# Comparing Scale Parameter Estimators for Gaussian Process Regression: Cross Validation and Maximum Likelihood

Motonobu Kanagawa

EURECOM, Sophia Antipolis, France

Joint work with
Masha Naslidnyk (University College London)
Toni Karvonen (University of Helsinki)
Maren Mahsereci (University of Tuebingen)

Probabilistic Numerics Spring School 2024, Southampton UK

# Outline

# Gaussian Process (GP) Interpolation

- Given training data $(x_1, f(x_1), \dots, (x_N, f(x_N))$, estimate the unknown data-generating function $f(x)$.

- By specifying a GP prior $f \sim GP(0, k)$ as with kernel $k$, GP interpolation produces the GP posterior of $f$ with posterior mean $m_N(x)$ and posterior covariance $k_N(x, x')$:

$$m_N(x) := k(x, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} f(\mathbf{x}), \tag{1a}$$

$$k_N(x, x') := k(x, x') - k(x, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} k(x', \mathbf{x}), \tag{1b}$$

where

$$\mathbf{x} := [x_1, \dots, x_N]^\top, \qquad f(\mathbf{x}) := [f(x_1), \dots, f(x_N)]^\top \in \mathbb{R}^N,$$

$$k(x, \mathbf{x}) := [k(x, x_1), \dots, k(x, x_N)]^\top \in \mathbb{R}^N, \quad k(\mathbf{x}, \mathbf{x}) := (k(x_i, x_j)) \in \mathbb{R}^{N \times N}.$$

# Gaussian Process (GP) Interpolation

- Posterior mean $m_N(x)$ is used for predicting $f(x)$.

- Posterior variance $k_N(x) := k_N(x, x)$ is used for uncertainty quantification (UQ) of $f(x)$.

- Many applications: Probabilistic numerics, Bayesian optimization, etc.

# Gaussian Process (GP) Interpolation

- Posterior mean $m_N(x)$ is used for predicting $f(x)$.

- Posterior variance $k_N(x) := k_N(x, x)$ is used for uncertainty quantification (UQ) of $f(x)$.

- Many applications: Probabilistic numerics, Bayesian optimization, etc.

## Challenge: Hyper Parameter Selection

- Both prediction and UQ performance depend on the choice of kernel parameters (or the kernel itself).

- This talk focuses on UQ performance.

# Credible Intervals

- For a constant $\alpha > 0$, a credible interval for $f(x)$ can be defined as

$$[m_N(x) - \alpha\sqrt{k_N(x)}, \quad m_N(x) + \alpha\sqrt{k_N(x)}].$$

(e.g., $\alpha \approx 1.96$ for a 95 % credible interval.)

- We want the interval to include the true $f(x)$, i.e.,

$$m_N(x) - \alpha\sqrt{k_N(x)} \leq \quad f(x) \quad \leq m_N(x) + \alpha\sqrt{k_N(x)}$$
$$\Updownarrow$$
$$-\alpha \leq \quad \frac{f(x) - m_N(x)}{\sqrt{k_N(x)}} \quad \leq +\alpha$$

# Asymptotically Well-calibrated Credible Intervals

$$-\alpha \leq \quad \frac{f(x) - m_N(x)}{\sqrt{k_N(x)}} \quad \leq +\alpha$$

- Therefore, for the interval to be well-calibrated, the posterior std $\sqrt{k_N}(x)$ and the prediction error $|f(x) - m_N(x)|$ should decrease to 0 at the same speed as $N \to \infty$.

# Asymptotically Well-calibrated Credible Intervals

$$-\alpha \leq \frac{f(x) - m_N(x)}{\sqrt{k_N(x)}} \leq +\alpha$$

- Therefore, for the interval to be well-calibrated, the posterior std $\sqrt{k_N}(x)$ and the prediction error $|f(x) - m_N(x)|$ should decrease to 0 at the same speed as $N \to \infty$.

- If $\sqrt{k_N(x)} \to 0$ faster than $|f(x) - m_N(x)| \to 0$, then the interval will not contain the true $f(x)$ as $N \to \infty$ (overconfident).

# Asymptotically Well-calibrated Credible Intervals

$$-\alpha \leq \frac{f(x) - m_N(x)}{\sqrt{k_N(x)}} \leq +\alpha$$

- Therefore, for the interval to be well-calibrated, the posterior std $\sqrt{k_N}(x)$ and the prediction error $|f(x) - m_N(x)|$ should decrease to 0 at the same speed as $N \to \infty$.

