02462 – Signals and data

Technical University of Denmark,
DTU Compute, Institut for Matematik og Computer Science.

Overview

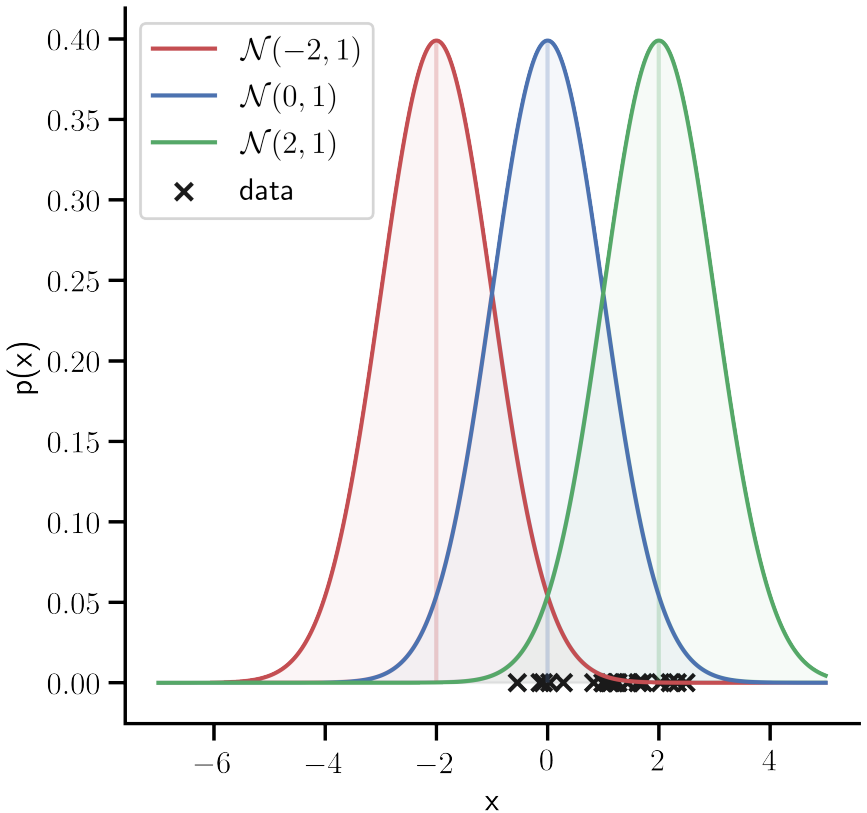
1. Inference
2. MAP

Inference

Inference

- Many problems fall into a category where,
  - we have collected *data* $\{x_n\}_{n=1}^N$.
  - we think the data is produced by a *probability distribution* $p(x|\theta)$.
  - we want to determine the *unknown parameter* θ.
- Estimating θ is known as *inference*.

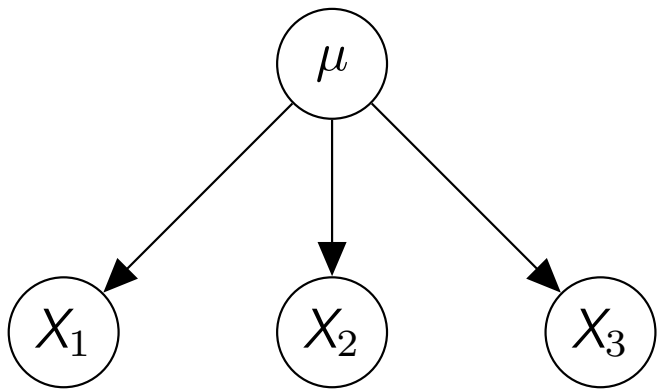
If we think the data is normal-distributed, which normal distribution then fits the data the best?



Building a Generative Story

How would we generate the data ourselves?

- First draw *random parameters*, e.g. μ.
- Draw *random data* using the parameters, e.g. draw from the normal distribution with mean μ.



