

MACHINE LEARNING IN BIOINFORMATICS

MODEL SELECTION AND REGULARIZATION

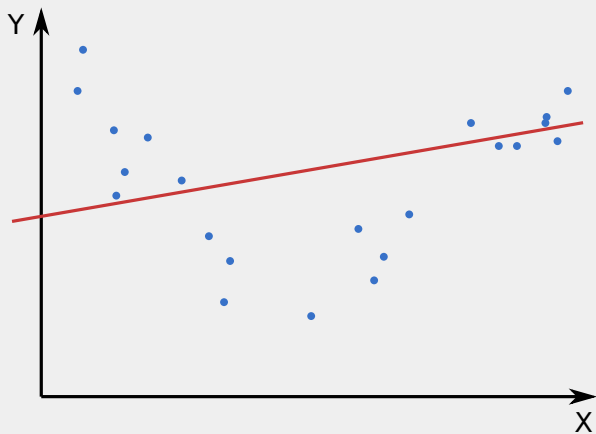
Philipp Benner
philipp.benner@bam.de

VP.1 - eScience
Federal Institute of Materials Research and Testing (BAM)

February 8, 2026

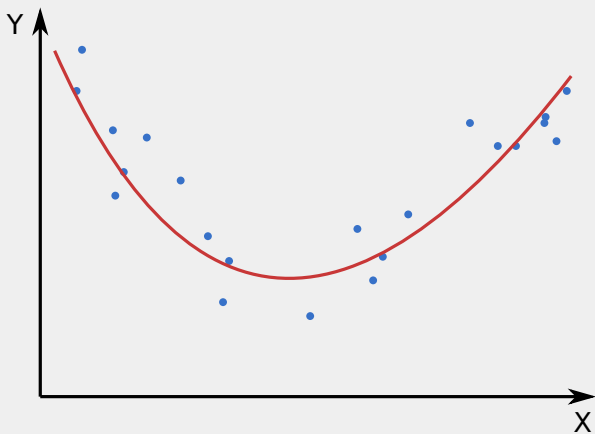
MODEL SELECTION PROBLEM

MODEL SELECTION



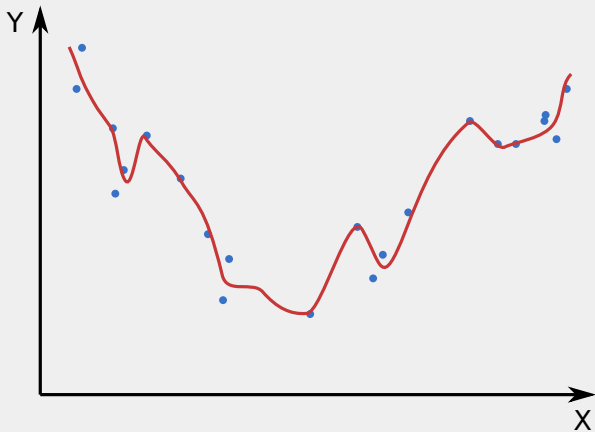
Linear model class

MODEL SELECTION



Quadratic model class

MODEL SELECTION



Polynomial model class

BIAS-VARIANCE DECOMPOSITION AND TRADEOFF

BIAS-VARIANCE DECOMPOSITION

- Let \mathbf{Y} , \mathbf{X} and ϵ be random variables such that $\mathbf{Y} = f(\mathbf{X}) + \epsilon$, with $\mathbb{E}[\epsilon] = 0$ and $\text{var}[\epsilon] = \sigma^2$
- Assume that \hat{f}_D has been estimated on some training data $D = (X, y)$, where X is a matrix of n observations from \mathbf{X} and y a vector of n observations from \mathbf{Y}
- Let $\bar{f}(x) = \mathbb{E}_D[\hat{f}_D(x)]$, then at a query point x we have

$$\mathbb{E}_{\mathbf{Y}, D}[(\mathbf{Y} - \hat{f}_D(x))^2] = \underbrace{[\bar{f}(x) - f(x)]^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}_D[\bar{f}(x) - \hat{f}_D(x)]^2}_{\text{Variance}} + \sigma^2$$

- bias: Is there a bias towards a particular kind of solution (e.g. linear model)? (inductive bias)
- variance: How much does the estimated model change if you train on a different data set? (overfitting)

BIAS-VARIANCE DECOMPOSITION: DERIVATION I

- We study the mean squared prediction error at a fixed x :

$$\mathbb{E}_{\mathbf{Y}, D} [(\mathbf{Y} - \hat{f}_D(x))^2]$$

where $\mathbf{Y} = f(x) + \epsilon$, $\mathbb{E}[\epsilon] = 0$, $\text{Var}(\epsilon) = \sigma^2$.

