ROBUST EMPIRICAL BAYES FOR GAUSSIAN PROCESSES

Masha Naslidnyk
University College London

Gaussian Process Summer School, 13th of September 2023

CHALLENGE

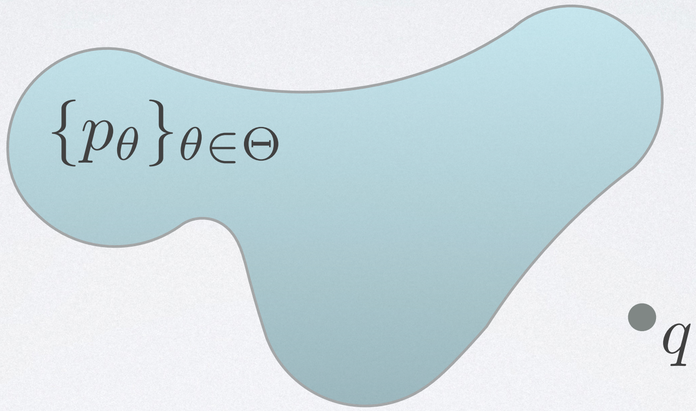
— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

Domain expertise
Bespoke models
Small, well-curated datasets

Black-box problems
Flexible models
Large, complex datasets

CHALLENGE

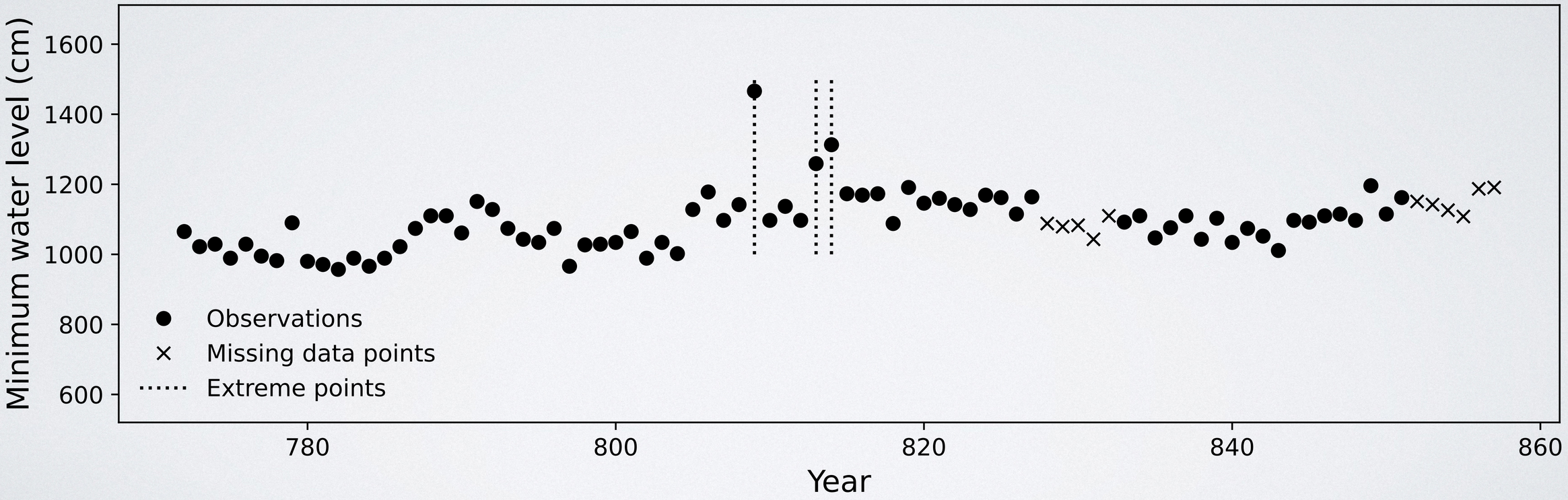
— Paradigm shift in Bayesian inference exacerbates **model misspecification** —



- Under misspecification, uncertainty quantification becomes brittle, especially in nonparametric models like **Gaussian Processes** (GPs).

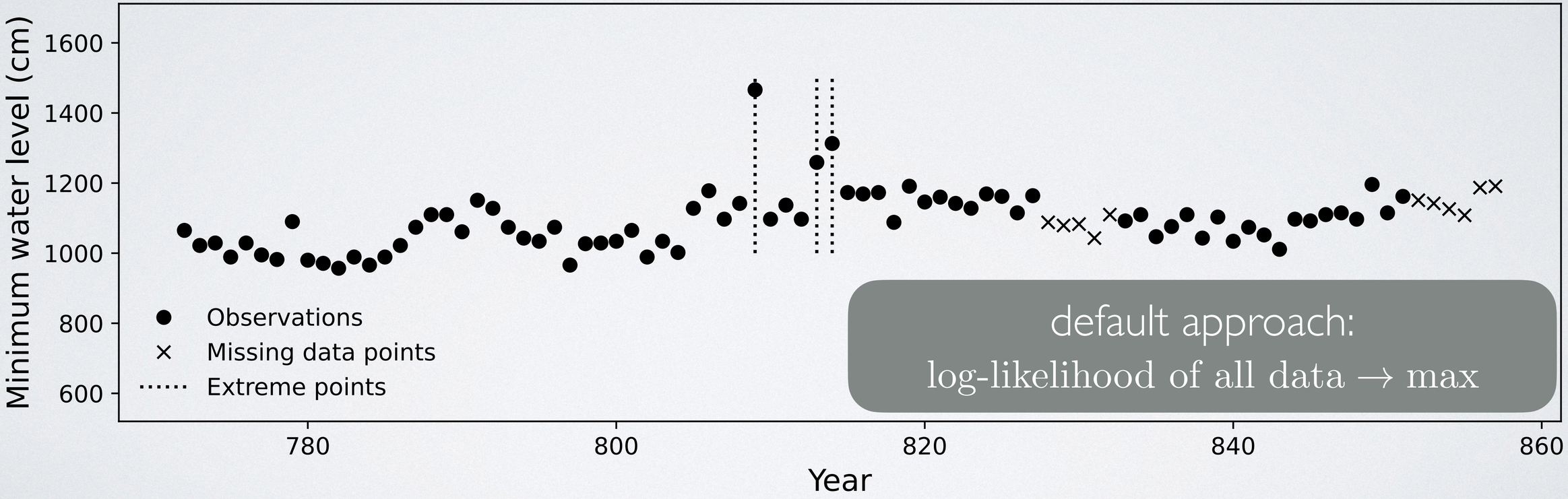
CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —



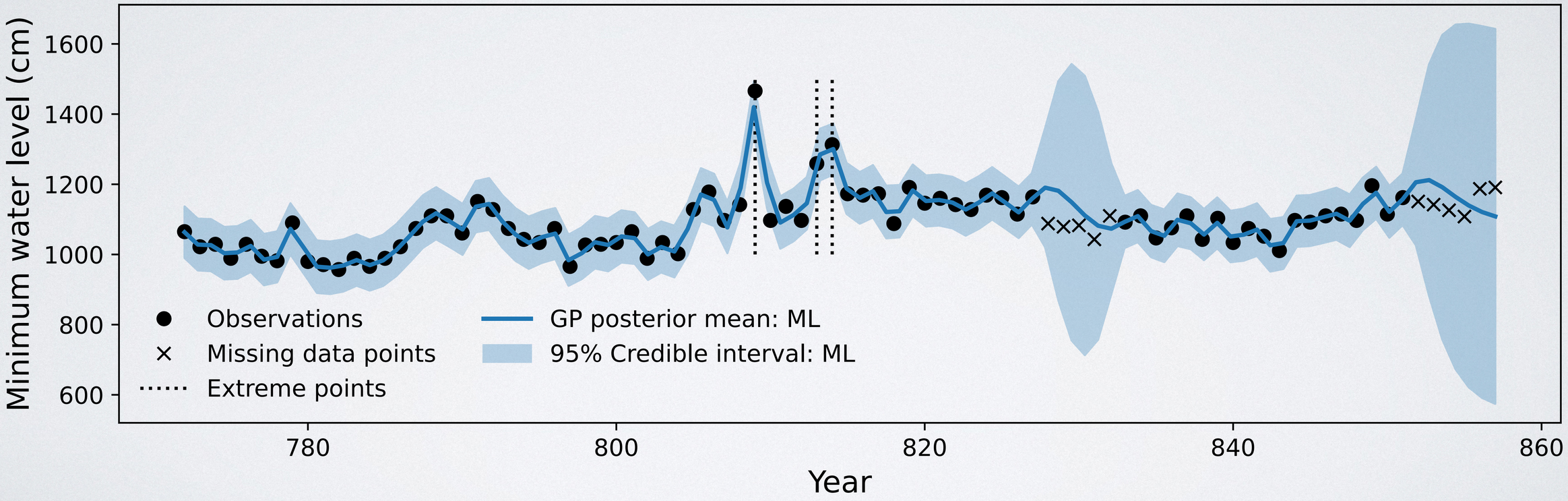
CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —



CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —



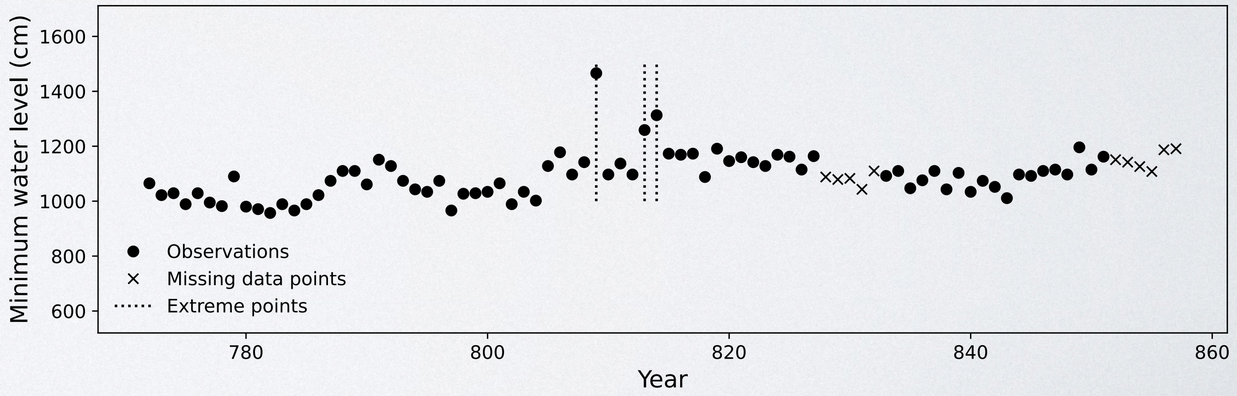
TYPES OF MISSPECIFICATION

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

TYPES OF MISSPECIFICATION

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

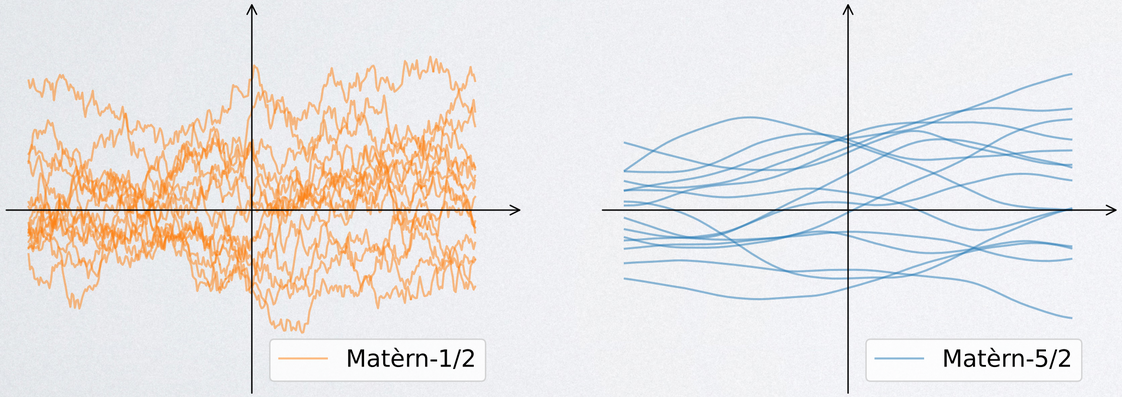
📌 Likelihood misspecification



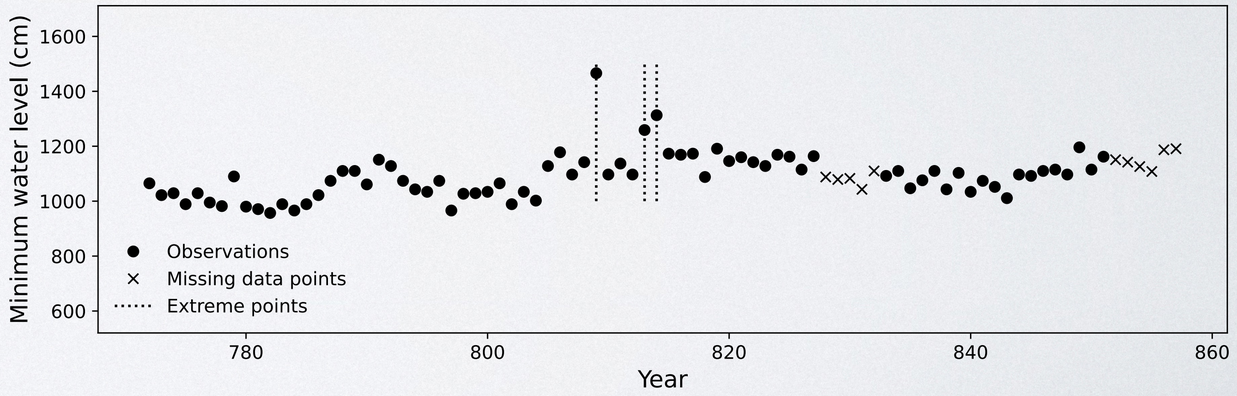
TYPES OF MISSPECIFICATION

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

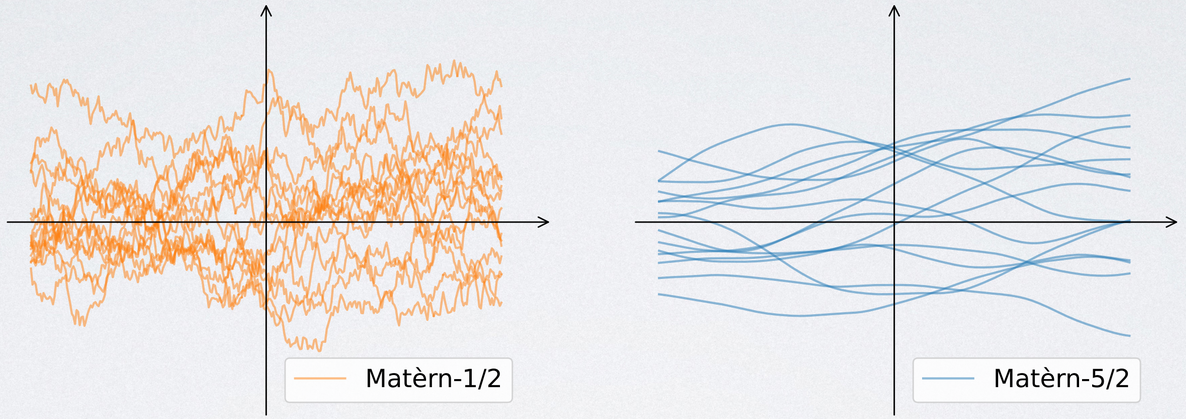
Smoothness misspecification



Likelihood misspecification



SMOOTHNESS MISSPECIFICATION



- Promising empirical results for leave-one-out cross-validation!

$$\sum_i \text{log-likelihood of datapoint } i \text{ given the rest of the data} \to \max$$

(instead of $\text{log-likelihood of all data} \to \max$)

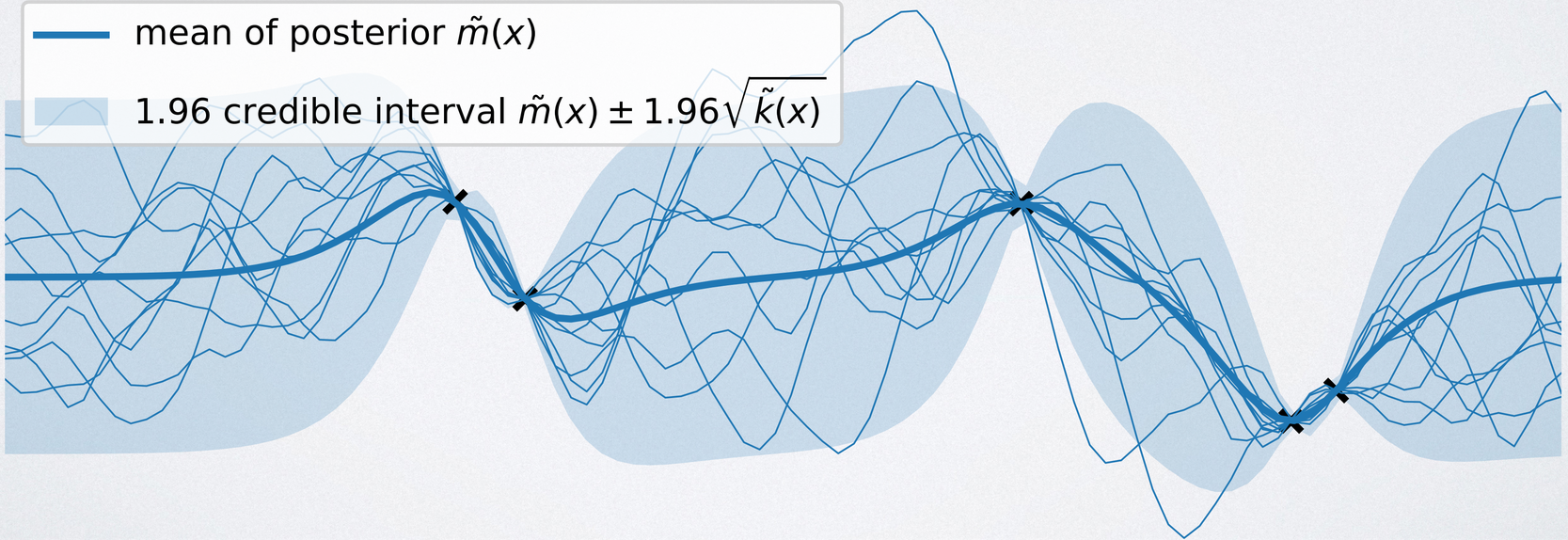
Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper- parameters of Gaussian processes with model misspecification.

CREDIBLE INTERVALS

- α-credible interval quantifies uncertainty of the model:

$$[\tilde{m}(x) - \alpha\sqrt{\tilde{k}(x)}, \tilde{m}(x) + \alpha\sqrt{\tilde{k}(x)}]$$

(for α=1.96, 95% of samples will lie within the interval)

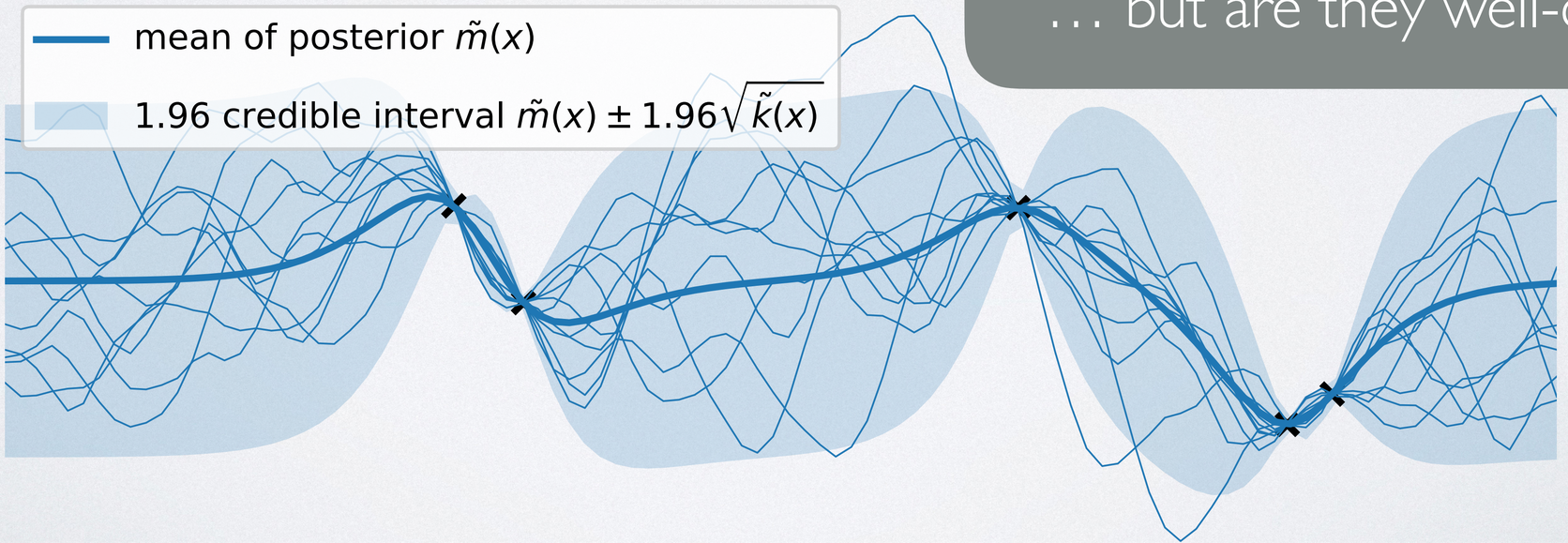


CREDIBLE INTERVALS

- α-credible interval quantifies uncertainty of the model:

$$[\tilde{m}(x) - \alpha\sqrt{\tilde{k}(x)}, \tilde{m}(x) + \alpha\sqrt{\tilde{k}(x)}]$$

… but are they well-calibrated?

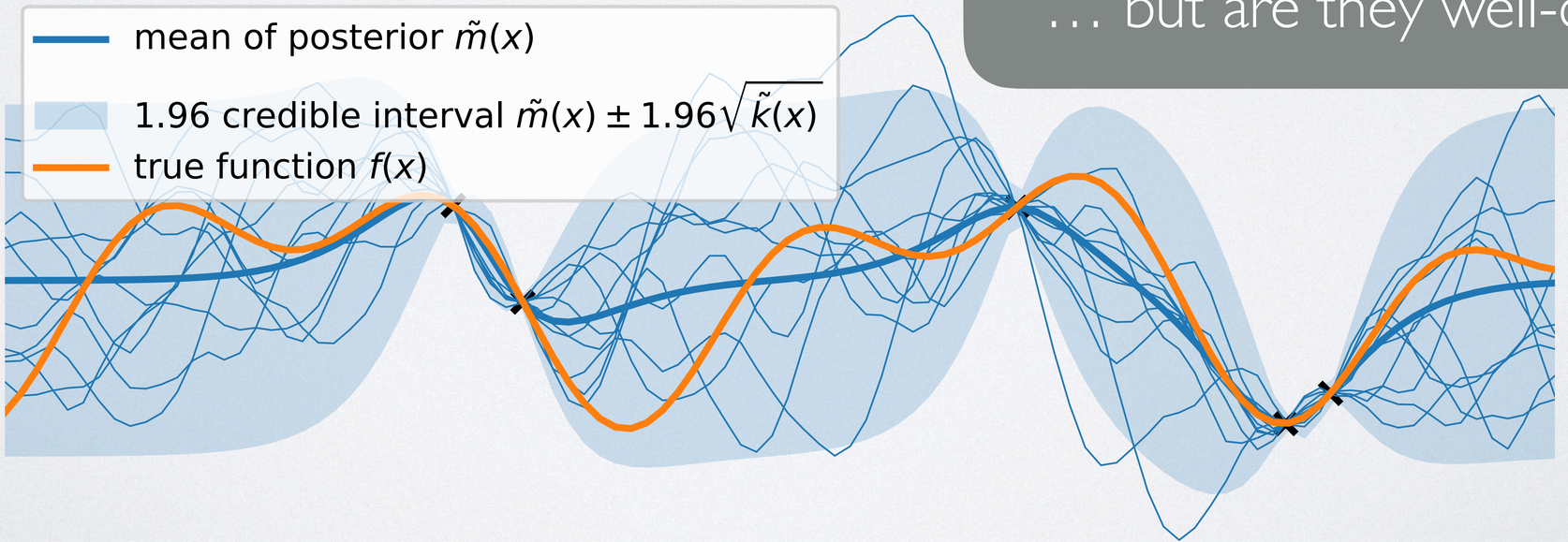


CREDIBLE INTERVALS

- α-credible interval quantifies uncertainty of the model:

$$[\tilde{m}(x) - \alpha\sqrt{\tilde{k}(x)}, \tilde{m}(x) + \alpha\sqrt{\tilde{k}(x)}]$$

… but are they well-calibrated?