Example: Linear Regression

Linear regression tries to find a function f mapping input x to targets y of the form,

$$f(\boldsymbol{x}; \boldsymbol{w}) = w_0 + w_1 x$$

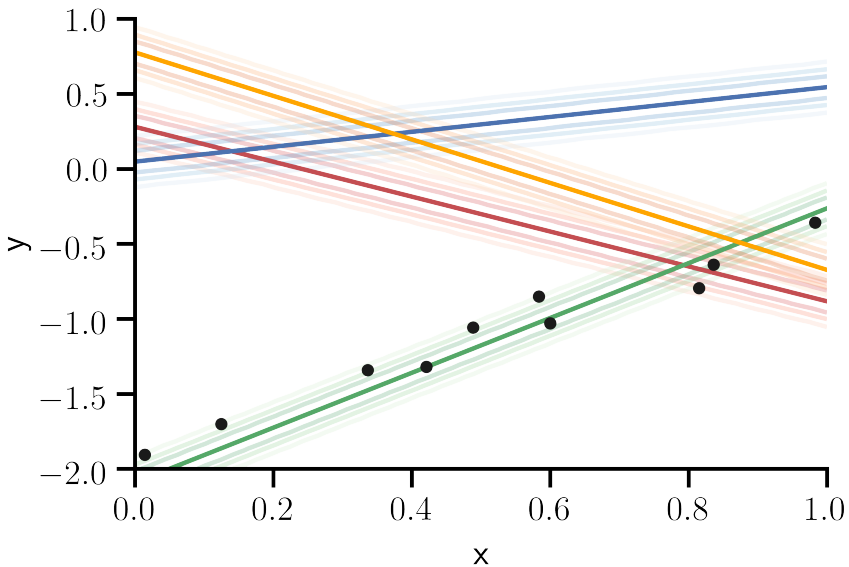
Probabilistic linear regression tries to take into account that the model is not exact by adding a noise term,

$$y = f(x; \boldsymbol{w}) + \epsilon.$$

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is random noise.
- because of the noise, y becomes a random variable too,

$$p(y|x, \boldsymbol{w}) = \mathcal{N}(y; f(x; \boldsymbol{w}), \sigma^2)$$

Bayesian linear regression completes the model with a prior $p(\boldsymbol{w})$.



Generative story sample a line $\boldsymbol{w}$, then compute the mean $f(\boldsymbol{x}; \boldsymbol{w})$ and sample the value of y by adding noise.

The Posterior

- For parameters θ, we call $p(\theta)$ the *prior*.
- For data $\mathcal{D}$, we define $p(\mathcal{D}|\theta)$ as the *likelihood*.
- Following the generative story, we can *sample hypothetical datasets*.
- What parameter θ is most likely to generate the data we observe $\mathcal{D}$?

Posterior

Using Bayes' theorem, we can find the probability of the parameters θ given the data $\mathcal{D}$,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \qquad \text{(posterior)}$$

- Computing the posterior is known as *Bayesian inference*.

