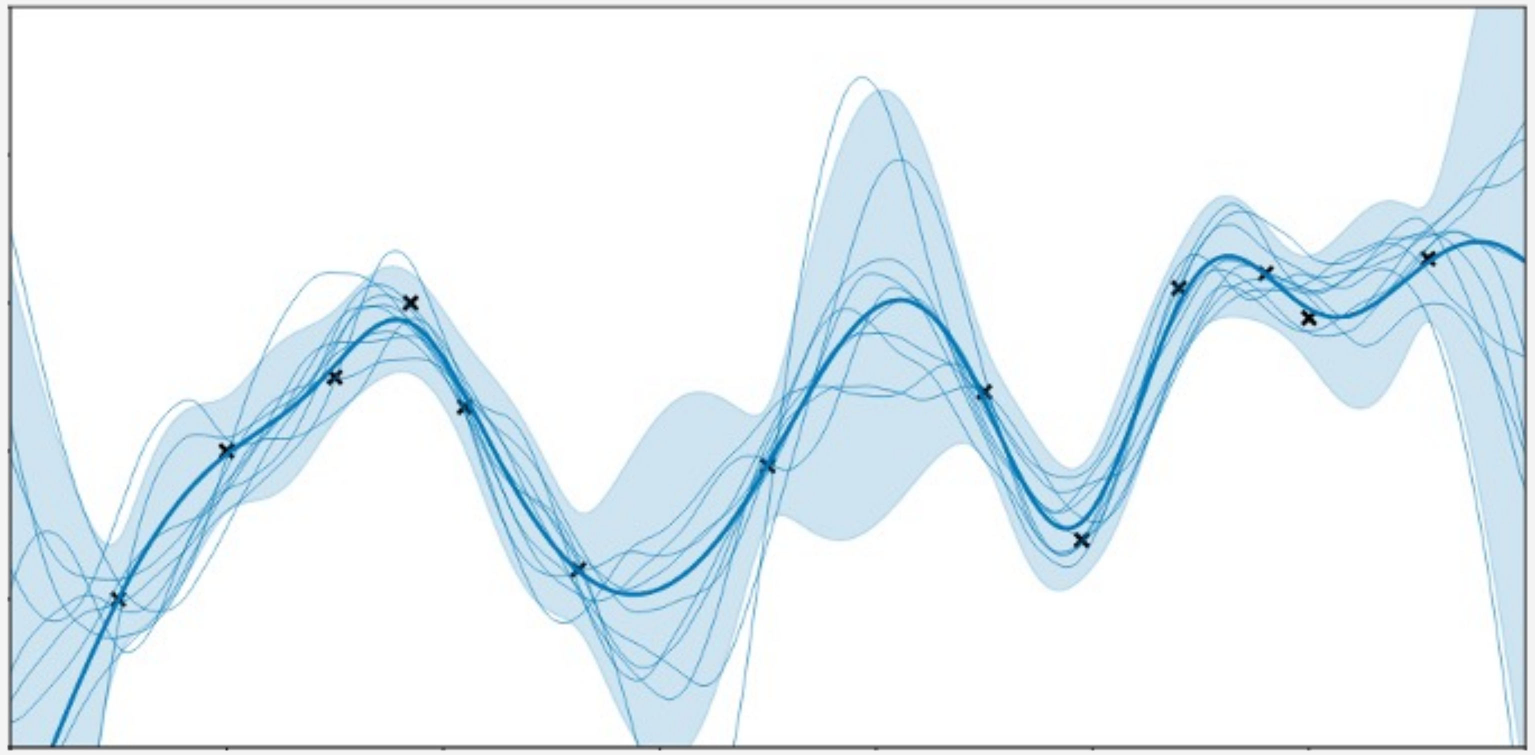
Maximum Likelihood Estimation in Gaussian Process Regression is Ill-Posed

Toni Karvonen and Chris J. Oates



Samuel Oliveira, Huey Sun, Anabel Yong

Machine Learning Seminar – Week 6
19th February 2024

Gaussian Processes

- Define a distribution over functions

- Fully defined by a prior mean function and a covariance kernel.

- **GP Learning**: learn the hyperparameters of the mean function and covariance kernel.

Hyper-Parameter Estimation

- MLE Type 2 (Evidence/Marginal Likelihood).

- Cross-Validation.

- MAP Estimation.

MLE Type-2

- Most common hyper-parameter optimization process.
- No additional degrees of freedom
- **Research Question**: How do the ML estimates vary as a function of the input data?

Paper Overview

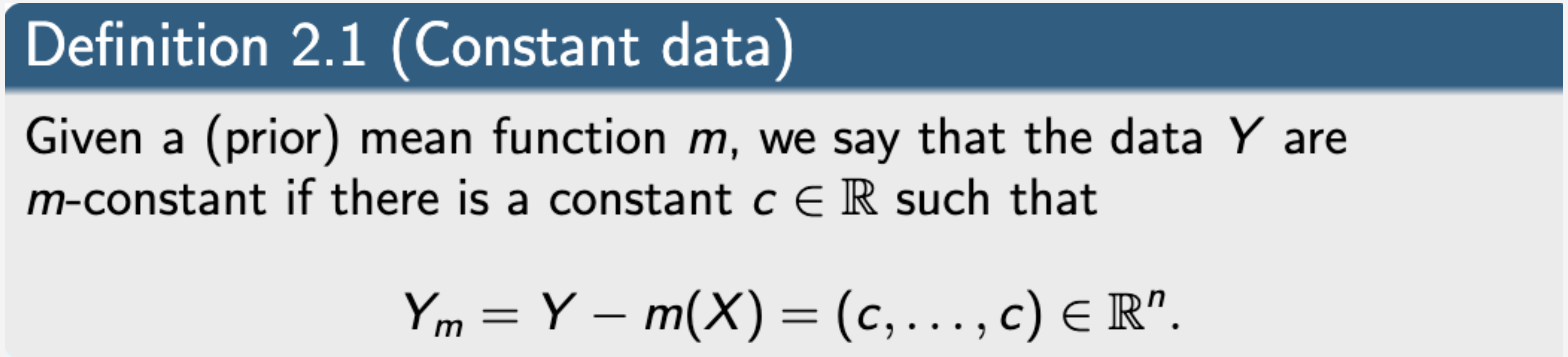
Maximum Likelihood Estimation in Gaussian Process Regression is Ill-Posed

- Theoretical conditions for MLE in GPs to be ill-posed.
- Extension of these conditions to other methods:
    - CV
    - MAP
- Use of regularisation to fix ill-posedness.

Theoretical Settings

- **Noiseless** data.
  - No observation noise, data represents truthful observations of an underlying function.
- **Non-asymptotic** setting.
  - Finite number of datapoints.
- Estimation of one single **lengthscale** parameter.

M-Constant Dataset

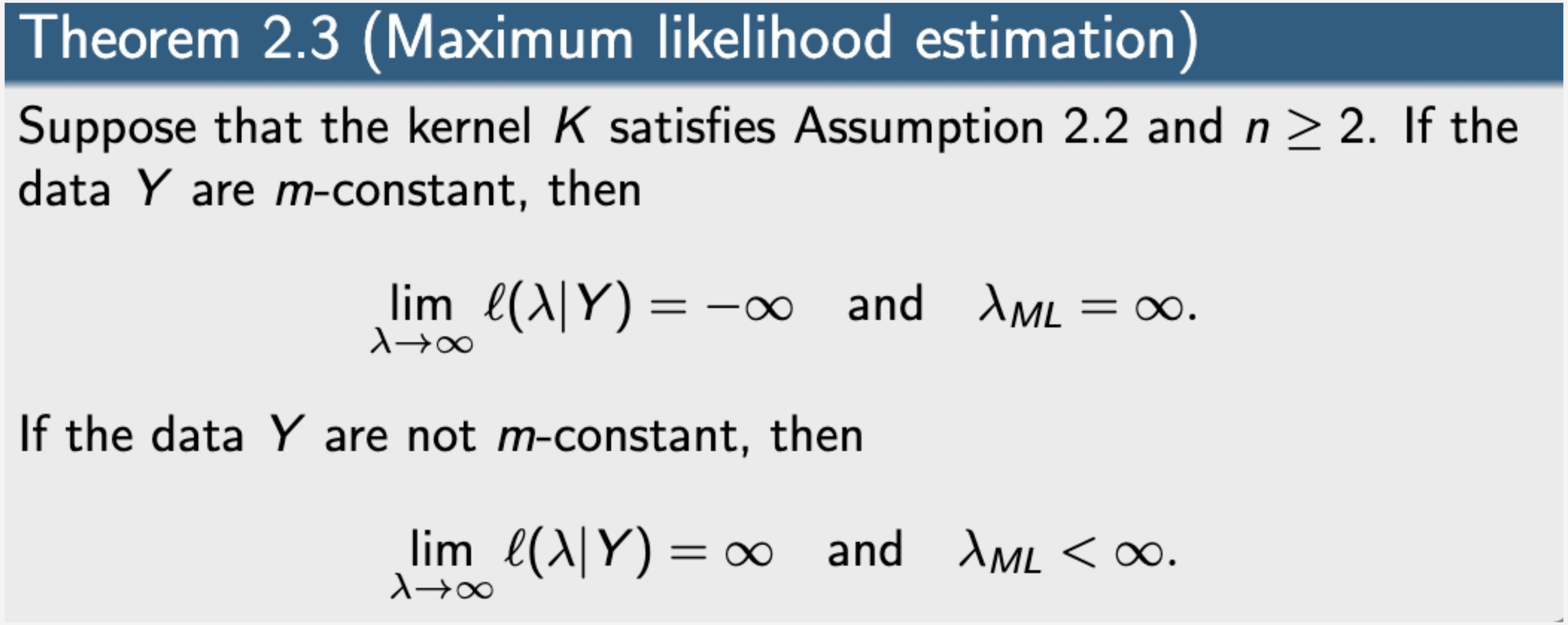


Definition 2.1 (Constant data)

