Linear Regression

Advanced Statistical Inference

Simone Rossi

Linear regression

Objectives for today

1. Review of linear regression
2. Understand the probabilistic interpretation of (regularized) loss minimization

Break

3. Introduction of Bayesian linear regression
4. Compute the posterior distribution and make predictions
5. Model selection and other properties of Bayesian linear regression

Break

6. Class exercise on Bayesian inference for coin toss

A quick recap on probability

Consider two continuous random variables x and y

- Sum rule (marginalization):

$$p(x) = \int p(x, y) \mathrm{d}y$$

- Product rule (conditioning):

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x)$$

- Bayes' rule:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Consider a random vector $\boldsymbol{x}$ with D components ($\boldsymbol{z} \in \mathbb{R}^D$)

- Chain rule:

$$p(\boldsymbol{z}) = p(z_1, z_2, \ldots, z_D) = p(z_1)p(z_2 \mid z_1)p(z_3 \mid z_1, z_2) \cdots p(z_D \mid z_1, z_2, \ldots, z_{D-1})$$

If z_i are independent, then

$$p(z_1 \mid z_2, \ldots, z_{D-1}) = p(z_1)$$

and the chain rule becomes

$$p(\boldsymbol{z}) = p(z_1)p(z_2) \cdots p(z_D) = \prod_{d=1}^{D} p(z_d)$$

Definitions

- **Input**, features, covariates: $\boldsymbol{x} \in \mathbb{R}^D$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix} \in \mathbb{R}^{N \times D} \quad \text{or with a bias term} \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x}_N \end{bmatrix} \in \mathbb{R}^{N \times (D+1)}$$

- **Output**, target, response: $y \in \mathbb{R}$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

- **Dataset**: $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$

Regression

- **Objective**: Learn a function $f : \mathbb{R}^D \to \mathbb{R}$

Linear models implement a linear combination of (basis) functions

$$f(\boldsymbol{x}) = \sum_{d=1}^{D} w_d \varphi_d(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\varphi}(\boldsymbol{x})$$

- **Parameters**: $\boldsymbol{w} = [w_1, \dots, w_D]^\top$
- **Basis functions**: $\boldsymbol{\varphi}(\boldsymbol{x}) = [\varphi_1(\boldsymbol{x}), \dots, \varphi_D(\boldsymbol{x})]^\top$

> **!** Important
>
> Any model that can be written as a linear combination of parameters (**not** the input) is a linear model

Linear regression

For simplicity, let's consider linear functions

$$f(\boldsymbol{w}, \boldsymbol{x}) = \sum_{d=1}^{D} w_d x_d = \boldsymbol{w}^\top \boldsymbol{x}$$

- **Objective**: Find $\boldsymbol{w}$ that minimizes the error

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (y_n - f(\boldsymbol{w}, \boldsymbol{x}_n))^2 = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

Least squares solution

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

- **Gradient**: $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = -\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$
- **Solution**: $\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Exercise

Implement the least squares solution for linear regression using Cholesky decomposition and back-substitution (ref revision of linear algebra)

Probabilistic interpretation of linear regression

Minimizing the loss is equivalent to maximizing the likelihood of the data under a Gaussian noise model

$$\exp(-\gamma \mathcal{L}_i) \propto \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \gamma^{-1}\boldsymbol{I}\right)$$

(Implicit) Assumption: Data is generated by a linear model with **Gaussian noise** $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent across samples (with $\sigma^2 = \gamma^{-1}$).