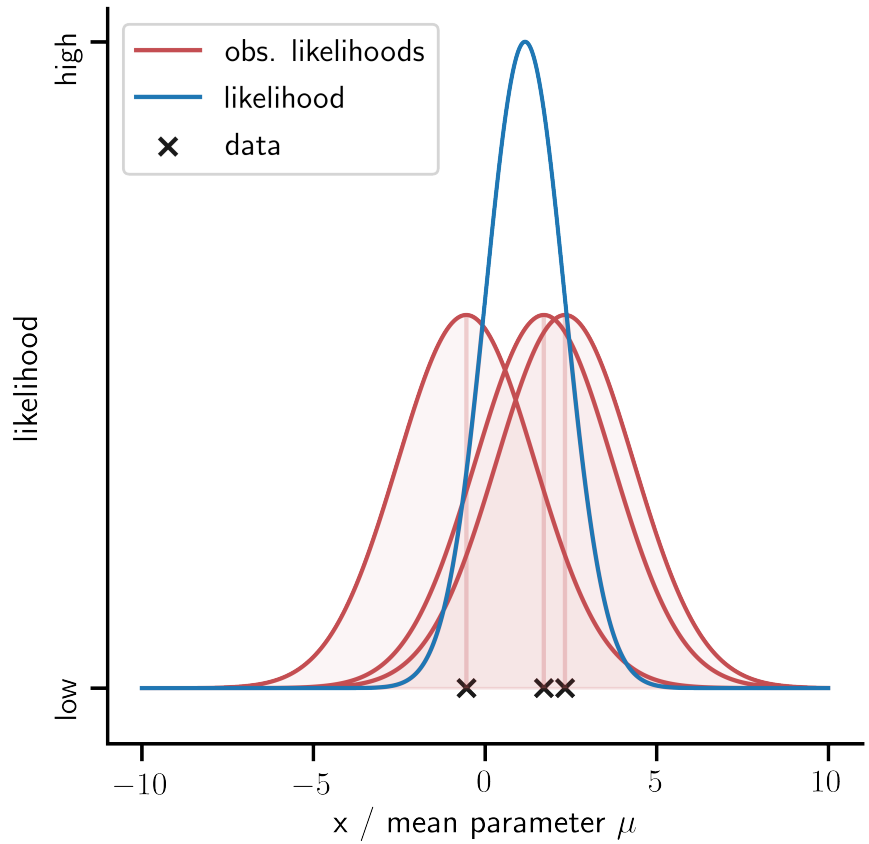
Likelihood

- the *likelihood* is the probability of the fixed data $\mathcal{D}$ for a specific θ,

$$L(\theta) = p(\mathcal{D}|\theta)$$

- it measures how *likely* it is to generate data if the parameter is θ.
- Think of it as *evidence* in favor of picking a particular θ!
- The evidence of a dataset is the combined evidence of its independent observations,

$$L(\theta) = p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \underbrace{p(x_n|\theta)}_{\substack{\text{observation} \\ \text{likelihood}}} .$$



Example if $L(\mu) = \mathcal{N}(x|\mu, \sigma^2_{\text{fixed}})$, each observation likelihood is high when μ is close to the datapoint x. The likelihood is high when μ is close to all the data.

[0] here assuming independent x_n

The Objective Prior

The prior $p(\boldsymbol{\theta})$ sets the generative story in motion — each sample presents a possible way the data could have been generated.

- A choice of prior implies that some models are more likely than others.
- If we think *all models are equally likely*, we have to use a prior that gives them equal weight:

$$p(\boldsymbol{\theta}) \propto 1 \qquad \text{(objective prior)}$$

- This makes the posterior proportional to the likelihood,

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}) \cdot 1}{p(\mathcal{D})} \propto p(\mathcal{D}|\boldsymbol{\theta})$$

We are disregarding everything but the data evidence.

The Most Probable Parameter — Objective Version

An intuitive inference strategy is to pick the parameter that is *most probable*. If the prior is conservative, this is the same as finding the parameter with *highest likelihood*!

1. Find the likelihood function $L(\boldsymbol{\theta}) = p(\mathcal{D}|\theta)$.
2. Pick the model with the highest likelihood $\theta_{\text{ML}} = \arg\max_\theta p(\mathcal{D}|\theta)$.

We can convert this into a *model loss* by using the *negative log-likelihood*.

The Maximum Likelihood Estimator

The MAP estimate of the parameter θ based on data $\mathcal{D}$ is found by solving,

$$\theta_{\text{ML}} = \arg\min \underbrace{-\ln p(\mathcal{D}|\theta)}_{\text{log-likelihood}}.$$

MAP

The Problem with Being Objective

Problem *You want to figure out the probability of a rainy day. You collect 10 days of data. Every day it rains. Maximum likelihood gives you* 100% *chance of rain.*

- assuming that every model is *equally likely* is just as much an *assumption*.
- *Example* do you think slope $w_1 = 10^{80}$ is just as likely as $w_1 = 1$ in all cases?

For *Bayesian linear regression*, a common choice of prior is

$$p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \sigma^2_{\text{prior}} \boldsymbol{I}). \qquad \text{(normal prior)}$$

What assumptions does this prior make?

slope is normal it is more likely that the slope is flat than steep.

intercept is normal it is more likely that the function value is close to 0 than far from 0.