Given a (prior) mean function m, we say that the data Y are m-constant if there is a constant $c \in \mathbb{R}$ such that

$$Y_m = Y - m(X) = (c, \ldots, c) \in \mathbb{R}^n.$$

Maximum Likelihood Estimation



Theorem 2.3 (Maximum likelihood estimation)

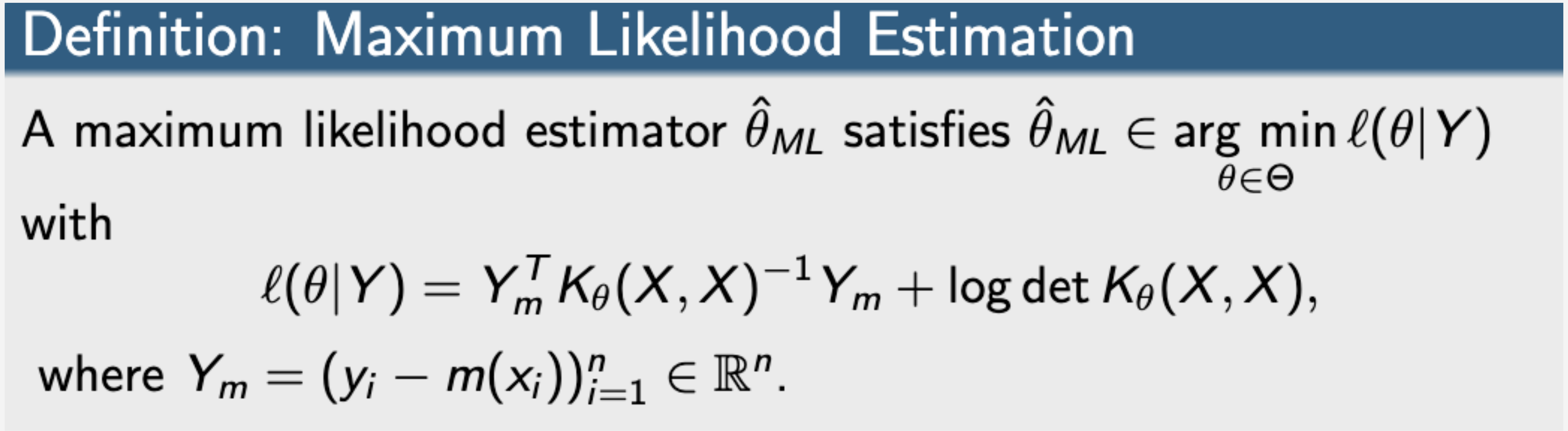
Suppose that the kernel K satisfies Assumption 2.2 and $n \geq 2$. If the data Y are m-constant, then

$$\lim_{\lambda \to \infty} \ell(\lambda | Y) = -\infty \quad \text{and} \quad \lambda_{ML} = \infty.$$

If the data Y are not m-constant, then

$$\lim_{\lambda \to \infty} \ell(\lambda | Y) = \infty \quad \text{and} \quad \lambda_{ML} < \infty.$$

Trade-offs in the loss function



Definition: Maximum Likelihood Estimation

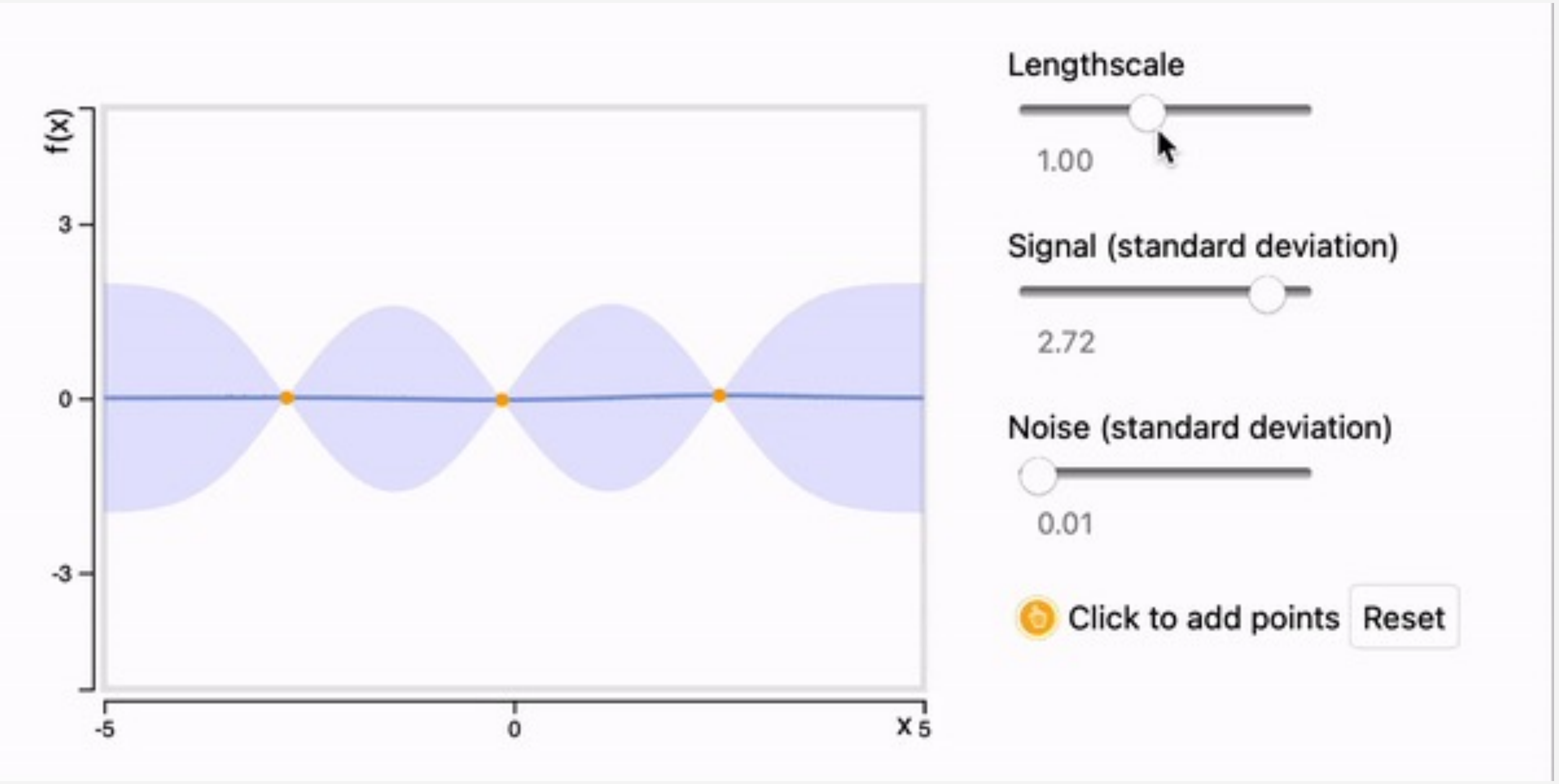
A maximum likelihood estimator $\hat{\theta}_{ML}$ satisfies $\hat{\theta}_{ML} \in \arg\min_{\theta \in \Theta} \ell(\theta | Y)$

with

$$\ell(\theta|Y) = Y_m^T K_\theta(X,X)^{-1} Y_m + \log\det K_\theta(X,X),$$

where $Y_m = (y_i - m(x_i))_{i=1}^n \in \mathbb{R}^n$.

Intuition



Lengthscale

1.00

Signal (standard deviation)

2.72

Noise (standard deviation)

0.01

Figure 1: Visualization of m-constant data as lengthscale varies, *Deisenroth et al.* [2]

Trade-offs in the Log Likelihood Function

- We can represent the kernel function as $\Phi(\frac{x-y}{\lambda})$.

