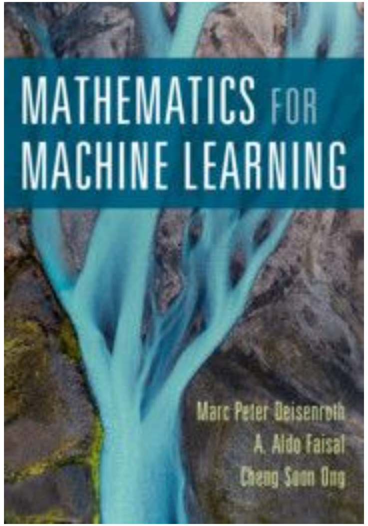


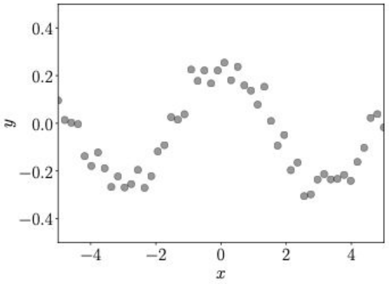

# Chapter 9. Linear Regression

San Diego Machine Learning

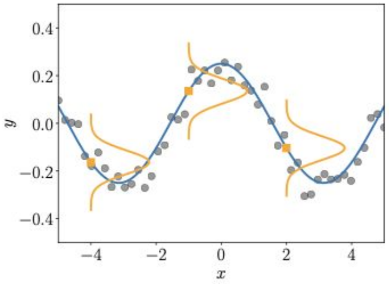
Ryan Chesler

# Objective

- Find a function that not only models the training data but also generalizes to new inputs
- We assume some noise, but we are trying to find the underlying function




(a) Regression problem: observed noisy function values from which we wish to infer the underlying function that generated the data.




(b) Regression solution: possible function that could have generated the data (blue) with indication of the measurement noise of the function value at the corresponding inputs (orange distributions).

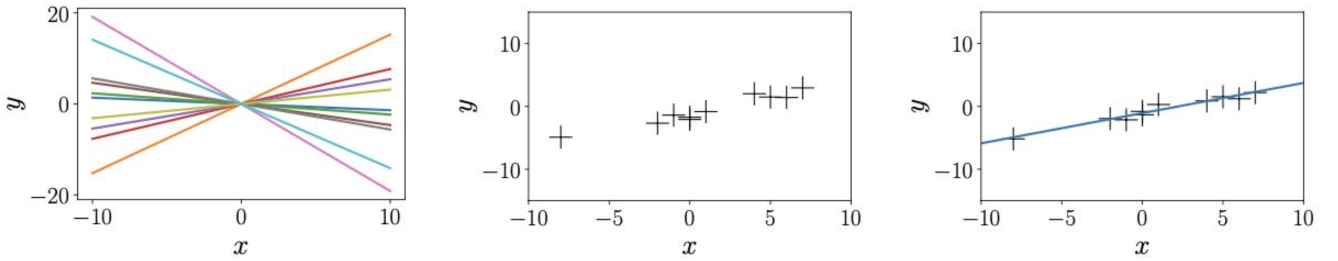
# Modeling

- Model Type and Parameterization
  - Function class, eg. polynomial and to what degree
- How to find the optimal parameters
  - Loss function and deciding how it should be optimized
- Overfitting
  - How well does the model perform on new inputs
- Relationship between loss functions and parameter priors
- Uncertainty modeling
  - Confidence bounds to represent areas of uncertainty

# 9.1 Problem Formulation

- Find parameters θ that “work well” for the data

$$
\begin{aligned}
& p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \boldsymbol{x}^\top \boldsymbol{\theta},\, \sigma^2\right) \\
\iff & y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma^2\right),
\end{aligned}
$$




(a) Example functions (straight lines) that can be described using the linear model in (9.4).

(b) Training set.

(c) Maximum likelihood estimate.

# 9.2 Parameter Estimation

Training set (x1, y1)...(xn, yn)

$$= \prod_{n=1}^{N} \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right)$$

# 9.2.1 Maximum Likelihood Estimation

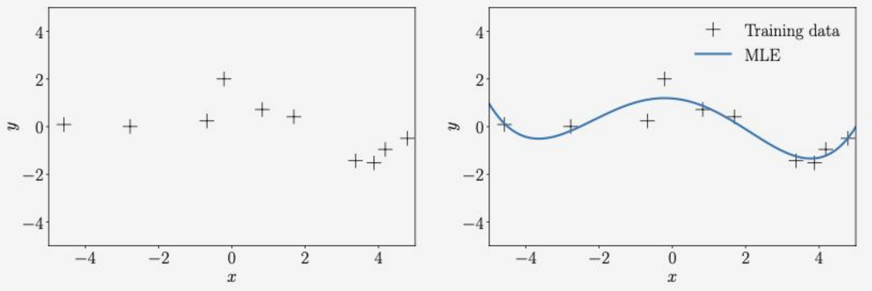
- Finding θ that maximizes the likelihood of y given x
- Negative Log-likelihood

$$-\sum_{i=1}^{n} \log p(y_i \mid x_i, \theta)$$

- Can find the closed form solution by computing the gradient and setting it to 0 and solving for θ