- **Step 1 – Substitute:**

$$\mathbb{E}_{\epsilon, D} [(f(x) + \epsilon - \hat{f}_D(x))^2]$$

- **Step 2 – Add and subtract the average prediction:**

$$f(x) + \epsilon - \hat{f}_D(x) = \underbrace{(f(x) - \bar{f}(x))}_{\text{bias term}} + \underbrace{(\bar{f}(x) - \hat{f}_D(x))}_{\text{variance term}} + \underbrace{\epsilon}_{\text{noise term}}$$

- **Step 3 – Square and take expectations:** Cross terms vanish because:

BIAS-VARIANCE DECOMPOSITION: DERIVATION II

- ▶ ϵ has mean 0 and is independent of $\hat{f}_D(x)$
- ▶ $\bar{f}(x) - \hat{f}_D(x)$ has mean 0 by definition

■ Result:

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}, D}[(\mathbf{Y} - \hat{f}_D(x))^2] &= \underbrace{[\bar{f}(x) - f(x)]^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_D[(\bar{f}(x) - \hat{f}_D(x))^2]}_{\text{variance}} + \underbrace{\sigma^2}_{\text{noise}} \\ &\quad + \cancel{\mathbb{E}_D[2(f(x) - \bar{f}(x))(\bar{f}(x) - \hat{f}_D(x))]} \\ &\quad + \cancel{[2(f(x) - \bar{f}(x))\epsilon]} \\ &\quad + \cancel{\mathbb{E}_D[2(\bar{f}(x) - \hat{f}_D(x))\epsilon]} \\ &= (\text{bias})^2 + \text{variance} + \text{noise}.\end{aligned}$$

ESTIMATION ERROR VS. PREDICTION ERROR

■ Estimation Error:

- ▶ Goal: Estimate an unknown fixed quantity (e.g., parameter θ) based on data
- ▶ Decomposition of expected error:

$$\mathbb{E}_X \left[\|\hat{\theta}(X) - \theta\|^2 \right] = \text{Bias}^2 + \text{Variance}$$

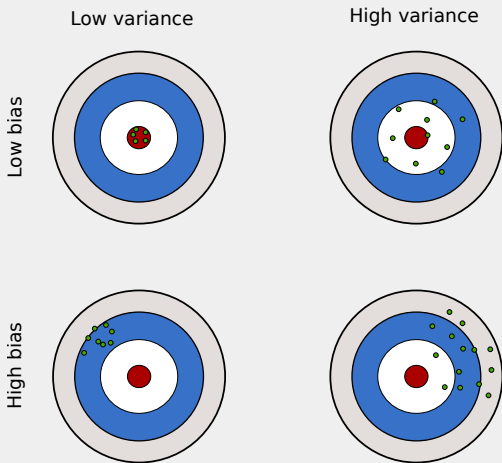
■ Prediction Error:

- ▶ Goal: Predict a future or unseen outcome Y from input X , based on a learned model \hat{f}
- ▶ Decomposition of expected prediction error:

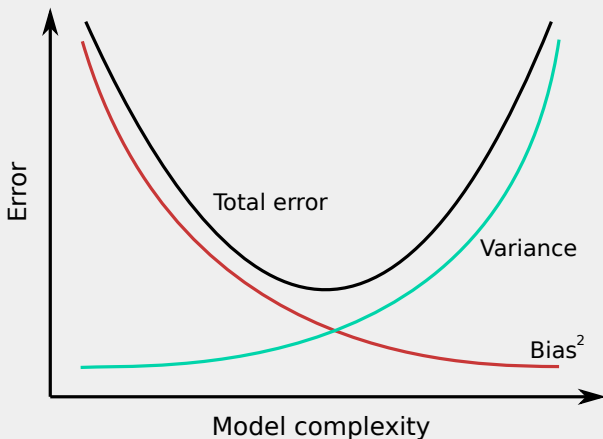
$$\mathbb{E}_{\mathbf{Y}, D} \left[(\mathbf{Y} - \hat{f}_D(X))^2 \right] = \text{Bias}^2 + \text{Variance} + \sigma^2$$

