Intro to Gaussian Processes

Data Science in Electron Microscopy

Philipp Pelz

2024-01-09

https://github.com/ECLIPSE-Lab/WS24_DataScienceForEM

Introduction to Gaussian Processes 1

- Gaussian processes provide a mechanism for directly reasoning about the high-level properties of functions that could fit our data.

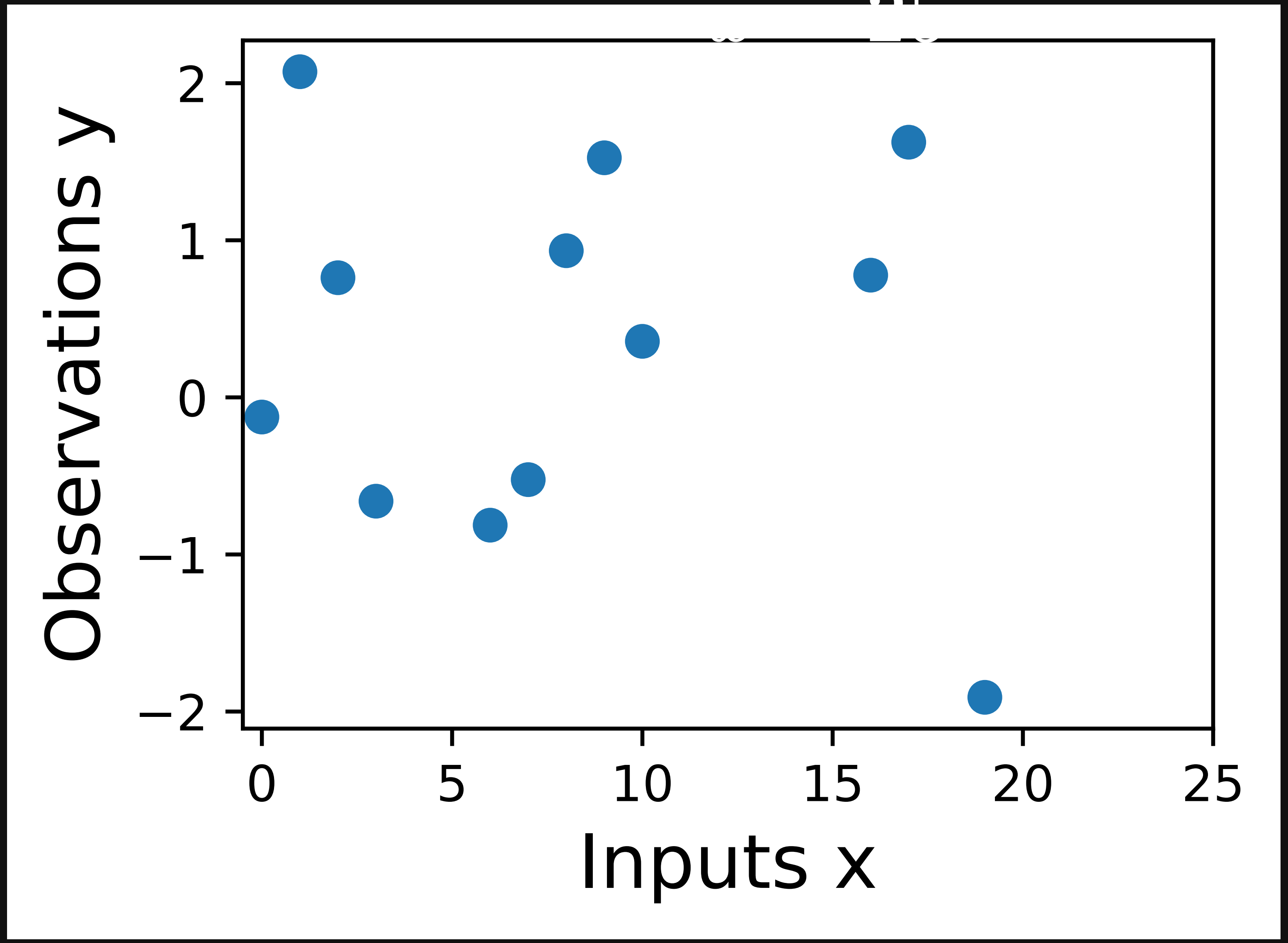
- may have a sense of whether these functions are quickly varying, periodic, involve conditional independencies, or translation invariance.

- Gaussian processes: easily incorporate these properties into our model, by directly specifying a Gaussian distribution over the function values that could fit our data.

Introduction to Gaussian Processes 2

- Suppose we observe the following dataset, of regression targets (outputs), , indexed by inputs, .
- example: targets could be changes in carbon dioxide concentrations, inputs y could be the times at which these targets have been recorded

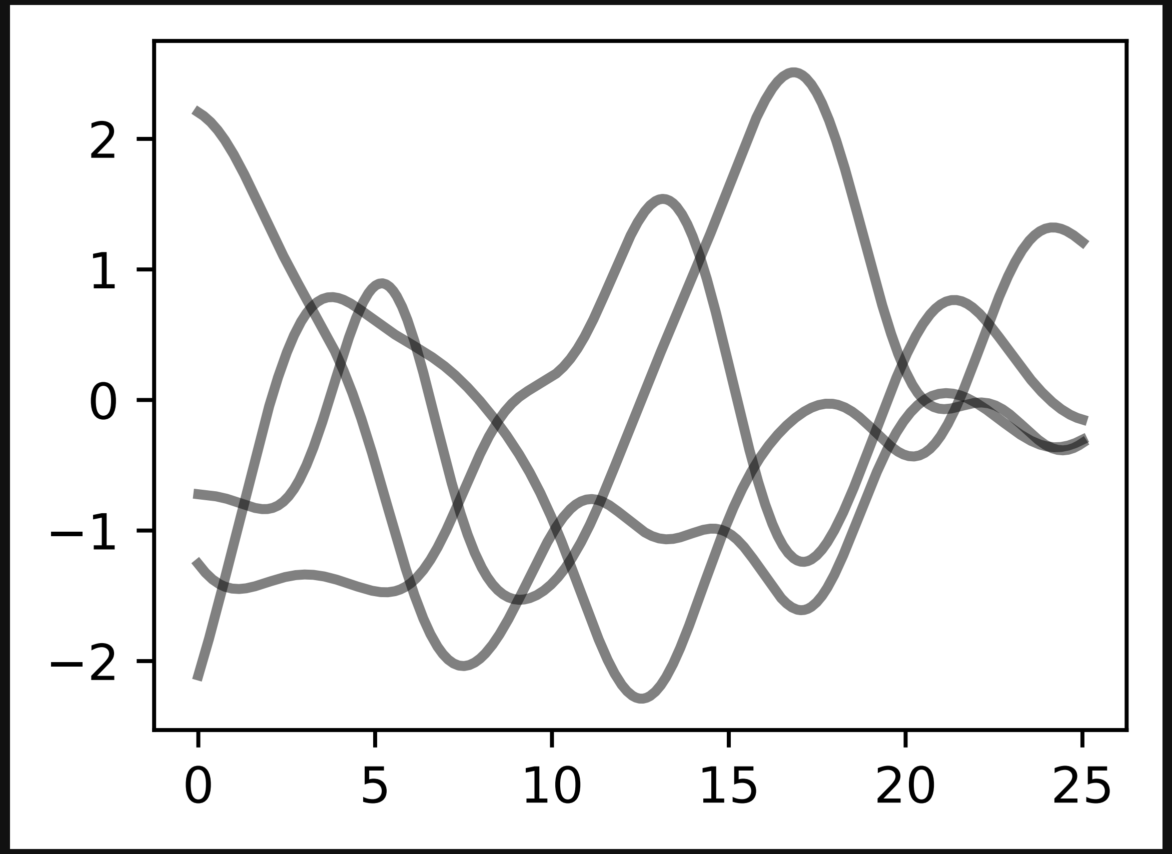
- What are some features of the data? How quickly does it seem to varying? Do we have data points collected at regular intervals, or are there missing inputs? How would you imagine filling in the missing regions, or forecasting up until ?



Introduction to Gaussian Processes 3

- start by specifying a prior distribution over what types of functions we might believe to be reasonable.

