Machine Learning

Sören Laue

Universität Hamburg

Bias-Variance Tradeoff

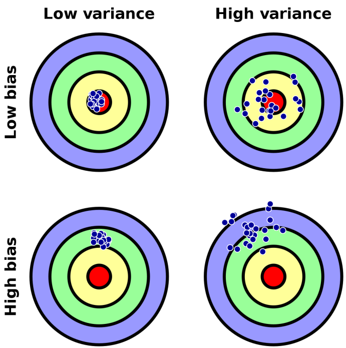
- we would like to have a model that captures the training data accurately but also generalizes well to unseen data
- we have seen that this is usually not possible
- high variance models can capture training data arbitrarily good but might overfit (high model complexity)
- high bias models have a small model complexity but might underfit the data
- need to trade off bias vs. variance

Bias-Variance Tradeoff – Least Squares Regression

- we have training points, a regression method, and a loss function (least squares here)
- we are interested in the generalization error, i.e., the *expected* prediction error we make on some unseen data point x
- assume f^* is the true function/signal
- f_n is the prediction model returned by the algorithm, i.e., the function from our model class $\mathcal{F}$ that we have learned based on n (random) training points
- $\mathbb{E}\left[(f_n(x) - f^*(x))^2\right] = \underbrace{\mathbb{E}\left[(f_n(x) - \mathbb{E}[f_n(x)])^2\right]}_{\text{variance term}} + \underbrace{(\mathbb{E}[f_n(x)] - f^*(x))^2}_{\text{bias term}}$
- variance term = variance of the random variable $f_n(x)$
- bias term = how much $\mathbb{E}[f_n(x)]$ and $f^*(x)$ deviate

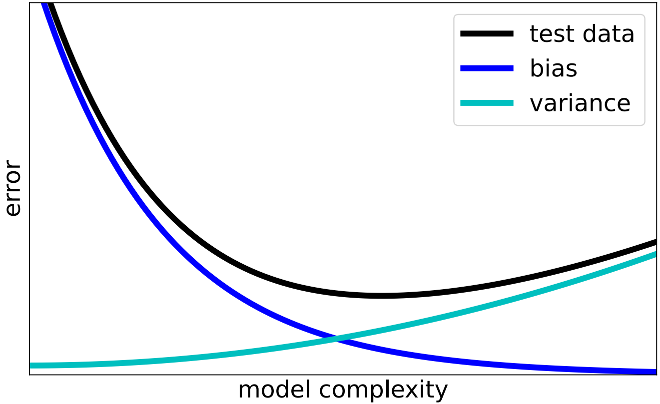
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bias-Variance Tradeoff

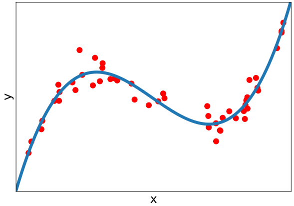


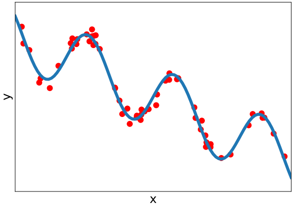
- equation for expected prediction error, i.e., bias-variance decomposition was for least squares loss function
- similar equations exist also for other loss functions

Bias-Variance Tradeoff



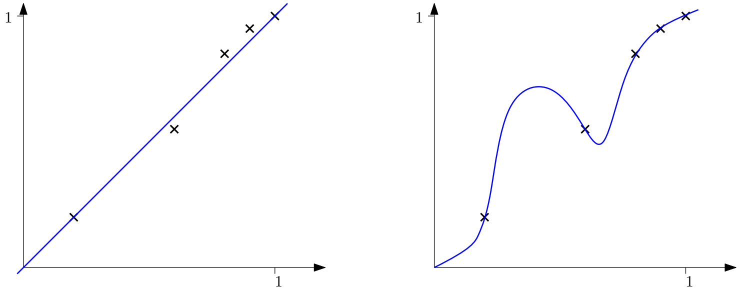
Basis Functions and Bias-Variance Tradeoff





- Just use many basis functions and we can model **anything.**
- e.g., polynomials (Weierstrass, 1865), piece-wise linear functions, wavelets, ... are uniform approximators
- Did we solve ML? Are we done now? Just use tons of basis functions. Make $\mathcal{F}$ as big and complex as possible.
- **bias will go to 0, but variance will go up, and so will generalization error**

Remember this slide?