- ▶ σ^2 : Irreducible noise from inherent randomness in the data

BIAS-VARIANCE DECOMPOSITION

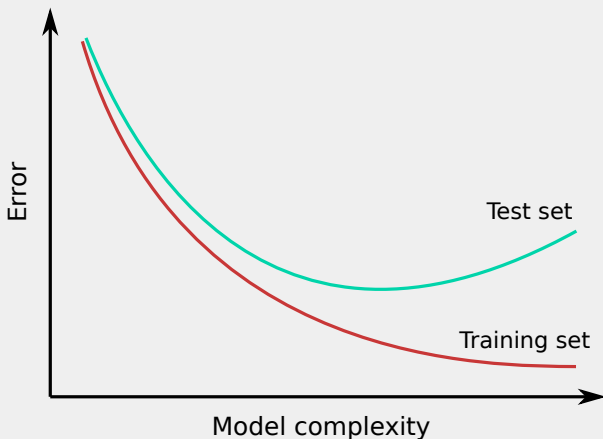


BIAS-VARIANCE DECOMPOSITION



^oNote that here we average over multiple data sets. On a single data set we might observe bumps when increasing model complexity

BIAS-VARIANCE DECOMPOSITION



^oNote that here we average over multiple data sets. On a single data set we might observe bumps when increasing model complexity

BIAS-VARIANCE DECOMPOSITION - LESSONS LEARNED

- Every model comes with a bias
- More complex models have a smaller bias but larger variance
- A bias is required to reduce the variance, but introducing a good bias requires domain knowledge
- Classical statistics often uses unbiased estimators, which is nowadays often questioned
- Keep in mind: There is no free lunch!¹

¹The *no free lunch theorem* [Wolpert and Macready, 1997] tells us that there exists no generic model that works well on all domains, but we need to tailor our models to the data at hand in order to introduce a model bias, which reduces variance.

JAMES STEIN ESTIMATOR

THE ESTIMATION PROBLEM

We observe a random vector:

$$X = (\theta_1 + \varepsilon_1, \dots, \theta_p + \varepsilon_p), \quad \varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2)$$

- Each component X_i is a single noisy observation of θ_i
- Only **one observation per dimension** - no replication
- Goal: Estimate the unknown mean vector $\theta = (\theta_1, \dots, \theta_p)$

THE ESTIMATION PROBLEM

- Observations: $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$, independently for $i = 1, \dots, p$
- The likelihood of θ given X is:

$$L(\theta; X) \propto \exp\left(-\frac{1}{2\sigma^2}\|X - \theta\|^2\right)$$

- Maximizing the likelihood \Rightarrow *Maximum Likelihood Estimator (MLE)*:

$$\hat{\theta}_{\text{MLE}} = X$$

Evaluating Estimators: Squared Error Loss

To quantify how well an estimator performs, we use the loss function:

$$L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$$

and define its **risk** (expected loss):

$$R(\hat{\theta}, \theta) = \mathbb{E}_{\theta} \left[\|\hat{\theta}(X) - \theta\|^2 \right]$$

- Under squared error loss, the risk can be decomposed into:

$$\text{Risk} = \|\text{Bias}(\hat{\theta})\|^2 + \text{Var}(\hat{\theta})$$

- This bias-variance decomposition helps us understand tradeoffs in estimator performance.

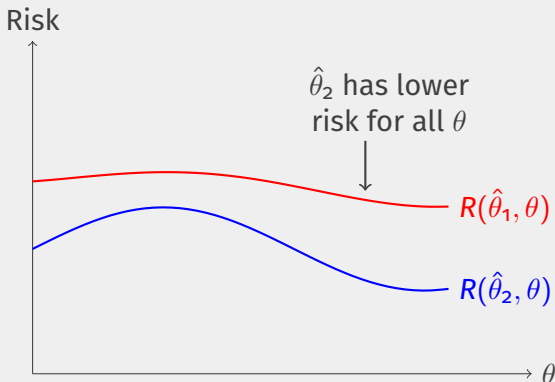
Risk and Admissibility

- The **risk** of an estimator is its expected loss:

$$R(\hat{\theta}, \theta) = \mathbb{E}_{X \sim P_{\theta}} \left[\|\hat{\theta}(X) - \theta\|^2 \right]$$

- An estimator is **admissible** if no other estimator has uniformly lower risk (i.e., lower for all θ , strictly lower for some).

