Lecture 5: GPs and Streaming regression

- Gaussian Processes
- Information gain
- Confidence intervals

Recall: Non-parametric regression

- Input space $\mathcal{X} \subset \mathbb{R}^n$, target space $\mathcal{Y} = \mathbb{R}$
- Input matrix $\mathbf{X}$ of size $m \times n$, target vector $\mathbf{y}$ of size $m \times 1$
- Feature mapping $\phi : \mathcal{X} \mapsto \mathbb{R}^d$
- Assumption: $\mathbf{y} = \mathbf{\Phi}\mathbf{w}$
- Minimize

$$J_\lambda(\mathbf{w}) = \frac{1}{2}(\mathbf{\Phi}\mathbf{w} - \mathbf{y})^\top(\mathbf{\Phi}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \qquad \lambda \geq 0$$

- The solution is

$$\mathbf{w} = (\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I}_n)^{-1}\mathbf{\Phi}^\top\mathbf{y} = \mathbf{\Phi}^\top(\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda\mathbf{I}_m)^{-1}\mathbf{y}$$

Recall: Kernel regression

- Let $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$ and $\mathbf{k}(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{\Phi}^\top$ be the kernel matrix/vector:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_m) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & & \vdots \\ k(\mathbf{x}_m, \mathbf{x}_1) & k(\mathbf{x}_m, \mathbf{x}_2) & \dots & k(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix} \qquad \mathbf{k}(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) \\ k(\mathbf{x}, \mathbf{x}_2) \\ \vdots \\ k(\mathbf{x}, \mathbf{x}_m) \end{bmatrix}$$

- The predictions for the input data are given by

$$\hat{\mathbf{y}} = \mathbf{\Phi}\mathbf{\Phi}^\top (\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda \mathbf{I}_m)^{-1} \mathbf{y} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I}_m)^{-1} \mathbf{y}$$

- The prediction for a new input point $\mathbf{x}$ is given by

$$\hat{f}(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{\Phi}^\top (\mathbf{K} + \lambda \mathbf{I}_m)^{-1} \mathbf{y} = \mathbf{k}(\mathbf{x})(\mathbf{K} + \lambda \mathbf{I}_m)^{-1} \mathbf{y}$$

Recall: Bayesian view of regression

- Consider noisy observations $y = f(\mathbf{x}) + \epsilon = \phi(\mathbf{x})^\top \mathbf{w} + \epsilon$
- Recall Bayes' rule: posterior $= \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$

$$P_\phi(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{P_\phi(\mathbf{y}|\mathbf{X}, \mathbf{w}) P(\mathbf{w})}{P_\phi(\mathbf{y}|\mathbf{X})}$$

$\Rightarrow$ Marginal likelihood is independent of weights $\mathbf{w}$

- With Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$P_\phi(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} P_\phi(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \phi(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2}{2\sigma^2}\right) = \mathcal{N}_m\left(\mathbf{\Phi}\mathbf{w}, \sigma^2 \mathbf{I}_m\right)$$

Posterior distribution on parameters

- With Gaussian prior on parameters $\mathbf{w} \sim \mathcal{N}_d(0, \Sigma_{\mathbf{w}})$

$$
\begin{aligned}
P_\phi(\mathbf{w}|\mathbf{y}, \mathbf{X}) &\propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{\Phi w}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\mathbf{w}^\top \Sigma_{\mathbf{w}}^{-1} \mathbf{w}}{2}\right) \\
&= \exp\left(-\frac{\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi w} - \mathbf{w}^\top\mathbf{\Phi y} + \mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Phi w} + \sigma^2\mathbf{w}^\top\Sigma_{\mathbf{w}}^{-1}\mathbf{w}}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi w} - \mathbf{w}^\top\mathbf{\Phi y} + \mathbf{w}^\top(\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma^2\Sigma_{\mathbf{w}}^{-1})\mathbf{w}}{2\sigma^2}\right) \\
&\propto \exp\left((\mathbf{w} - \mathbf{b})^\top \mathbf{A}^{-1}(\mathbf{w} - \mathbf{b})\right)
\end{aligned}
$$

where $\mathbf{A}^{-1} = \sigma^{-2}(\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma^2\Sigma_{\mathbf{w}}^{-1})$ and $\mathbf{b} = (\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma^2\Sigma_{\mathbf{w}}^{-1})^{-1}\mathbf{\Phi}^\top\mathbf{y}$

$\Rightarrow$ The posterior distribution is Gaussian!

Predictive distribution

- The pointwise posterior predictive distribution is a normal distribution

$$\tilde{f}(\mathbf{x})|\mathbf{x}_1, \ldots, \mathbf{x}_m, y_1, \ldots, y_m \sim \mathcal{N}\left(\hat{f}(\mathbf{x}), s^2(\mathbf{x})\right)$$

of expectation

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \phi(\mathbf{x})^\top(\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma^2\Sigma_{\mathbf{w}}^{-1})^{-1}\mathbf{\Phi}^\top\mathbf{y} \\ &= \phi(\mathbf{x})^\top\Sigma_{\mathbf{w}}\mathbf{\Phi}^\top(\mathbf{\Phi}\Sigma_{\mathbf{w}}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_m)^{-1}\mathbf{y}\end{aligned}$$

and variance

$$\begin{aligned}s^2(\mathbf{x}) &= \sigma^2\phi(\mathbf{x})^\top(\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma^2\Sigma_{\mathbf{w}}^{-1})^{-1}\phi(\mathbf{x}) \\ &= \phi(\mathbf{x})^\top\Sigma_{\mathbf{w}}\phi(\mathbf{x}) - \phi(\mathbf{x})^\top\Sigma_{\mathbf{w}}\mathbf{\Phi}^\top(\mathbf{\Phi}^\top\Sigma_{\mathbf{w}}\mathbf{\Phi} + \sigma^2\mathbf{I}_m)^{-1}\mathbf{\Phi}\Sigma_{\mathbf{w}}\phi(\mathbf{x})\end{aligned}$$

