\mathbf{f}_*)} d\mathbf{f}_* \\ &= \int q(\mathbf{f}_*) \log \left[p(\mathbf{y}_* | \boldsymbol{\theta}) \frac{p(\mathbf{f}_* | \mathbf{y}_*, \boldsymbol{\theta})}{q(\mathbf{f}_*)} \right] d\mathbf{f}_* \\ &= \log p(\mathbf{y}_* | \boldsymbol{\theta}) \int q(\mathbf{f}_*) d\mathbf{f}_* \\ &\quad + \int q(\mathbf{f}_*) \log \frac{p(\mathbf{f}_* | \mathbf{y}_*, \boldsymbol{\theta})}{q(\mathbf{f}_*)} d\mathbf{f}_* \\ &= \log p(\mathbf{y}_* | \boldsymbol{\theta}) - \underbrace{\left[- \int q(\mathbf{f}_*) \log \frac{p(\mathbf{f}_* | \mathbf{y}_*, \boldsymbol{\theta})}{q(\mathbf{f}_*)} d\mathbf{f}_* \right]}_{KL(q(\mathbf{f}_*) || p(\mathbf{f}_* | \mathbf{y}_*, \boldsymbol{\theta}))} \end{aligned}$$

Predictive posterior