# Polynomial Regression

- Often straight lines are not sufficient to represent a real function
- "Linear" only means linear in parameters, it is still possible to represent nonlinear functions
- We can transform the inputs to a new nonlinear space like adding higher powers
- From X to Φ space
- X, x^2, x^3
- Can fit the same linear model on top of this transformed space to get a non-linear outcome in the original space

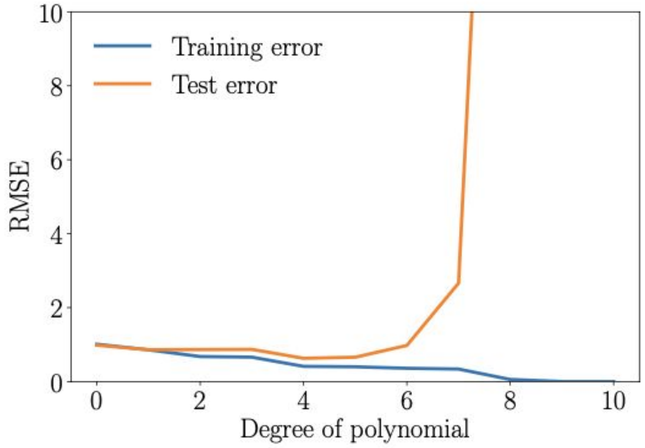


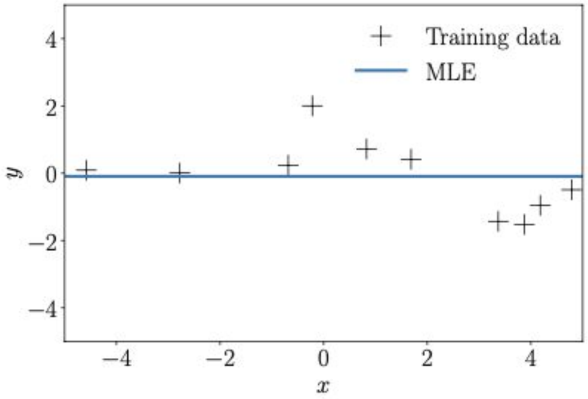
(a) Regression dataset.
(b) Polynomial of degree 4 determined by maximum likelihood estimation.

# 9.2.2 Overfitting

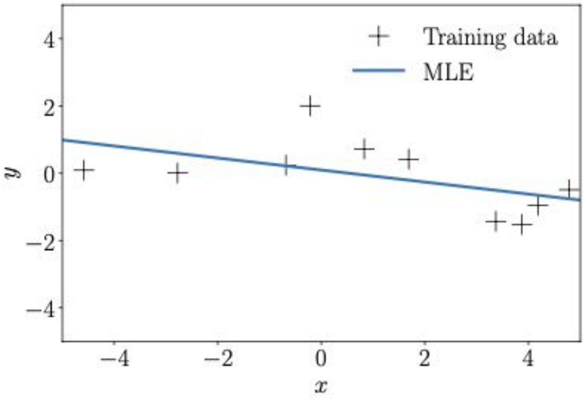
- Root Mean Squared Error, remove assumption about the noise and measure error in the original units