Maximum likelihood estimation

Maximum likelihood estimation is solving

$$\arg\max_{\boldsymbol{w}} \prod_{i=1}^{N} \underbrace{\mathcal{N}\left(y_i \mid \boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2\right)}_{p(y_i \mid \boldsymbol{w}, \boldsymbol{x}_i)}$$

$p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = \prod_{i=1}^{N} p(y_i \mid \boldsymbol{w}, \boldsymbol{x}_i)$ is the **likelihood** of the data given the model

Tip

We *never* maximize the likelihood directly, but the log-likelihood

$$\arg\max_{\boldsymbol{w}} \sum_{i=1}^{N} \log \mathcal{N}\left(y_i \mid \boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2\right)$$

Because log is monotonic and concave, the optimum value is the same and numerically more stable

Likelihood is not a probability

- The likelihood function is not a probability distribution.
- The likelihood can take any non-negative value, not just values between 0 and 1.
- It represents the density of the data ($\boldsymbol{y}$) given the model ($\boldsymbol{w}$).

Key insight:

- **Probability**: Fix the parameters, vary the data. It answers "what data might we see?"
- **Likelihood**: Fix the data, vary the parameters. It answers "which parameters best explain the data?"

Properties of maximum likelihood estimation

- **Consistency**: As $N \to \infty$, the MLE converges to the true parameter value

> ⚠ Note
>
> The consistency property makes sense if the model is correct (e.g. the data is generated by a linear model with Gaussian noise). But this is an assumption that is often not met in practice.

- **Unbiasedness**: The expected value of the MLE is *unbiased*

$$\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})}[\hat{\boldsymbol{w}}] = \boldsymbol{w}$$

Proof that MLE is unbiased

The proof is quite easy

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})}[\hat{\boldsymbol{w}}] &= \int \hat{\boldsymbol{w}} p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{y} \\
&= \int (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \, \mathrm{d}\boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \int \boldsymbol{y} \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \, \mathrm{d}\boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{w}
\end{aligned}$$

Regularization

Ridge regression adds a penalty term to the loss function

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

- **Objective**: Find $\boldsymbol{w}$ that minimizes the error while keeping the parameters small
- **Solution**: $\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$

Probabilistic interpretation of regularization

- **Assumption**: Data is generated by a linear model with Gaussian noise independent across samples

Same trick as before (exponential of the negative loss)

$$\begin{aligned}\exp(-\gamma\mathcal{L}) &= \exp\left(-\frac{\gamma}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 - \frac{\gamma}{2}\lambda\|\boldsymbol{w}\|_2^2\right) = \\ &= \exp\left(-\frac{\gamma}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2\right)\exp\left(-\frac{\gamma}{2}\lambda\|\boldsymbol{w}\|_2^2\right) = \\ &\propto \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \gamma^{-1}\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{0}, (\gamma\lambda)^{-1}\boldsymbol{I}\right)\end{aligned}$$

Minimizing the loss is equivalent to maximizing the product of two Gaussian distributions (likelihood and prior).

We are getting closer to Bayesian inference!

Bayesian linear regression

Bayesian inference

Bayesian inference allows to "transform" a prior distribution over the parameters into a posterior **after** observing the data

Bayes' rule :

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X})p(\boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

- **Prior**: $p(\boldsymbol{w})$

  - Encodes our beliefs about the parameters **before** observing the data
- **Likelihod**: $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X})$
  - Encodes our model of the data
- **Posterior**: $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$
  - Encodes our beliefs about the parameters **after** observing the data (e.g. conditioned on the data)
- **Evidence**: $p(\boldsymbol{y} \mid \boldsymbol{X})$
  - Normalizing constant, ensures that $\int p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})\, \mathrm{d}\boldsymbol{w} = 1$

Bayesian linear regression - Likelihood and prior

Modeling observation as noisy realization of a linear combination of the features As before, we assume a Gaussian likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$$

For the prior, we use a Gaussian distribution over the model parameters

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \boldsymbol{S})$$

In practice, we often use a diagonal covariance matrix $\boldsymbol{S} = \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}$

When can we compute the posterior?

> ℹ Definition
>
> A prior is **conjugate** to a likelihood if the posterior is in the same family as the prior.

Only a few conjugate priors exist, but they are very useful.