- show several sample functions from a Gaussian process. Does this prior look reasonable? we are not looking for functions that fit our dataset, but instead for specifying reasonable high-level properties of the solutions, such as how quickly they vary with inputs. Note that we will see code for reproducing all of the plots in this notebook, in the next notebooks on priors and inference.



Sample prior functions that we may want to represent with our model.

Introduction to Gaussian Processes 3.5

LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

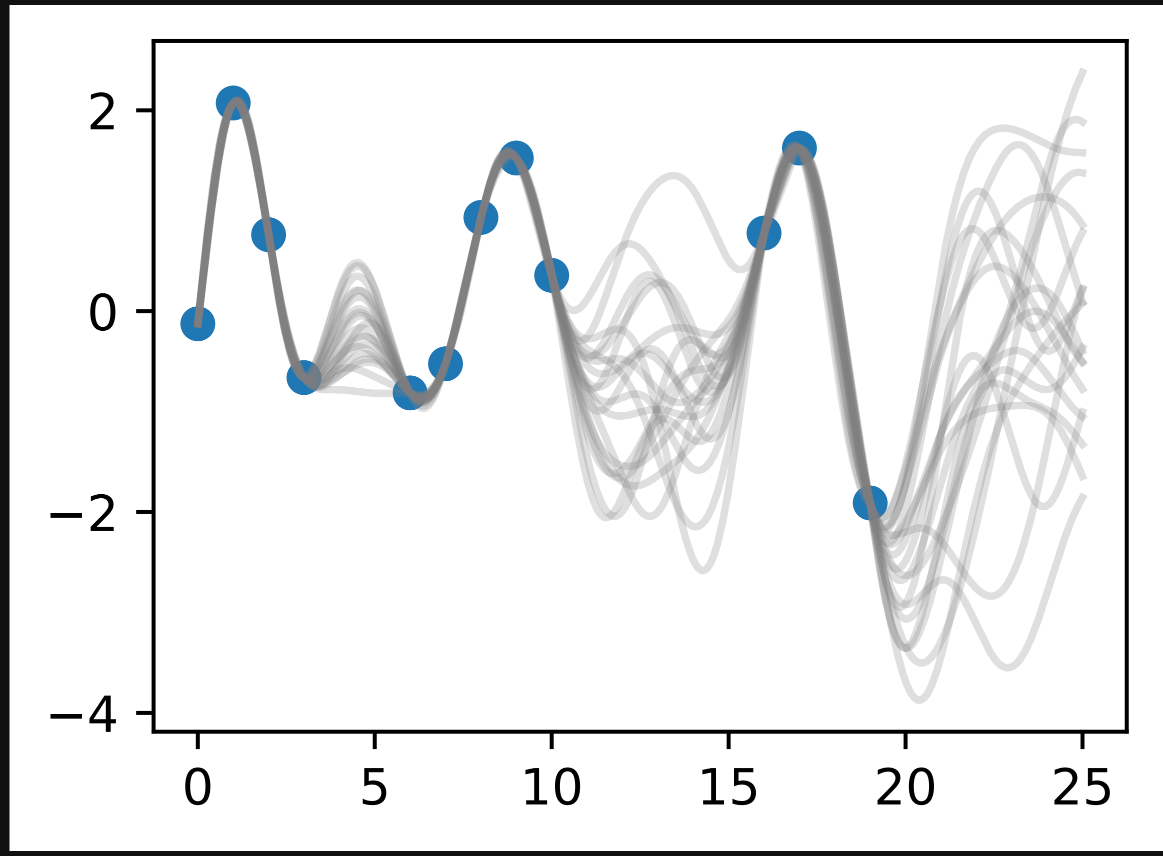
POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

Bayes theorem

Introduction to Gaussian Processes 4

Once we condition on data, we can use this prior to infer a posterior distribution over functions that could fit the data. Here we show sample posterior functions.



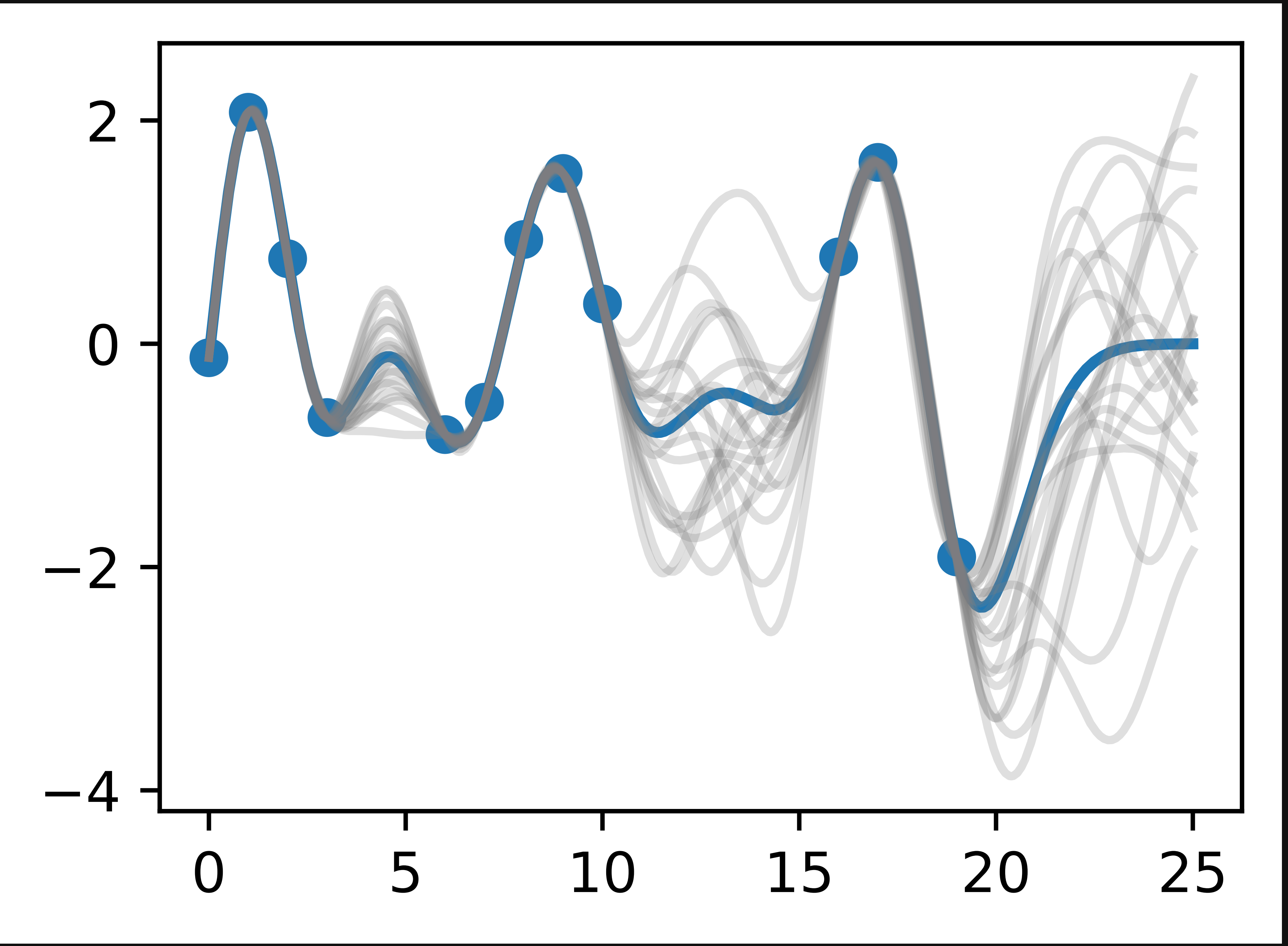
Sample posterior functions, once we have observed the data.

- each of these functions are entirely consistent with our data, perfectly running through each observation.

- In order to use these posterior samples to make predictions, we can average the values of every possible sample function from the posterior, to create the curve below, in thick blue.

- Note that we do not actually have to take an infinite number of samples to compute this expectation; as we will see later, we can compute the expectation in closed form.