EVALUATING ESTIMATORS



- An estimator that is tuned to perform best at a specific parameter value can have very low risk at that point, but much higher risk elsewhere.

SURPRISING RESULT

James-Stein Theorem [James et al., 1961]

For $p \geq 3$, the MLE $\hat{\theta}_{\text{MLE}} = X$ is **inadmissible** under squared error loss.

- There exists an estimator with strictly lower risk than the MLE for all θ
- This result shocked statisticians: the unbiased estimator is not optimal!

The James-Stein estimator is defined as:

$$\hat{\theta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right) X$$

- Shrinks X toward the origin
- Introduces bias, but reduces variance
- Dominates MLE in risk for $p \geq 3$

POSITIVE-PART JAMES-STEIN ESTIMATOR

To avoid over-shrinking:

$$\hat{\theta}_{JS}^+ = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right)_+ X$$

where $(a)_+ = \max(a, 0)$

- Further reduces risk
- Dominates the standard James-Stein estimator
- Still inadmissible, but better performance

KEY TAKEAWAYS I

- We observe one noisy measurement per parameter:
 $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$
- As dimension p increases, the number of parameters increases — but the information per parameter stays the same
- The MLE is unbiased but has high variance in high dimensions
- Shrinkage estimators (like James–Stein) reduce variance at the cost of bias, improving overall estimation
- Bayes and empirical Bayes estimators can be admissible and offer further improvement

KEY TAKEAWAYS II

Conclusion

Without regularization, estimating many parameters from few observations leads to poor performance.

Historical Impact [Efron and Morris, 1977]

Before 1961, the maximum likelihood estimator (MLE) was widely believed to be the best estimator one could hope for, especially if it was unbiased. James and Stein's theorem showed that for $p \geq 3$, the MLE is **inadmissible** under squared error loss: there exists another estimator with strictly lower risk for all θ . This surprising result revolutionized statistical thinking and inspired the development of shrinkage methods, empirical Bayes, and modern regularization techniques.

WHICH ESTIMATOR IS ADMISSIBLE? I

- The MLE, James–Stein, and positive-part James–Stein estimators are all **inadmissible** (in sufficiently high dimension).
- **Admissible:** No other estimator has uniformly lower risk (lower for all θ , strictly lower for some).

WHICH ESTIMATOR IS ADMISSIBLE? II

Admissibility of Bayes Estimator

[Berger, 2013, Lehmann and Casella, 1998]

Let $\pi(\theta)$ be any **proper** prior distribution and consider squared error loss:

$$L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2.$$

Then the Bayes estimator

$$\hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\theta \mid X = x]$$

is **admissible**.

- Admissibility means no other estimator dominates it uniformly, but does *not* imply low risk for all θ .

WHICH ESTIMATOR IS ADMISSIBLE? III

- If the prior is concentrated far from the true θ , the estimator may have low risk near the prior mean but high risk elsewhere.
- **Example:** For $\theta \sim \mathcal{N}(\mu_0, \tau^2 I_p)$ and $X | \theta \sim \mathcal{N}(\theta, \sigma^2 I_p)$,

$$\hat{\theta}_{\text{Bayes}} = \mu_0 + \frac{\tau^2}{\tau^2 + \sigma^2} (X - \mu_0)$$

is admissible for any $\tau^2 > 0$. Small τ^2 causes strong shrinkage toward μ_0 .

- The degenerate case $\tau^2 = 0$ is not a proper prior and admissibility is not guaranteed.