$\rightarrow$ using Sherman-Morrison

Reinterpreting regularization

- Recall kernel regression predictions:

$$\hat{f}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \lambda \mathbf{I}_m)^{-1}\mathbf{y}$$

- Using prior $\Sigma_\mathbf{w} = \frac{\sigma^2}{\lambda}\mathbf{I}_d$, the predictive mean rewrites as:

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \phi(\mathbf{x})^\top \Sigma_\mathbf{w} \mathbf{\Phi}^\top (\mathbf{\Phi}\Sigma_\mathbf{w}\mathbf{\Phi}^\top + \sigma^2 \mathbf{I}_m)^{-1}\mathbf{y} \\ &= \phi(\mathbf{x})^\top \frac{\sigma^2}{\lambda}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\frac{\sigma^2}{\lambda}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_m\right)^{-1}\mathbf{y} \\ &= \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \lambda \mathbf{I}_m)^{-1}\mathbf{y}\end{aligned}$$

$\Rightarrow$ λ encodes some prior on weights $\mathbf{w}$

Reinterpreting regularization (cont'd)

- Still using $\Sigma_{\mathbf{w}} = \frac{\sigma^2}{\lambda}\mathbf{I}_d$, the predictive variance rewrites as:

$$
\begin{aligned}
s^2(\mathbf{x}) &= \phi(\mathbf{x})^\top \Sigma_{\mathbf{w}} \phi(\mathbf{x}) - \phi(\mathbf{x})^\top \Sigma_{\mathbf{w}} \mathbf{\Phi}^\top (\mathbf{\Phi}^\top \Sigma_{\mathbf{w}} \mathbf{\Phi} + \sigma^2 \mathbf{I}_m)^{-1} \mathbf{\Phi} \Sigma_{\mathbf{w}} \phi(\mathbf{x}) \\
&= \phi(\mathbf{x})^\top \frac{\sigma^2}{\lambda} \phi(\mathbf{x}) - \phi(\mathbf{x})^\top \frac{\sigma^2}{\lambda} \mathbf{\Phi}^\top \left(\mathbf{\Phi}^\top \frac{\sigma^2}{\lambda} \mathbf{\Phi} + \sigma^2 \mathbf{I}_m \right)^{-1} \mathbf{\Phi} \frac{\sigma^2}{\lambda} \phi(\mathbf{x}) \\
&= \frac{\sigma^2}{\lambda} k_\lambda(\mathbf{x}, \mathbf{x}) \qquad \text{with} \\
k_\lambda(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top \left(\mathbf{K} + \lambda \mathbf{I}_m \right)^{-1} \mathbf{k}(\mathbf{x}')
\end{aligned}
$$

Summary

- Using prior $\Sigma_{\mathbf{w}} = \frac{\sigma^2}{\lambda}\mathbf{I}_d$, we have

$$\hat{f}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \lambda\mathbf{I}_m)^{-1}\mathbf{y}$$

$$s^2(\mathbf{x}) = \frac{\sigma^2}{\lambda}k_\lambda(\mathbf{x},\mathbf{x}) \qquad \text{with}$$

$$k_\lambda(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{k}(\mathbf{x})^\top\left(\mathbf{K} + \lambda\mathbf{I}_m\right)^{-1}\mathbf{k}(\mathbf{x}')$$

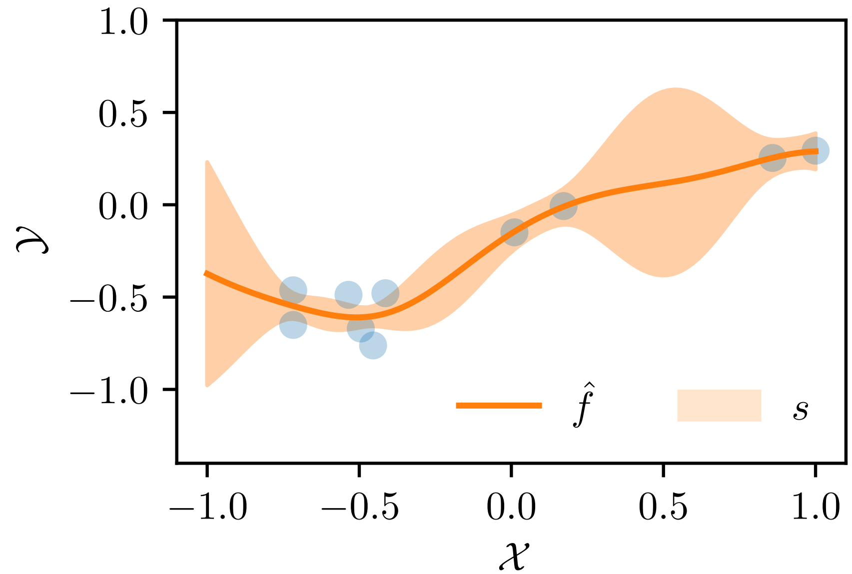
- Providing pointwise posterior prediction

$$\tilde{f}(\mathbf{x})|\mathbf{x}_1,\ldots,\mathbf{x}_m,y_1,\ldots,y_m \sim \mathcal{N}\left(\hat{f}(\mathbf{x}), s^2(\mathbf{x})\right)$$

$\Rightarrow$ What does it mean to use $\lambda \in \mathbb{R}_{\geq 0}$?