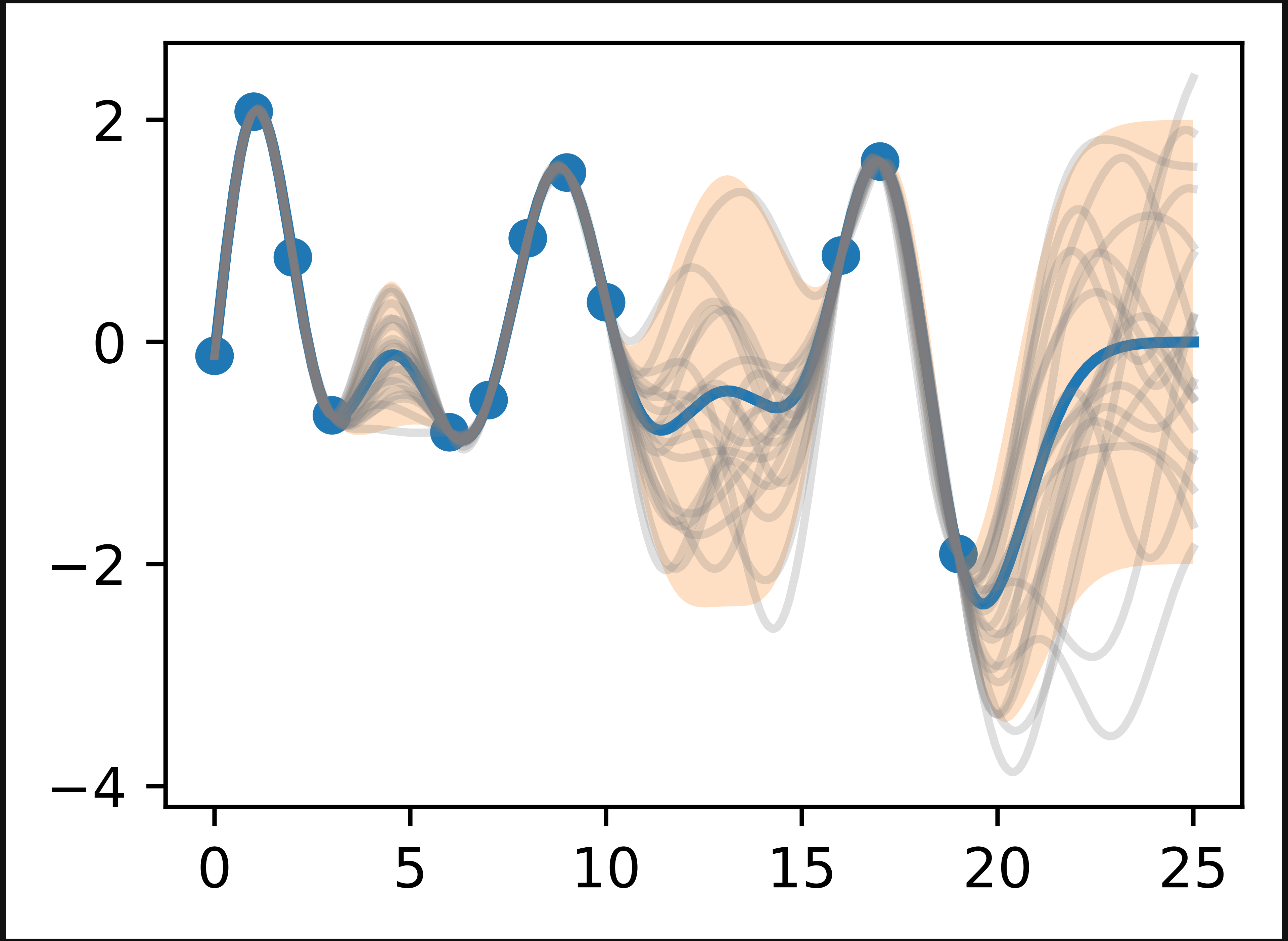
Introduction to Gaussian Processes 5



Posterior samples, alongside posterior mean, which can be used for point predictions, in blue.

- may also want a representation of uncertainty, so we know how confident we should be in our predictions.
- Intuitively: more variability in the sample posterior functions –> more uncertainty
- *epistemic uncertainty*, which is the *reducible uncertainty* associated with lack of information.
- acquire more data –> this type of uncertainty disappears, as there will be increasingly fewer solutions consistent with what we observe.
- Like with the posterior mean, we can compute the posterior variance (the variability of these functions in the posterior) in closed form.

Introduction to Gaussian Processes 6



Posterior samples, including 95% credible set.

- shade: two times the posterior standard deviation on either side of the mean, creating a *credible interval* that has a 95% probability of containing the true value of the function for any input x.
- plot looks somewhat cleaner if we remove the posterior samples, simply visualizing the data, posterior mean, and 95% credible set.
- Notice how the uncertainty grows away from the data, a property of epistemic uncertainty.

Introduction to Gaussian Processes 7

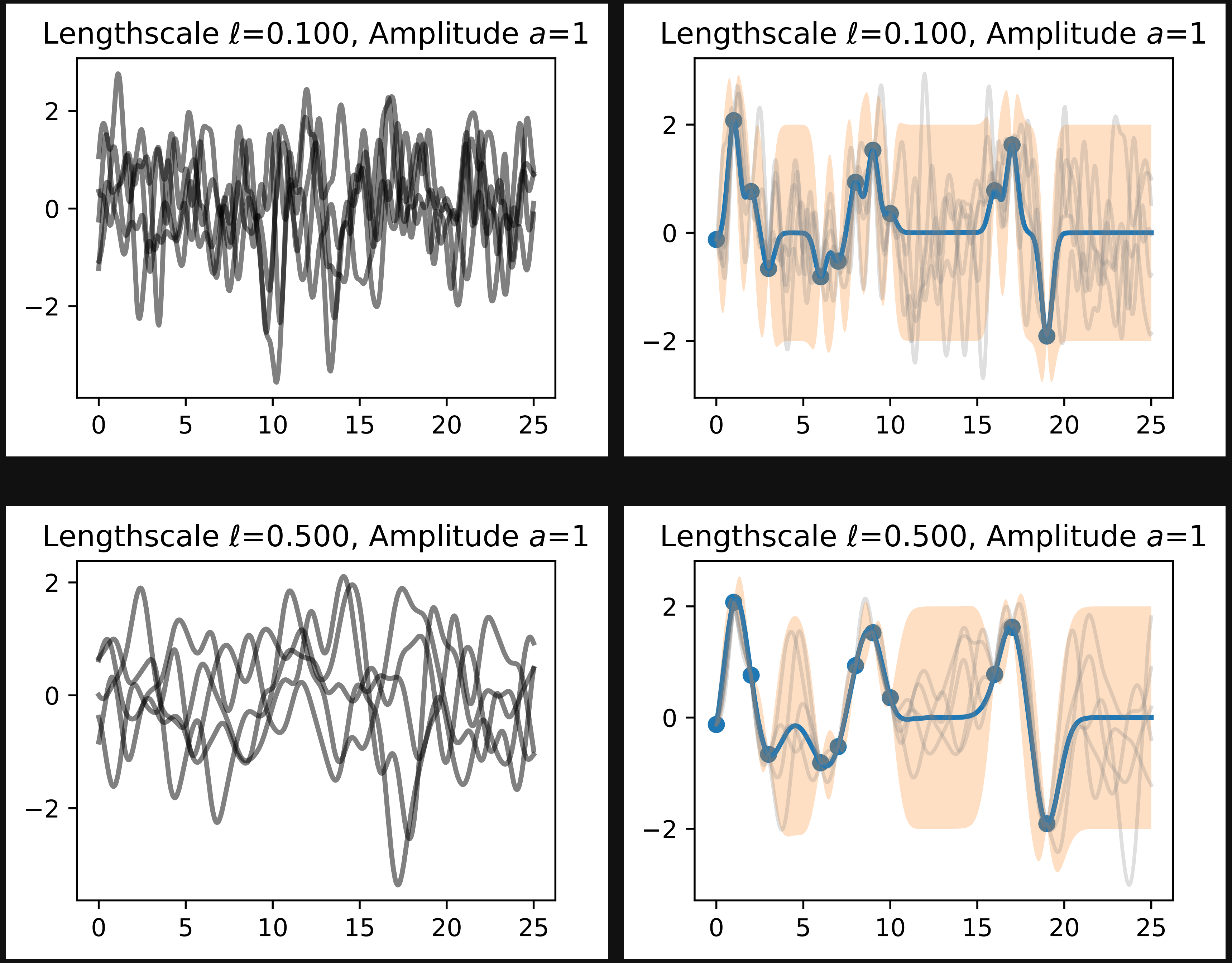
Introduction to Gaussian Processes 8

$$k_{\mathrm{RBF}}(x, x') = \mathrm{Cov}(f(x), f(x')) = a^2 \exp\Big(-\frac{1}{2\ell^2}||x$$

- The *length-scale* has a particularly pronounced effect on the predictions and uncertainty of a GP. At $||x - x$, the covariance between a pair of function values is $a^2 \exp(-0.5)$.
- At larger distances than ℓ, the values of the function values becomes nearly uncorrelated. This means that if we want to make a prediction at a point x_*, then function values with inputs x such that $||x - x'|| > \ell$ will not have a strong effect on our predictions.