Maximum Likelihood Principle

- assume we have the following random experiment:
- we have a coin that shows head with probability $p(heads) = \theta$ and tails with probability $p(tails) = 1 - \theta$.
- we throw the coin 10 times and 3 times heads comes up and 7 times tails
- what would most likely be the parameter θ?
- why would you compute it this way?
- let's follow the *maximum likelihood principle*

Maximum Likelihood Principle

- assumptions: all throws are independently and with the same coin
- our n data points $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$ are identically and independently distributed (iid)
- probability of seeing heads in throw i: $p(y^{(i)} = \mathit{heads} \mid \theta) = \theta$
- probability of seeing tails in throw i: $p(y^{(i)} = \mathit{tails} \mid \theta) = 1 - \theta$
- probability of any outcome:

$$\begin{aligned} p(y^{(1)}, y^{(2)}, \ldots, y^{(n)} \mid \theta) &= p(y^{(i)} \mid \theta) \cdot p(y^{(2)} \mid \theta) \ldots \cdot p(y^{(n)} \mid \theta) \\ &= \prod_i p(y^{(i)} \mid \theta) \end{aligned}$$

because they are independent random variables

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- probability of any outcome:

$$p(y^{(1)}, y^{(2)}, ..., y^{(n)} \mid \theta) = \prod_i p(y^{(i)} \mid \theta)$$

- in our case:

$$\begin{aligned} p(3 \text{ times head, } 7 \text{ times tails } \mid \theta) &= p(y^{(i)} = \mathit{heads} \mid \theta)^3 \cdot p(y^{(i)} = \mathit{tails} \mid \theta)^7 \\ &= \theta^3 \cdot (1-\theta)^7 \end{aligned}$$

- **maximum likelihood estimator (MLE)**: find the parameter θ that would make our observation most likely (that would maximize the probability to see our observation), i.e.,

$$\max_\theta \theta^3 \cdot (1-\theta)^7$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- $$\theta^* = \text{argmax}_\theta\, p(\theta) = \theta^3 \cdot (1-\theta)^7$$
- taking logarithm does not change the maximal point
$$\begin{aligned}\theta^* &= \text{argmax}_\theta \log(p(\theta)) \\ &= \text{argmax}_\theta\, 3 \cdot \log(\theta) + 7 \cdot \log(1-\theta)\end{aligned}$$
- maximum -> set derivative to 0
$$\frac{d \log(p)}{d\theta} = 3 \cdot \frac{1}{\theta} - 7 \cdot \frac{1}{1-\theta} \stackrel{!}{=} 0$$
- $$\begin{aligned}3(1-\theta^*) - 7\theta^* &= 0 \\ 3 - 10\theta^* &= 0 \\ \theta^* &= \frac{3}{10}\end{aligned}$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- $\theta^* = 0.3$ as expected
- we have a mathematical sound way of computing the best parameter θ that matches our intuition
- so far so good
- what happens if we throw the coin twice and we observe 1 heads and 1 tails?
- and what happens if we throw the coin twice and we observe 0 heads and 2 tails?
- we would predict $\theta^* = 0$, does this make sense?

Maximum A Posteriori (MAP)

- maximum likelihood principle can be too strict wrt to the observations
- instead of maximizing $p(\text{our observation} \mid \theta)$ we maximize

$$p(\theta \mid \text{our observation})$$

- in words: find the most probable parameter, given our observations
- this is called the **maximum a posteriori (MAP)** estimation
- so we treat θ as a random variable (Bayesian approach vs frequentist view)

Maximum A Posteriori (MAP)

- we have according to Bayes law

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

- follows easily from the definition of $p(A \mid B)$, since

$$p(A \mid B) = \frac{p(A, B)}{p(B)}$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum A Posteriori (MAP)

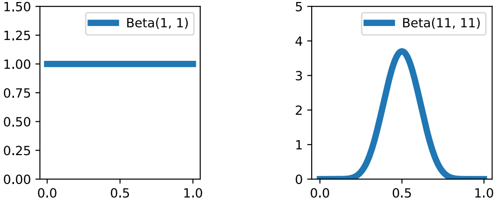
- so in our case we maximize

$$p(\theta \mid \text{our observation}) = \frac{p(\text{our observation} \mid \theta) \cdot p(\theta)}{p(\text{our observation})}$$

- $p(\theta)$ is some prior knowledge we have on the distribution of θ, aka **prior** (so once again we add some inductive bias here)