- If $\sqrt{k_N(x)} \to 0$ faster than $|f(x) - m_N(x)| \to 0$, then the interval will not contain the true $f(x)$ as $N \to \infty$ (overconfident).
- If $\sqrt{k_N(x)} \to 0$ slower than $|f(x) - m_N(x)| \to 0$, then the interval will get looser as $N \to \infty$ (underconfident).

# Asymptotically Well-calibrated Credible Intervals

$$-\alpha \leq \frac{f(x) - m_N(x)}{\sqrt{k_N(x)}} \leq +\alpha$$

- Therefore, for the interval to be well-calibrated, the posterior std $\sqrt{k_N}(x)$ and the prediction error $|f(x) - m_N(x)|$ should decrease to 0 at the same speed as $N \to \infty$.

- If $\sqrt{k_N(x)} \to 0$ faster than $|f(x) - m_N(x)| \to 0$, then the interval will not contain the true $f(x)$ as $N \to \infty$ (overconfident).
- If $\sqrt{k_N(x)} \to 0$ slower than $|f(x) - m_N(x)| \to 0$, then the interval will get looser as $N \to \infty$ (underconfident).

- If $\sqrt{k_N(x)} \to 0$ and $|f(x) - m_N(x)| \to 0$ at the same speed, we say the interval is asymptotically well-calibrated.

# The Issue

- The issue is that, the posterior variance $k_N(x)$ does not depend on observations $f(x_1), \ldots, f(x_N)$.

$$k_N(x) = k(x,x) - k(x,\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} k(x,\mathbf{x}),$$

- Thus, $\sqrt{k_N(x)}$ does not decay at the same rate as $|f(x) - m_N(x)|$ in general.

- Exception: when the prior $f \sim \mathcal{GP}(0,k)$ is correct (well-specified).
- But in general $f \sim \mathcal{GP}(0,k)$ cannot be perfectly correct (misspecified).

# Examples

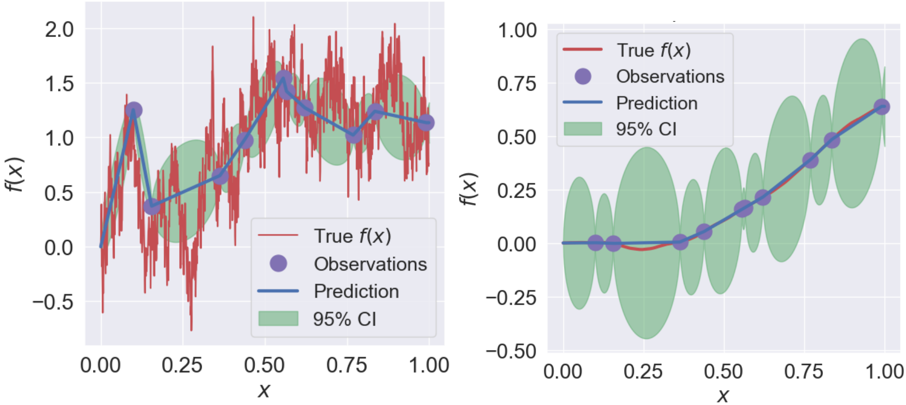



Figure 1: Left: when true $f$ is coarser. Right: when true $f$ is smoother.

# Scale Parameter

- To address the above issue, one must adapt the kernel $k$ to the data $f(x_1), \ldots, f(x_N)$.

- A simple way is to introduce a scale parameter $\sigma^2 > 0$, parameterize the kernel as

$$k_\sigma(x, x') := \sigma^2 k(x, x').$$

and estimate $\sigma^2$ from the data $f(x_1), \ldots, f(x_N)$.

# Scale Parameter

- To address the above issue, one must adapt the kernel $k$ to the data $f(x_1), \ldots, f(x_N)$.

- A simple way is to introduce a scale parameter $\sigma^2 > 0$, parameterize the kernel as

$$k_\sigma(x, x') := \sigma^2 k(x, x').$$

and estimate $\sigma^2$ from the data $f(x_1), \ldots, f(x_N)$.

- We discuss two estimators for the scale parameter $\sigma^2$: Maximum Likelihood (ML) and Cross Validation (CV).