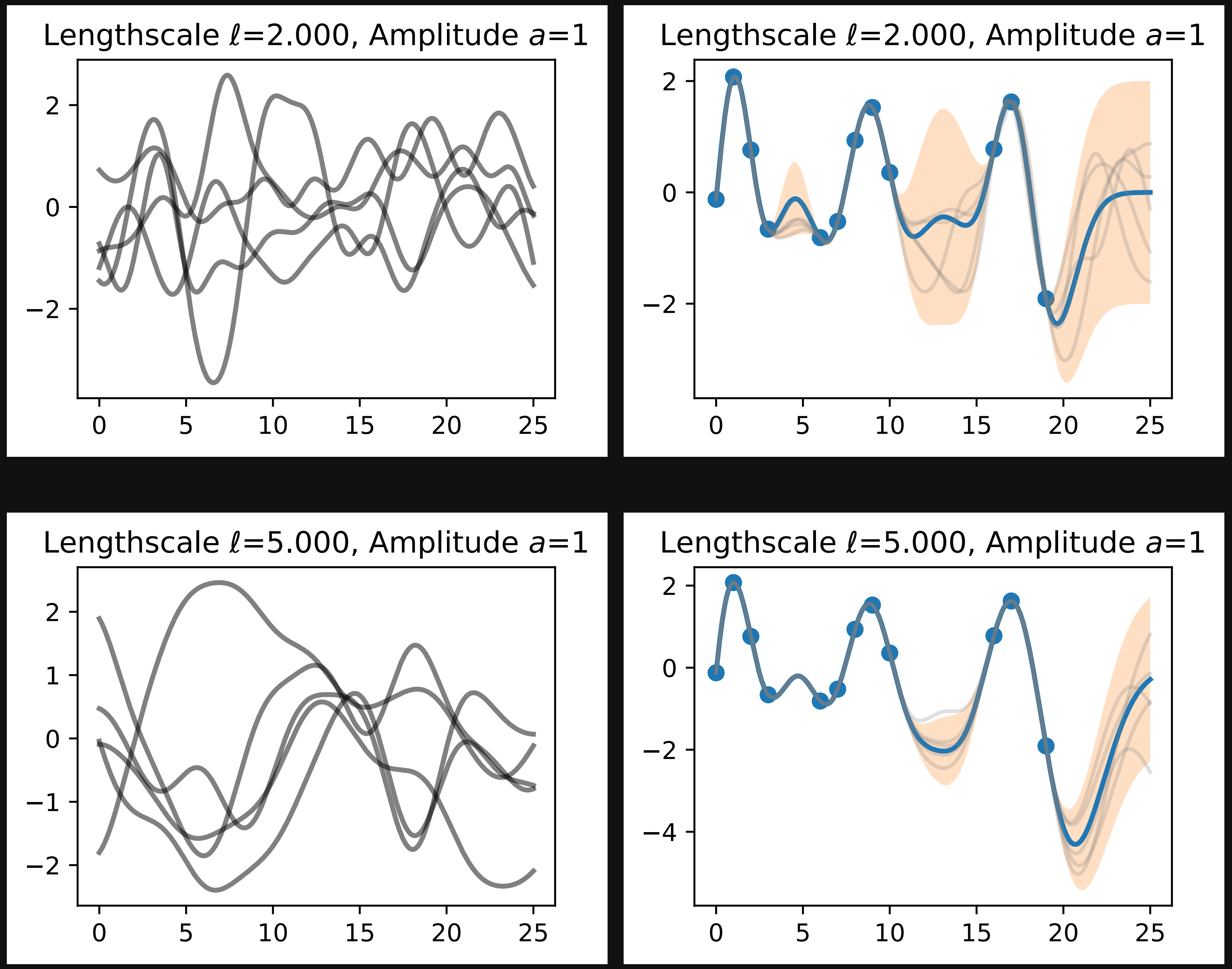
Introduction to Gaussian Processes 9

- how changing the lengthscale affects sample prior and posterior functions, and credible sets. The above fits use a length-scale of 2. Let’s now consider $\ell = 0.1, 0.5, 2, 5, 10$.



- A length-scale of 0.1 is very small relative to the range of the input domain we are considering, 25. For example, the values of the function at $x = 5$ and $= 10$ will have essentially no correlation at such a length-scale.
- On the other hand, for a length-scale of 10, the function values at these inputs will be highly correlated.
- Note that the vertical scale changes in the following figures.

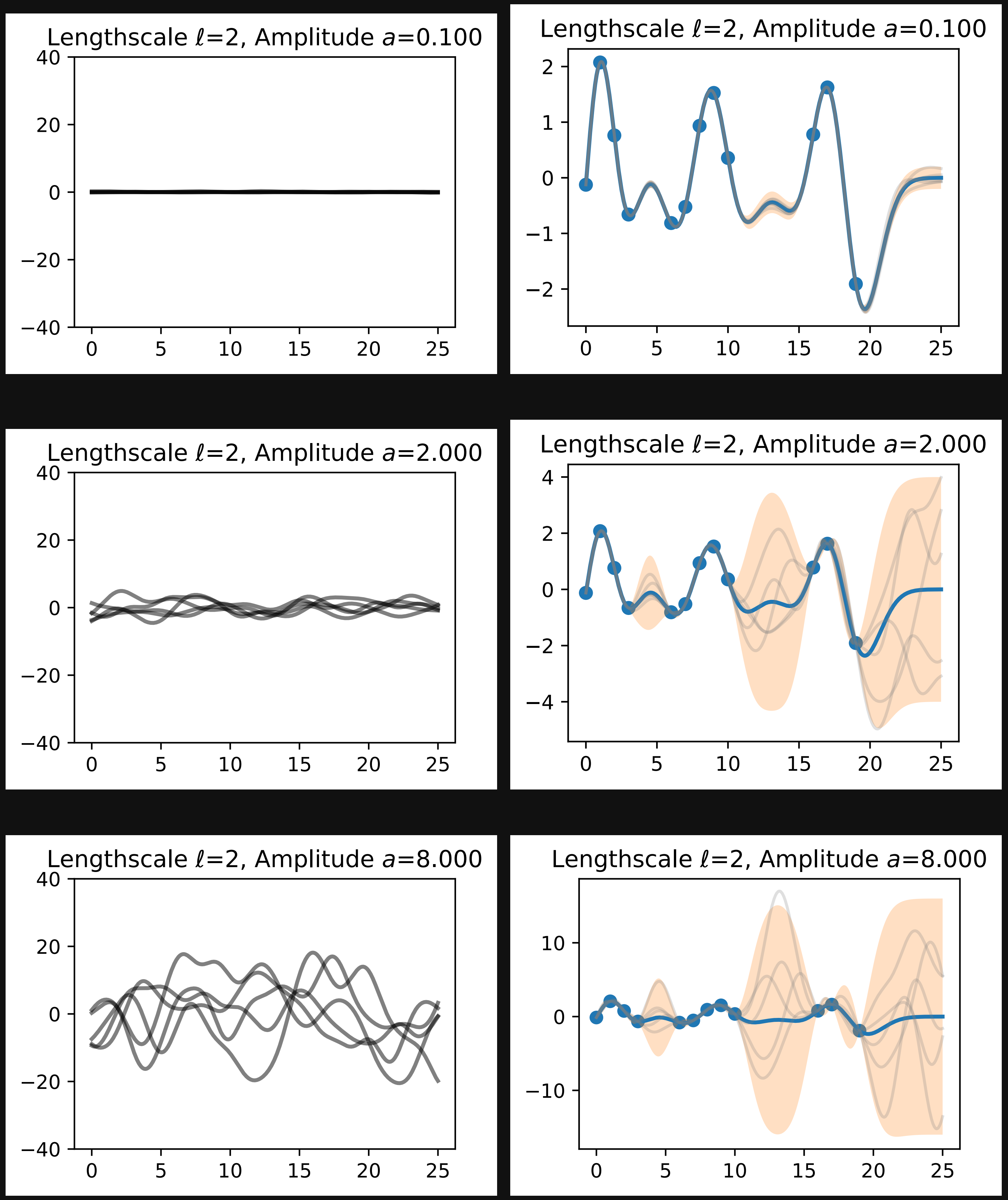
Introduction to Gaussian Processes 10



- length-scale of 0.1 is very small relative to the range of the input domain we are considering, 25.
- For example, the values of the function at $x = 5$ and $x = 10$ will have essentially no correlation at such a length-scale.
- On the other hand, for a length-scale of 10, the function values at these inputs will be highly correlated.
- Note that the vertical scale changes in the following figures.