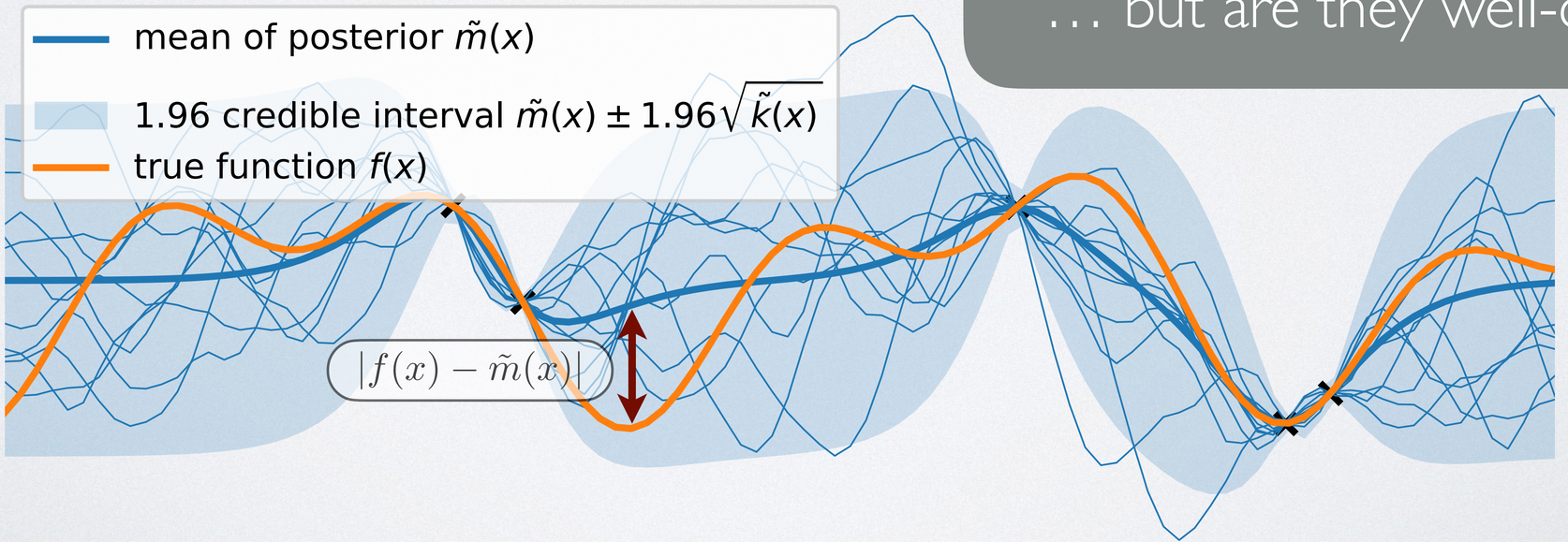


CREDIBLE INTERVALS

- α-credible interval quantifies uncertainty of the model:

$$[\tilde{m}(x) - \alpha\sqrt{\tilde{k}(x)}, \tilde{m}(x) + \alpha\sqrt{\tilde{k}(x)}]$$

… but are they well-calibrated?



CREDIBLE INTERVALS

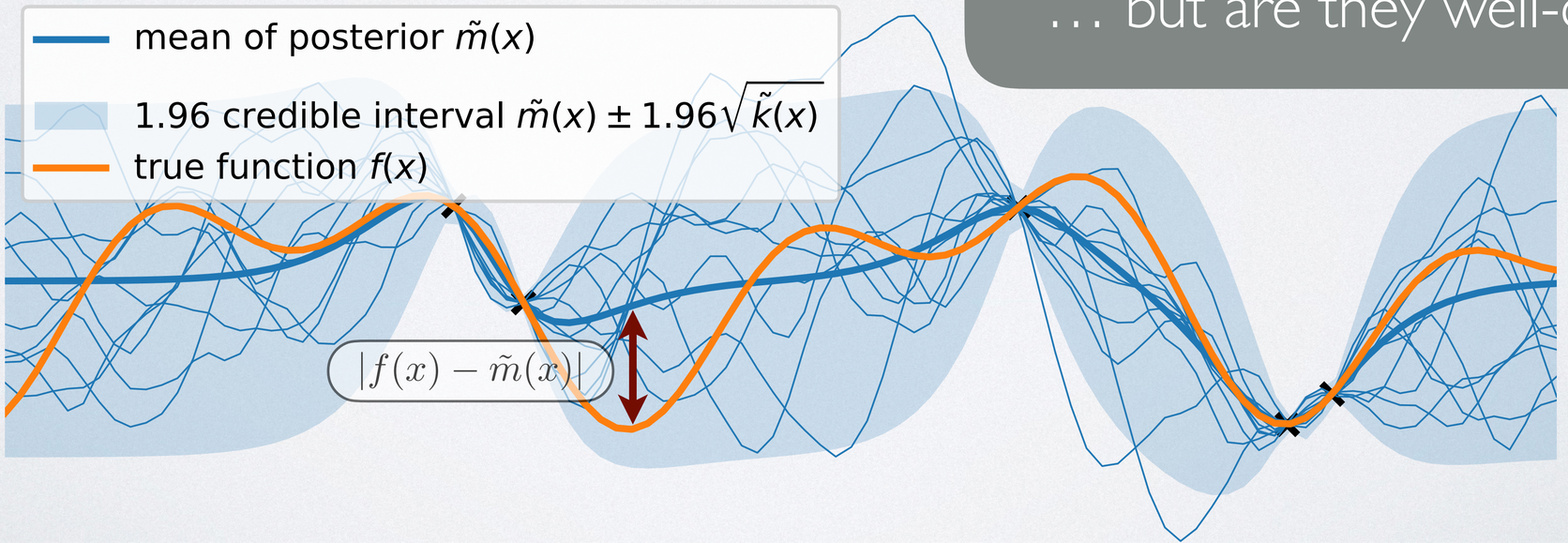
- Underconfidence:

$$\frac{|\tilde{m}(x) - f(x)|}{\sqrt{\tilde{k}(x)}} \to 0$$

- Overconfidence:

$$\frac{|\tilde{m}(x) - f(x)|}{\sqrt{\tilde{k}(x)}} \to \infty$$

… but are they well-calibrated?



CREDIBLE INTERVALS

- Underconfidence:

$$\frac{|\tilde{m}(x) - f(x)|}{\sqrt{\tilde{k}(x)}} \to 0$$

- Overconfidence:

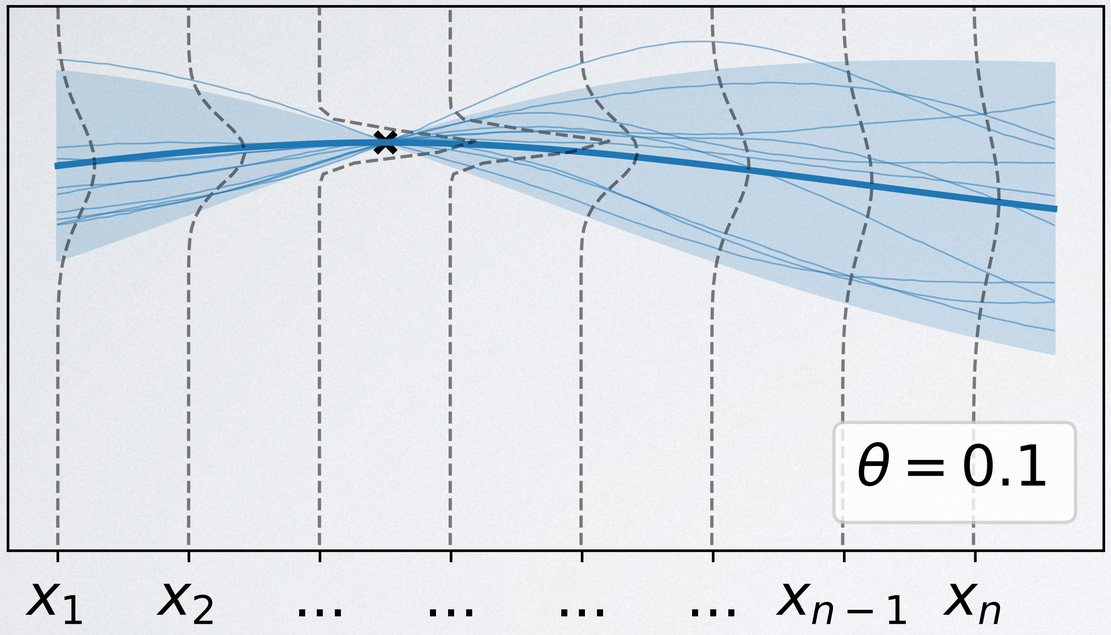
$$\frac{|\tilde{m}(x) - f(x)|}{\sqrt{\tilde{k}(x)}} \to \infty$$

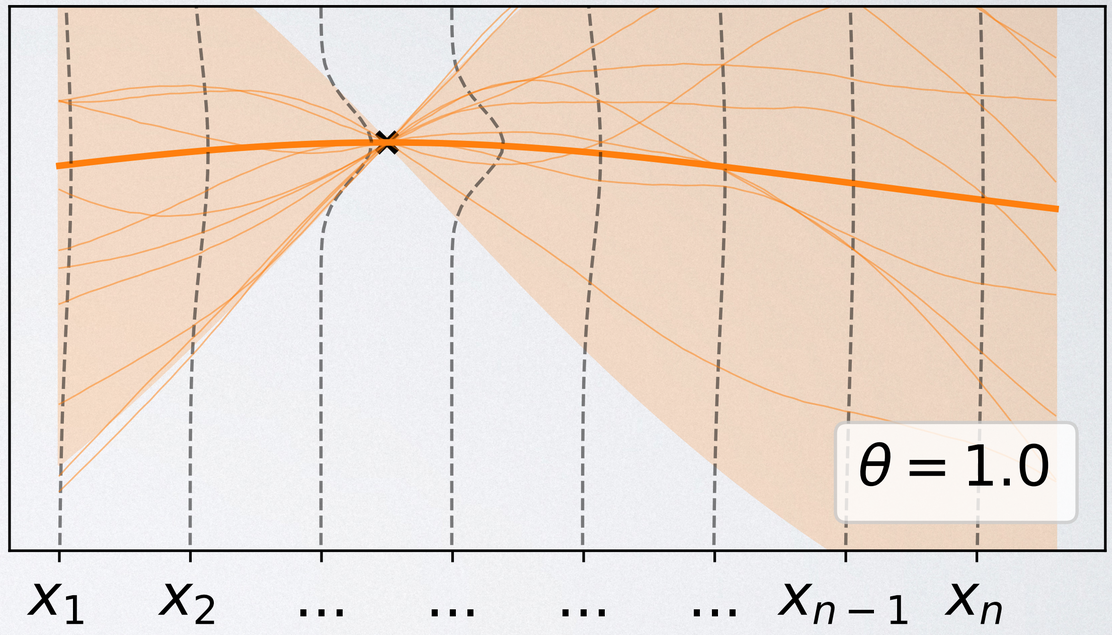
remember that posterior variance $\tilde{k}(x)$ does not depend on $f(x)$!

$$\tilde{k}(x, x') = k(x, x') - k(x, x_{1:n})^\top (k(x_{1:n}, x_{1:n}) + \sigma^2 \mathrm{Id}_n)^{-1} k(x', x_{1:n})$$

must choose hyper parameters of $k(x, x')$ well to prevent under/over confidence!