The Most Probable Parameter

In general, the most probable parameter corresponds to the maximum a posteriori solution (the *MAP*),

1. Find the posterior $p(\theta|\mathcal{D})$.
2. Pick the model with the highest posterior density $\theta_{\text{MAP}} = \arg\max_\theta p(\theta|\mathcal{D})$.

Joint and posterior are proportional when $\mathcal{D}$ is fixed,

$$p(\theta|\mathcal{D}) = \frac{p(\theta, \mathcal{D})}{p(\mathcal{D})} \propto p(\theta, \mathcal{D}).$$

so we can equivalently minimize the *negative logarithm of the joint density*.

The MAP Estimator

The MAP estimate of the parameter θ based on data $\mathcal{D}$ is found by solving,

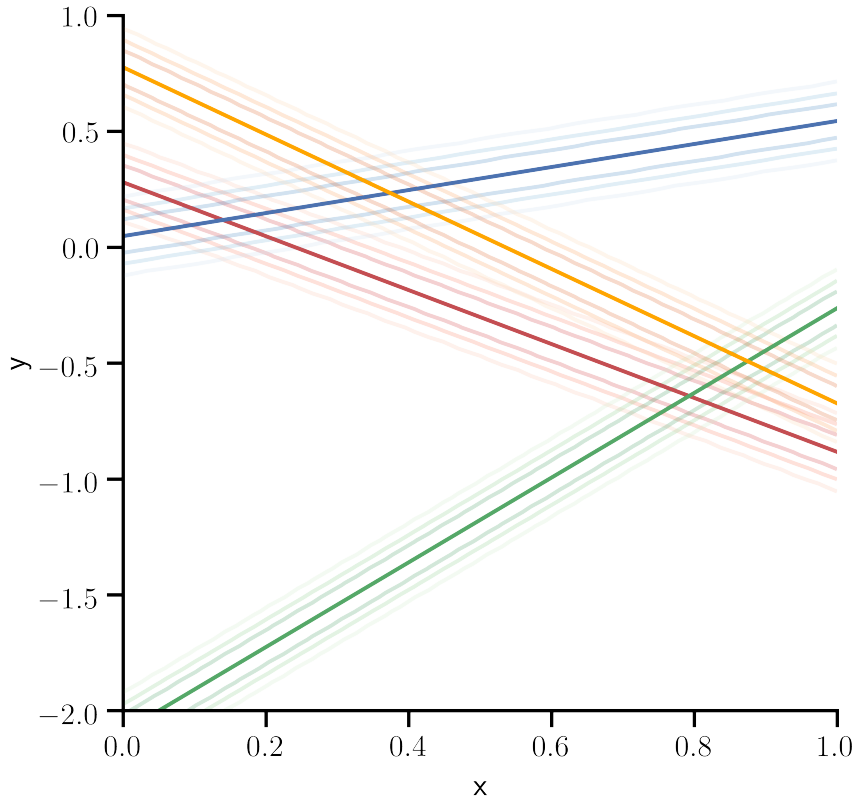
$$\theta_{\text{MAP}} = \arg\min -\ln p(\theta, \mathcal{D}) = \arg\min -\underbrace{\ln p(\mathcal{D}|\theta)}_{\text{log-likelihood}} - \underbrace{\ln p(\theta)}_{\text{log-prior}}.$$

The MAP θ_{MAP} is both *likely to be generated* in the generative story, and has data lending it *evidence*.