$$= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta})^2}$$

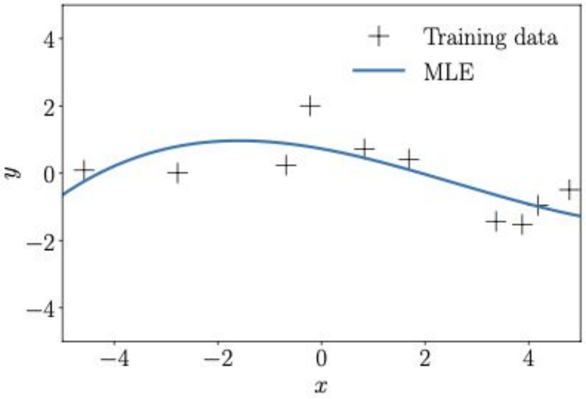




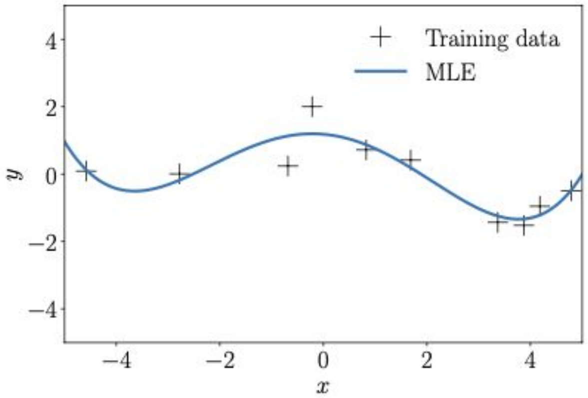

(a) $M = 0$



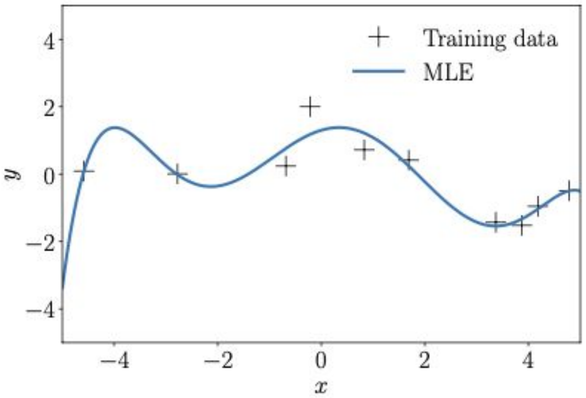

(b) $M = 1$



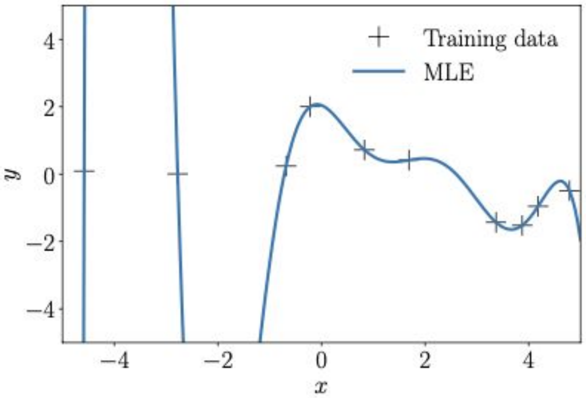

(c) $M = 3$




(d) $M = 4$



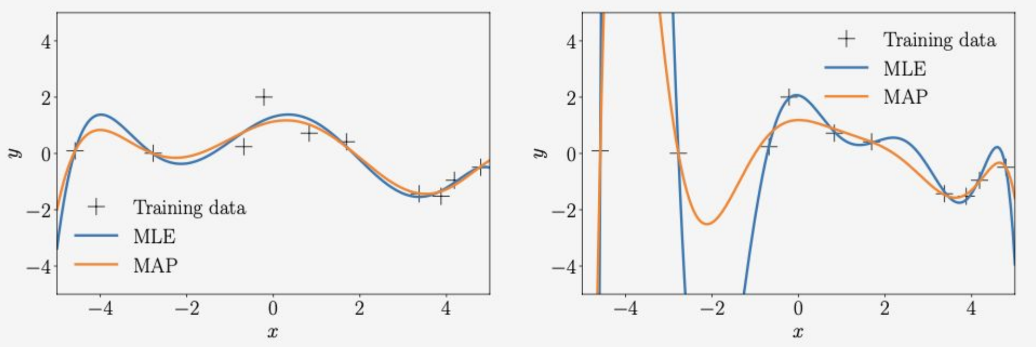

(e) $M = 6$




(f) $M = 9$

# 9.2.3 Maximum A Posteriori Estimation

- Instead of just finding the most likely, find the most likely given some prior about the parameters
- Gaussian Prior on a single parameter encodes an expectation that the values lie in the interval [-2, 2]
- Maximizing the posterior distribution
- Similar to maximum likelihood but adding a term for the log-prior




(a) Polynomials of degree 6.

(b) Polynomials of degree 8.