THE AMPLITUDE PARAMETER

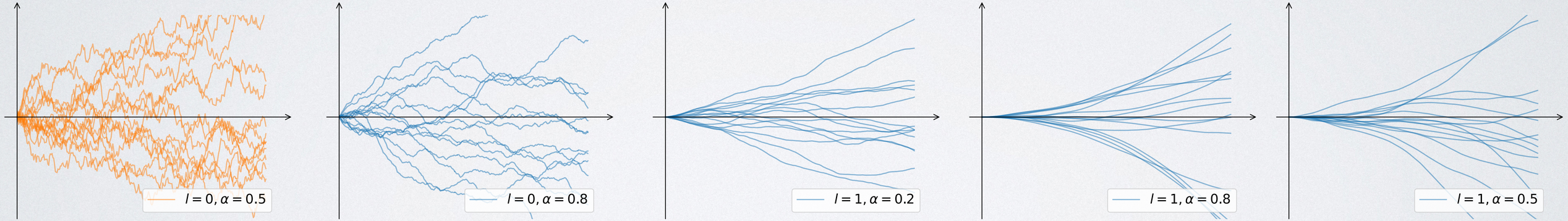




$$k_\theta(x, x') = \theta k(x, x')$$

LOO-CV ESTIMATOR IS SENSITIVE TO SMOOTHNESS

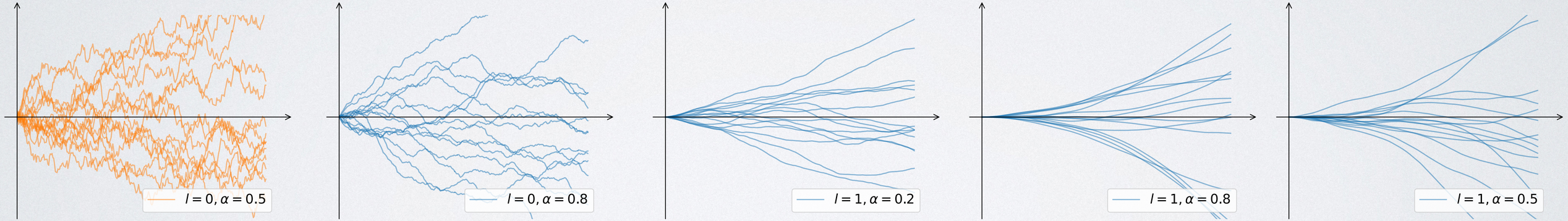
for the Brownian motion kernel in the noiseless setting, CV can yield asymptotically well-calibrated uncertainty estimates for a broader class of functions f than the ML estimator!



Naslidnyk, M., Kanagawa, M., Karvonen, T., and Mahsereci, M. (2023). Comparing Scale Parameter Estimators for Gaussian Process Regression: Cross Validation and Maximum Likelihood.

LOO-CV ESTIMATOR IS SENSITIVE TO SMOOTHNESS

for the Brownian motion kernel in the noiseless setting, CV can yield asymptotically well-calibrated uncertainty estimates for a broader class of functions f than the ML estimator!



… but it is not robust to outliers!

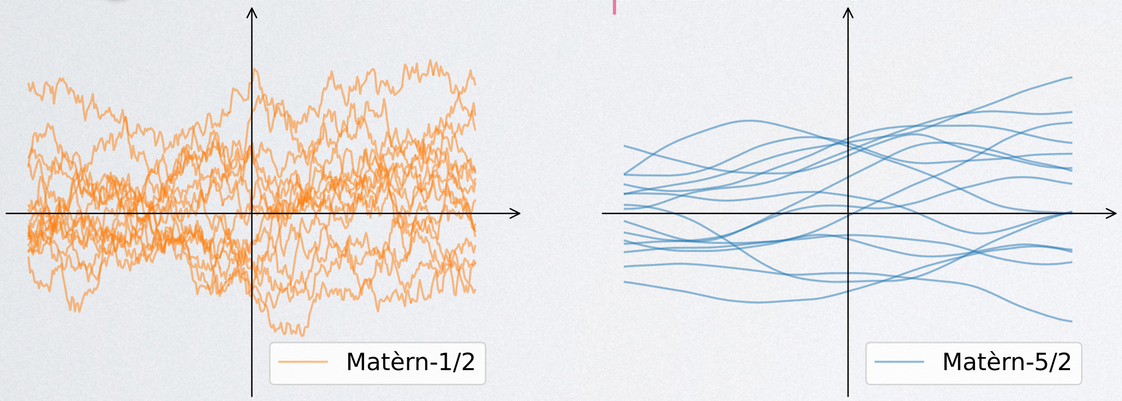
$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\tilde{m}_{\setminus i}(x_i))^2}{\tilde{k}_{\setminus i}(x_i,x_i)}$$

Naslidnyk, M., Kanagawa, M., Karvonen, T., and Mahsereci, M. (2023). Comparing Scale Parameter Estimators for Gaussian Process Regression: Cross Validation and Maximum Likelihood.

TYPES OF MISSPECIFICATION

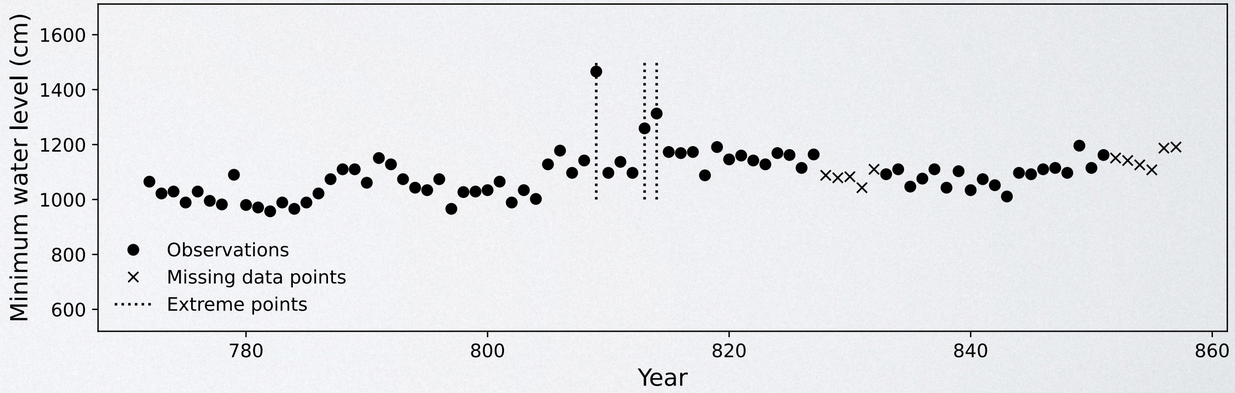
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Smoothness misspecification



LOO-CV Estimator is sensitive to smoothness.

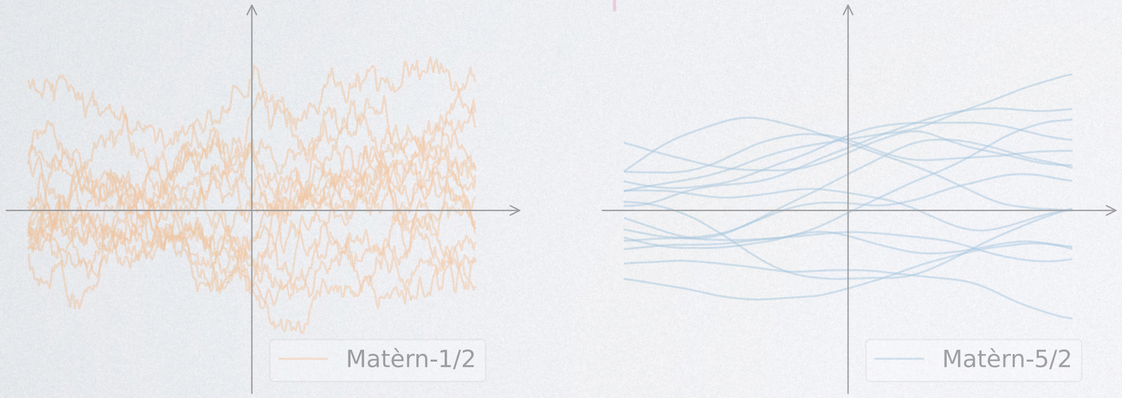
Likelihood misspecification



TYPES OF MISSPECIFICATION

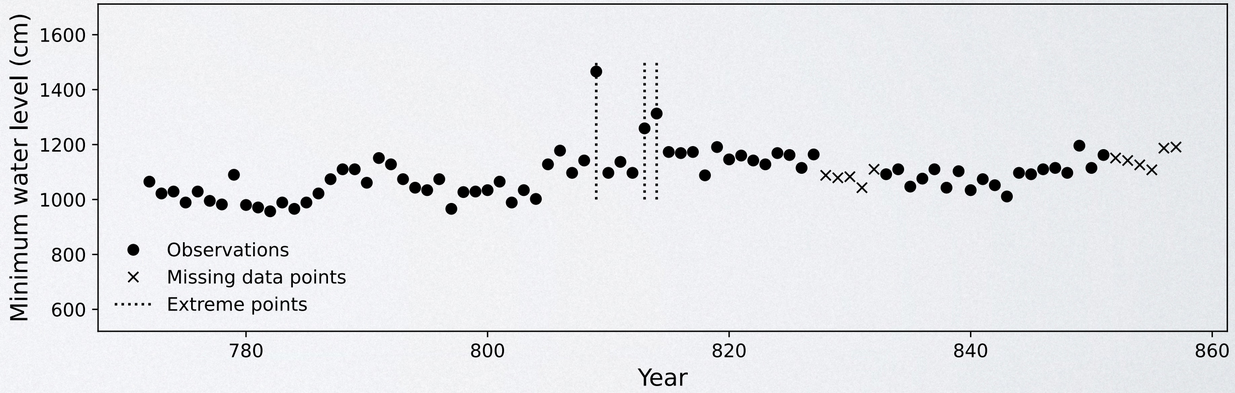
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Smoothness misspecification



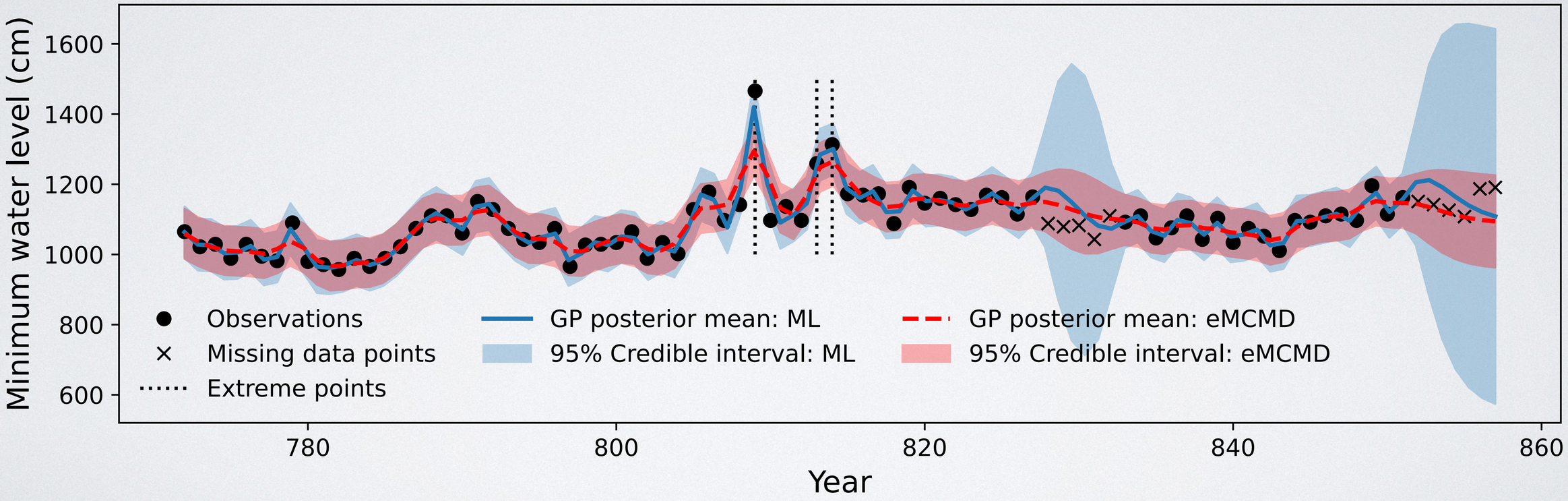
LOO-CV Estimator is sensitive to smoothness.