- Thus, our covariance function tends to $\Phi(0)$ as $\lambda \to \infty$.

Trade-offs in the Log Likelihood Function

- We can represent the kernel function as $\Phi(\frac{x-y}{\lambda})$.

- Thus, our covariance function tends to $\Phi(0)$ as $\lambda \to \infty$.

- If Φ satisfies certain "mild regularity conditions", then in the presence of m-constant data, $\lambda_{ML} = \infty$

- Proof involves Reproducing Kernel Hilbert Space techniques.

Trade-offs in the Log Likelihood Function

- We can represent the kernel function as $\Phi(\frac{x-y}{\lambda})$.

- Thus, our covariance function tends to $\Phi(\,0\,)$ as $\lambda \to \infty$.

- If Φ satisfies certain "mild regularity conditions", then in the presence of m-constant data, $\lambda_{ML} = \infty$

- Proof involves Reproducing Kernel Hilbert Space techniques.

Formally defining ill-posedness

An inference or estimation problem is well-posed if:

- a solution **exists**
- the solution is **unique**
- the solution depends on continuously on the data, i.e. *the* ***posterior is locally Lipschitz*** *in the data with respect to the Hellinger distance*

MLE’s ill-posedness

- As the lengthscale tends to infinity, the GP’s posterior covariance converges to 0.
    - No uncertainty estimation.
    - Small perturbations in the data result in drastic changes in the posterior.
- Thus, **MLE is ill-posed.**

What **does not** help?

This same phenomenon still occurs when we:

- Use cross-validation instead of MLE.
- Also estimate the prior mean.
- Estimate a scaling parameter σ alongisde the lengthscale.

What **does** help?

Regularization via MAP

Placing a hyper-prior on the lengthscale parameter changes our method to MAP estimation, where we find the max of:

$$\log p(\lambda|Y) = -\frac{1}{2}\ell(\lambda|Y) + \log p(\lambda) + const.$$

As lengthscale grows, MAP estimate of lambda will be finite even for m-constant data (preventing ill-posedness).

What **does** help?

Regularization via added observation noise

- Ill-posedness is **well-posed** if observed data is assumed to have added Gaussian noise with variance δ^2.
  - Modified log likelihood function is:

$$\ell(\lambda|Y) = Y_m^T(K_\lambda(X,X) + \delta^2 I_n)^{-1} Y_m + \log\det(K_\lambda(X,X) + \delta^2 I_n)$$

- However, inference is corrupted by artificial noise.

Pitfalls

However, even when we add observation noise, the covariance matrix can still be very close to **singular** if the lengthscale is too large.

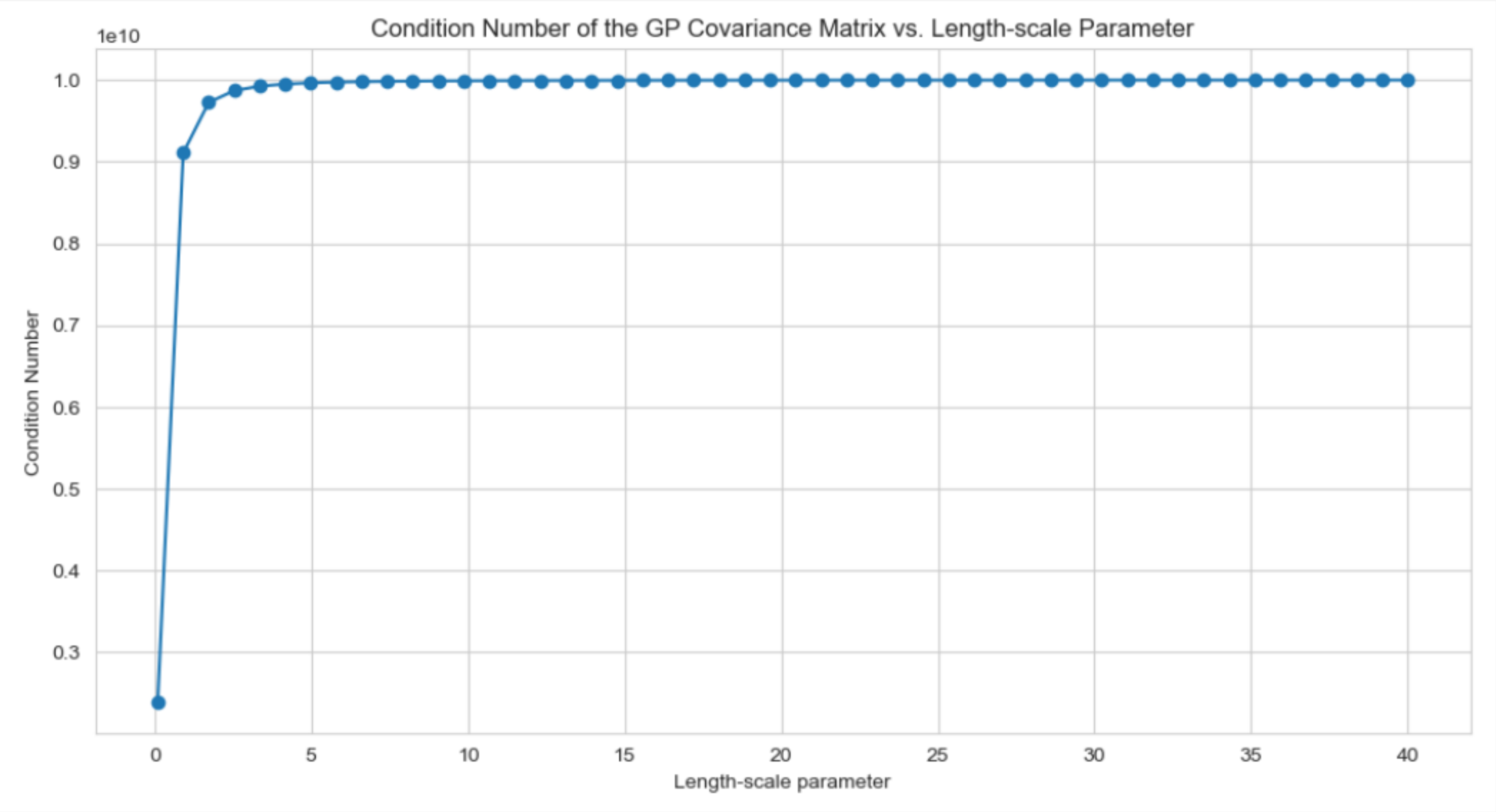


Figure 2: Covariance matrix's condition number as a function of lengthscale parameter.

Practical Recommendations

How to use these results in practice?

✗ Set an upper bound for the parameter estimate.

- Likely to simply obtain that value, which might be somewhat arbitrary.

✓ Check if data is m-constant before training the GP.

- If so, warn the user that training will diverge.

Future Work

- Obtain theoretical results on the behaviour of MLE for **“close” to m-constant** datasets.

- Study estimation of parameters for other types of kernels (e.g. **non-stationary**).

References

[1] Karvonen T, Oates CJ. Maximum Likelihood Estimation in Gaussian Process Regression is Ill-Posed. 2023.

[2] Deisenroth M, Luo Y, van der Wilk M, A Practical Guide to Gaussian Processes, 2020
https://infallible-thompson-49de36.netlify.app/

[3] Johannes B, Minnen D, Singh S, Sung Oh Hwang, Johnston N. Variational image compression with a scale hyperprior.

[4] Introduction to Gaussian Process Regression https://juanitorduz.github.io/gaussian_process_reg/