COMPLEXITY MEASURES

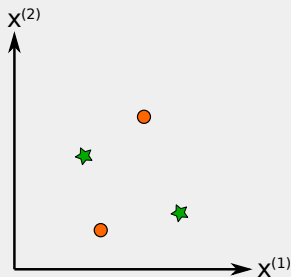
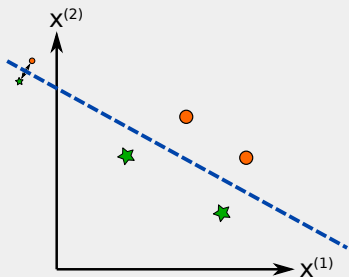
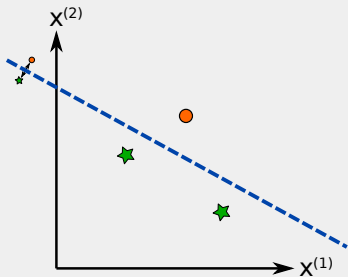
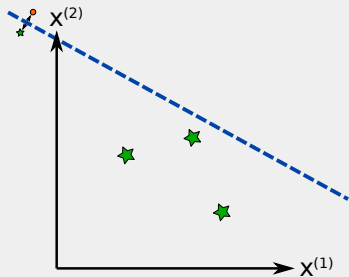
VC-Dimension (Vapnik Chervonenkis)

Let \mathbb{F}_p be a set of classifiers on an n -dimensional input space. The VC-dimension $VC(\mathbb{F}_p)$ is defined as the maximum number of points that can be correctly classified by at least one member of \mathbb{F}_p .

■ Examples:

- ▶ Linear classifier on \mathbb{R}^p : $VC = p + 1$
- ▶ SVM with RBF kernel: $VC = \infty$
- ▶ Neural network with n_e edges, n_v nodes and sigmoid activation function: $\Omega(n_e^2) < VC < \mathcal{O}(n_e^2 n_v^2)$
[Shalev-Shwartz and Ben-David, 2014, Section 20.4]

COMPLEXITY OF CLASSIFIERS - VC DIMENSION



Degrees of Freedom (DF) [Efron, 1986]

The **degrees of freedom** of an estimate $\hat{y} = \hat{f}(X)$ is defined as

$$\text{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(\hat{y}_i, y_i) = \frac{1}{\sigma^2} \text{tr cov}(\hat{y}, y),$$

where

- X denotes a fixed set of n covariates of dimension p
- $y = (y_1, \dots, y_n)$ is a vector of n observations from

$$\mathbf{Y} = f(X) + \epsilon$$

for some function f , assuming $\mathbb{E}[\epsilon] = 0$ and $\text{var}[\epsilon] = \sigma^2$

¹df is normalized by the magnitude of the aleatory uncertainty (σ^2)

MEASURES OF MODEL COMPLEXITY - DF

- Degrees of freedom for the OLS estimate:

$$\begin{aligned}\text{df}(\hat{y}) &= \frac{1}{\sigma^2} \text{tr} \text{cov}(\hat{y}, y) \\ &= \frac{1}{\sigma^2} \text{tr} \text{cov} \left(X(X^\top X)^{-1} X^\top y, y \right) \\ &= \frac{1}{\sigma^2} \text{tr} \left(X(X^\top X)^{-1} X^\top \right) \text{cov}(y, y) \\ &= \text{tr} \left(X(X^\top X)^{-1} X^\top \right) \\ &= p\end{aligned}$$

- $\text{df}(\hat{y}) = p$, i.e. the number of parameters, assuming independent feature vectors (i.e. columns of X)
- This result holds for $p < n$

$X(X^\top X)^{-1} X^\top$ is the hat matrix $H \in \mathbb{R}^{n \times n}$, hence $\text{df}(\hat{y}) = \text{rank}(H)$

- Ridge regression is defined as

$$\hat{\theta} = \arg \min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2$$

for some regularization strength $\lambda \geq 0$

- The ridge estimator has

$$\text{df}(\hat{y}) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$$

degrees of freedom, where $(d_j)_j$ are the singular values of X

- Increasing λ decreases model complexity

- There is some criticism about used DF as measure of model complexity [Janson et al., 2015]
- In some cases, we also need X to be random [Luan et al., 2021]
- We will see other measures when turning to model selection

MODEL SELECTION

MODEL SELECTION APPROACHES

- A measure of accuracy or fit, such as the mean squared error (MSE), is not enough: Increasing model complexity will always lead to a better fit
- Estimating a model requires to minimize both
 - ▶ **in-sample-error** (loss on training data), and
 - ▶ **out-of-sample-error** (generalization error)
- Cross-validation (CV) estimates generalization error on left-out samples²
- Traditional statistics: Combine measure of accuracy (in-sample-error) with a penalty for complexity

²Heavy hyperparameter tuning using CV can lead to overfitting and requires to select a final holdout set