- ML: maximizing the marginal likelihood.
- CV: maximizing the held-out (log) likelihood, averaged over CV splits.

# ML and CV Estimators for the Scale Parameter

- For the scale parameter, both ML and CV estimators can be derived analytically (and can be computed in $O(N^3)$ time complexity).

**ML estimator**:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{f(\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} f(\mathbf{x})}{N} = \frac{1}{N}\sum_{n=1}^{N} \frac{[f(x_n) - m_{n-1}(x_n)]^2}{k_{n-1}(x_n)},$$

where $m_{n-1}$ and $k_{n-1}$ are the posterior mean and variance functions based on the first $n-1$ training observations $(x_1, f(x_1)), \ldots, (x_{n-1}, f(x_{n-1}))$.

**Leave-one-out (LOO) CV estimator**:

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N}\sum_{n=1}^{N} \frac{\left[f(x_n) - m_{\backslash n}(x_n)\right]^2}{k_{\backslash n}(x_n)},$$

where $m_{\backslash n}$ and $k_{\backslash n}$ are the posterior mean and variance functions with $(x_n, f(x_n))$ removed: $\{(x_1, f(x_1)), \ldots, (x_N, f(x_N))\} \backslash (x_n, f(x_n))$.

# ML and CV Estimators for the Scale Parameter

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \frac{[f(x_n) - m_{n-1}(x_n)]^2}{k_{n-1}(x_n)}, \quad \hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\left[f(x_n) - m_{\setminus n}(x_n)\right]^2}{k_{\setminus n}(x_n)},$$

- The CV estimator uses each data more evenly than the ML estimator?

**Questions:**

- How does this difference affect their asymptotic behaviours as $N \to \infty$?

- Do these estimates give asymptotically well-calibrated credible intervals?

- Which one is better in terms of UQ?

# (Closely) Related Work

- Most previous works consider well-specified settings where there exists a "true" scale parameter $\sigma_0^2$;
e.g., [Ying, 1991, Zhang, 2004, Bachoc et al., 2017, Bachoc et al., 2020].

- However, in general, there exists no such "true $\sigma_0^2$".

- In such misspecified settings, both ML and CV estimators' asymptotic properties are not well understood.

- [Karvonen et al., 2020, Wang, 2021] derive upper bounds (and lower bounds in some cases) for the ML estimator (assuming deterministic $f$).
- To our knowledge, there is no theoretical result for the CV estimator.
- [Bachoc, 2013, Petit et al., 2022] empirically compare the ML and CV estimators under different model misspecification settings.

# Outline

# Our Contributions

- This work analyses the asymptotic behaviours of the CV and ML estimators of the scale parameter.

- This is the first analysis for the CV estimator.
- We also obtain new results for the ML estimator.

# Our Contributions

- This work analyses the asymptotic behaviours of the CV and ML estimators of the scale parameter.

- This is the first analysis for the CV estimator.
- We also obtain new results for the ML estimator.

- To facilitate the analysis, we focus on the following setting.

1) **Noise-free observations (Interpolation)**: $f(x_1), f(x_2), \dots, f(x_N)$.

2) **Quasi-uniform input points**: $0 \leq x_1 < x_2 < \cdots < x_N \leq 1$.

# Problem Setting

**3) Brownian motion prior**: the kernel is

$$k(x, x') = \min(x, x'), \quad x, x' \in [0, 1].$$

– Then $f \sim \mathcal{GP}(0, k)$ is a Brownian motion.

– The smoothness of the model is $1/2$ (in terms of differentiability).

# Problem Setting

**4) True $f$'s smoothness is $0 < \ell + H < 2$. ($\ell \in \{0, 1\}, 0 < H < 1$.)**:

- When $\ell = 0$ and $0 < H < 1$: We assume $f$ to be a fractional Brownian motion with Hurst parameter $H$. We write $f_{FBM} \sim \mathcal{GP}(0, k_{0,H})$.[1]

— Smoothness $H$. When $H = 1/2$, this is a standard Brownian motion.

[1] $k_{0,H}(x, x') = \frac{1}{2}(|x|^{2H} + |x'|^{2H} - |x - x'|^{2H})$

# Problem Setting

**4) True $f$'s smoothness is $0 < \ell + H < 2$. ($\ell \in \{0, 1\}, 0 < H < 1$.):**

- When $\ell = 0$ and $0 < H < 1$: We assume $f$ to be a fractional Brownian motion with Hurst parameter $H$. We write $f_{FBM} \sim \mathcal{GP}(0, k_{0,H})$.[1]

— Smoothness $H$. When $H = 1/2$, this is a standard Brownian motion.