Appendix 1: Visualizing m-constant GP's

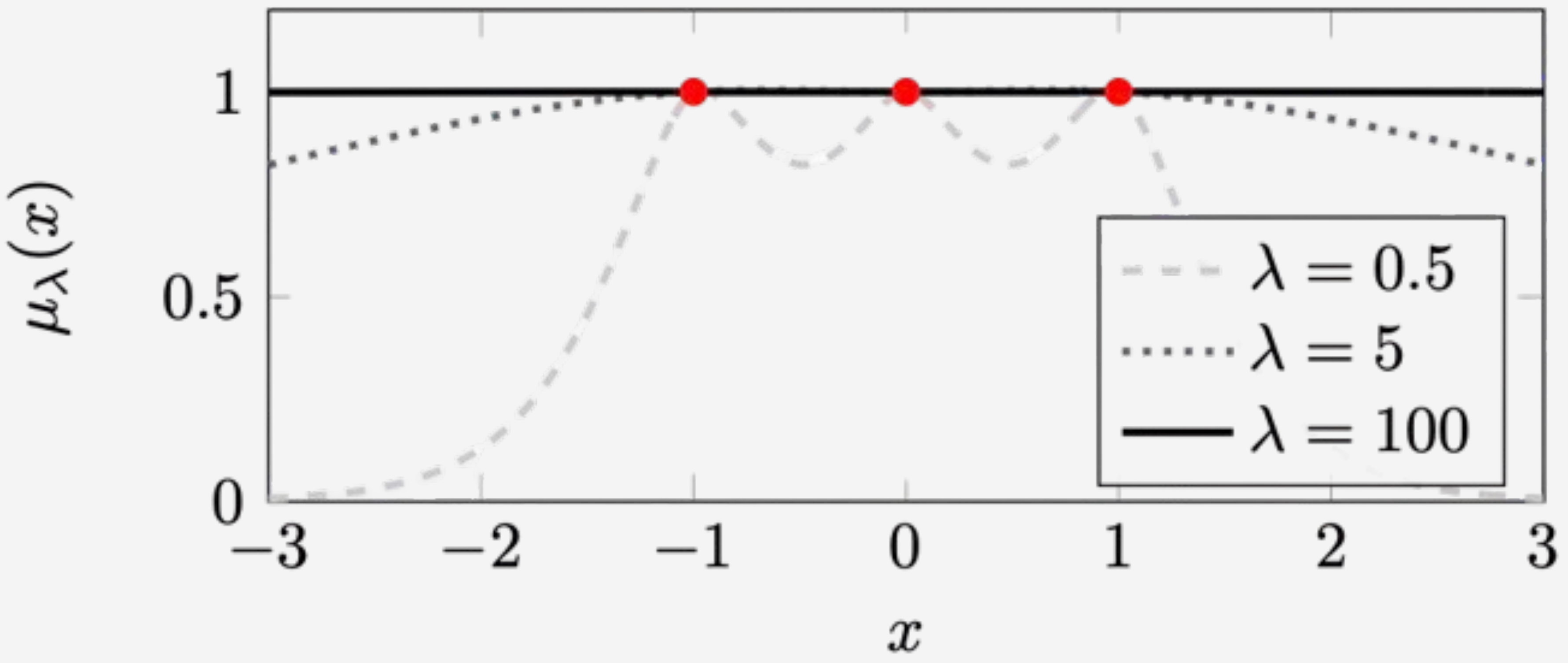
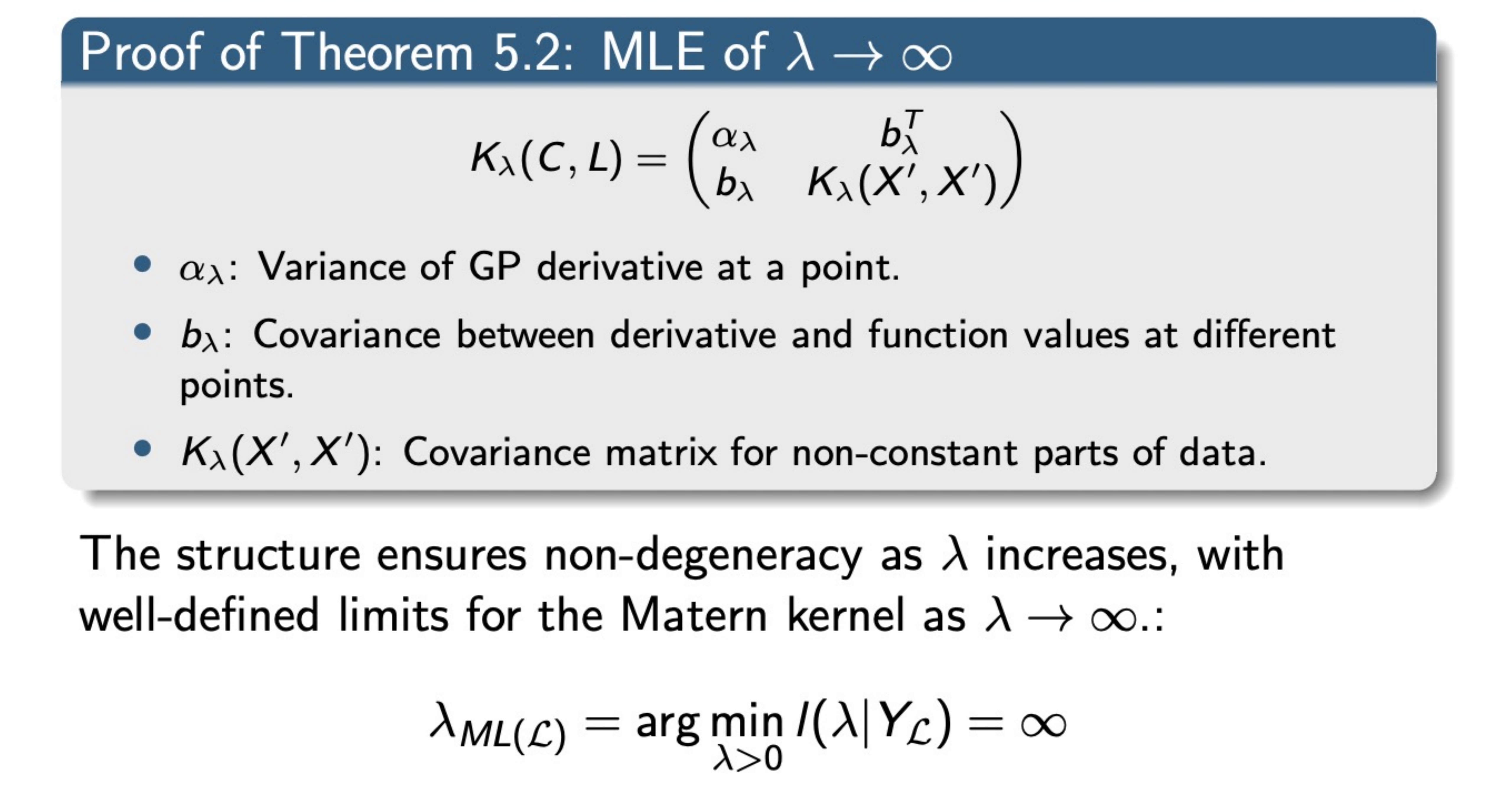


Figure A1: GP posterior mean across various lengthscales, for m-constant data. Karvonen et al. [1]

Appendix 2: Generalizations and Extensions

- Linear Information and General Kernels
- Product Kernels and Multiple Lengthscales
- Infinitely Smooth Kernels

Appendix 2.1: Linear Information and General Kernels



Proof of Theorem 5.2: MLE of $\lambda \to \infty$

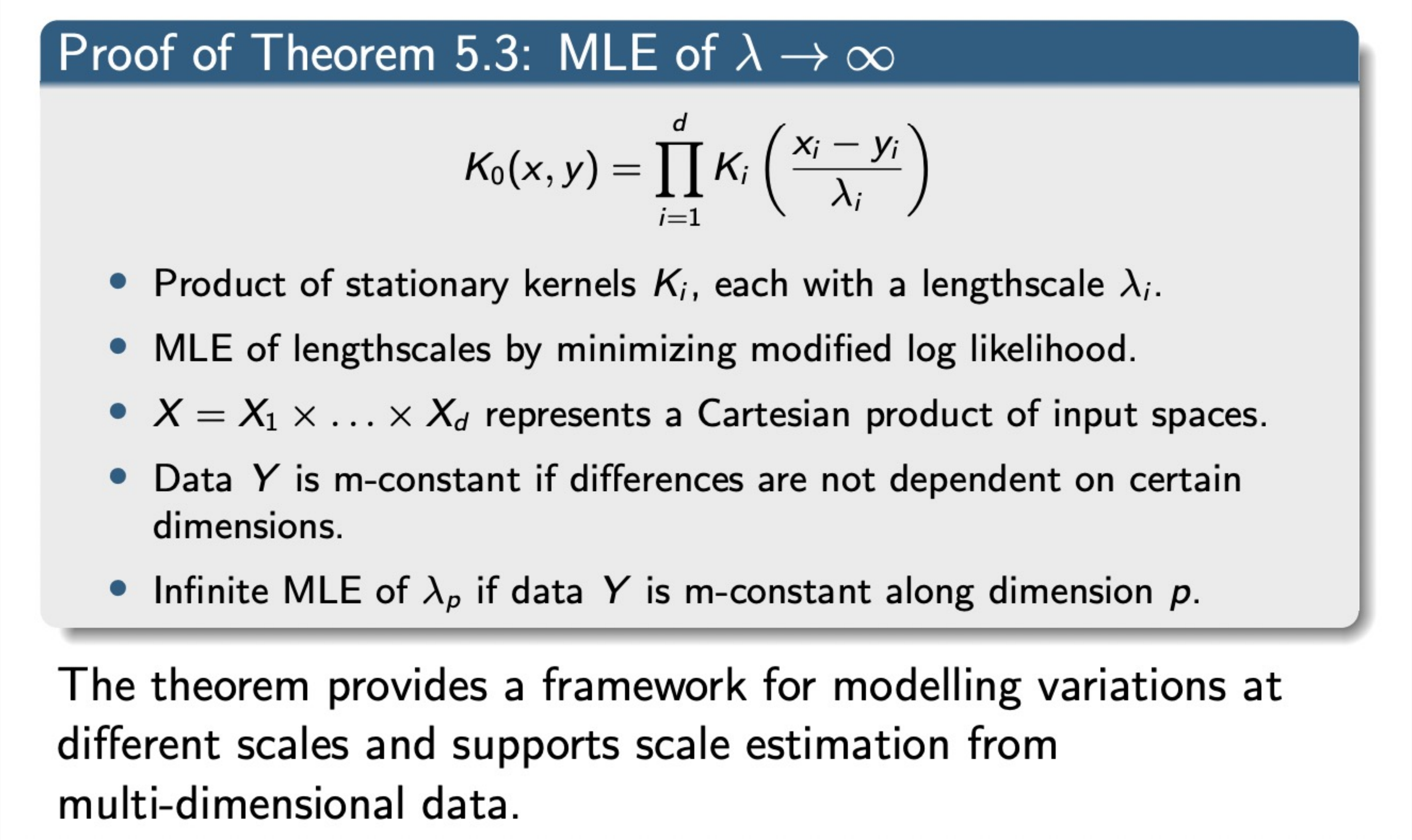
$$K_\lambda(C, L) = \begin{pmatrix} \alpha_\lambda & b_\lambda^T \\ b_\lambda & K_\lambda(X', X') \end{pmatrix}$$