Likelihood misspecification

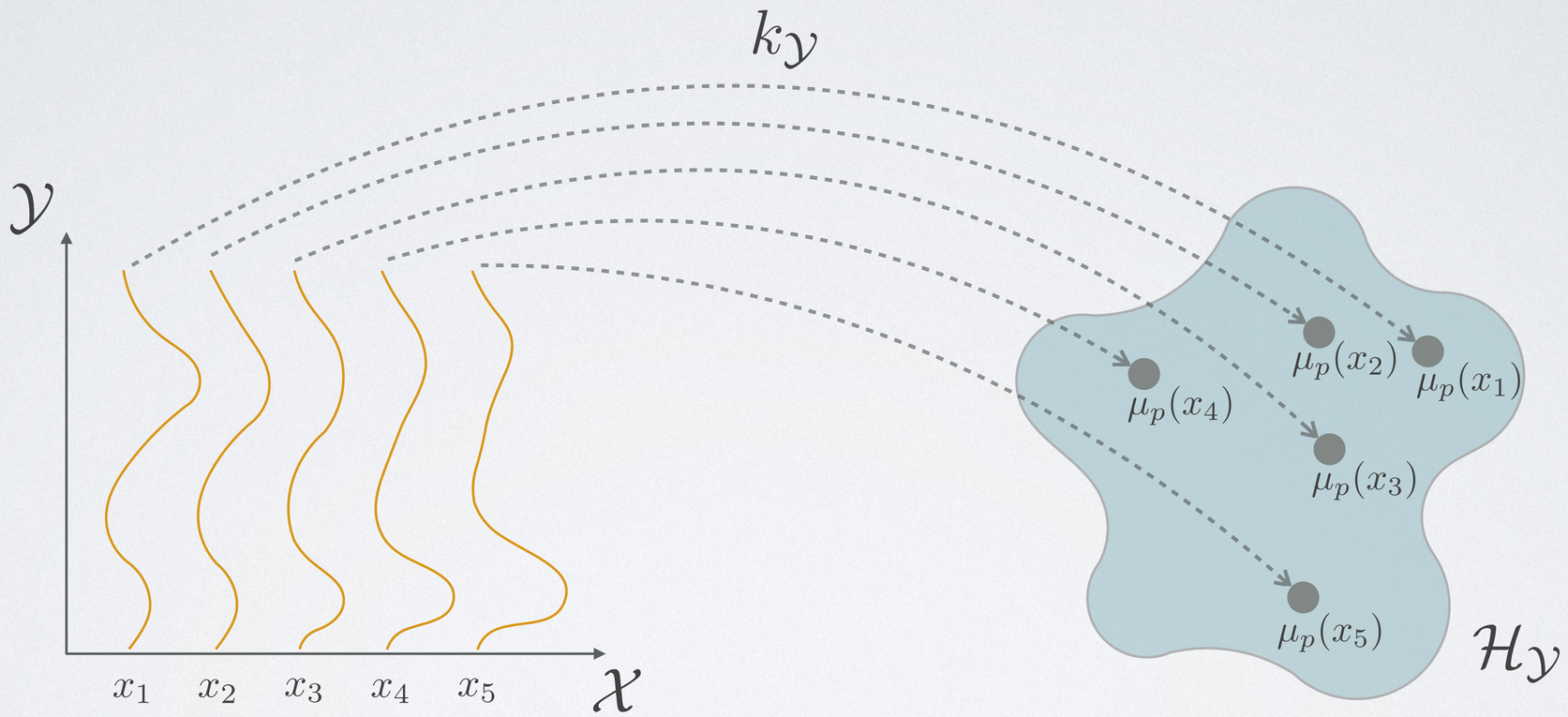


Robust **distance-based methods** are a practical remedy, but previously proposed distances between **conditional distributions** are intractable!

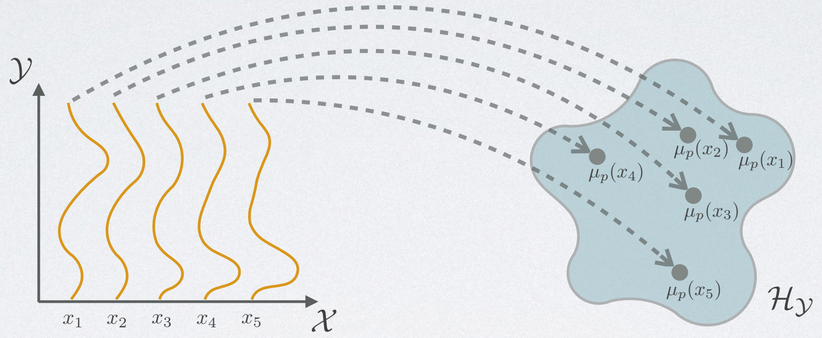
BACK TO THE MOTIVATING EXAMPLE



EMBEDDING CONDITIONAL DISTRIBUTIONS



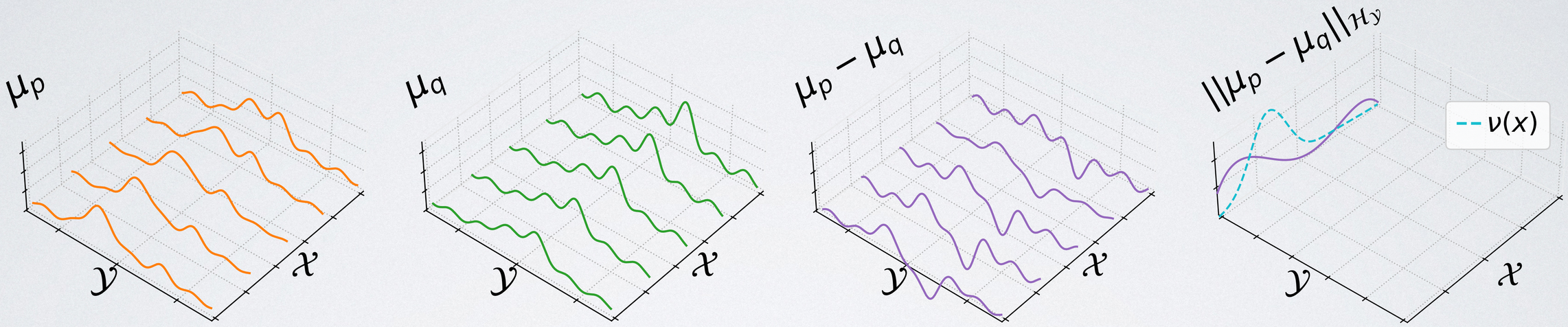
EMBEDDING CONDITIONAL DISTRIBUTIONS



- For any $x \in \mathcal{X}$, $p(x, \cdot)$ is a **probability measure** on $\mathcal{Y}$: $\mathbb{P}(Y \in A | X = x) = \int_A p(x, \mathrm{d}y)$
- p can be mapped to a function $\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, its **conditional kernel mean embedding**,

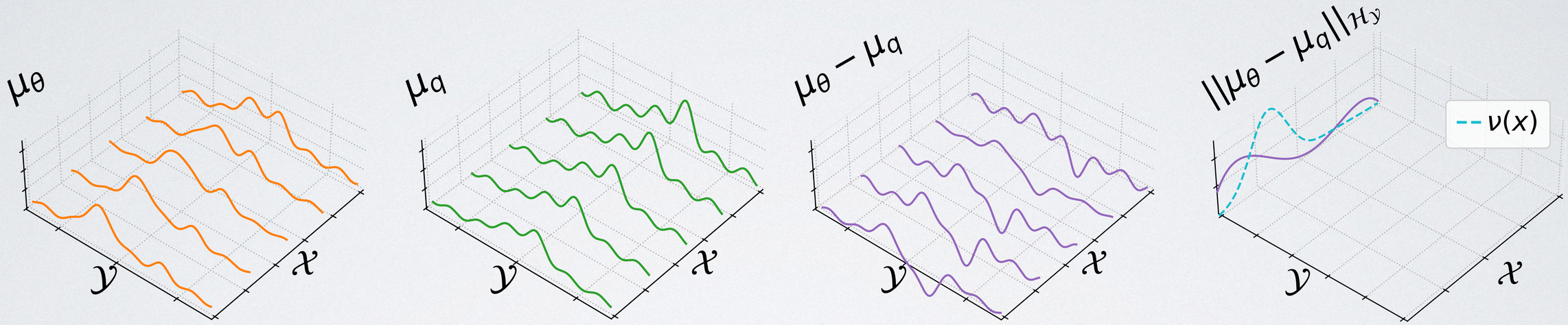
$$\mu_p(x)(y) = \mathbb{E}_{Y \sim p(x,\cdot)}[k_\mathcal{Y}(Y, y)] = \int_\mathcal{Y} k_\mathcal{Y}(y', y) p(x, \mathrm{d}y')$$

EXPECTED MEAN CONDITIONAL MEAN DISCREPANCY (EMCMD)



$$\mathbb{D}_\nu(\mu_p, \mu_q) = \mathbb{E}_{X\sim\nu}\|\mu_p(X) - \mu_q(X)\|_{\mathcal{H}_\mathcal{Y}}$$

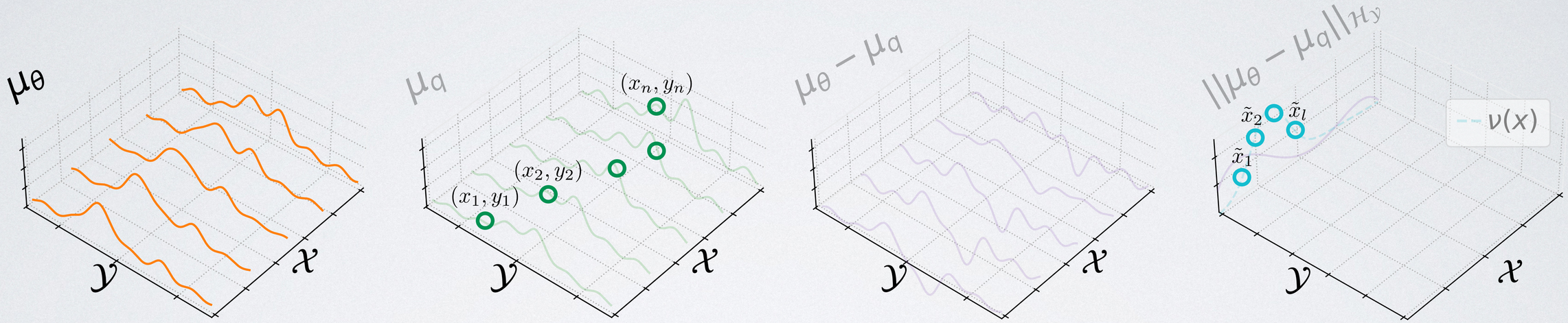
EMCMD-BASED PARAMETER ESTIMATION



For a model p_θ and the true data-generating distribution q, we denote $\mu_\theta = \mu_{p_\theta}$ and target

$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_\nu(\mu_\theta, \mu_q)$$

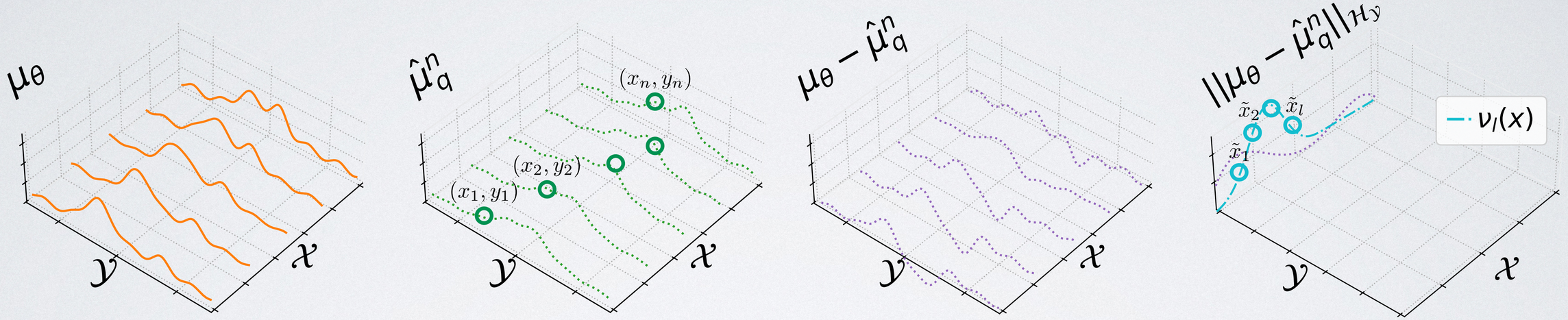
EMCMD-BASED PARAMETER ESTIMATION



For a model p_θ and the true data-generating distribution q, we denote $\mu_\theta = \mu_{p_\theta}$ and target

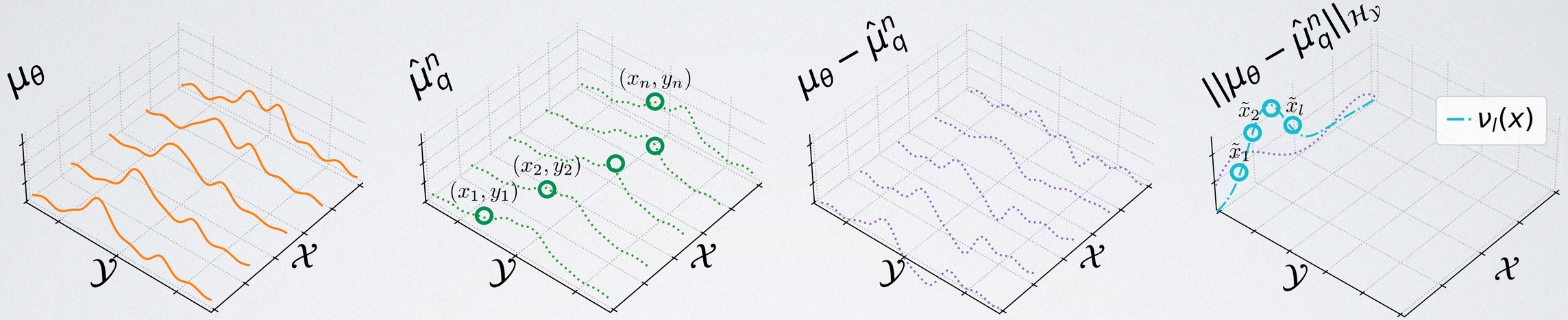
$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_\nu(\mu_\theta, \mu_q)$$