Pointwise posterior distribution

- At each point $\mathbf{x} \in \mathcal{X}$, we have a distribution $\mathcal{N}\left(\hat{f}(\mathbf{x}), s^2(\mathbf{x})\right)$
- We can sample from these $\tilde{f}(\mathbf{x}) \sim \mathcal{N}\left(\hat{f}(\mathbf{x}), s^2(\mathbf{x})\right)$



Joint distribution

- Suppose you *query* your model at locations $\mathbf{X}_*$
- Extend the prior to include query points:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} | \mathbf{X}, \mathbf{X}_* \sim \mathcal{N}_{m+m_*}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{X},\mathbf{X}} & \mathbf{K}_{\mathbf{X},\mathbf{X}_*} \\ \mathbf{K}_{\mathbf{X}_*,\mathbf{X}} & \mathbf{K}_{\mathbf{X}_*,\mathbf{X}_*} \end{bmatrix}\right)$$

$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}_m(\mathbf{f}, \sigma^2 \mathbf{I}_m)$$

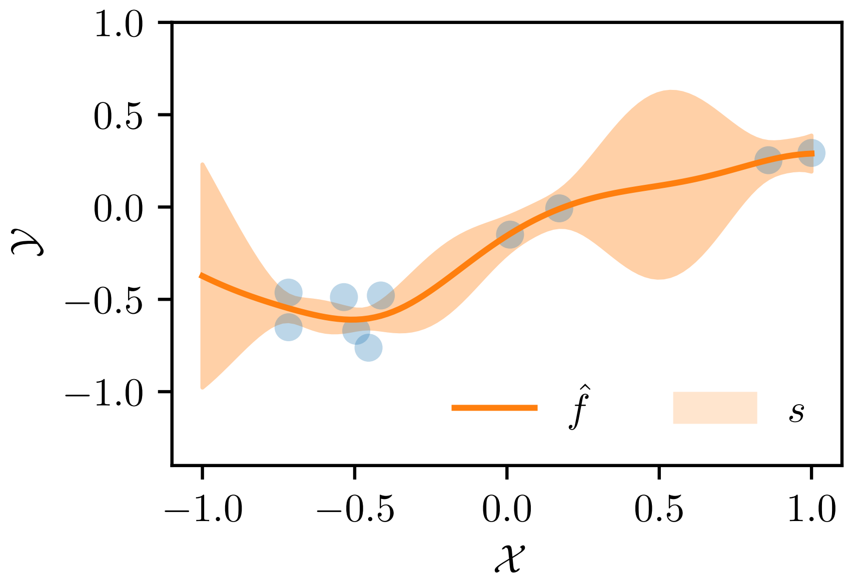
- Using joint normality of $\mathbf{f}_*$ and $\mathbf{y}$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}_{m+m_*}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma^2 \mathbf{I}_m & \mathbf{K}_{\mathbf{X},\mathbf{X}_*} \\ \mathbf{K}_{\mathbf{X}_*,\mathbf{X}} & \mathbf{K}_{\mathbf{X}_*,\mathbf{X}_*} \end{bmatrix}\right)$$

Gaussian Process (GP)

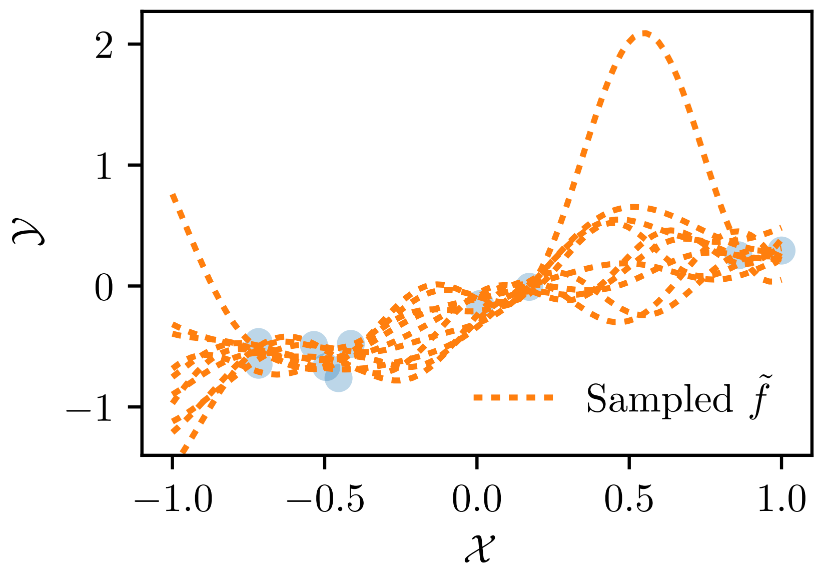
- By considering the covariance between *every points in the space*, we get a distribution over functions!
- Posterior distribution on f:

$$P[f|\mathbf{X},\mathbf{y}] \sim \mathcal{N}_{|\mathcal{X}|}\left(\left[\hat{f}(\mathbf{x})\right]_{\mathbf{x}\in\mathcal{X}}, \frac{\sigma^2}{\lambda}\left[k_\lambda(\mathbf{x},\mathbf{x}')\right]_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\right)$$



Sampling from a Gaussian Process

- Generalization of normal probability distribution to the function space
  - From a normal distribution we sample variables
  - From a GP we sample *functions*!