- α_λ: Variance of GP derivative at a point.
- b_λ: Covariance between derivative and function values at different points.
- $K_\lambda(X', X')$: Covariance matrix for non-constant parts of data.

The structure ensures non-degeneracy as λ increases, with well-defined limits for the Matern kernel as $\lambda \to \infty$.:

$$\lambda_{ML(\mathcal{L})} = \arg\min_{\lambda > 0} l(\lambda | Y_{\mathcal{L}}) = \infty$$

Appendix 2.2: Product Kernels and Multiple Lengthscales



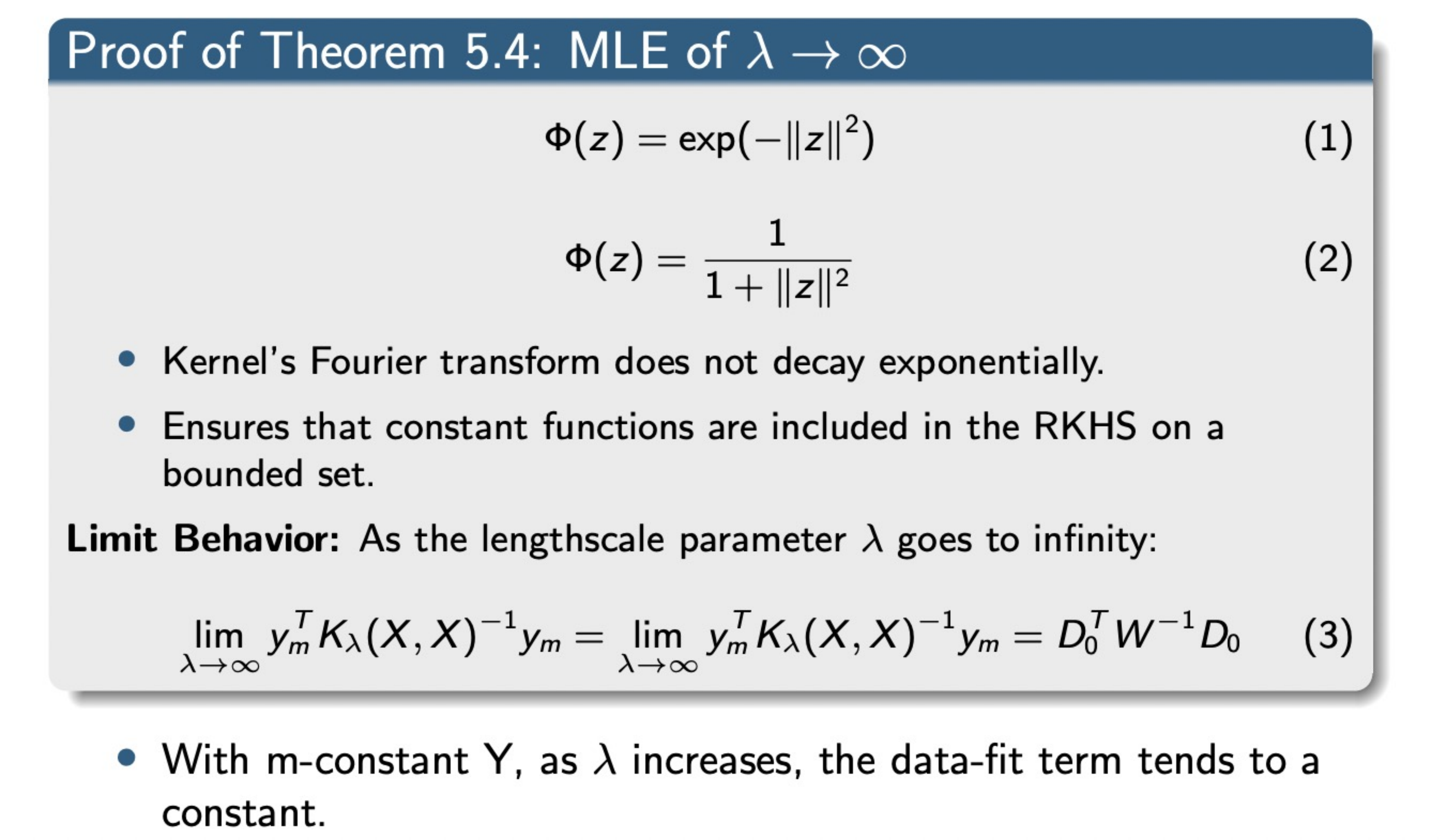
Proof of Theorem 5.3: MLE of $\lambda \to \infty$

$$K_0(x, y) = \prod_{i=1}^{d} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$$

- Product of stationary kernels K_i, each with a lengthscale λ_i.
- MLE of lengthscales by minimizing modified log likelihood.
- $X = X_1 \times \ldots \times X_d$ represents a Cartesian product of input spaces.
- Data Y is m-constant if differences are not dependent on certain dimensions.
- Infinite MLE of λ_p if data Y is m-constant along dimension p.

The theorem provides a framework for modelling variations at different scales and supports scale estimation from multi-dimensional data.

Appendix 2.3: Infinitely Smooth Kernels



Proof of Theorem 5.4: MLE of $\lambda \to \infty$

$$\Phi(z) = \exp(-\|z\|^2) \tag{1}$$

$$\Phi(z) = \frac{1}{1 + \|z\|^2} \tag{2}$$

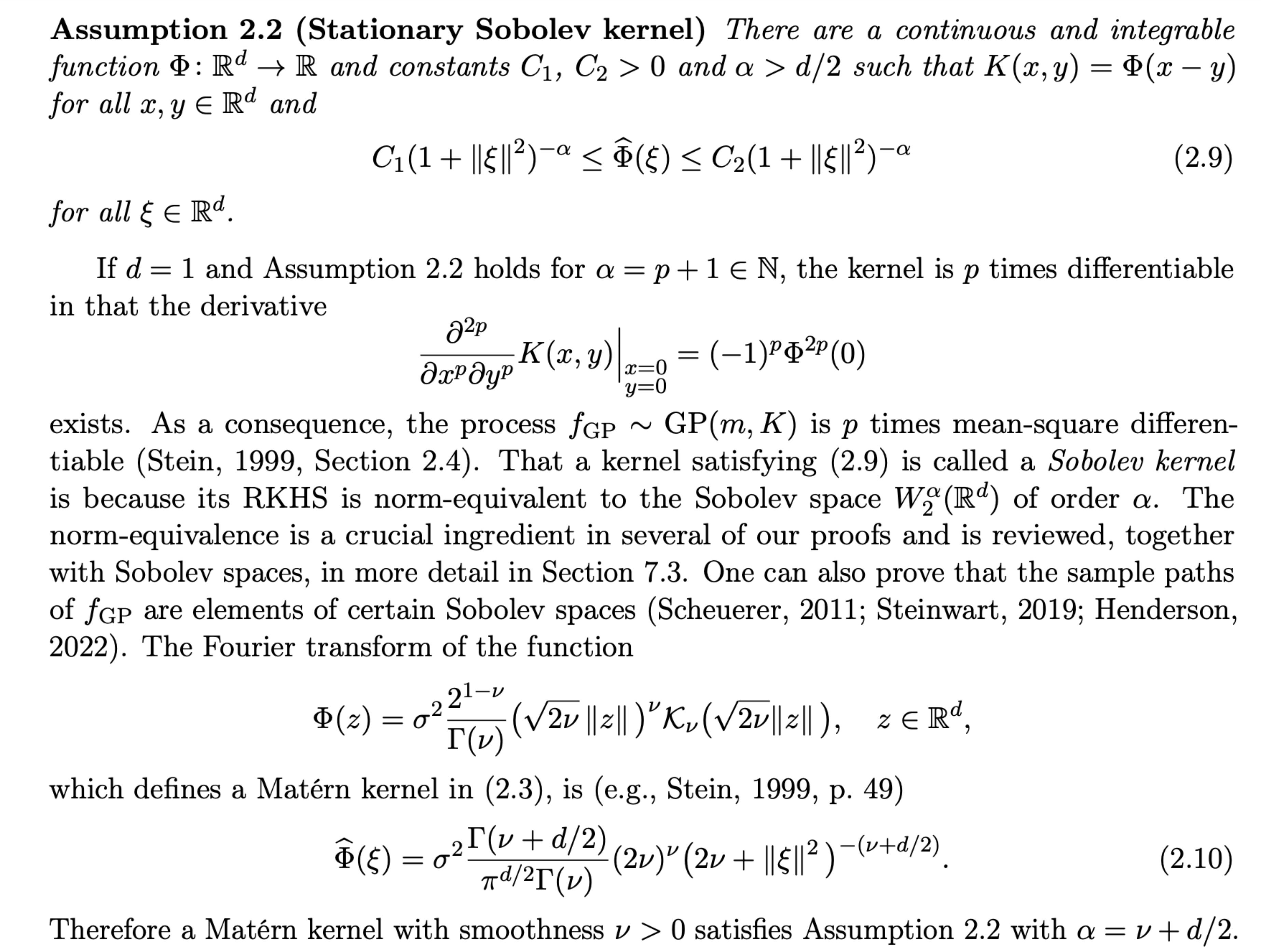
- Kernel's Fourier transform does not decay exponentially.
- Ensures that constant functions are included in the RKHS on a bounded set.