EMCMD-BASED PARAMETER ESTIMATION



$$\hat{\mu}_q^n(x)(\cdot) = k_\mathcal{X}(x, x_{1:n})^\top \left(k_\mathcal{X}(x_{1:n}, x_{1:n}) + n\lambda \mathrm{Id}_n\right)^{-1} k_\mathcal{Y}(y_{1:n}, \cdot), \qquad \nu_l = \frac{1}{l}\sum_{i=1}^{l} \delta_{\tilde{x}_i}$$

EMCMD-BASED PARAMETER ESTIMATION

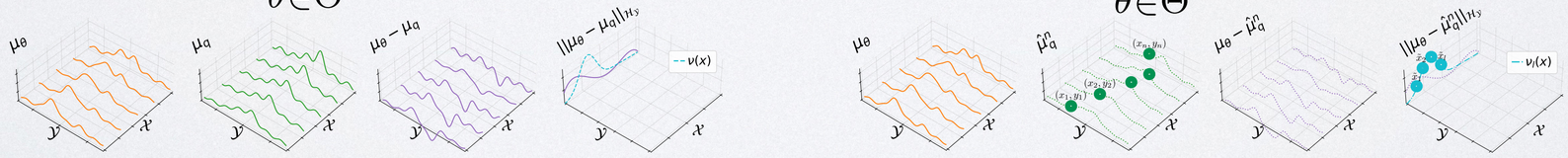


$$\hat{\mu}_q^n(x)(\cdot) = k_\mathcal{X}(x, x_{1:n})^\top \left(k_\mathcal{X}(x_{1:n}, x_{1:n}) + n\lambda \mathrm{Id}_n\right)^{-1} k_\mathcal{Y}(y_{1:n}, \cdot), \qquad \nu_l = \frac{1}{l}\sum_{i=1}^{l} \delta_{\tilde{x}_i}$$

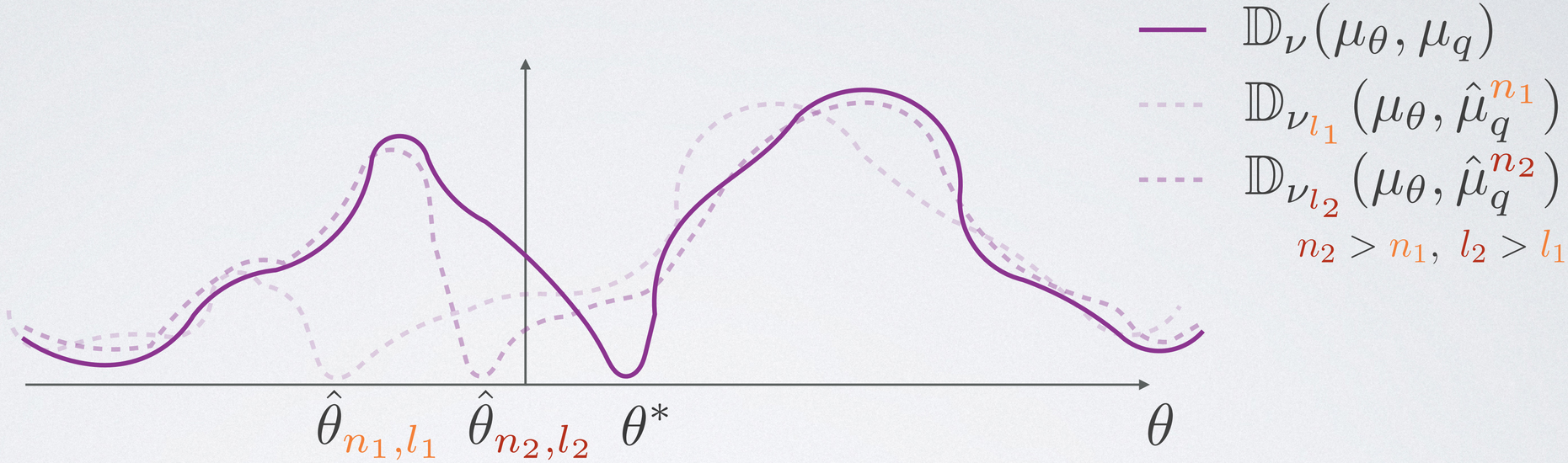
$$\hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{D}_{\nu_l}\left(\mu_\theta, \hat{\mu}_q^n\right)$$

THEORETICAL GUARANTEES

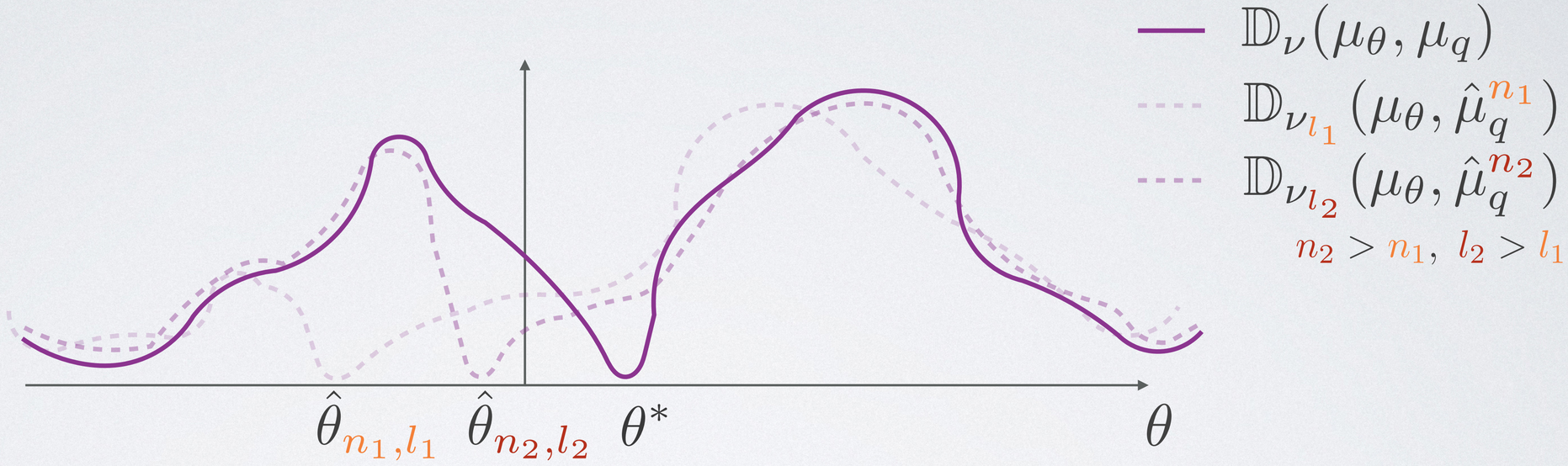
$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_{\nu}(\mu_\theta, \mu_q) \longrightarrow \hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n)$$



CONSISTENCY AND GENERALISATION



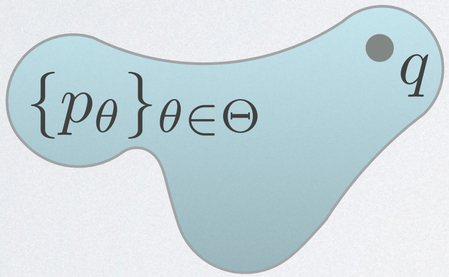
CONSISTENCY AND GENERALISATION

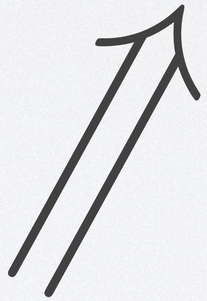


- Consistency: $\mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n) \xrightarrow{P} \mathbb{D}_\nu(\mu_\theta, \mu_q)$ e.f., uniformly in θ; $\hat{\theta}_{n,l} \xrightarrow{P} \theta^*$ e.f.
- Generalisation bound: $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_q) \xrightarrow{P} \mathbb{D}_\nu(\mu_{\theta^*}, \mu_q)$ e.f.

CONSISTENCY AND GENERALISATION

If $q = p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} 0$ e.f.





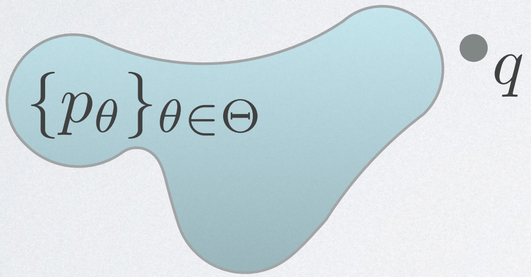
- Generalisation bound: $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_q) \xrightarrow{P} \mathbb{D}_\nu(\mu_{\theta^*}, \mu_q)$ e.f.

ROBUSTNESS

almost

If $q = p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} 0$ e.f.

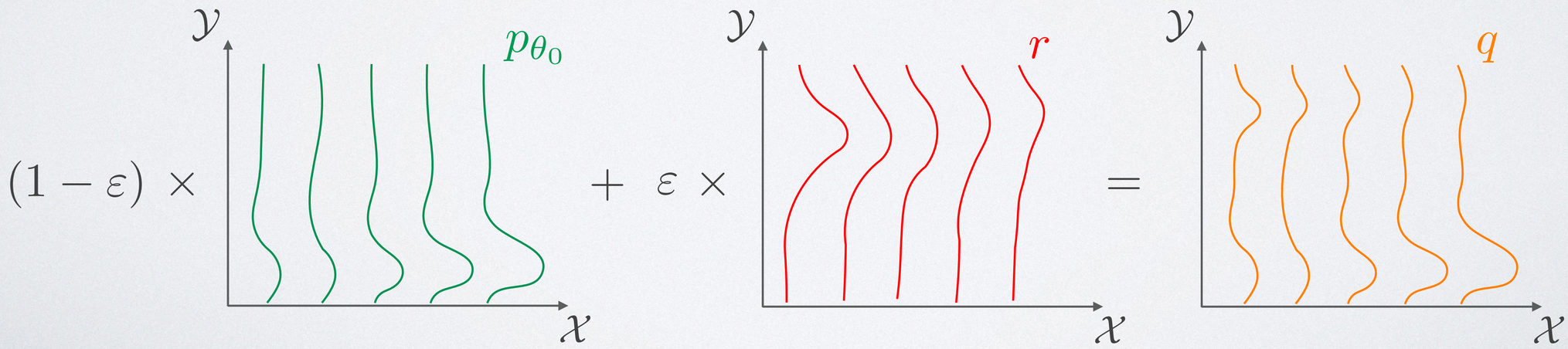
almost



ROBUSTNESS

If almost $q = p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P}$ almost 0 e.f.