Sampling from a Gaussian Process: How to

- Observe that

$$\mathcal{N}_{|\mathcal{X}|}\left(\left[\hat{f}(\mathbf{x})\right]_{\mathbf{x}\in\mathcal{X}}, \frac{\sigma^2}{\lambda}\left[k_\lambda(\mathbf{x},\mathbf{x}')\right]_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\right)$$

defines a $|\mathcal{X}|$-dimensional multivariate Gaussian distribution

$\rightarrow$ What if $|\mathcal{X}| = \infty$ (e.g. $\mathcal{X} = [-1, 1]$)?

- We can consider a discrete, finite, set $\mathbb{X} \subset \mathcal{X}$ and sample from

$$\mathcal{N}_{|\mathbb{X}|}\left(\left[\hat{f}(\mathbf{x})\right]_{\mathbf{x}\in\mathbb{X}}, \frac{\sigma^2}{\lambda}\left[k_\lambda(\mathbf{x},\mathbf{x}')\right]_{\mathbf{x},\mathbf{x}'\in\mathbb{X}}\right)$$

- This will result in a function $\tilde{f}$ evaluated at every $\mathbf{x} \in \mathbb{X}$

Learning the hyperparameters

- If we assume that $\Sigma_{\mathbf{w}} = \mathbf{I}_d$, then we have $\lambda = \sigma^2$
- Let $\boldsymbol{\theta}$ denote the kernel hyperparameters (e.g. ρ)
- In practice you may not know the noise and the *optimal* kernel hypers
- Recall: multivariate normal density

$$P(\mathbf{y}|\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m)^{-1}\mathbf{y}\right)}{\sqrt{(2\pi)^D|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m|}}$$

- Maximize the marginal likelihood $\mathcal{L} = \log P(\mathbf{y}|\boldsymbol{\theta})$:

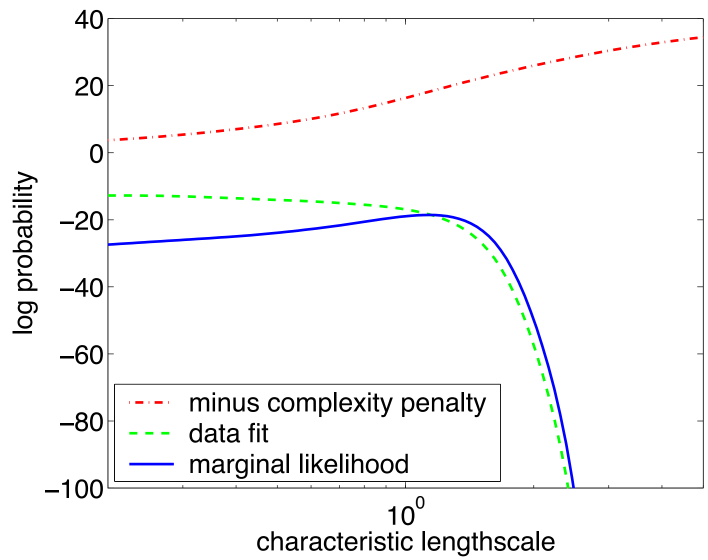
$$\mathcal{L} = -\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m)^{-1}\mathbf{y} - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m|$$

Anatomy of marginal likelihood

- Marginal likelihood:

$$\mathcal{L} = \log P(\mathbf{y}|\boldsymbol{\theta}) \propto -\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}_m|$$

- 1st term: quality of predictions; 2nd term: model complexity
- Trade-off (from Rasmussen & Williams, 2006):



Gradient-based optimization

- Compute gradients:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{1}{2}\mathbf{y}^\top (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}_m)^{-1} \frac{\partial (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}_m)}{\partial \theta_i} (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}_m)^{-1} \mathbf{y} \\ - \frac{1}{2} \mathrm{Tr} \left((\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}_m)^{-1} \frac{\partial (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}_m)}{\partial \theta_i} \right)$$

- Minimize the negative
- Non-convex optimization task

Summary

- Normal priors on the weights distribution $\rightarrow$ Gaussian Process
- Regularization $\rightarrow$ prior on the weights covariance
- GP provides a posterior distribution on functions
  - Expectation: kernel regression model
  - Covariance $\rightarrow$ confidence intervals
- Sample discretized functions from a GP

Typical supervised setting

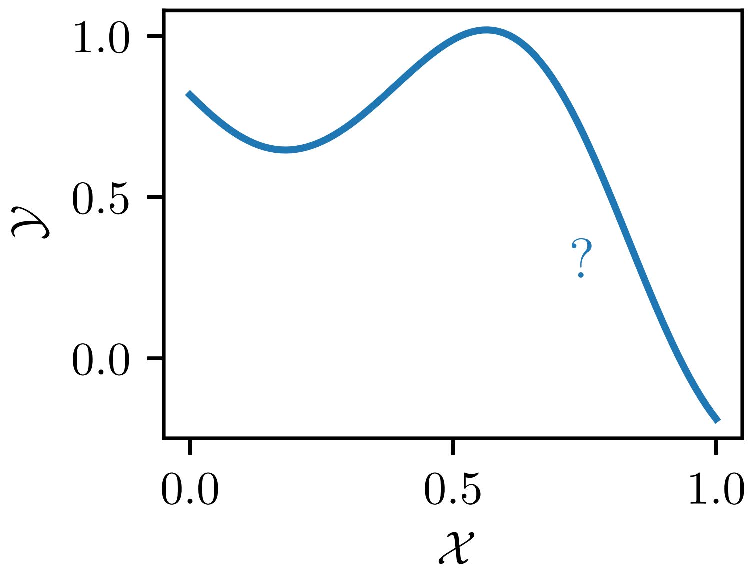
- Have dataset $(\mathbf{X}, \mathbf{y})$ of previously acquired data
- Learn model $\mathbf{w}$ on $(\mathbf{X}, \mathbf{y})$
- Apply model $\mathbf{w}$ to provide predictions at new data points
- Samples $(\mathbf{X}, \mathbf{y})$ are often assumed to be i.i.d.