MODEL SELECTION APPROACHES - LOO-CV

- Leave-one-out Cross-Validation (LOO-CV) at iteration $i = 1, 2, \dots, n$:
 - ▶ Compute estimate on data set without the i -th sample
 - ▶ Compute prediction error on the i -th sample
- Report the average prediction error over all n samples
- PRESS statistic (predicted residual error sum of squares):

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{-i})^2$$

where \hat{y}_{-i} is the prediction for the i -th sample where the model has been estimated on all but the i -th sample

MODEL SELECTION APPROACHES - PRESS

- LOO-CV is very costly for large data sets and complex models
- k -fold CV with $k = 5$ or $k = 10$ is often used in practice
- For (ridge) linear regression with mean squared error we can efficiently compute LOO-CV [Cook, 1977]

$$\begin{aligned}\text{PRESS} &= \sum_{i=1}^n (y_i - \hat{y}_{-i})^2 \\ &= \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{(1 - H_{ii})^2}\end{aligned}$$

- The matrix

$$H = X(X^T X + \lambda I)^{-1} X^T$$

is called the hat matrix, because it puts a hat on y , i.e. $\hat{y} = Hy$

MODEL SELECTION APPROACHES - GCV

- Models like smoothing splines or kernel regressions still have a hat matrix, but it is costly or undefined in closed form.
- H may be dense, very large, or not explicitly known.
- Generalized Cross-Validation (GCV) uses the approximation

$$H_{ii} \approx \text{tr}(H)/n$$

- The GCV therefore is defined as

$$\text{GCV} = \frac{\|y - \hat{y}\|^2}{\left(1 - \frac{\text{tr}(H)}{n}\right)^2}$$

- GCV only requires $\text{tr}(H)$, which can often be approximated (e.g. using randomized trace estimation).

MODEL SELECTION APPROACHES

- LOO-CV is computationally very expensive
- k -fold CV is cheaper, but uses a large fraction of the data for testing
- Model performance could be better if this data was used for training
- Overfitting if we use CV for testing too many models (requires final hold out data)
- Can we do model selection by using all data for training?

MODEL SELECTION APPROACHES - DF

- Assume again the following model

$$\mathbf{Y} = f(X) + \epsilon$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed set of n predictors and $\mathbf{Y} \in \mathbb{R}^n$

- Setup is very similar to the bias-variance decomposition, but X is now fixed
- Let $\mathbf{Y}_t \in \mathbb{R}^n$ a vector of n independent observations and $\hat{f}_{\mathbf{Y}_t}$ an estimate on the training set (X, \mathbf{Y}_t) , then [Efron, 1986]

$$\underbrace{\mathbb{E}_{\mathbf{Y}, \mathbf{Y}_t} \left\| \mathbf{Y} - \hat{f}_{\mathbf{Y}_t}(X) \right\|_2^2}_{\text{expected prediction error}} = \underbrace{\mathbb{E}_{\mathbf{Y}_t} \left\| \mathbf{Y}_t - \hat{f}_{\mathbf{Y}_t}(X) \right\|_2^2}_{\text{expected training error}} + 2\sigma^2 \text{df}(\hat{f})$$

- This motivates the following model selection criterium [Mallows, 2000]

$$\underbrace{\left\| y_t - \hat{f}_{y_t}(X) \right\|_2^2}_{\text{training error}} + \underbrace{2\sigma^2 \text{df}(\hat{f})}_{\text{complexity penalty}}$$

- The more complex a model, the larger the penalty
- If two models fit the data equally well, we select the simpler one (Occam's razor)

MODEL SELECTION APPROACHES - BAYES APPROACH

- Assume we have a set of models $(m_i)_i$
- In a probabilistic setting we evaluate the probability of a model m_i given data x , i.e. using Bayes theorem

$$\text{pr}(m_i | x) = \frac{\text{pr}(x | m_i)\text{pr}(m_i)}{\sum_j \text{pr}(x | m_j)\text{pr}(m_j)} = \frac{\text{pr}(x | m_i)\text{pr}(m_i)}{\text{pr}(x)}$$