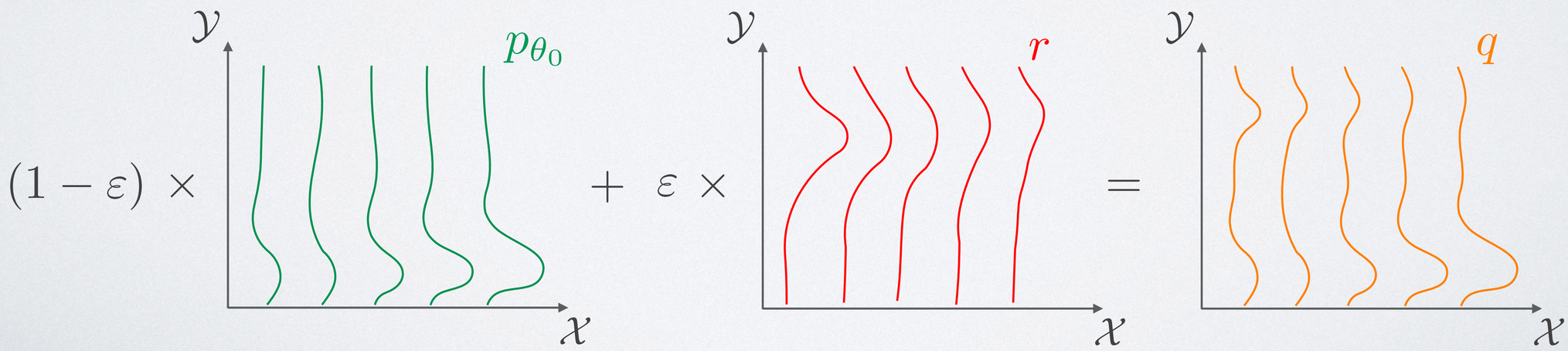
ε-Huber contamination, $\varepsilon \in [0, 1]$:



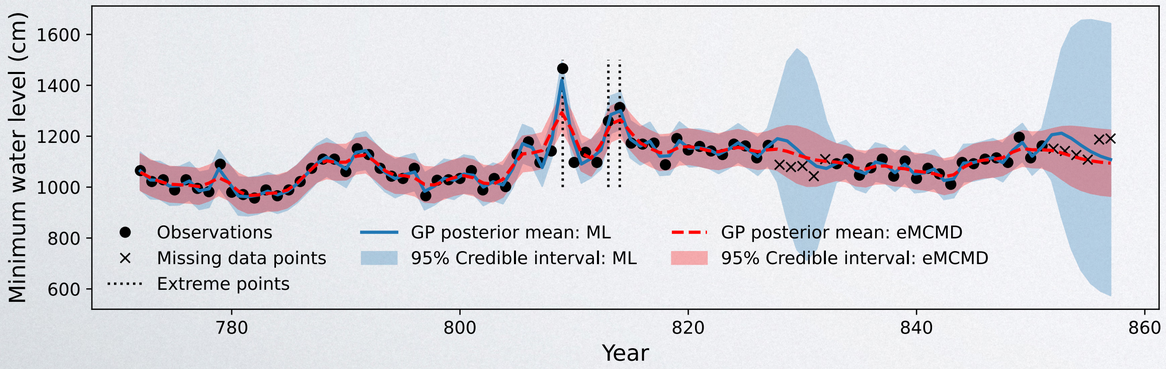
ROBUSTNESS

If $q = (1-\varepsilon)p_{\theta_0} + \varepsilon r$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} C_0\varepsilon$ e.f.

ε-Huber contamination, $\varepsilon \in [0,1]$:

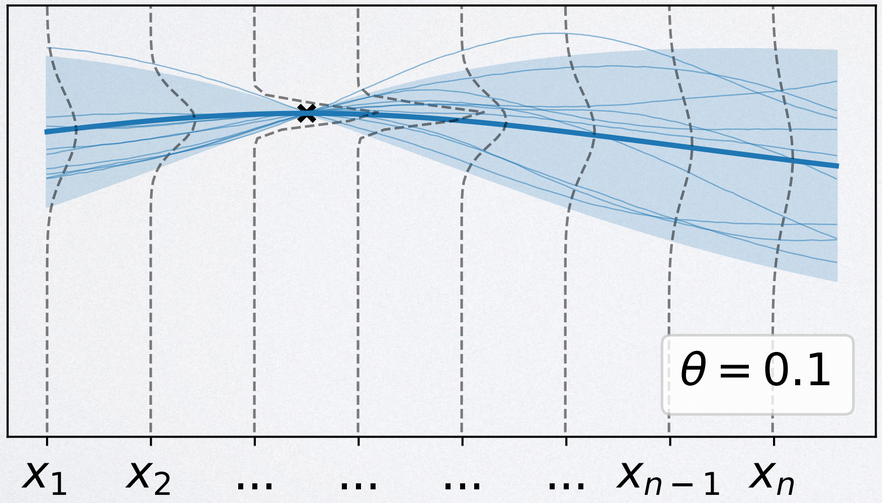


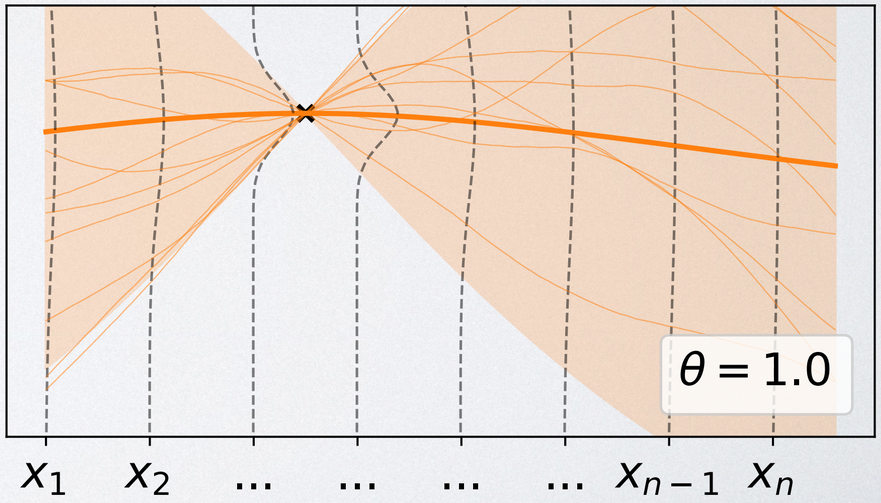
NUMERICAL EXPERIMENTS: ROBUST EMPIRICAL BAYES FOR GAUSSIAN PROCESSES



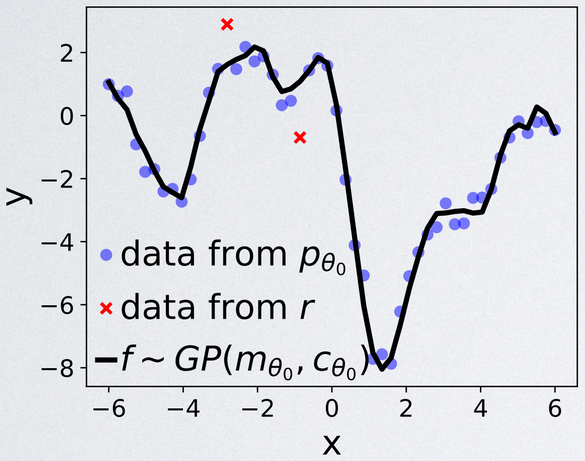
SETTING

- p_θ is a marginal of a Gaussian process $\mathcal{GP}(0, c_\theta)$
- $c_\theta(x, x') = \theta c(x, x')$



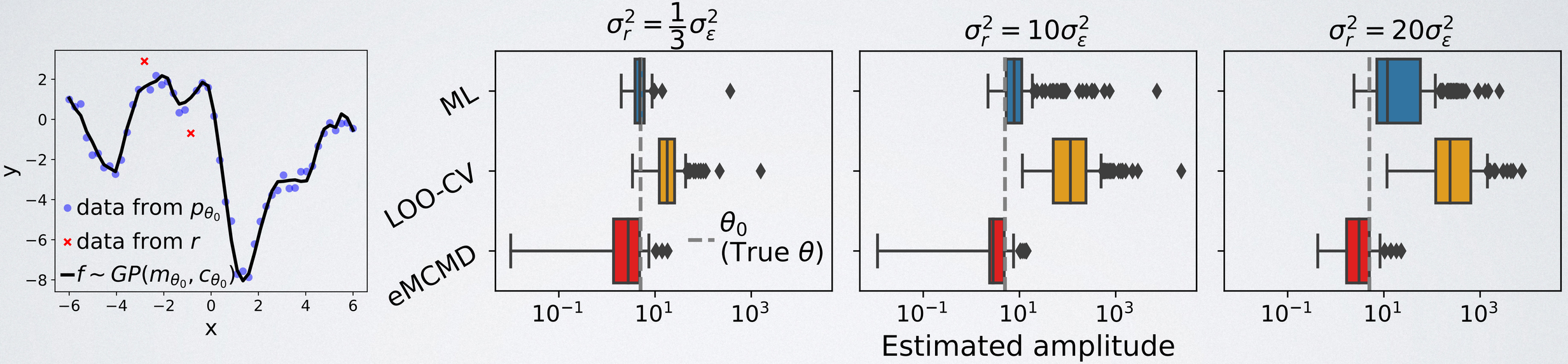


GP REGRESSION
WITH STUDENT-T MEASUREMENT ERROR



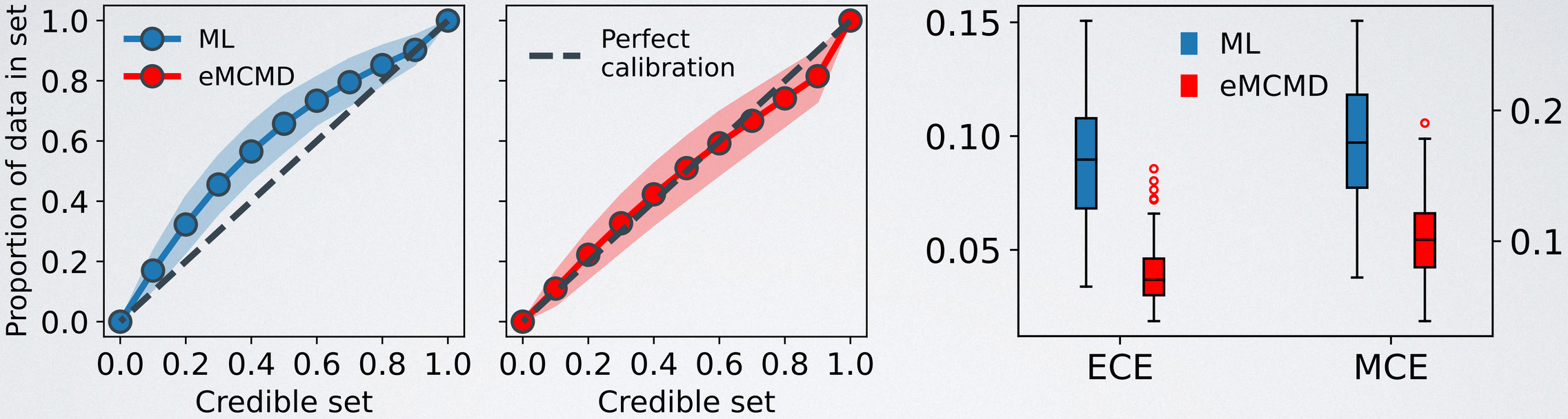
- Controlled environment: true function is a sample from a GP.
- 95% of data have Gaussian noise
- 5% of data have student-t noise
- Modeller assumes the noise is Gaussian.

GP REGRESSION WITH STUDENT-T MEASUREMENT ERROR



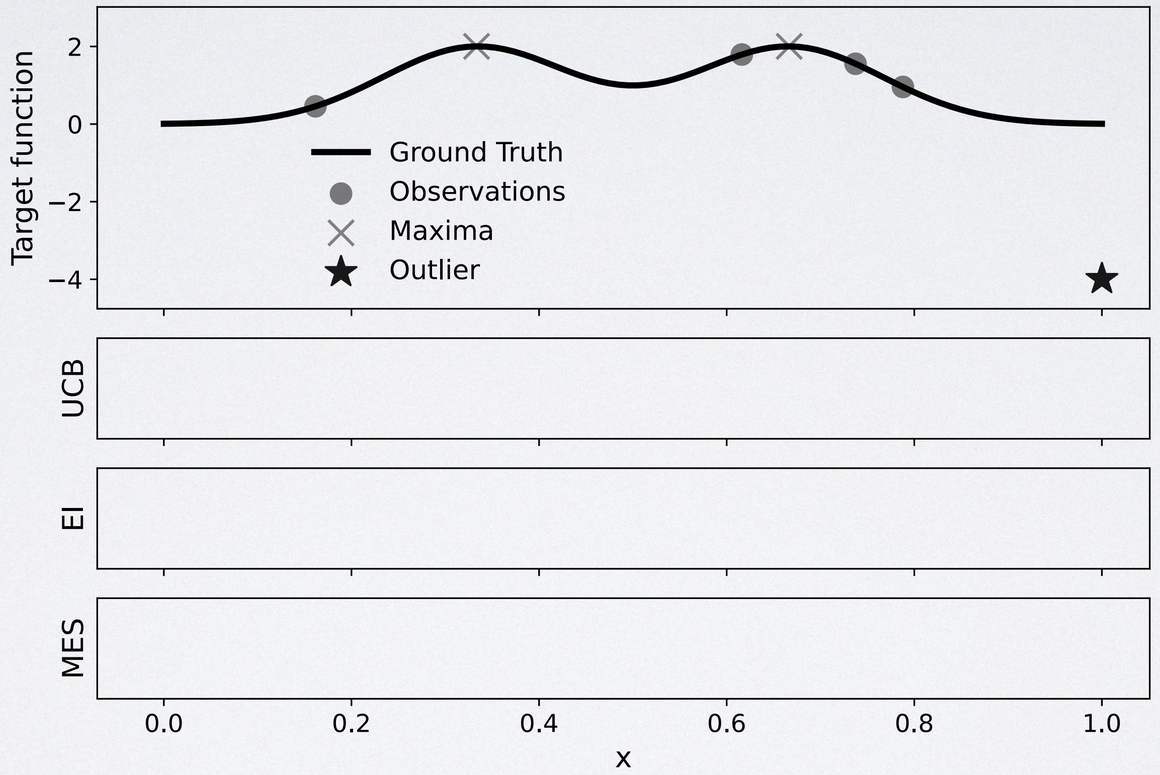
- 95% of data have noise $\sim \sigma_\varepsilon^2 \mathcal{N}(0, 1)$
- 5% of data have noise $\sim \sigma_\varepsilon^2 \text{Student-t}(3)$
- **ML and LOO-CV are not robust to outliers; eMCMD is!**