- When $\ell = 1$ and $0 < H < 1$: We define $f$ as the integral of $f_{FBM} \sim \mathcal{GP}(0, k_{0,H})$ (i.e., an integrated fractional Brownian motion).

$$f_{iFBM}(x) = \int_0^x f_{FBM}(t)dt.$$

— Smoothness $1 + H$.

[1] $k_{0,H}(x, x') = \frac{1}{2}(|x|^{2H} + |x'|^{2H} - |x - x'|^{2H})$

# Outline

# Asymptotics of CV and ML Estimators ($\ell = 0$, $0 < H < 1$)

Theorem (when $\ell = 0$ and $0 < H < 1$):

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{1-2H}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{1-2H}).$$

- CV and ML have the same asymptotic properties.

# Asymptotics of CV and ML Estimators ($\ell = 0$, $0 < H < 1$)

**Theorem (when $\ell = 0$ and $0 < H < 1$):**

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{1-2H}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{1-2H}).$$

- CV and ML have the same asymptotic properties.

When $0 < H < 1/2$: True $f$ is rougher than the GP prior.

- $1 - 2H > 0 \implies \hat{\sigma}^2_{\mathrm{CV}}, \hat{\sigma}^2_{\mathrm{ML}}$ increase as $N$ increases.
- $\implies$ Correcting the small posterior var $k_N(x)$ (overconfidence)

- Recall that the posterior variance with the scale parameter is $\hat{\sigma}^2 k_N(x)$

# When the True Function is Rougher ($\ell + H = 0.2$)

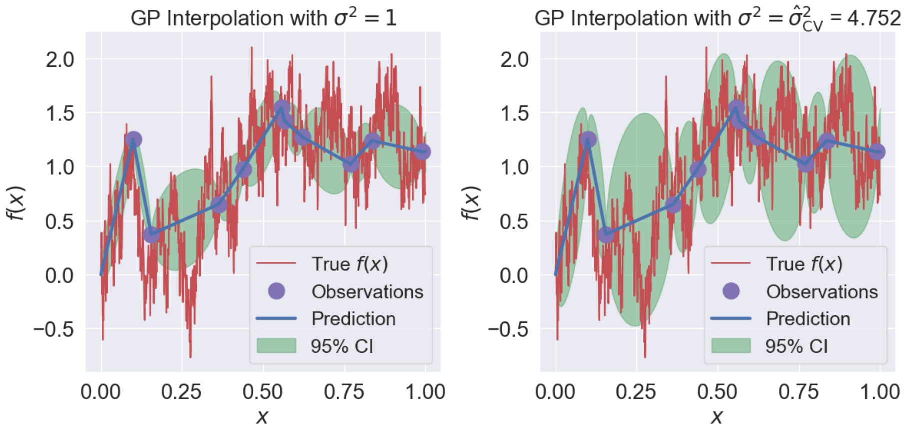



Figure 2: Left: no scale parameter; Right: with the CV estimator. (ML yields similar credible intervals.)

# Asymptotics of CV and ML Estimators ($\ell = 0$, $0 < H < 1$)

Theorem (when $\ell = 0$ and $0 < H < 1$):

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{1-2H}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{1-2H}).$$

When $H = 1/2$: The GP prior (Brownian motion) is well-specified.

- $1 - 2H = 0$. $\Longrightarrow$ $\hat{\sigma}^2_{\mathrm{CV}}, \hat{\sigma}^2_{\mathrm{ML}}$ converge to th "true value" $\sigma_0^2$ as $N \to \infty$ increases.

# Asymptotics of CV and ML Estimators ($\ell = 0$, $0 < H < 1$)

Theorem (when $\ell = 0$ and $0 < H < 1$):

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{1-2H}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{1-2H}).$$

When $H = 1/2$: The GP prior (Brownian motion) is well-specified.

- $1 - 2H = 0$. $\Longrightarrow$ $\hat{\sigma}^2_{\mathrm{CV}}, \hat{\sigma}^2_{\mathrm{ML}}$ converge to th "true value" $\sigma_0^2$ as $N \to \infty$ increases.