- We compare two models m_i and m_j using

$$\frac{\text{pr}(m_i | x)}{\text{pr}(m_j | x)} = \frac{\frac{\text{pr}(x | m_i)\text{pr}(m_i)}{\text{pr}(x)}}{\frac{\text{pr}(x | m_j)\text{pr}(m_j)}{\text{pr}(x)}} = \frac{\text{pr}(x | m_i)\text{pr}(m_i)}{\text{pr}(x | m_j)\text{pr}(m_j)}$$

because $\text{pr}(x)$ drops

MODEL SELECTION APPROACHES - BAYES FACTOR

- With a uniform prior over models we arrive at the Bayes factor [Kass and Raftery, 1995]

$$\frac{\text{pr}(x | m_i)}{\text{pr}(x | m_j)}$$

- Hence, in Bayesian model selection, we evaluate a model m based on its *marginal likelihood*

$$\text{pr}(x | m) = \int_{\theta} \text{pr}(x | \theta, m) \text{pr}(\theta | m) d\theta$$

where θ are the model parameters

- The marginal likelihood is often difficult to evaluate, even numerically!

MODEL SELECTION APPROACHES - BIC

- The marginal likelihood is tractable only for very simple models
- As an alternative, we use approximations of the marginal likelihood
- The **Bayes information criterion (BIC)** is such an approximation. Let x contain n samples and assume that $n \gg p$, then

$$\text{pr}(x | m) \approx \exp \left\{ -\frac{1}{2} \text{BIC}(x; m) \right\}$$
$$\text{BIC}(x; m) = -2 \log \text{pr}(x | \hat{\theta}, m) + p \log(n)$$

where $\hat{\theta}$ refers to the maximum likelihood estimate and p to the number of parameters

MODEL SELECTION APPROACHES - BIC

- Let \mathbf{Y} and ϵ be two random variables such that $\mathbf{Y} = f(X) + \epsilon$
- Let $f_{\hat{\theta}}$ denote a maximum likelihood estimate on some training data
- For $\epsilon \sim \text{Normal}(0, \sigma^2)$ the BIC is related to the mean squared error with complexity penalty

$$\begin{aligned}\text{BIC}(\mathbf{x}; m) &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - f_{\hat{\theta}}(x_i))^2 + p \log(n) + C_n \\ &\propto \frac{1}{\sigma^2} \|\mathbf{y} - f_{\hat{\theta}}(\mathbf{x})\|_2^2 + p \log(n)\end{aligned}$$

where C_n is a constant depending on n , which can be dropped for model comparison

MODEL SELECTION APPROACHES - FIC

- BIC assumes $n \gg p$ and therefore depends only on the number of parameters
- Fisher Information Approximation (FIA) [Ly et al., 2017]:

$$\begin{aligned}\text{pr}(\mathbf{x} \mid m) &\approx \exp \{-\text{FIA}(\mathbf{x}; m)\} \\ \text{FIA}(\mathbf{x}; m) &= \underbrace{-\log \text{pr}(\mathbf{x} \mid \hat{\theta}, m) + \frac{p}{2} \log \left(\frac{n}{2\pi} \right) + \log C_m}_{\text{BIC like term}} \\ C_m &= \underbrace{\int_{\theta} \sqrt{\det \mathcal{I}_m(\theta)} d\theta}_{\text{Geometric complexity}}\end{aligned}$$

where \mathcal{I}_m denotes the *Fisher information matrix*

- C_m is essential if $n \gg p$ is not given [Cheema and Sugiyama, 2020]

HOW DO WE CONTROL MODEL COMPLEXITY?

■ Regularization (e.g. ridge regression):

- ▶ Constrain the feasible set of parameter values
- ▶ Keep the number of parameters in the model constant, but allow them to become zero

■ Number of parameters:

- ▶ A good approximation of model complexity if $n > p$
- ▶ For $n < p$ we saw that the optimization problem has many solutions
 - In deep neural networks, the gradient descent method can act similar to a regularizer
 - Model complexity can decrease when adding more parameters (double descent)

REGULARIZATION

l_k -PENALIZED REGRESSION

Objective function

$\omega(\theta) = -\log \text{pr}_\theta(\mathbf{y})$ (maximum likelihood), or

$\omega(\theta) = \|\mathbf{y} - \mathbf{X}\theta\|_2^2$ (linear regression)