- for instance θ can follow a beta distribution, i.e.,

$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$



Maximum A Posteriori (MAP)

- since $p(\text{our observation})$ does not depend on θ we maximize

$$p(\theta \,|\, \text{our observation}) \sim p(\text{our observation} \,|\, \theta) \cdot p(\theta)$$

- for $\mathrm{Beta}(\alpha, \beta)$ the probability density function is

$$\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

where $B(\alpha, \beta)$ is a scaling factor depending on α and β

- so we maximize

$$\theta^{n_{\text{heads}}} \cdot (1 - \theta)^{n_{\text{tails}}} \cdot \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum A Posteriori (MAP)

- so we maximize

$$\theta^{n_{\text{heads}}} \cdot (1-\theta)^{n_{\text{tails}}} \cdot \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- $$\frac{1}{B(\alpha,\beta)} \cdot \theta^{n_{\text{heads}}+\alpha-1} \cdot (1-\theta)^{n_{\text{tails}}+\beta-1}$$

- or we maximize the logarithm

$$\log\left(\frac{1}{B(\alpha,\beta)} \cdot \theta^{n_{\text{heads}}+\alpha-1} \cdot (1-\theta)^{n_{\text{tails}}+\beta-1}\right)$$

- $$\log\left(\frac{1}{B(\alpha,\beta)}\right) + (n_{\text{heads}}+\alpha-1)\log(\theta) + (n_{\text{tails}}+\beta-1)\log(1-\theta)$$

- compute derivative and set to 0

$$\frac{n_{\text{heads}}+\alpha-1}{\theta} - \frac{n_{\text{tails}}+\beta-1}{1-\theta} \stackrel{!}{=} 0$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum A Posteriori (MAP)

- $$\frac{n_{\text{heads}} + \alpha - 1}{\theta} - \frac{n_{\text{tails}} + \beta - 1}{1 - \theta} \overset{!}{=} 0$$

- $$(n_{\text{heads}} + \alpha - 1)(1 - \theta) - (n_{\text{tails}} + \beta - 1)\theta = 0$$

- $$n_{\text{heads}} + \alpha - 1 = \theta \cdot (n_{\text{heads}} + \alpha - 1 + n_{\text{tails}} + \beta - 1)$$

- $$\theta^* = \frac{n_{\text{heads}} + \alpha - 1}{n_{\text{heads}} + \alpha - 1 + n_{\text{tails}} + \beta - 1}$$

Maximum A Posteriori (MAP)

- $$\theta^* = \frac{n_{\text{heads}} + \alpha - 1}{n_{\text{heads}} + \alpha - 1 + n_{\text{tails}} + \beta - 1}$$
- for $n_{\text{heads}} = 0$ and $n_{\text{tails}} = 2$ and $\alpha = \beta = 11$, we obtain
$$\theta^* = \frac{10}{20 + 2} \approx 0.45$$
- it is like having seen our observation and *additionally* $\alpha - 1$ heads and $\beta - 1$ tails
- (beta-distribution is a conjugate prior to the Binomial distribution)
- maximum likelihood estimator (MLE) would give $\theta^* = 0$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MLE and MAP – Summary

- task: given some observations, what is the best parameter θ
- we have seen two (mathematically grounded) approaches that match our intuition
- maximum likelihood estimator (MLE): maximize

$$p(\text{observations} \mid \theta)$$

- estimates θ only based on observations
- maximum a posteriori estimator (MAP): maximize

$$p(\theta \mid \text{observation}) = \frac{p(\text{observation} \mid \theta) \cdot p(\theta)}{p(\text{observation})}$$

- estimates θ based on observations and prior knowledge
- treats parameter itself as a random variable (shift to Bayesian view)
- ‘softens’ the impact of the observations

MLE and MAP

- which one (MLE or MAP) is the *right* one?
- both are correct – in general neither is better than the other
- we have seen, when we have little data MAP seems to be a better choice
- will give a more satisfying answer to this soon

MLE and MAP and ML

- let's apply MLE and MAP to ML problems
- given observations / (training) data what are the best parameters w