- as the length-scale increases the ‘wiggliness’ of the functions decrease, and our uncertainty decreases.
- If the length-scale is small, the uncertainty will quickly increase as we move away from the data, as the datapoints become less informative about the function values.

Introduction to Gaussian Processes 11



- now vary the amplitude parameter, holding the length-scale fixed at 2.
- Note the vertical scale is held fixed for the prior samples, and varies for the posterior samples, so you can clearly see both the increasing scale of the function, and the fits to the data.

Introduction to Gaussian Processes 12

$$k_{\mathrm{RBF}}(x, x') = \mathrm{Cov}(f(x), f(x')) = a^2 \exp\left(-\frac{1}{2\ell^2}||x\right.$$

- amplitude parameter affects the scale of the function, but not the rate of variation. . . .
- generalization performance of our procedure will depend on having reasonable values for these hyperparameters. . . .
- Values of $\ell = 2$ and $a = 1$ appeared to provide reasonable fits, while some of the other values did not.
- Fortunately, there is a robust and automatic way to specify these hyperparameters, using what is called the *marginal likelihood*, which we will return to in the notebook on inference.

So what is a GP, really?

- GP: any collection of function values $f(x_1), \ldots, f(x_n)$, indexed by any collection of inputs $x_1, \ldots, x_n$ has a joint multivariate Gaussian distribution.

- mean vector μ of this distribution is given by a *mean function*, which is typically taken to be a constant or zero.
- covariance matrix of this distribution is given by the *kernel* evaluated at all pairs of the inputs x.
-

$$\begin{bmatrix} f(x) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\mu, \begin{bmatrix} k(x,x) & k(x,x_1) & \ldots & k(x,x_n) \\ k(x_1,x) & k(x_1,x_1) & \ldots & k(x_1,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x) & k(x_n,x_1) & \ldots & k(x_n,x_n) \end{bmatrix}\right) \quad (1)$$

- Equation (1) specifies a GP prior. We can compute the conditional distribution of $f(x)$ for any x given $f(x_1), \ldots$, the function values we have observed.
- This conditional distribution is called the *posterior*, and it is what we use to make predictions.

Introduction to Gaussian Processes 14

In particular,

$$f(x)|f(x_1), \ldots, f(x_n) \sim \mathcal{N}(m, s^2)$$

where

$$m = k(x, x_{1:n})k(x_{1:n}, x_{1:n})^{-1} f(x_{1:n})$$

$$s^2 = k(x, x) - k(x, x_{1:n})k(x_{1:n}, x_{1:n})^{-1} k(x, x_{1:n})$$

where $k(x, x_{1:n})$ is a $1 \times n$ vector formed by evaluating $k(x, x_i)$ for $i = 1, \ldots, n$ and $k(x_{1:n}, x_{1:n})$ is an $n \times n$ matrix formed by evaluating $k(x_i, x_j)$ for $i, j = 1, \ldots, n$. is what we can use as a point predictor for any x, and s^2 what we use for uncertainty:

- if we want to create an interval with a 95% probability that $f(x)$ is in the interval, we would use $m \pm 2s$.
- predictive means and uncertainties for all the above figures were created using these equations.
- observed data points were given by $f(x_1), \ldots, f(x_n)$ and chose a fine grained set of x points to make predictions.

Introduction to Gaussian Processes 15

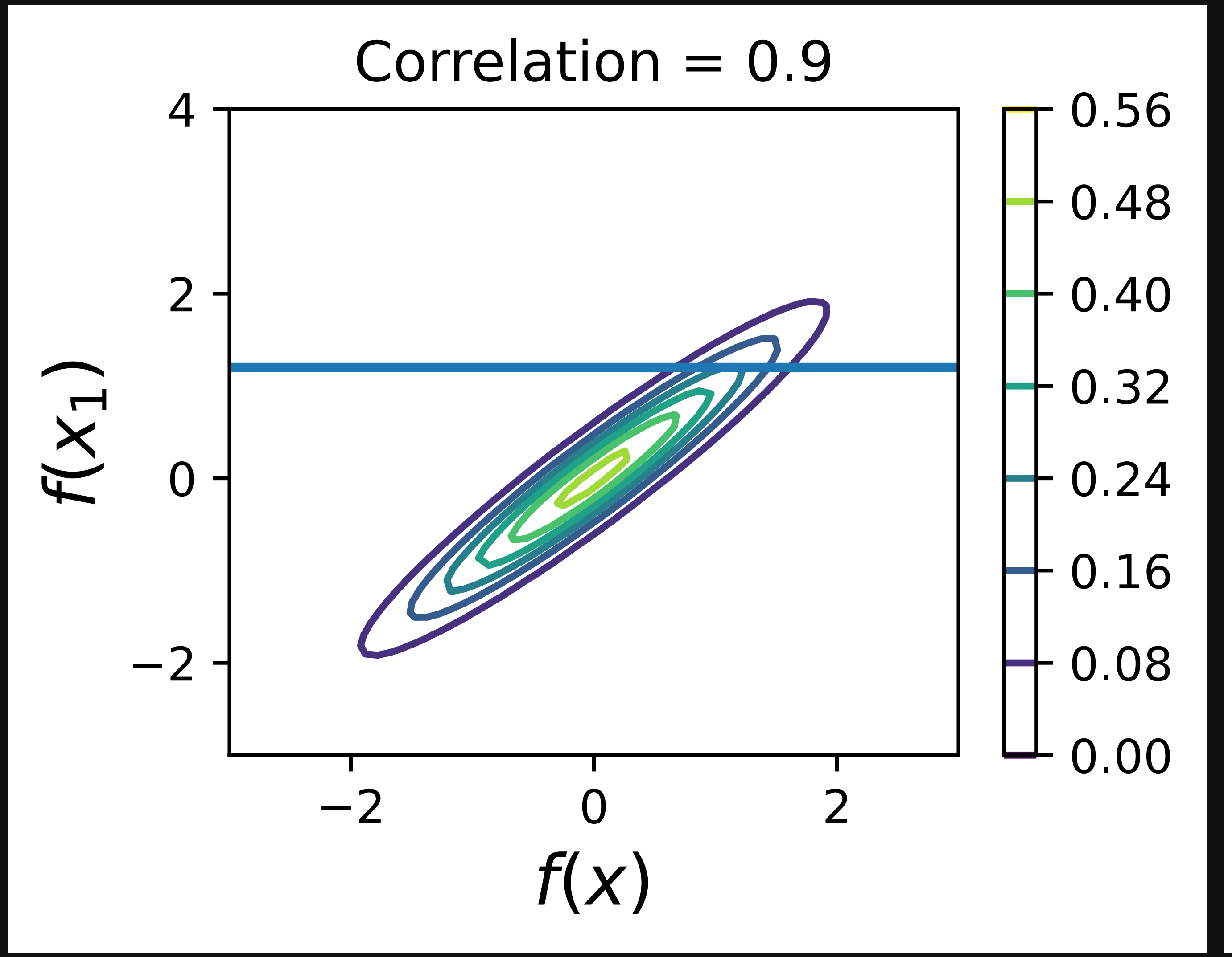
- suppose we observe a single datapoint, $f(x_1)$, and we want to determine the value of $f(x)$ at some x.
- $f(x)$ described by Gaussian process –> joint distribution over $(f(x), f(x_1))$ is Gaussian:

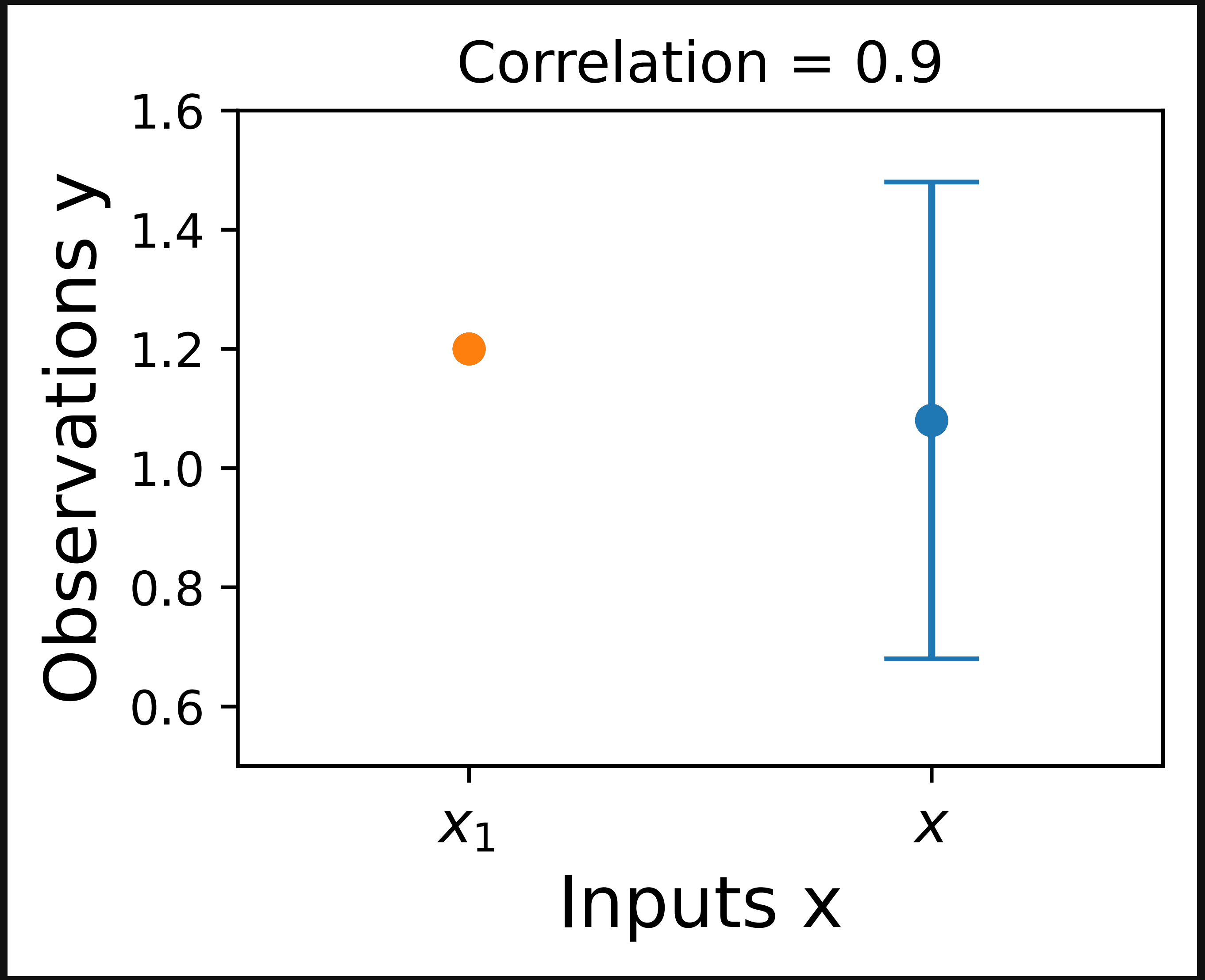
$$\begin{bmatrix} f(x) \\ f(x_1) \end{bmatrix} \sim \mathcal{N}\left(\mu, \begin{bmatrix} k(x,x) & k(x,x_1) \\ k(x_1,x) & k(x_1,x_1) \end{bmatrix}\right)$$

- off-diagonal expression $k(x,x_1) = k(x_1,x)$ tells us how correlated the function values will be – how strongly determined $f(x)$ will be from $f(x_1)$.
- have seen already that if we use a large length-scale, relative to the distance between x and x_1, $||x - x_1||$, then the function values will be highly correlated.
- visualize the process of determining $f(x)$ from $f(x_1)$ both in the space of functions, and in the joint distribution over $f(x_1), f(x)$.
- initially consider an x such that $k(x,x_1) = 0.9$, and $k(x,x) = 1$, meaning that the value of $f(x)$ is moderately correlated with the value of $f(x_1)$.
- In the joint distribution, the contours of constant probability will be relatively narrow ellipses.

Introduction to Gaussian Processes 16

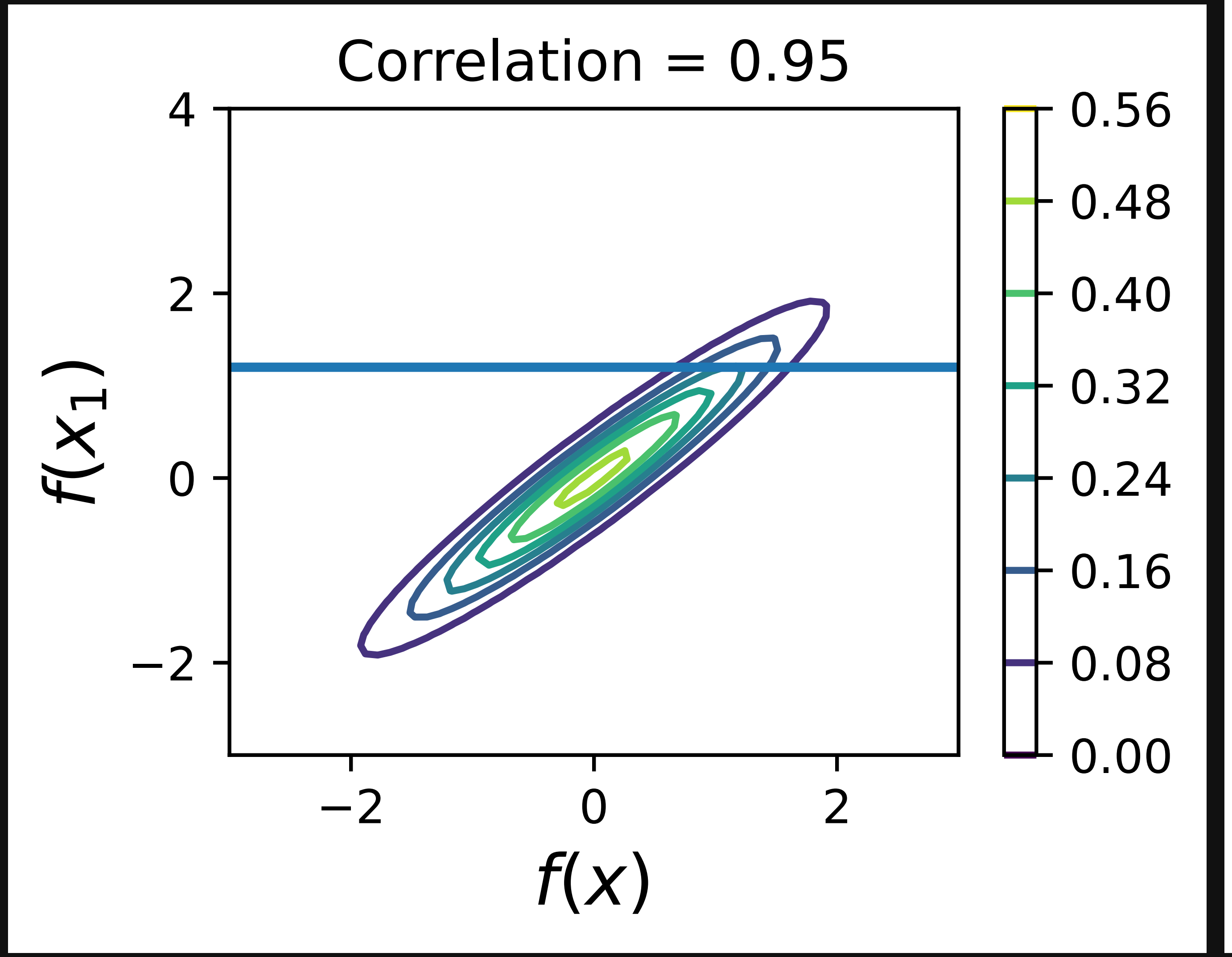
- Suppose we observe $f(x_1) = 1.2$. To condition on this value of $f(x_1)$, we can draw a horizontal line at 1.2 on our plot of the density, and see that the value of $f(x)$ is mostly constrained to $[0.64, 1.52]$.
- We have also drawn this plot in function space, showing the observed point $f(x_1)$ in orange, and 1 standard deviation of the Gaussian process predictive distribution for $f(x)$ in blue, about the mean value of 1.08.