When $1/2 < H < 1$: True $f$ is smoother than the GP prior.

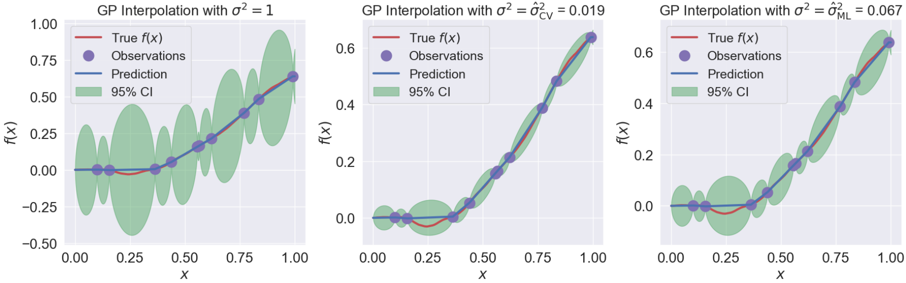
- $1 - 2H < 0$. $\Longrightarrow$ $\hat{\sigma}^2_{\mathrm{CV}}, \hat{\sigma}^2_{\mathrm{ML}}$ decrease as $N \to \infty$.
- Correcting the large posterior var $k_N(x)$ (underconfidence).

# Asymptotics of CV and ML Estimators ($1 < \ell + H \leq 1.5$)

**Theorem** (when $\ell = 1$ and $0 < H \leq 0.5$):

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{-1-2H}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{-1}).$$

- CV adapts to the true smoothness $1 + H$ (correcting underconfidence).
- ML does not adapt (underconfidence not corrected $\Rightarrow$ conservative credible intervals).



# When the True Function is Smoother ($\ell + H = 1.5$)

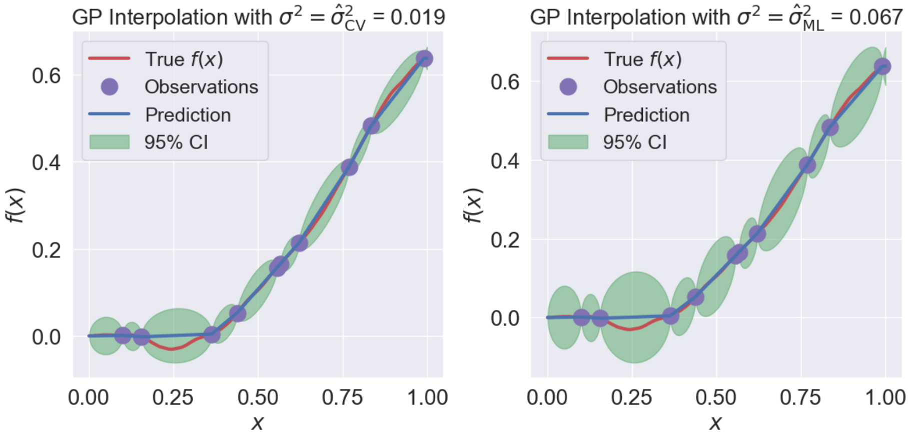



Figure 3: Left: with the CV estimator; Right: with the ML estimator.

# Asymptotics of CV and ML Estimators ($1.5 < \ell + H < 2$)

Theorem (when $\ell = 1$ and $0.5 < H < 1$):

$$\mathbb{E}[\hat{\sigma}^2_{\mathrm{CV}}] = \Theta(N^{-2}), \qquad \mathbb{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \Theta(N^{-1}).$$

- CV also no longer adapts (while yielding tighter intervals than ML).

- These asymptotic rates of CV and ML have been empirically confirmed (for different smoothness settings); please see our paper if interested.

# Outline

# Consequences to Credible Intervals

- For $\hat{\sigma} = \hat{\sigma}_{\mathrm{CV}}$ or $\hat{\sigma} = \hat{\sigma}_{\mathrm{ML}}$, a credible intervals may be given as

$$[m_N(x) - \alpha\hat{\sigma}\sqrt{k_N(x)}, \quad m_N(x) + \alpha\hat{\sigma}\sqrt{k_N(x)}].$$

(e.g., $\alpha \approx 1.96$ for a 95 % credible interval.)