Streaming setting

- Start with (possibly empty) dataset $(\mathbf{X}_0, \mathbf{y}_0)$
- For each time step $t = 1, 2, \ldots$:
  - Fit model $\mathbf{w}_t$ on $\mathbf{X}_{t-1}$ and $\mathbf{y}_{t-1}$
  - Acquire a new sample $(\mathbf{x}_t, y_t)$ (possibly using $\mathbf{w}_t$)
  - Define $\mathbf{X}_t = [\mathbf{x}_i]_{i=1\ldots t}$, $\mathbf{y}_t = [y_i]_{i=1\ldots t}$
- Samples may be dependent of model (not i.i.d.)

$\rightarrow$ Samples influence model / Samples depend on model

Application: Online function optimization



- Unknown function $f : \mathcal{X} \mapsto \mathbb{R}$
- Sequentially select locations $(\mathbf{x}_t)_{t \geq 1}$ at which to observe the function y_t
- Try to maximize/minimize the observations

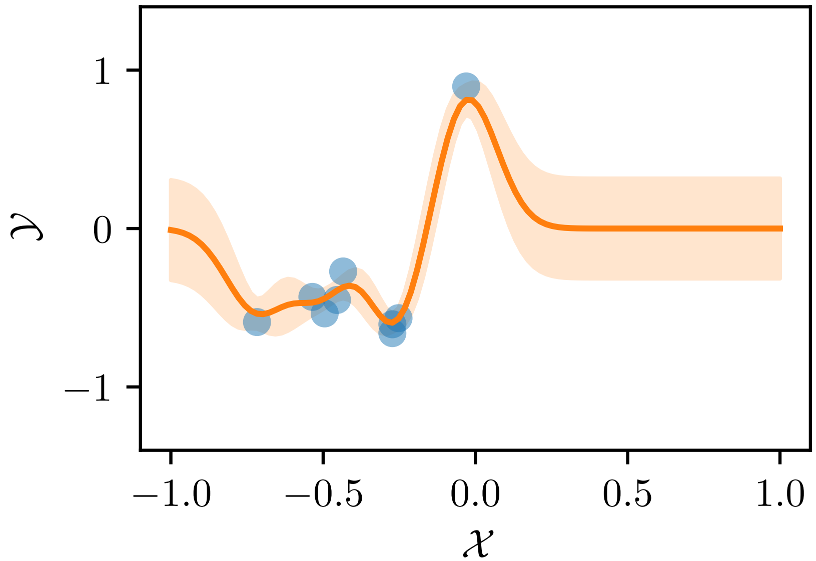
Example

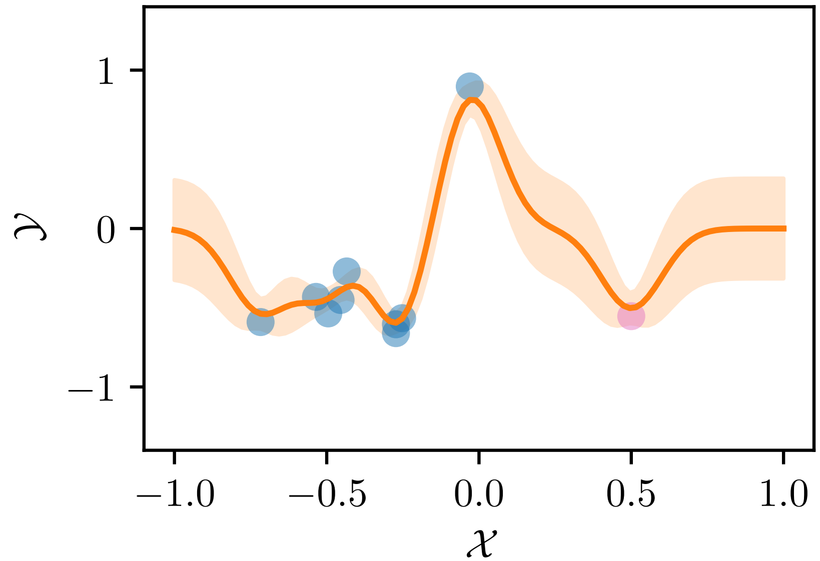
What is the optimal treatment dosage for a given disease?

- For each time step $t = 1, 2, \ldots$:
  - New patient t comes in
  - We decide on treatment dosage $\mathbf{x}_t$
  - We observe the patient's response to treatement, y_t
- Our goal is to cure patients as effectively as possible
- What you don't want: give bad dosages that *were known to be bad*
- What you want:
  - give *good* dosages
  - try *informative* dosages

Information

- An informative sample improves the model by reducing its uncertainty





How do we quantity the reduction of uncertainty
after observing $y_1, \ldots, y_t$ at locations $\mathbf{x}_1, \ldots, \mathbf{x}_t$?