Limit Behavior: As the lengthscale parameter λ goes to infinity:

$$\lim_{\lambda \to \infty} y_m^T K_\lambda(X, X)^{-1} y_m = \lim_{\lambda \to \infty} y_m^T K_\lambda(X, X)^{-1} y_m = D_0^T W^{-1} D_0 \tag{3}$$

- With m-constant Y, as λ increases, the data-fit term tends to a constant.

Appendix 4: What are said "mild regularity conditions"?



Assumption 2.2 (Stationary Sobolev kernel) *There are a continuous and integrable function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ and constants C_1, $C_2 > 0$ and $\alpha > d/2$ such that $K(x,y) = \Phi(x-y)$ for all $x, y \in \mathbb{R}^d$ and*

$$C_1(1+\|\xi\|^2)^{-\alpha} \leq \widehat{\Phi}(\xi) \leq C_2(1+\|\xi\|^2)^{-\alpha} \tag{2.9}$$

for all $\xi \in \mathbb{R}^d$.

If $d = 1$ and Assumption 2.2 holds for $\alpha = p + 1 \in \mathbb{N}$, the kernel is p times differentiable in that the derivative

$$\frac{\partial^{2p}}{\partial x^p \partial y^p} K(x,y)\Big|_{\substack{x=0\\y=0}} = (-1)^p \Phi^{2p}(0)$$

exists. As a consequence, the process $f_{\mathrm{GP}} \sim \mathrm{GP}(m, K)$ is p times mean-square differentiable (Stein, 1999, Section 2.4). That a kernel satisfying (2.9) is called a *Sobolev kernel* is because its RKHS is norm-equivalent to the Sobolev space $W_2^\alpha(\mathbb{R}^d)$ of order α. The norm-equivalence is a crucial ingredient in several of our proofs and is reviewed, together with Sobolev spaces, in more detail in Section 7.3. One can also prove that the sample paths of f_{GP} are elements of certain Sobolev spaces (Scheuerer, 2011; Steinwart, 2019; Henderson, 2022). The Fourier transform of the function

$$\Phi(z) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \big(\sqrt{2\nu}\,\|z\|\big)^\nu \mathcal{K}_\nu\big(\sqrt{2\nu}\|z\|\big), \quad z \in \mathbb{R}^d,$$

which defines a Matérn kernel in (2.3), is (e.g., Stein, 1999, p. 49)

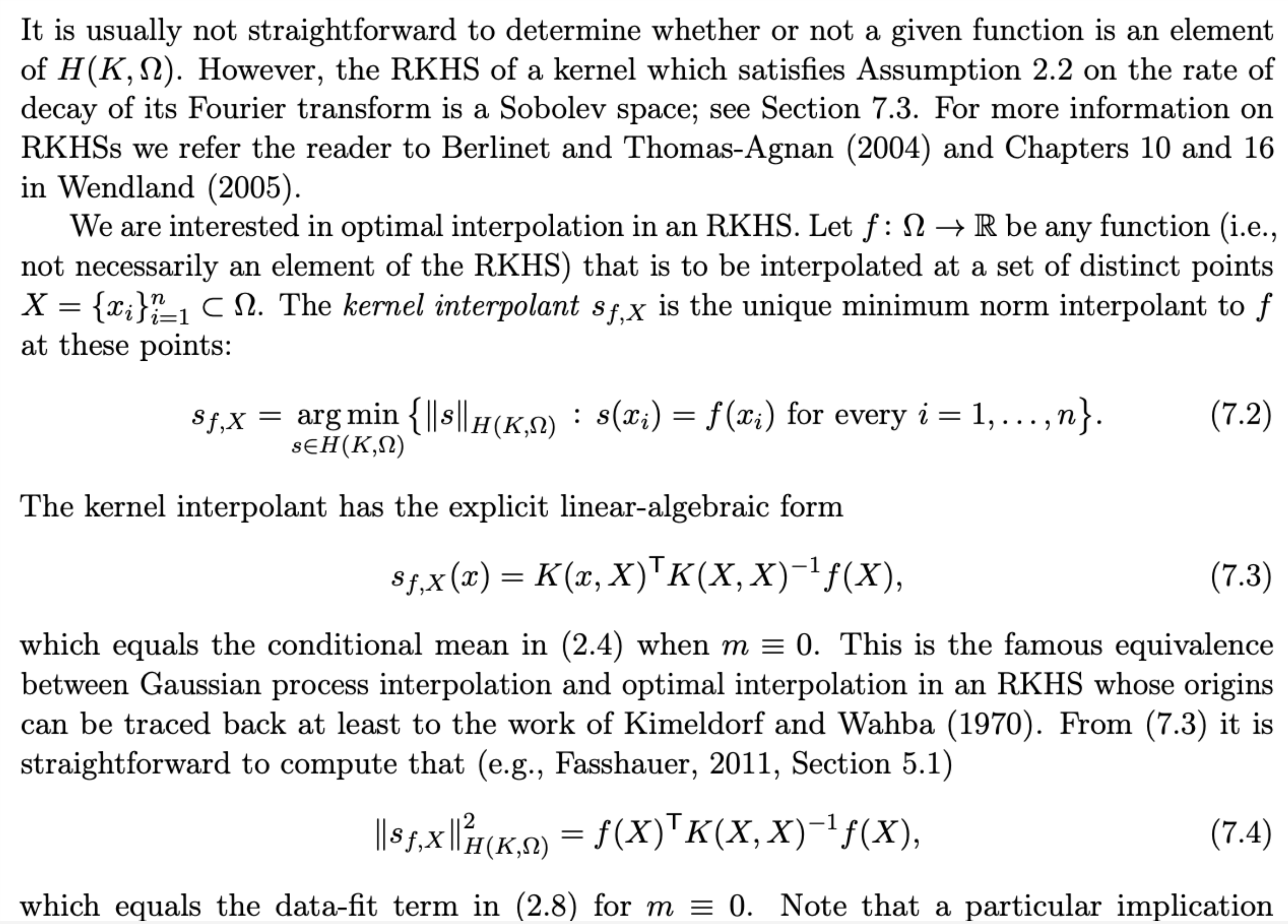
$$\widehat{\Phi}(\xi) = \sigma^2 \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)} (2\nu)^\nu \big(2\nu + \|\xi\|^2\big)^{-(\nu+d/2)}. \tag{2.10}$$

Therefore a Matérn kernel with smoothness $\nu > 0$ satisfies Assumption 2.2 with $\alpha = \nu + d/2$.

In short, we want our kernel Φ to have a polynomially decaying Fourier transform.

Also, our kernel must be stationary and positive definite

Appendix 4b: Why is this relevant?



It is usually not straightforward to determine whether or not a given function is an element of $H(K,\Omega)$. However, the RKHS of a kernel which satisfies Assumption 2.2 on the rate of decay of its Fourier transform is a Sobolev space; see Section 7.3. For more information on RKHSs we refer the reader to Berlinet and Thomas-Agnan (2004) and Chapters 10 and 16 in Wendland (2005).