Linear Regression Continued: 0 observations

$$p(\boldsymbol{w}|\mathcal{D}_0) = p(\boldsymbol{w})$$

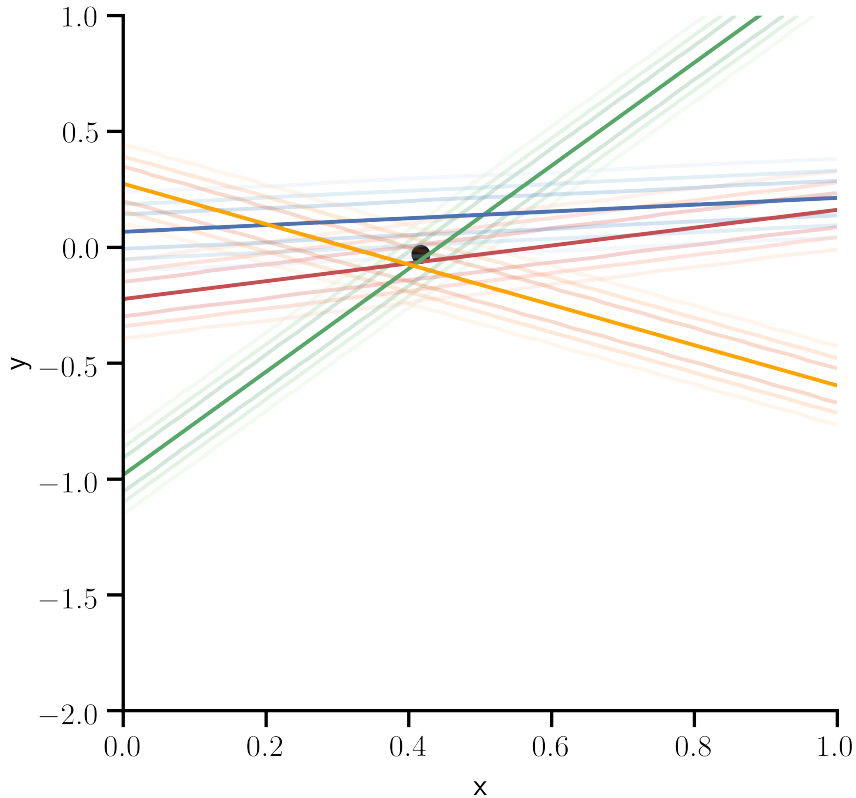
- Without data, the posterior is just the prior $p(\boldsymbol{w})$.
- With one data point, the posterior is limited to lines passing through the point.
- With two points, the posterior concentrates on lines passing through both.



Linear Regression Continued: 1 observation

$$p(\boldsymbol{w}|\mathcal{D}_1) = \frac{p(y_1, x_1|\boldsymbol{w})p(\boldsymbol{w})}{p(y_1, x_1)}$$

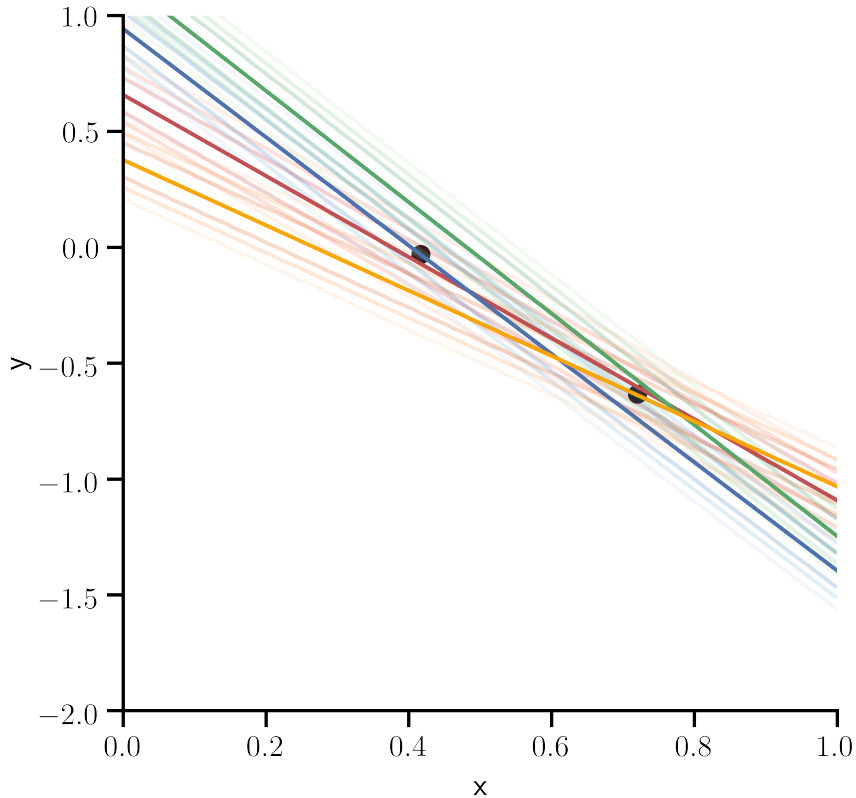
- Without data, the posterior is just the prior $p(\boldsymbol{w})$.
- With one data point, the posterior is limited to lines passing through the point.
- With two points, the posterior concentrates on lines passing through both.