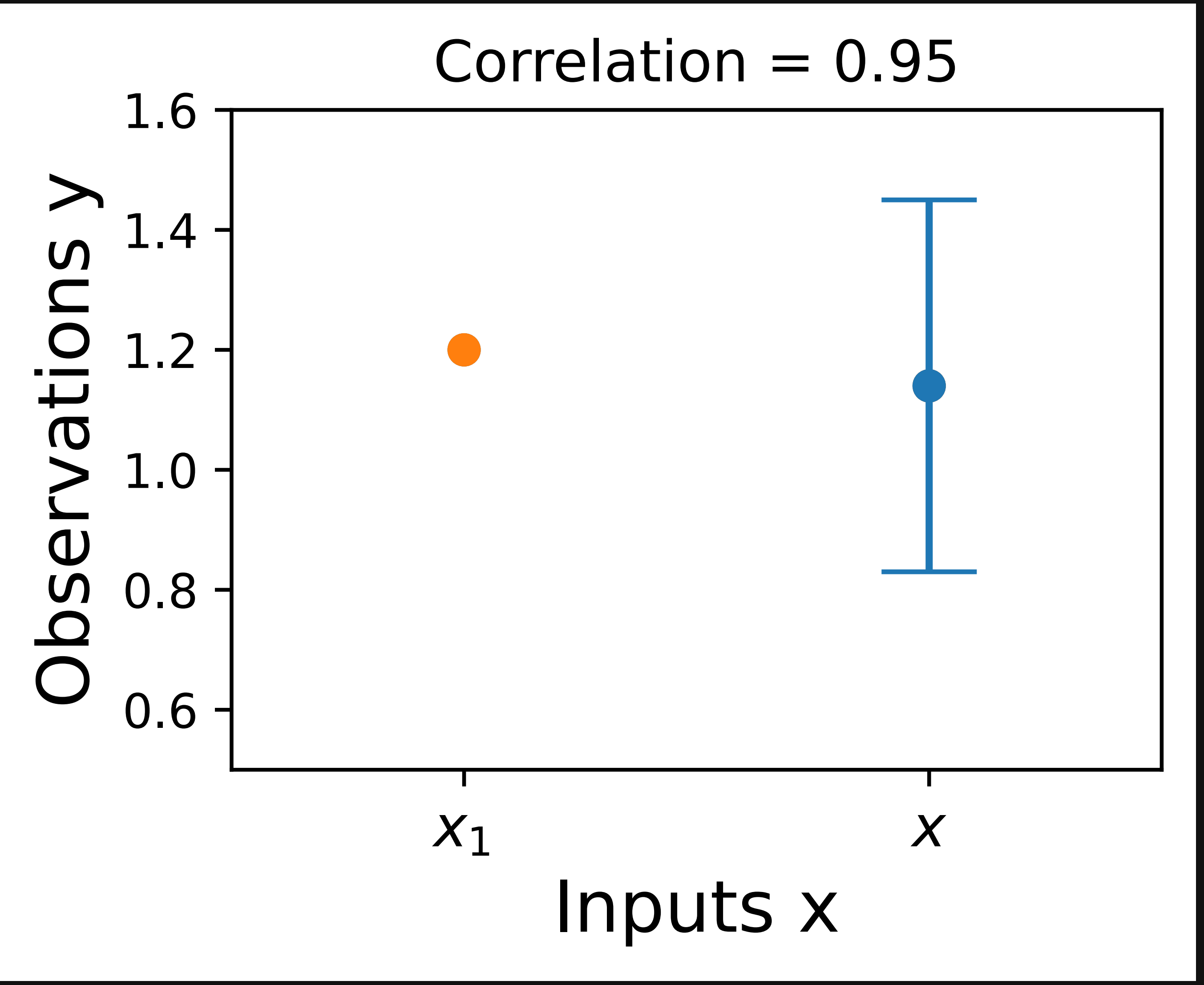




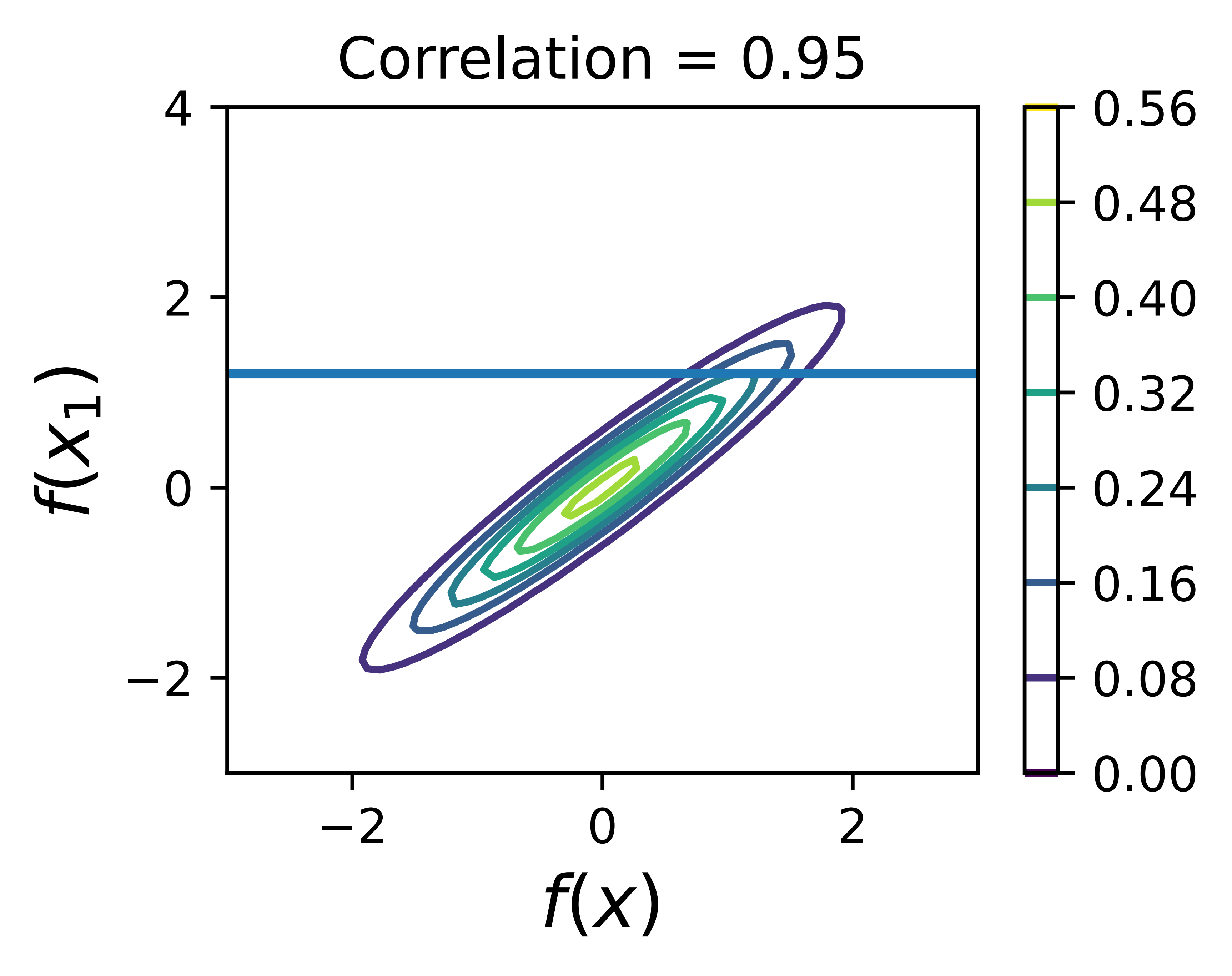
Introduction to Gaussian Processes 17

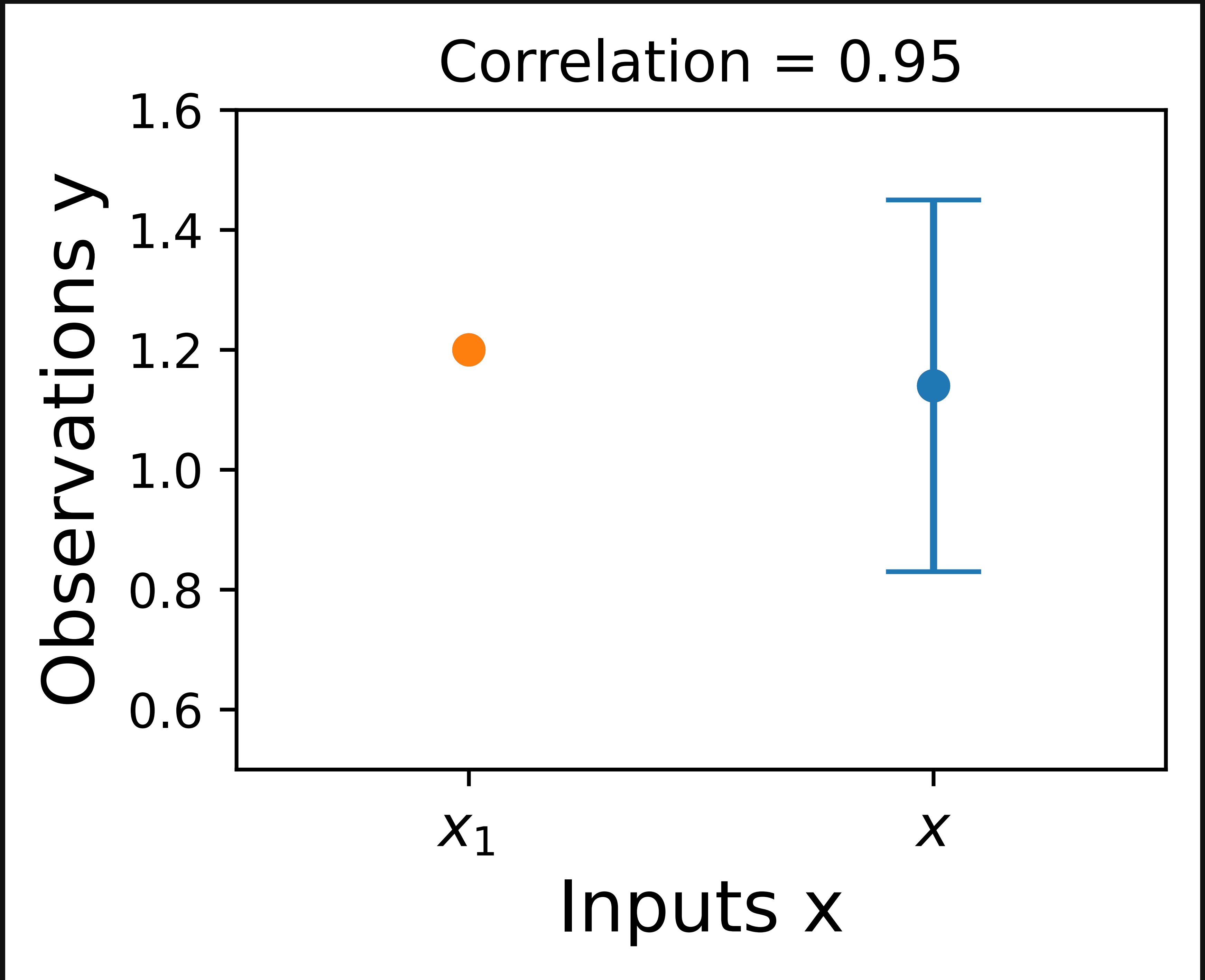
- suppose we have a stronger correlation, $k(x, x_1) = 0.95$.
- the ellipses have narrowed further, and the value of $f(x)$ is even more strongly determined by $f(x_1)$.
- Drawing a horizontal line at 1.2, we see the contours for $f(x)$ support values mostly within $[0.83, 1.45]$.
- show the plot in function space, with one standard deviation about the mean predictive value of 1.14.





Introduction to Gaussian Processes 18





- posterior mean predictor of our Gaussian process is closer to , because there is a stronger correlation. 1.2
- also uncertainty (the error bars) have somewhat decreased.
- Despite strong correlation between function values, uncertainty still quite large, because we have only observed a single data point!

Introduction to Gaussian Processes 19

- This procedure can give us a posterior on $f(x)$ for any x, for any number of points we have observed.
- Suppose we observe $f(x_1), f(x_2)$.
- visualize the posterior for $f(x)$ at a particular $x = x'$ in function space.
- exact distribution for $f(x)$ is given by the above equations. $f(x)$ is Gaussian distributed, with mean

$$m = k(x, x_{1:3})k(x_{1:3}, x_{1:3})^{-1} f(x_{1:3})$$

and variance

$$s^2 = k(x, x) - k(x, x_{1:3})k(x_{1:3}, x_{1:3})^{-1} k(x, x_{1:3})$$

- we have been considering *noise free* observations.
- easy to include observation noise. If we assume that the data are generated from a latent noise free function $f(x)$ plus iid Gaussian noise $\epsilon(x) \sim \mathcal{N}(0, \sigma^2)$ with variance , then our covariance function simply becomes $k(x_i, x_j) -$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Summary 1

- typical machine learning: we specify a function with some **free parameters** (such as a neural network and its weights), and we **focus on estimating those parameters**, which may not be interpretable.

- Gaussian process: **reason about distributions over functions directly**, which enables us to **reason about the high-level properties of the solutions**.

- properties are controlled by a covariance function (kernel), which often has a few highly interpretable hyperparameters.
- hyperparameters include the **length-scale**, which controls how rapidly (how wiggily) the functions are. Another hyperparameter is the **amplitude**, which controls the vertical scale over which our functions are varying.