We are interested in optimal interpolation in an RKHS. Let $f\colon \Omega \to \mathbb{R}$ be any function (i.e., not necessarily an element of the RKHS) that is to be interpolated at a set of distinct points $X = \{x_i\}_{i=1}^n \subset \Omega$. The *kernel interpolant* $s_{f,X}$ is the unique minimum norm interpolant to f at these points:

$$s_{f,X} = \operatorname*{arg\,min}_{s \in H(K,\Omega)} \left\{ \|s\|_{H(K,\Omega)} \, : \, s(x_i) = f(x_i) \text{ for every } i = 1, \ldots, n \right\}. \tag{7.2}$$

The kernel interpolant has the explicit linear-algebraic form

$$s_{f,X}(x) = K(x,X)^\mathsf{T} K(X,X)^{-1} f(X), \tag{7.3}$$

which equals the conditional mean in (2.4) when $m \equiv 0$. This is the famous equivalence between Gaussian process interpolation and optimal interpolation in an RKHS whose origins can be traced back at least to the work of Kimeldorf and Wahba (1970). From (7.3) it is straightforward to compute that (e.g., Fasshauer, 2011, Section 5.1)

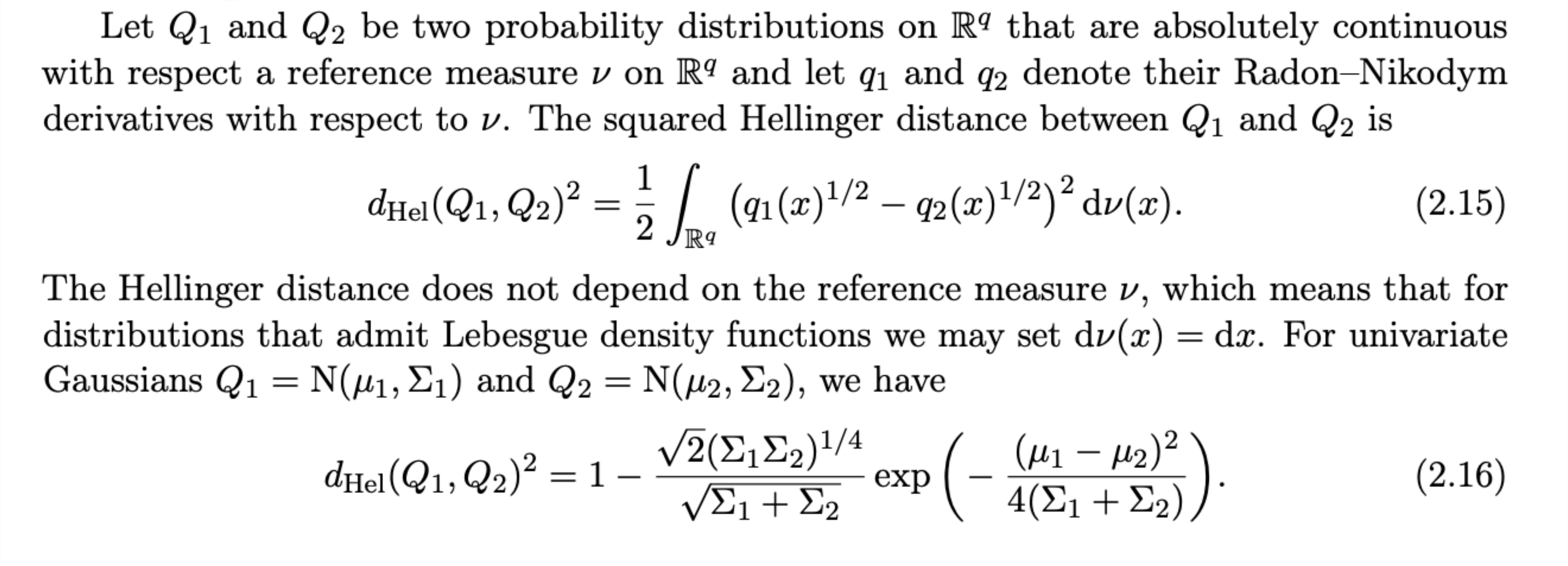
$$\|s_{f,X}\|_{H(K,\Omega)}^2 = f(X)^\mathsf{T} K(X,X)^{-1} f(X), \tag{7.4}$$

which equals the data-fit term in (2.8) for $m \equiv 0$. Note that a particular implication

We can use the equivalence between Gaussian process interpolation and optimal interpolation in an RKHS in our proofs

This allows us to bound the data fit term!

Appendix 5a: What was that Hellinger distance stuff?



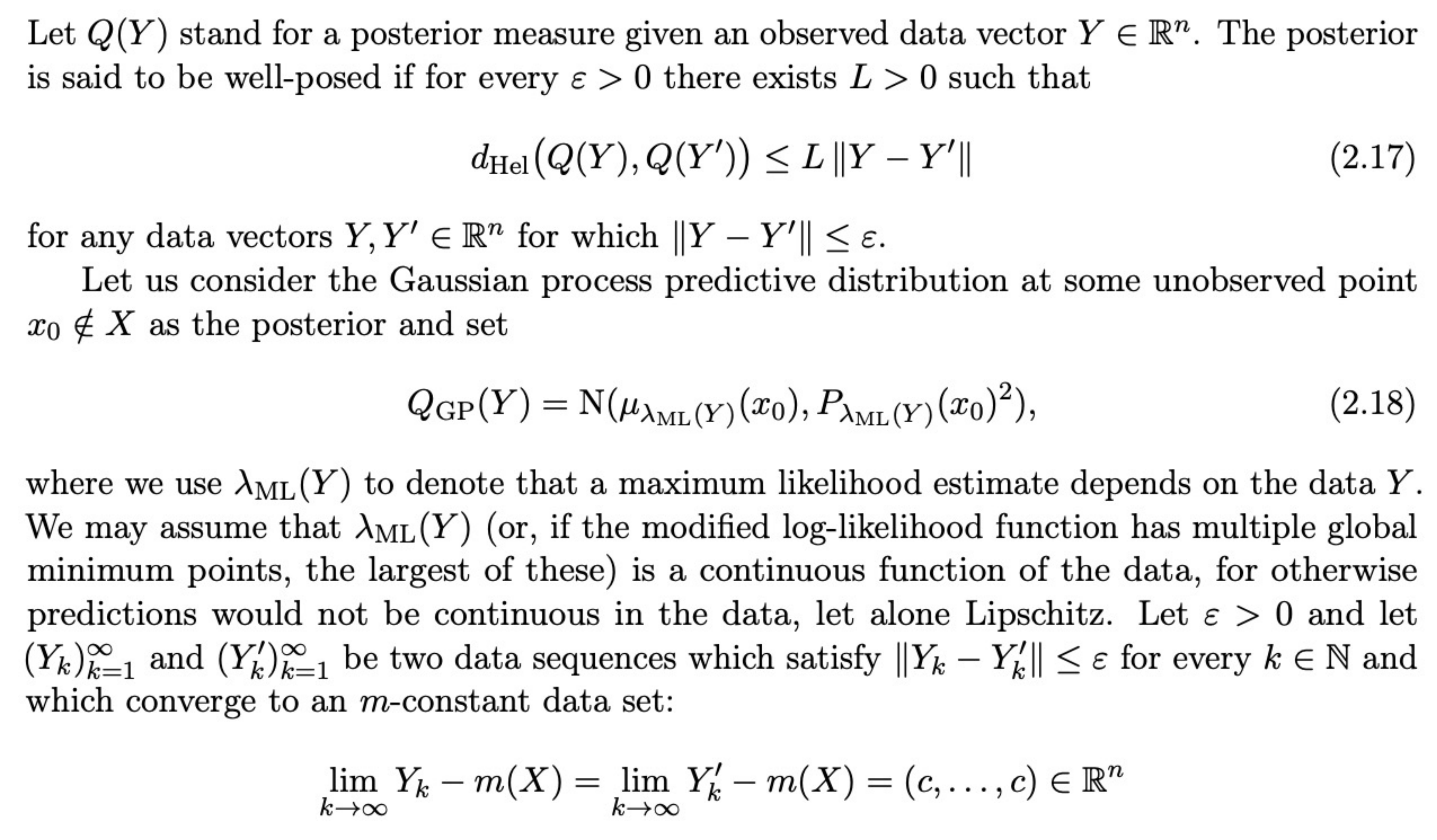
Let Q_1 and Q_2 be two probability distributions on $\mathbb{R}^q$ that are absolutely continuous with respect a reference measure ν on $\mathbb{R}^q$ and let q_1 and q_2 denote their Radon–Nikodym derivatives with respect to ν. The squared Hellinger distance between Q_1 and Q_2 is

$$d_{\mathrm{Hel}}(Q_1, Q_2)^2 = \frac{1}{2} \int_{\mathbb{R}^q} \left(q_1(x)^{1/2} - q_2(x)^{1/2} \right)^2 \mathrm{d}\nu(x). \tag{2.15}$$

The Hellinger distance does not depend on the reference measure ν, which means that for distributions that admit Lebesgue density functions we may set $\mathrm{d}\nu(x) = \mathrm{d}x$. For univariate Gaussians $Q_1 = \mathrm{N}(\mu_1, \Sigma_1)$ and $Q_2 = \mathrm{N}(\mu_2, \Sigma_2)$, we have

$$d_{\mathrm{Hel}}(Q_1, Q_2)^2 = 1 - \frac{\sqrt{2}(\Sigma_1 \Sigma_2)^{1/4}}{\sqrt{\Sigma_1 + \Sigma_2}} \exp\left(- \frac{(\mu_1 - \mu_2)^2}{4(\Sigma_1 + \Sigma_2)} \right). \tag{2.16}$$

Appendix 5b: How does it relate to ill-posedness?