- This interval is asymptotically well-calibrated if

$$|f(x) - m_N(x)| \quad \text{and} \quad \hat{\sigma}\sqrt{k_N(x)}$$

decrease to 0 as $N \to \infty$ at the same rate.

# Consequences to Credible Intervals

- For $\hat{\sigma} = \hat{\sigma}_{\mathrm{CV}}$ or $\hat{\sigma} = \hat{\sigma}_{\mathrm{ML}}$, a credible intervals may be given as

$$[m_N(x) - \alpha\hat{\sigma}\sqrt{k_N(x)}, \quad m_N(x) + \alpha\hat{\sigma}\sqrt{k_N(x)}].$$

(e.g., $\alpha \approx 1.96$ for a 95 % credible interval.)

- This interval is asymptotically well-calibrated if

$$|f(x) - m_N(x)| \quad \text{and} \quad \hat{\sigma}\sqrt{k_N(x)}$$

decrease to 0 as $N \to \infty$ at the same rate.

- For CV and ML, can this hold? (and if so, when?)

- To study this question, we consider the ratio

$$\frac{|f(x) - m_N(x)|}{\hat{\sigma}\sqrt{k_N(x)}}.$$

- It is necessary that this neither converge to 0 nor diverge to $\infty$ as $N \to \infty$.

# Consequences to Credible Intervals $(0 < \ell + H < 1)$

- Recall that true $f$ is assumed to be $f \sim \mathcal{GP}(0, k_{\ell,H})$ (smoothness $\ell + H$).

**Theorem (when $\ell = 0$ and $0 < H < 1$)**

$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} k_N(x)} = \Theta(1),$$

$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} k_N(x)} = \Theta(1),$$

- Both CV and ML are asymptotically well-calibrated (on average)

# Consequences to Credible Intervals ($1 < \ell + H \leq 1.5$)

Theorem (when $\ell = 1$ and $0 < H \leq 1/2$)

$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\text{CV}} k_N(x)} = \Theta(1),$$
$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\text{ML}} k_N(x)} = \Theta(N^{-2H}),$$

- CV is asymptotically well-calibrated.

- ML is asymptotically underconfident.

# Consequences to Credible Intervals ($1.5 < \ell + H < 2$)

**Theorem (when $\ell = 1$ and $1/2 < H \leq 1$)**

$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} k_N(x)} = \Theta(N^{1-2H}),$$

$$\sup_{x \in [0,1]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} k_N(x)} = \Theta(N^{-2H}),$$

- Both are asymptotically underconfident.

- The ratio decreases more slowly with CV than ML (CV is relatively less conservative).

# Outline

# Conclusion and Future Directions

- Under specific settings (one-dimension, Brownian motion prior, scale parameter estimation), we have shown, for the first time, quantitatively that **the CV estimator is better than the ML estimator in terms of UQ.**

- Future work should study whether similar results hold in other settings (e.g., smoother priors, other hyperparameters, noisy observations).

- Other results (e.g., when the true function is deterministic) are also available in our paper:

Comparing Scale Parameter Estimators for Gaussian Process Regression: Cross Validation and Maximum Likelihood
Naslidnyk, M., Kanagawa, M., Karvonen, T., Mahsereci, M. (2023). arXiv preprint arXiv:2307.07466.

Bachoc, F. (2013).
Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification.
*Computational Statistics & Data Analysis*, 66:55–69.

Bachoc, F., Betancourt, J., Furrer, R., and Klein, T. (2020).
Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes.
*Electronic Journal of Statistics*, 14(1):1962–2008.

Bachoc, F., Lagnoux, A., and Nguyen, T. M. N. (2017).
Cross-validation estimation of covariance parameters under fixed-domain asymptotics.
*Journal of Multivariate Analysis*, 160:42–67.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C. J., and Särkkä, S. (2020).
Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions.

*SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

Petit, S., Bect, J., Feliot, P., and Vazquez, E. (2022).
Parameter selection in Gaussian process interpolation: An empirical study of selection criteria.
*arXiv:2107.06006v4*.

Wang, W. (2021).
On the inference of applying Gaussian process modeling to a deterministic function.
*Electronic Journal of Statistics*, 15(2):5014–5066.

Ying, Z. (1991).
Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process.
*Journal of Multivariate Analysis*, 36(2):280–296.

Zhang, H. (2004).
Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics.

*Journal of the American Statistical Association*, 99(465):250–261.