CALIBRATION DIAGRAMS FOR WIND-FARM MODELLING



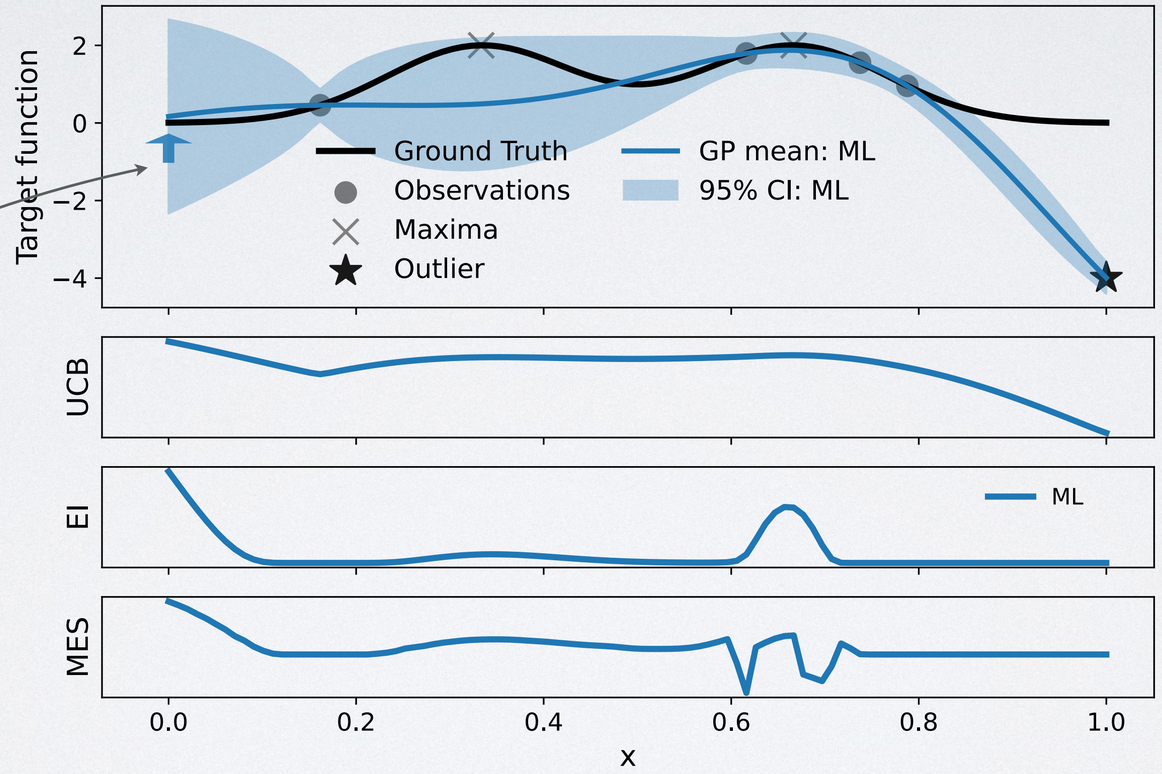
- Input: location of wind farms and wind angle. Output: energy produced. Noise is assumed Gaussian, but isn't likely to be—so the model is misspecified.
- **ML is underconfident; eMCMD is better calibrated.**

MISSPECIFIED BAYESIAN OPTIMISATION



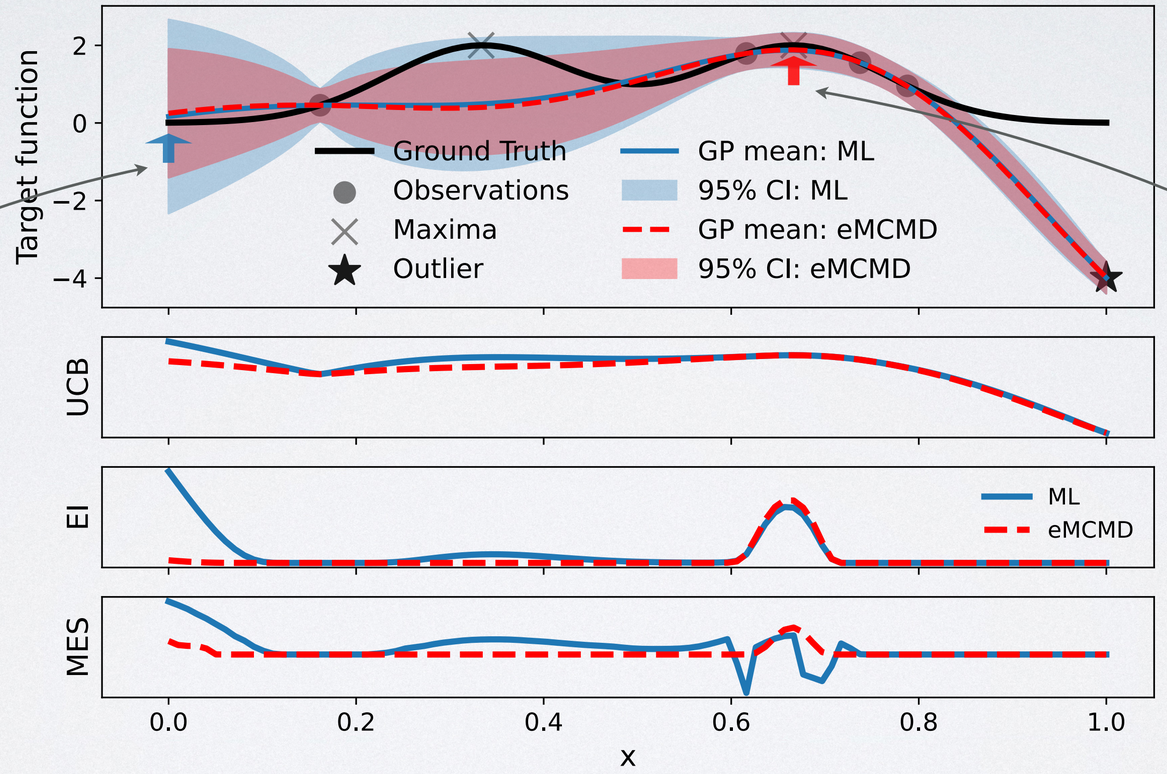
Task: locate the maximum of a function.

MISSPECIFIED BAYESIAN OPTIMISATION



even an outlier that does **not** affect the maximum can hinder optimisation…

MISSPECIFIED BAYESIAN OPTIMISATION

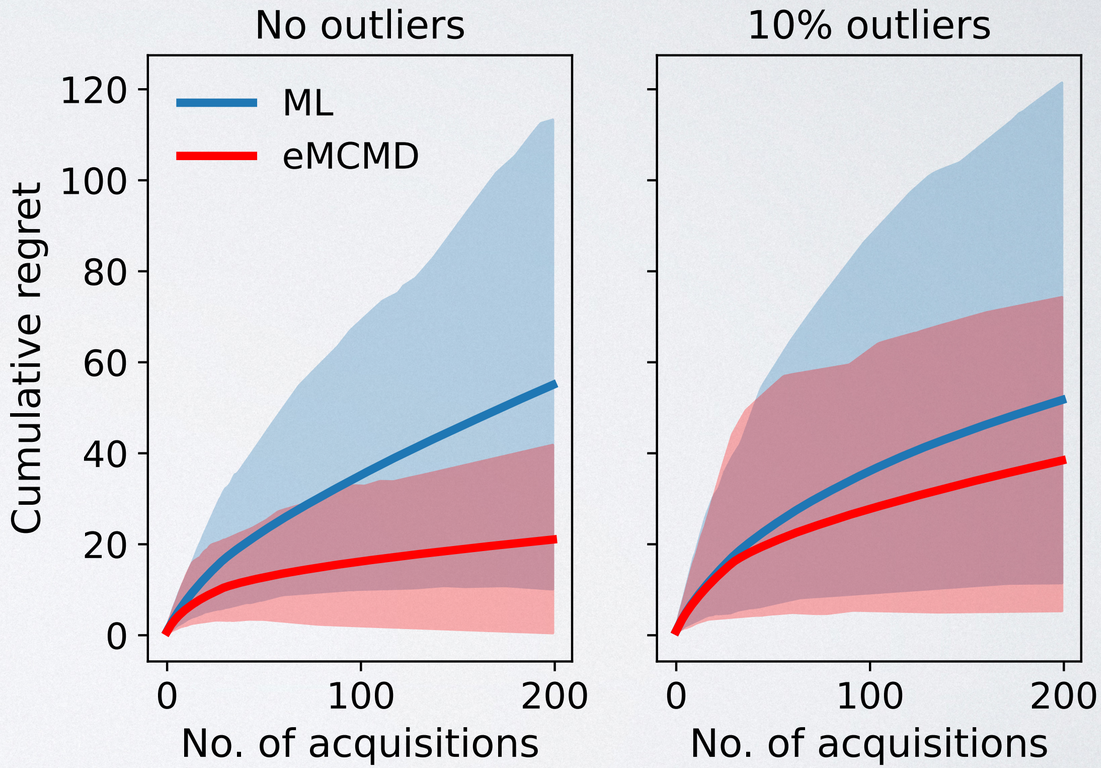


even an outlier that does **not** affect the maximum can hinder optimisation…

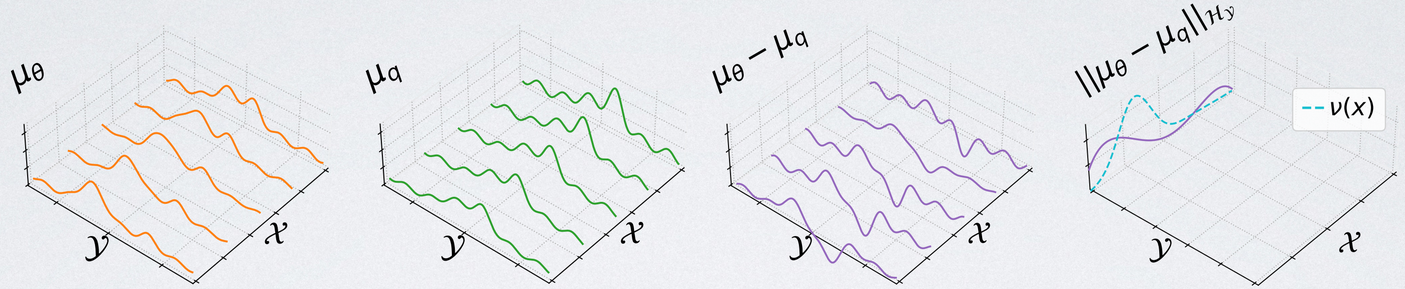
eMCMD, unlike ML, is robust to outliers, so this doesn't happen!

MISSPECIFIED BAYESIAN OPTIMISATION

- Hartmann 6D function.
- Outliers don't change the maximum.
- Still, **eMCMD reaches the maximum faster than ML...**
- **...even when no outliers are added, since there is still misspecification.**



CONCLUSION AND FINAL REMARKS



- Introduced eMCMD, $\mathbb{D}_\nu(\mu_p, \mu_q) = \mathbb{E}_{X\sim\nu}\|\mu_p(X) - \mu_q(X)\|_{\mathcal{H}_\mathcal{Y}}$
- Proved optimality of the approximation $\hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n)$
- Optimality results hold even if we need to estimate *both* μ_θ and μ_q
- What's next? Robust (generalised) Bayesian inference!

THANK YOU!

QUESTIONS?