Let $Q(Y)$ stand for a posterior measure given an observed data vector $Y \in \mathbb{R}^n$. The posterior is said to be well-posed if for every $\varepsilon > 0$ there exists $L > 0$ such that

$$d_{\mathrm{Hel}}\big(Q(Y), Q(Y')\big) \leq L \, \|Y - Y'\| \tag{2.17}$$

for any data vectors $Y, Y' \in \mathbb{R}^n$ for which $\|Y - Y'\| \leq \varepsilon$.

Let us consider the Gaussian process predictive distribution at some unobserved point $x_0 \notin X$ as the posterior and set

$$Q_{\mathrm{GP}}(Y) = \mathrm{N}(\mu_{\lambda_{\mathrm{ML}}(Y)}(x_0), P_{\lambda_{\mathrm{ML}}(Y)}(x_0)^2), \tag{2.18}$$

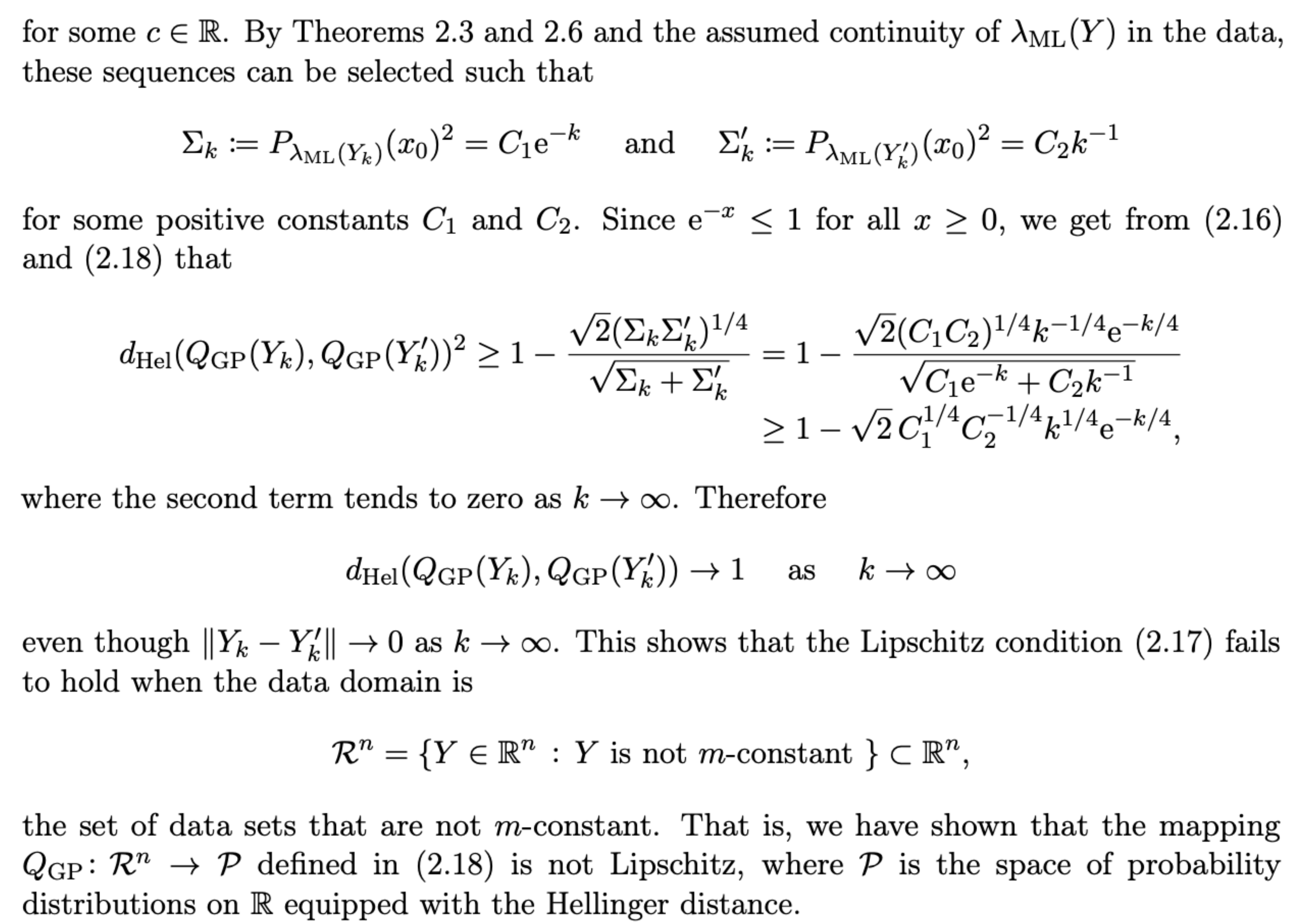
where we use $\lambda_{\mathrm{ML}}(Y)$ to denote that a maximum likelihood estimate depends on the data Y. We may assume that $\lambda_{\mathrm{ML}}(Y)$ (or, if the modified log-likelihood function has multiple global minimum points, the largest of these) is a continuous function of the data, for otherwise predictions would not be continuous in the data, let alone Lipschitz. Let $\varepsilon > 0$ and let $(Y_k)_{k=1}^{\infty}$ and $(Y_k')_{k=1}^{\infty}$ be two data sequences which satisfy $\|Y_k - Y_k'\| \leq \varepsilon$ for every $k \in \mathbb{N}$ and which converge to an m-constant data set:

$$\lim_{k \to \infty} Y_k - m(X) = \lim_{k \to \infty} Y_k' - m(X) = (c, \ldots, c) \in \mathbb{R}^n$$

Definition of Lipschitz continuous

Define two sequences that converge to a m-constant data set

Appendix 5b: How does it relate to ill-posedness? (cont.)



for some $c \in \mathbb{R}$. By Theorems 2.3 and 2.6 and the assumed continuity of $\lambda_{\mathrm{ML}}(Y)$ in the data, these sequences can be selected such that

$$\Sigma_k := P_{\lambda_{\mathrm{ML}}(Y_k)}(x_0)^2 = C_1 \mathrm{e}^{-k} \quad \text{and} \quad \Sigma_k' := P_{\lambda_{\mathrm{ML}}(Y_k')}(x_0)^2 = C_2 k^{-1}$$

for some positive constants C_1 and C_2. Since $\mathrm{e}^{-x} \leq 1$ for all $x \geq 0$, we get from (2.16) and (2.18) that

$$\begin{aligned} d_{\mathrm{Hel}}(Q_{\mathrm{GP}}(Y_k), Q_{\mathrm{GP}}(Y_k'))^2 \geq 1 - \frac{\sqrt{2}(\Sigma_k \Sigma_k')^{1/4}}{\sqrt{\Sigma_k + \Sigma_k'}} &= 1 - \frac{\sqrt{2}(C_1 C_2)^{1/4} k^{-1/4} \mathrm{e}^{-k/4}}{\sqrt{C_1 \mathrm{e}^{-k} + C_2 k^{-1}}} \\ &\geq 1 - \sqrt{2}\, C_1^{1/4} C_2^{-1/4} k^{1/4} \mathrm{e}^{-k/4}, \end{aligned}$$

where the second term tends to zero as $k \to \infty$. Therefore

$$d_{\mathrm{Hel}}(Q_{\mathrm{GP}}(Y_k), Q_{\mathrm{GP}}(Y_k')) \to 1 \quad \text{as} \quad k \to \infty$$

even though $\|Y_k - Y_k'\| \to 0$ as $k \to \infty$. This shows that the Lipschitz condition (2.17) fails to hold when the data domain is

$$\mathcal{R}^n = \{ Y \in \mathbb{R}^n : Y \text{ is not } m\text{-constant } \} \subset \mathbb{R}^n,$$

the set of data sets that are not m-constant. That is, we have shown that the mapping $Q_{\mathrm{GP}} \colon \mathcal{R}^n \to \mathcal{P}$ defined in (2.18) is not Lipschitz, where $\mathcal{P}$ is the space of probability distributions on $\mathbb{R}$ equipped with the Hellinger distance.

Be tricky with selection of the sequences

This is a consequence of the covariance equaling 0, as any data that is not m-constant will have a Hellinger distance of 1