A little bit of information theory

- *Mutual information* between underlying function f and observations $y_1, \ldots, y_t$ at locations $\mathbf{x}_1, \ldots, \mathbf{x}_t$:

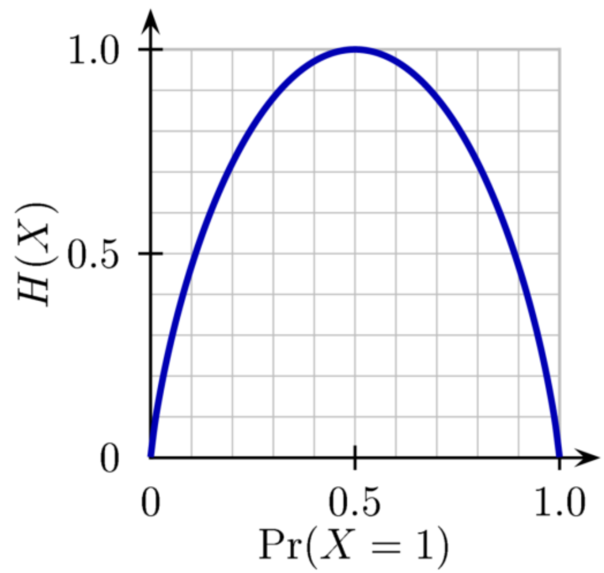
$$I(y_1, \ldots, y_t; f) = \underbrace{H(y_1, \ldots, y_t)}_{\text{Marginal entropy}} - \underbrace{H(y_1, \ldots, y_t|f)}_{\text{Conditional entropy}}$$

- It quantifies the "amount of information" obtained about one random variable, through the other random variable
- *Entropy* is the "amount of information" held by a random variable
- $H(Y|X) = 0$ if and only if the value of Y is completely determined by the value of X
- $H(Y|X) = H(Y)$ if and only if Y and X are independent random variables

Amount of information

$$H(X) = \mathbb{E}[-\ln P_X] = \sum_{i=1}^{n} p(x_i) \log_b (p(x_i))$$

Example: Tossing a coin



- A fair coin has maximal entropy: it is the less predictable!
- Every new sample that you get changes your *model* the most

Mutual information

- Entropy of $X \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$H(X) = \mathbb{E}\left[-\ln \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}}\right] = \frac{D}{2} + \frac{D}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}|$$

$\rightarrow$ Using $\mathbb{E}\left[\mathbf{a}^\top \mathbf{M}^{-1}\mathbf{a}\right] = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{a}^\top \mathbf{M}^{-1}\mathbf{a}\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{a}\mathbf{a}^\top \mathbf{M}^{-1}\right)\right] = D$

- Recall the joint distribution between m observations $\mathbf{y}$ and m_* query points $\mathbf{X}_*$:

$$\begin{bmatrix}\mathbf{y} \\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}_{m+m_*}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma^2\mathbf{I}_m & \mathbf{K}_{\mathbf{X},\mathbf{X}_*} \\ \mathbf{K}_{\mathbf{X}_*,\mathbf{X}} & \mathbf{K}_{\mathbf{X}_*,\mathbf{X}_*}\end{bmatrix}\right)$$

- Then $I(y_1, \ldots, y_t; f) = \frac{1}{2}\ln|\mathbf{K}_t + \sigma^2\mathbf{I}_t|$

Another decomposition of mutual information

- Recall that $y_1, \ldots, y_t | f(x_1), \ldots, f(x_t) \sim \mathcal{N}_t \left((f(x_1), \ldots, f(x_t)), \sigma^2 \mathbf{I}_t \right)$

$\rightarrow$ Using that $y_i = f(x_i) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- Pluging-in the conditional entropy of a multivariate normal distribution:

$$
\begin{aligned}
I(y_1, \ldots, y_t; f) &= H(y_1, \ldots, y_t) - \frac{1}{2} \ln |2\pi e \sigma^2 \mathbf{I}_t| \\
&= H(y_1, \ldots, y_t) - \frac{t}{2} \ln 2\pi e - \frac{t}{2} \sigma^2
\end{aligned}
$$

- What is the marginal entropy $H(y_1, \ldots, y_t)$?

Entropy of observations

- Recursively decompose

$$
\begin{aligned}
H(y_1, \ldots, y_t) &= H(y_1, \ldots, y_{t-1}) + H(y_t | y_1, \ldots, y_{t-1}) \\
&= H(y_1, \ldots, y_{t-1}) + \frac{1}{2} \ln \left[2\pi e \left(\sigma^2 + \frac{\sigma^2}{\lambda} k_{\lambda,t-1}(x_t, x_t) \right) \right] \\
&\vdots \\
&= H(y_1) + H(y_2|y_1) + \cdots + \frac{1}{2} \ln \left[2\pi e \left(\sigma^2 + \frac{\sigma^2}{\lambda} k_{\lambda,t-1}(x_s, x_s) \right) \right] \\
&= \sum_{s=1}^{t} \frac{1}{2} \ln \left[2\pi e \sigma^2 \left(1 + \frac{1}{\lambda} k_{\lambda,s-1}(x_s, x_s) \right) \right]
\end{aligned}
$$

→ Still using that $y_i = f(x_i) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$