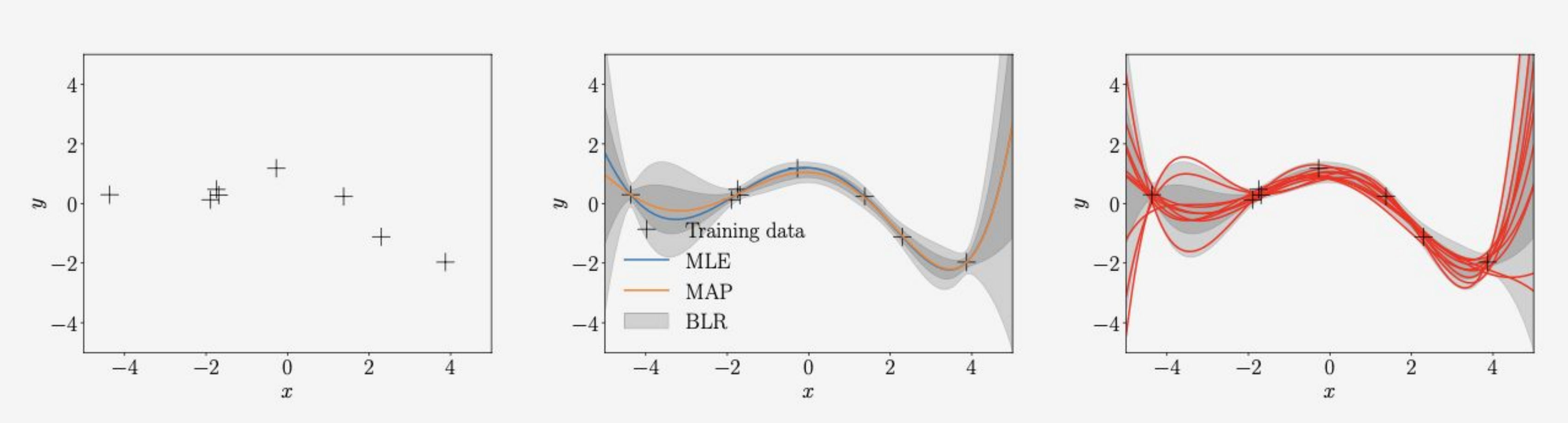
# 9.2.4 MAP Estimation as Regularization

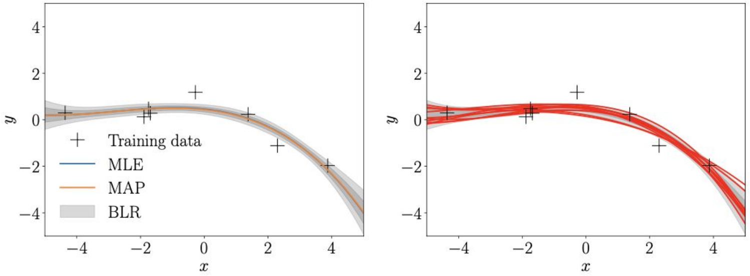
- Lambda determines “strictness” of regularization
- Can use different norms to constrain in different ways. Smaller p-norms lead to sparser solutions
  - Useful for variable selection
  - P = 1 is called LASSO

$$\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

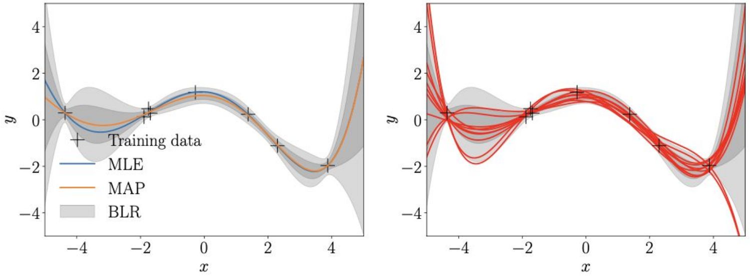
# 9.3 Bayesian Linear Regression

- The same linear model we have discussed in previous sections
- Need a prior like in MAP
- Not very interested in the parameters of Φ
- Average over all plausible parameter settings when we make predictions

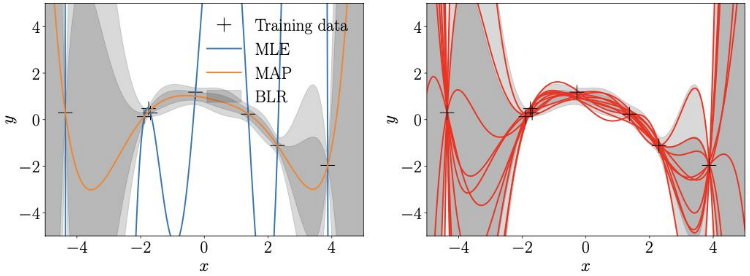





(a) Posterior distribution for polynomials of degree $M = 3$ (left) and samples from the posterior over functions (right).




(b) Posterior distribution for polynomials of degree $M = 5$ (left) and samples from the posterior over functions (right).



# Summary

- Maximum Likelihood - single point estimate of the parameters Φ that maximize likelihood of y given x. No prior
- Maximum A Posteriori - single point estimate that includes some prior that constrains the parameters
- We can represent nonlinear functions even with linear models
- Bayesian Linear Models - instead of looking for the most likely parameters returns a distribution based on all plausible parameters based on some prior