Linear Regression Continued: 2 observations

$$p(\boldsymbol{w}|\mathcal{D}_2) = \frac{p(y_2, x_2|\boldsymbol{w})p(y_1, x_1|\boldsymbol{w})p(\boldsymbol{w})}{p(y_2, x_2, y_1, x_1)}$$

- Without data, the posterior is just the prior $p(\boldsymbol{w})$.
- With one data point, the posterior is limited to lines passing through the point.
- With two points, the posterior concentrates on lines passing through both.



Linear Regression Continued: MAP

returning to the linear regression case with observations $\{(x_n, y_n)\}_{n=1}^N$,

$$-\ln p(\mathcal{D}|\boldsymbol{w}) = -\sum_{n=1}^{N} \ln \mathcal{N}(y_n; f(x_n; \boldsymbol{w}), \sigma^2) = \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - f(x_n; \boldsymbol{w}))^2}_{\text{least squares loss!}} + \underbrace{\frac{N}{2} \ln 2\pi\sigma^2}_{\text{constant}}$$

Frequently, *loss functions are just negative log-likelihoods in disguise!*
To find the MAP, we also need to include the log-prior,

$$-\ln p(\boldsymbol{w}) = -\ln \mathcal{N}(\boldsymbol{w}; \mathbf{0}, \sigma^2_{\text{prior}} \boldsymbol{I}) = \underbrace{\frac{1}{2\sigma^2_{\text{prior}}} \sum_{d=1}^{D} w_d^2}_{\text{L2 regularizer!}} + \underbrace{\frac{1}{2} \ln 2\pi\sigma^2_{\text{prior}}}_{\text{constant}} . \qquad (1)$$

Priors are regularizers in disguise! Regularizers bias the solution towards more reasonable values; L2 encourages functions with small slope and intercept.

Linear Regression Continued: MAP

returning to the linear regression case with observations $\{(x_n, y_n)\}_{n=1}^N$,

$$-\ln p(\mathcal{D}|\boldsymbol{w}) = -\sum_{n=1}^{N} \ln \mathcal{N}(y_n; f(x_n; \boldsymbol{w}), \sigma^2) = \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - f(x_n; \boldsymbol{w}))^2}_{\text{least squares loss!}} + \underbrace{\frac{N}{2} \ln 2\pi\sigma^2}_{\text{constant}}$$

Frequently, *loss functions are just negative log-likelihoods in disguise!*
To find the MAP, we also need to include the log-prior,

$$-\ln p(\boldsymbol{w}) = -\ln \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \sigma^2_{\text{prior}} \boldsymbol{I}) = \underbrace{\frac{1}{2\sigma^2_{\text{prior}}} \sum_{d=1}^{D} w_d^2}_{\text{L2 regularizer!}} + \underbrace{\frac{1}{2} \ln 2\pi\sigma^2_{\text{prior}}}_{\text{constant}}. \tag{1}$$

Priors are regularizers in disguise! Regularizers bias the solution towards more reasonable values; L2 encourages functions with small slope and intercept.

Maximum Likelihood and Regularization

- If we let σ^2_{prior} become very large, the regularization will disappear.
- What remains is the log-likelihood which defines the *maximum likelihood stimator*,

$$\theta_{\text{ML}} = \arg\min - \ln p(\mathcal{D}|\theta).$$

- the *prior* $p(\theta)$ shifts the posterior towards itself.