→ Also using the uncertainty about the location of the *true* f

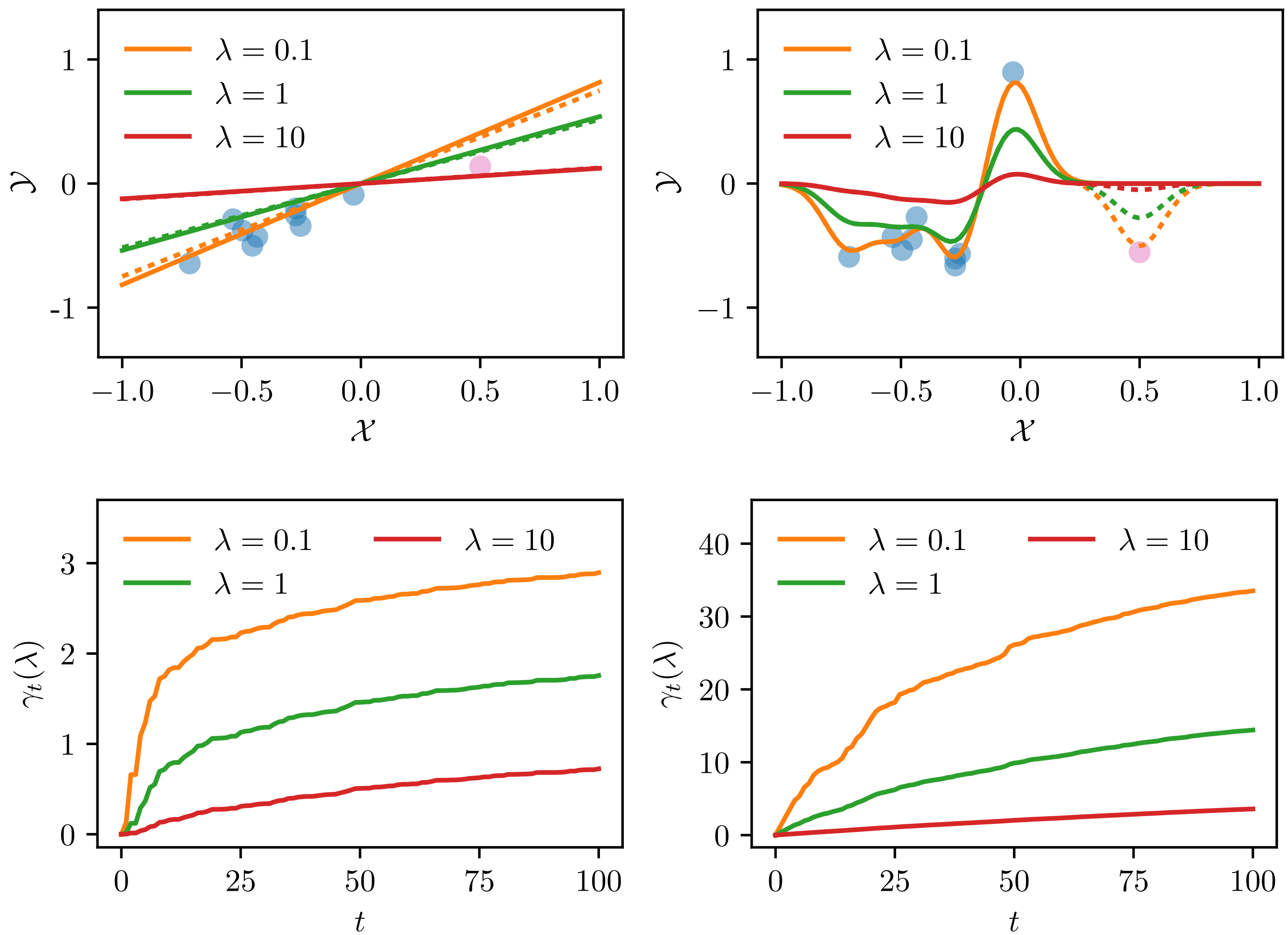
Information gain

$$I(y_1,\ldots,y_t;f) = \sum_{s=1}^{t} \frac{1}{2}\ln\left[2\pi e\sigma^2\left(1+\frac{1}{\lambda}k_{\lambda,s-1}(x_s,x_s)\right)\right] - \frac{t}{2}\ln(2\pi e) - \frac{t}{2}\sigma^2$$
$$= \sum_{s=1}^{t} \frac{1}{2}\ln\left[1+\frac{1}{\lambda}k_{\lambda,s-1}(x_s,x_s)\right]$$

- Information gain $\gamma_t(\lambda) = I(y_1,\ldots,y_t;f)$: reduction of uncertainty on f after observing $y_1,\ldots,y_t$
- Information gain is inversely proportionnal to λ

$\rightarrow$ Limiting changes in function limits the contribution of samples

Information gain vs regularization



Summary

- Streaming regression: sequentially gather (potentially non-i.i.d) samples
- Information gain measures the total information that could result from adding a new sample (observation) to the model
- Information gain is controlled by the *information sharing* capability of the kernel
- Information gain is controlled by the changes in model admitted by regularization
- Information gain will play a part in confidence intervals

Confidence intervals

- Given that we have gathered t samples under the streaming setting, what kind of guarantees can we have on the resulting model?
- More specifically, could we guarantee that

$$|f(\mathbf{x}) - \hat{f}_t(\mathbf{x})| \leq \text{something}$$

  simultaneously for all $\mathbf{x} \in \mathcal{X}$ and for all $t \geq 0$?
- Motivations:
  - Control bad behaviours in critical applications
  - Help selecting the next sample location
    * Maximize/minimize function?
    * Maximize model improvement?
  - Derive sound algorithms

Result (Maillard, 2016)

- Under the assumption of subgaussian noise...

$$|f(\mathbf{x}) - \hat{f}_t(\mathbf{x})| \leq \sqrt{\frac{k_{\lambda,t}(\mathbf{x},\mathbf{x})}{\lambda}}\left[\sqrt{\lambda}\|f\|_{\mathcal{K}} + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda)}\right]$$

- With probability higher than $1 - \delta$
- Simultaneously for all $t \geq 0$, for all $\mathbf{x} \in \mathcal{X}$

$\Rightarrow$ Observe that the error bound scales with the information gain!