- **representing many different functions** that can fit the data, and combining them all together into a predictive distribution, is a **distinctive feature of Bayesian methods**.

- greater amount of variability between possible solutions far away from the data –> uncertainty intuitively grows as we move from the data.

Summary 2

- **Gaussian process represents a distribution over functions by specifying a multivariate normal (Gaussian) distribution over all possible function values**.

- possible to **easily manipulate Gaussian distributions** to find the distribution of one function value based on the values of any set of other values.

- **observe a set of points** –> **condition on these points** and **infer a distribution** over what the value of the function might look like at any other input.

- How we model the correlations between these points is determined by the covariance function and is what defines the generalization properties of the Gaussian process.

- GPs easy to work with, have many applications, and help us understand and develop other model classes, like neural networks.

Exercises

1. What is the difference between epistemic uncertainty versus observation uncertainty?
2. Besides rate of variation and amplitude, what other properties of functions might we want to consider, and what would be real-world examples of functions that have those properties?
3. The RBF covariance function we considered says that covariances (and correlations) between observations decrease with their distance in the input space (times, spatial locations, etc.). Is this a reasonable assumption? Why or why not?
4. Is a sum of two Gaussian variables Gaussian? Is a product of two Gaussian variables Gaussian? If (a,b) have a joint Gaussian distribution, is a|b (a given b) Gaussian? Is a Gaussian?
5. Repeat the exercise where we observe a data point at $f(x_1) = 1.2$, but now suppose we additionally observe $f(x_2) =$. Let $k(x, x_1) = 0.9$, and $k(x, x_2) = 0.8$. Will we be more or less certain about the value of $f(x)$, than when we had only observed $f(x_1)$? What is the mean and 95% credible set for our value of $f(x)$ now?
6. Do you think increasing our estimate of observation noise would increase or decrease our estimate of the length-scale of the ground truth function?
7. As we move away from the data, suppose the uncertainty in our predictive distribution increases to a point, then stops increasing. Why might that happen?