Maximum Likelihood Principle

- assume we have data $\left(x^{(i)}, y^{(i)}\right)_{i=1}^{n}$ and the label of the data is generated by the linear function $x^\top w$ (true signal) and some noise ε, i.e.,

$$y = x^\top w + \varepsilon$$

- here y is a random variable
- assume the noise ε follows a normal distribution, i.e.,

$$\varepsilon \sim N(0, \sigma^2)$$

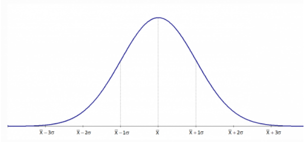
- so we have

$$y \,|\, x, w \sim N(x^\top w, \sigma^2)$$

- so, if we know w and σ^2 we can compute the probability of y

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle



- for $N(\mu, \sigma^2)$ the probability density function is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

- if we know μ and σ^2 we can compute the probability of the outcome y
- so in our case

$$p(y \,|\, x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y-x^\top w}{\sigma}\right)^2\right)$$

Maximum Likelihood Principle

- we assume all data points $\left(x^{(i)}, y^{(i)}\right)_{i=1}^{n}$ are *independently and identically distributed* (iid)
- independent: $p(y^{(i)}, y^{(j)} \mid x, w) = p(y^{(i)} \mid x, w) \cdot p(y^{(j)} \mid x, w)$
- identically distributed: they all follow the same distribution $N(x^\top w, \sigma^2)$
- given w, σ^2, and $X = \left(x^{(1)}, x^{(2)}, \dots, x^{(n)}\right)$ what is the probability to see the observation / labels $y = (y^{(1)}, y^{(2)}, \dots, y^{(n)})$?
-
$$\begin{aligned} p(y \mid X, w) &= \prod_{i=1}^{n} p(y^{(i)} \mid X, w) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right) \end{aligned}$$

Maximum Likelihood Principle

- maximum likelihood estimator (MLE): find the parameter w such that the likelihood of the observations is maximized
- $$\operatorname{argmax}_w p(y \mid X, w)$$
- $$\operatorname{argmax}_w \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - \left(x^{(i)}\right)^\top w}{\sigma}\right)^2\right)$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- $$\operatorname{argmax}_w \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)$$
- or maximize the logarithm
$$\operatorname{argmax}_w \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)\right)$$
- $$\operatorname{argmax}_w \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)\right)$$

Maximum Likelihood Principle

- $$\operatorname{argmax}_w \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)\right)$$

- $$\operatorname{argmax}_w \sum_{i=1}^{n} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)\right)\right)$$

- $$\operatorname{argmax}_w \sum_{i=1}^{n} \log\left(\exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right)\right)$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- $$\text{argmax}_w \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2$$
- $$\text{argmax}_w -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - (x^{(i)})^\top w\right)^2$$
- $$\text{argmin}_w \sum_{i=1}^{n}\left(y^{(i)} - (x^{(i)})^\top w\right)^2$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum Likelihood Principle

- $$\text{argmin}_w \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w\right)^2$$
- $$\text{argmin}_w \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}^{(i)}\right)$$

  with

  $$l(y, \hat{y}) = (y - \hat{y})^2$$

  and

  $$\hat{y} = x^\top w$$
- this is empirical risk minimization with the squared loss function

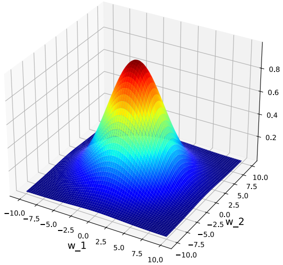
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Equivalence Empirical Risk Minimization and Maximum Likelihood Principle

empirical risk minimization	maximum likelihood
minimize	maximize
sum	product
risk / loss function	noise distribution
l_2-loss	Gaussian distribution
l_1-loss	Laplacian distribution
$\vdots$	

Maximum A Posteriori (MAP)

- that was MLE, let's look at MAP now
- assume we have some prior knowledge/distribution for parameter w
- e.g., w follows a multivariate normal distribution $N(0, \tau^2 \mathbb{I})$



- its probability density function is

$$p(w) \sim \frac{1}{(2\pi\tau^2)^{d/2}} \cdot \exp\left(-\frac{1}{2\tau^2} \|w\|_2^2\right)$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum A Posteriori (MAP)

- MAP: maximize probability of parameter w given the observations ($p(w \mid X, y)$), i.e., choose w that maximizes the posterior probability

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

- $$\operatorname{argmax}_w p(w \mid X, y) = \operatorname{argmax}_w p(y \mid X, w) \cdot p(w)$$

- $$\operatorname{argmax}_w \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma}\right)^2\right) \quad \cdot \quad \frac{1}{(2\pi\tau^2)^{d/2}} \exp\left(-\frac{1}{2\tau^2}\|w\|_2^2\right)$$