Examples:

- Gaussian likelihood and Gaussian prior ⇒ Gaussian posterior
- Binomial likelihood and Beta prior ⇒ Beta posterior

Full table available on wikipedia

Why is this useful?

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

- **Generally** the posterior is **intractable** to compute
  - We don't the form of the posterior $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$
  - The evidence $p(\boldsymbol{y} \mid \boldsymbol{X})$ is an integral
    - without closed form solution
    - high-dimensional and computationally intractable to compute numerically

. . .

- **Analytical solution** thanks to conjugacy:
  - We know the form of the posterior
  - We know the form of the normalization constant
  - We don't need to compute the evidence, just some algebra to get the posterior

---

Back to our model, the posterior must be Gaussian $\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

Ignoring constant terms in $\boldsymbol{w}$:

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right) \\ &= \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right) \end{aligned}$$

From the likelihood and prior, we can write the posterior as

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w}) &\propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 - \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{S}^{-1}\boldsymbol{w}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{S}^{-1}\right)\boldsymbol{w} - \frac{2}{\sigma^2}\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y}\right)\right) \end{aligned}$$

. . .

From previous slide,

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)$$

We can identify the posterior mean and covariance

Posterior covariance

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{S}^{-1}\right)^{-1}$$

Posterior mean

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{y}$$

How to make predictions?

The posterior distribution $p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$ gives us the uncertainty about the parameters. **How can we use it to make predictions?**

. . .

The predictive distribution is the distribution of the target variable $y_\star$ given the input $\boldsymbol{x}_\star$

Obtained by **marginalizing** the parameters using the posterior

$$p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{X}) = \int p(y_\star \mid \boldsymbol{w}, \boldsymbol{x}_\star) p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{w}$$

. . .

For linear regression, the predictive distribution is Gaussian

$$p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(y_\star \mid \boldsymbol{\mu}^\top \boldsymbol{x}_\star, \boldsymbol{x}_\star^\top \boldsymbol{\Sigma}\boldsymbol{x}_\star + \sigma^2)$$

Bayesian linear regression with basis functions

The same approach can be used with non-linear basis functions

Transform the input $\boldsymbol{x}$ using a non-linear function $\varphi(\boldsymbol{x})$

$$\boldsymbol{x} \to \varphi(\boldsymbol{x}) = [\varphi_1(\boldsymbol{x}), \ldots, \varphi_D(\boldsymbol{x})]^\top$$

For convenience, define $\boldsymbol{\Phi} = [\varphi(\boldsymbol{x}_1), \ldots, \varphi(\boldsymbol{x}_N)]^\top$

Posterior

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top \boldsymbol{y} \quad \text{and} \quad \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{S}^{-1}\right)^{-1}$$

Predictive distribution

$$p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(y_\star \mid \boldsymbol{\mu}^\top \varphi(\boldsymbol{x}_\star), \varphi(\boldsymbol{x}_\star)^\top \boldsymbol{\Sigma} \varphi(\boldsymbol{x}_\star) + \sigma^2)$$

Where we used polynomial basis functions $\varphi_i(\boldsymbol{x}) = x^i$: $f(\boldsymbol{w}, \boldsymbol{x}) = \sum_{i=0}^{K} w_i x^i$

Analysis of Bayesian linear regression

Connection with ridge regression and maximum a posteriori estimation

Maximum a posteriori estimation (MAP) computes the mode of the posterior distribution $\arg\max p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$. For Gaussians, it is the same as the mean

$$\arg\min \mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

$$\arg\max p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

If $\lambda = \sigma_{\boldsymbol{y}}^2 / \sigma_{\boldsymbol{w}}^2$, the **ridge regression** solution is equivalent to the MAP solution with a Gaussian prior

Effect of the prior

- Prior encodes our beliefs about the parameters before observing the data.
- Prior effect diminishes with more data
- When we don't have much data, the prior can have a strong effect on the posterior

. . .