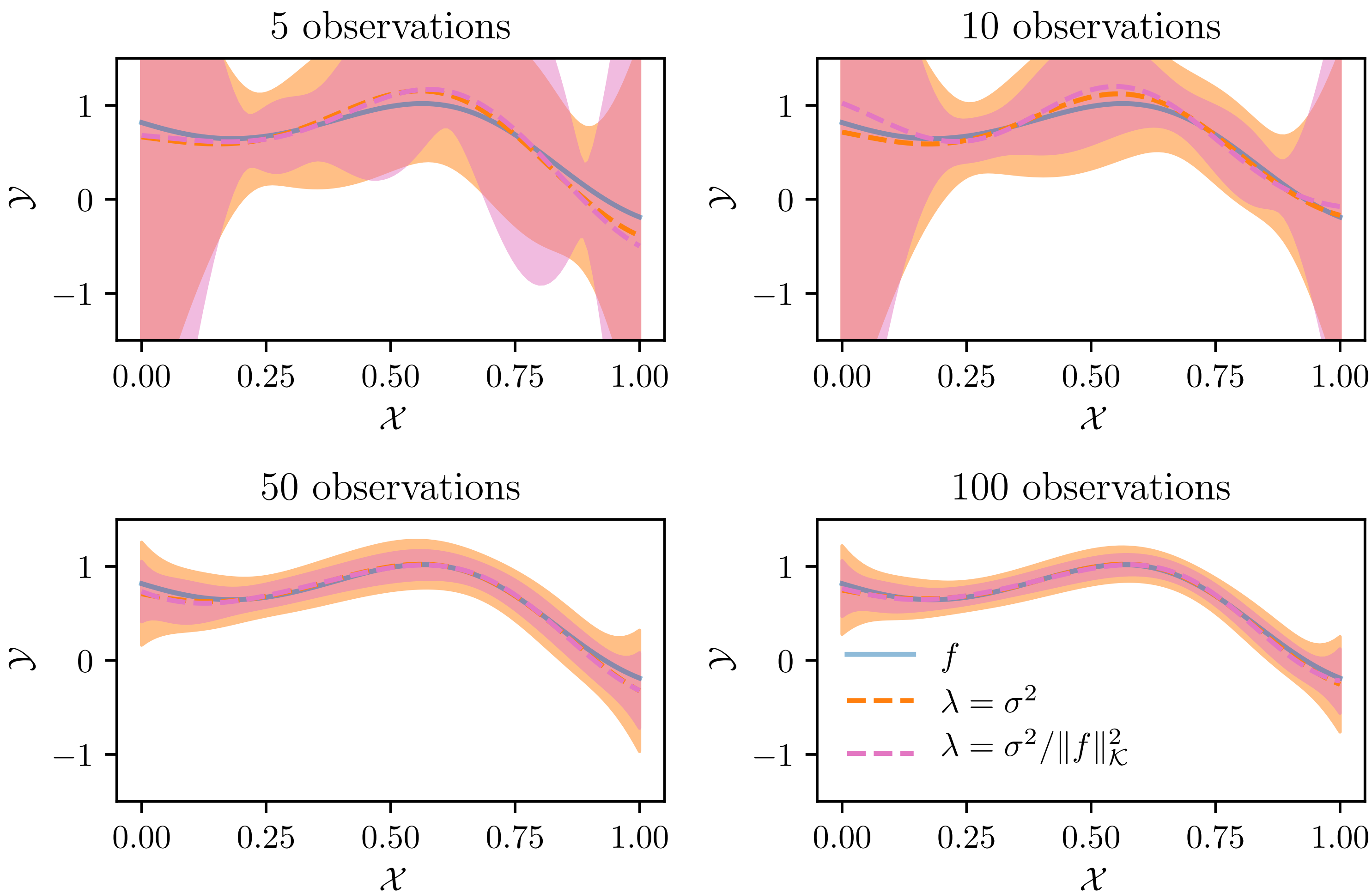
Subgaussian noise

- A real-valued random variable X is σ^2-subgaussian if $\mathbb{E}\left[e^{\gamma X}\right] \leq e^{\gamma^2\sigma^2/2}$

$\rightarrow$ The Laplace transform of X is dominated by the Laplace transform of a random variable sampled from $\mathcal{N}(0, \sigma^2)$

- Require that the tails of the noise distribution are dominated by the tails of a Gaussian distribution
- For example, true for
  - Gaussian noise
  - Bounded noise

Confidence envelope



Summary

- It is possible to have guarantees even with non-i.i.d. data
- The predition error depends on
  - How well your model shares information across observations
  - How well your model is adapted to the function
  - How noisy the observations are
  - Regularization $\rightarrow$ prior on $\Sigma_{\mathbf{w}}$
- Confidence intervals/envelopes will be really useful for deriving algorithms (we will see that later)