- this leads to

$$\min_w \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w\right)^2 + \lambda \|w\|_2^2$$

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Maximum A Posteriori (MAP)

- or maximize the logarithm

$$\operatorname{argmax}_w \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y^{(i)} - (x^{(i)})^\top w}{\sigma} \right)^2 \right) \cdot \frac{1}{(2\pi\tau^2)^{d/2}} \exp\left(-\frac{1}{2\tau^2} \|w\|_2^2 \right) \right)$$

- simplifying yields

$$\operatorname{argmax}_w -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w \right)^2 - \frac{1}{2\tau^2} \|w\|_2^2$$

- or equivalently

$$\operatorname{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w \right)^2 + \frac{1}{2\tau^2} \|w\|_2^2$$

Maximum A Posteriori (MAP)

- $$\mathrm{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w \right)^2 + \frac{1}{2\tau^2} \|w\|_2^2$$
- or equivalently
$$\mathrm{argmin}_w \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w \right)^2 + \frac{2\sigma^2}{2\tau^2} \|w\|_2^2$$
- or equivalently
$$\mathrm{argmin}_w \sum_{i=1}^{n} \left(y^{(i)} - (x^{(i)})^\top w \right)^2 + \lambda \|w\|_2^2$$
- this is regularized least squares regression (or more generally, regularized risk minimization)

Equivalence Regularized Risk Minimization and Maximum A Posteriori

- **MAP is equivalent to regularized risk minimization**
- **MLE is equivalent to empirical risk minimization**
- noise corresponds to the loss function
- prior distribution corresponds to regularizer
- variance parameters (σ^2, τ^2) correspond to regularization parameter λ
- two different views for the same problem
- we can now also answer the question when to use MLE or MAP
- different noise distributions and different priors for the parameter give rise to different ML models

Regression Models

- different noise distributions and different priors for the parameter give rise to different ML models
- let's look at different regression models based on different loss functions and different regularizer

Ridge Regression

$$\min_{w} \frac{1}{2n} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$$

- least squares regression with $\|.\|_2$-regularizer is also called **ridge regression**
- can be solved by solving a system of linear equations
- or by gradient descent

LASSO

- consider least squares regression with many features (either the data has many features of we use many basis functions)
- example: gene expression data $X \in \mathbb{R}^{n \times d}$, i.e., n data points each having d features
- each data point is the gene expression of d genes of one patient
- usually, only data of a few patients available
- but many genes
- n is usually a few hundred, d usually a few thousand
- label y - how severely has a patient developed a specific disease
- goal: find out which genes might have caused the disease
- if solved by least squares regression, how many possible optimal solutions w^* exist?

LASSO

- it is desirable to obtain a regressor w where many coefficients w_i are zero
- it is called **sparse solution**
- allows also for better iterpretability
- what regularizer should we use here?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

LASSO

- ideally, we would like to solve

$$\begin{array}{ll} \min_w & \|Xw - y\|^2 \\ \text{st} & \|w\|_0 \leq t \end{array}$$

- find an optimal solution that explains the data but only picks a t features
- this is an NP-hard problem