Question: How do we choose the prior?

1. Data type:

    - Real-values: Gaussian prior
    - Positive data: Log-normal prior, Gamma prior, etc.
    - 0-1 data: Beta prior
    - Data summing to 1: Dirichlet prior

2. Expert knowledge
3. Computational convenience

Model selection

We can now get a solution for the linear regression problem (Bayesian and not) but we have to choose the model

Questions:

- What is the best model for the data?
- How many basis functions should we use?
- How to avoid overfitting?

Attempted to choose the model based on the likelihood of the data. Is this a good idea?

NO!

- Higher complexity models will always have higher likelihood …
- … but they will also overfit the data and generalize poorly

Model selection - Cross-validation

- **Cross-validation**: Split the data into training and validation sets
- Solve the model with the training set and evaluate the performance on the validation set
- (Optional) Repeat the process for different splits of the data
- Choose the model that performs best on the validation set

Pros

- Simple
- Works well in practice

❗ Cons

- Computationally expensive
- Requires multiple runs
- Works poorly with small datasets
- **Violates** the **likelihood principle**

Likelihood principle

- **Likelihood principle**: All the information from the data is contained in the likelihood function
- **Consequence**: You should *not* base your inference on data that you *could have* observed but did not
- **Cross-validation** (and other frequentist methods) violate the likelihood principle

Note

- This is more of a philosophical point than a practical one.
- It's not a *rule of nature* but an argument that Bayesian methods are more *principled*.

Marginal likelihood

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

The **marginal likelihood** is the normalization constant of the posterior distribution

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \int p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) p(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}$$

- We are averaging the likelihood over **all possible values** of the parameters from the prior
- It tells us how likely the data is under the model

Let's be explicit about the model in the marginal likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\mathcal{M}}) = \int p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{\mathcal{M}}) p(\boldsymbol{w} \mid \boldsymbol{\mathcal{M}}) \, \mathrm{d}\boldsymbol{w}$$

where $\boldsymbol{\mathcal{M}}$ is the model (e.g. polynomial degree) and $p(\boldsymbol{w} \mid \boldsymbol{\mathcal{M}})$ is the prior over the parameters for the model $\boldsymbol{\mathcal{M}}$.

- Recipe for **Bayesian model selection**:

$$\widehat{\boldsymbol{\mathcal{M}}} = \arg\max_{\boldsymbol{\mathcal{M}}} p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\mathcal{M}})$$

This is also known as **Type II maximum likelihood** or **evidence maximization**

Model selection with Bayesian linear regression

Given:

- *Likelihood*: $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma_{\boldsymbol{y}}^2 \boldsymbol{I})$
- *Prior*: $p(\boldsymbol{w} \mid m) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$

The **marginal likelihood** is a Gaussian

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\mu}_p, \boldsymbol{X}\boldsymbol{\Sigma}_p \boldsymbol{X}^\top + \sigma_{\boldsymbol{y}}^2 \boldsymbol{I})$$

It does NOT depend on the parameters $\boldsymbol{w}$ but only on the model!

Why does it work? The Bayesian Occam's razor

Does $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\mathcal{M}})$ favor complex models? **No!**

- We *marginalize* over the parameters, not *maximize* them
- The marginal likelihood penalizes complex models that don't fit the data well

This is known as the **Bayesian Occam's razor**: the simplest model that explains the data is the best

Intuition

Apply chain rule to the marginal likelihood (drop all dependencies):

$$p(\boldsymbol{y}) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_2) \ldots p(y_N \mid y_1, \ldots, y_{N-1})$$

or equivalently

$$\log p(\boldsymbol{y}) = \log p(y_1) + \log p(y_2 \mid y_1) + \log p(y_3 \mid y_1, y_2) + \ldots + \log p(y_N \mid y_1, \ldots, y_{N-1})$$

- If the model is too complex, it will predict early data points well but later data points poorly