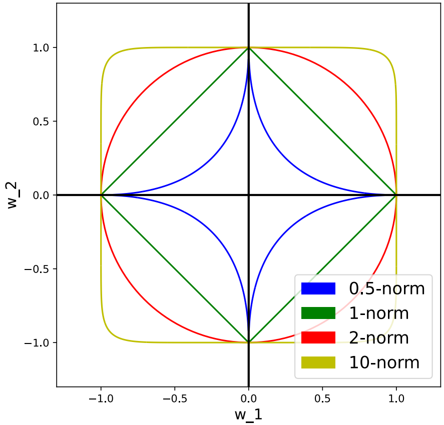
LASSO

- recall the p-norm, for $p > 0$

$$\|w\|_p = \left(\sum_{i=1}^{d} |w_i|^p \right)^{\frac{1}{p}}$$

- it is a norm for $p \geq 1$
- strictly speaking, it is not a norm for $p < 1$

LASSO



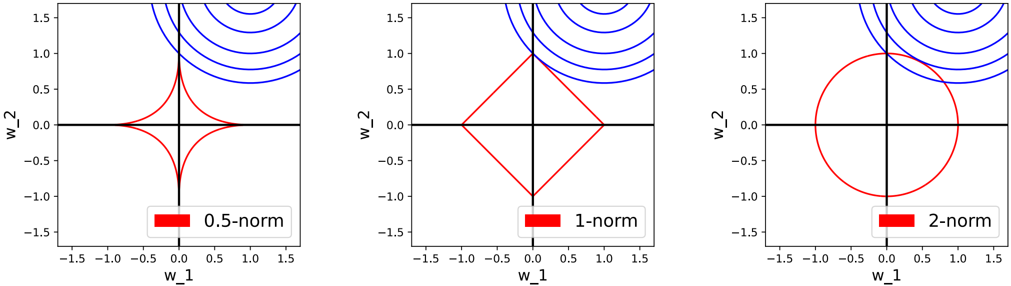
- vectors w with $\|w\|_p = 1$ for different values of p, aka unit balls

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

LASSO

- instead of using $\|.\|_0$, we use $\|.\|_1$
- in some sense, it is the closest convex norm to $\|.\|_0$
- it will produce sparse solutions

LASSO



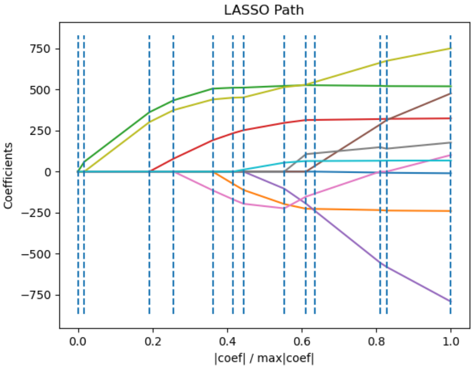
- contour lines of the loss function
- restrict w to have p-norm less than a constant t

LASSO

- hence, we solve the following regularized risk minimization problem
- $$\min_w \frac{1}{2n} \|Xw - y\|_2^2 + \lambda \|w\|_1$$
- this is called the **LASSO** (least absolute shrinkage and selection operator)
- performs feature selection
- can be solved using subgradient method (though more efficient methods exist)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

LASSO



- regularization path for the LASSO

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Elastic Net

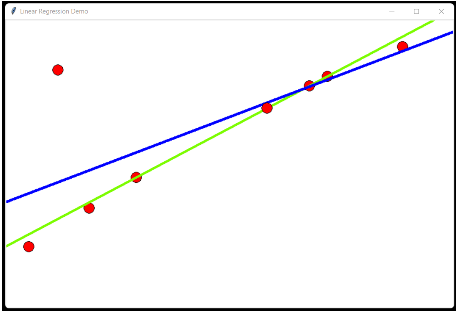
- instead of the LASSO, often the following is used

$$\min_w \frac{1}{2n} \|Xw - y\|_2^2 + \lambda \left(\alpha \|w\|_1 + \frac{1-\alpha}{2} \|w\|_2^2 \right)$$

- this is called the **elastic net**
- interpolates between ridge regression and LASSO
- often used for gene expression data

Robust Regression

- what to do when you have outlier in the data
- or equivalently, noise is not Gaussian distributed
- need a robust loss function that is insensitive to outliers



Robust Regression

- $$\min_w \frac{1}{n} \|Xw - y\|_1$$
- this is called **robust regression**
- can be solved using subgradient method (but more efficient methods exist)
- can also add different regularizer

Regression – Summary

- many different loss functions and many different regularizer exist and can be combined
- all have different characteristics and different applications scenarios
- can be combined with different basis functions
- you should be able to model almost any regression task and also solve it

Feature Scaling

- all features should be roughly on the same scale
- some methods are invariant to feature scaling, some are not
- regularizer are usually **not** invariant to feature scaling
- so always scale your features to be on the safe side
- also allows for better interpretability
- scale them such that they are all between $[0, 1]$ or $[-1, 1]$
- or normalize the data
- **remember**: use the same scaling method and scaling parameters also for the test data!

Feature Scaling

- normalizing data: let $X_{i,j} = (x_j^{(i)})$ be your data matrix
- first centering: center each feature, i.e., subtract the column mean from each column

$$X_{:,j}^{\text{centered}} = X_{:,j} - \bar{x}_j, \qquad \text{where } \bar{x}_j = \frac{1}{n}\sum_{i=0}^{n} X_{i,j}$$

- scale each column such that each column has 2-norm one

$$X_{:,j}^{\text{scaled}} = \frac{X_{:,j}^{\text{centered}}}{\left\| X_{:,j}^{\text{centered}} \right\|_2}$$