Regularized estimate with ℓ_k -norm penalty

$$\hat{\theta} = \begin{cases} \arg \min_{\theta} & \omega(\theta) \\ \text{subject to} & \|\theta\|_k^k = \Lambda \end{cases}$$

where

$$\|\theta\|_k = \left(\sum_{j=2}^p |\theta_j|^k \right)^{1/k}$$

²Remember that we do not regularize the bias or y-intercept θ_0

l_k -PENALIZED REGRESSION

Identify saddle points of Lagrangian

$$\mathcal{L}(\theta, \lambda) = \omega(\theta) + \lambda(\|\theta\|_k^k - \Lambda)$$

In practice, we do not work with Λ , but set λ such that the classification performance is optimal, i.e. we work with the Lagrangian

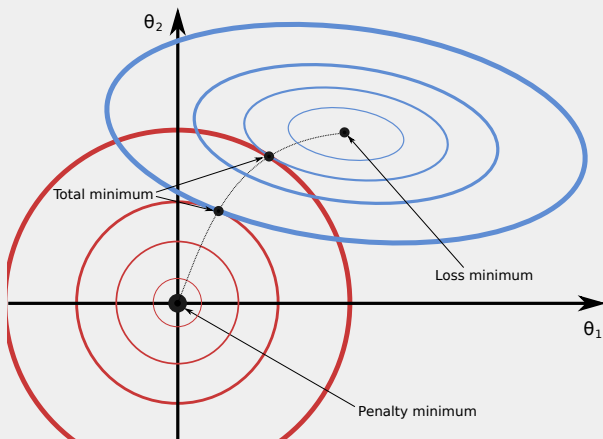
$$\hat{\theta}(\lambda) = \arg \min_{\theta} \omega(\theta) + \lambda \|\theta\|_k^k$$

At the optimum we must have

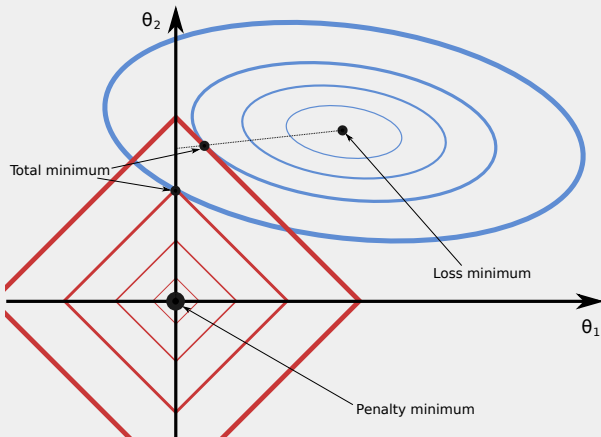
$$\nabla_{\theta} \omega(\theta) + \lambda \nabla_{\theta} \|\theta\|_k^k = \mathbf{0}$$

i.e. the gradients of $\omega(\theta)$ and $\lambda \|\theta\|_k^k$ must point to opposite directions

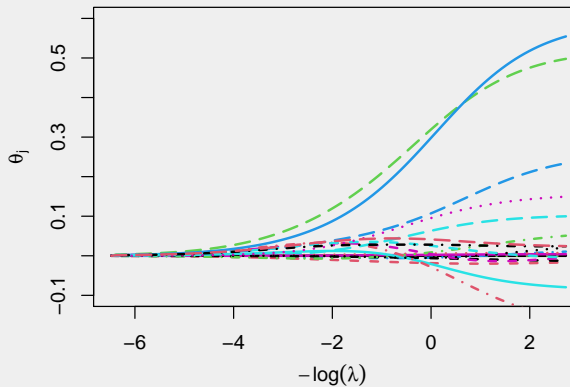
REGULARIZATION - K=2



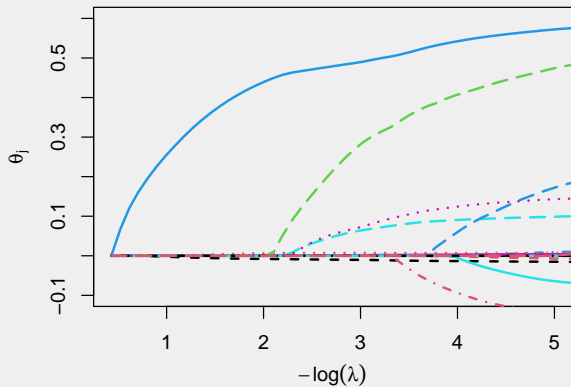
REGULARIZATION - K=1



REGULARIZATION PATHS - $K=2$

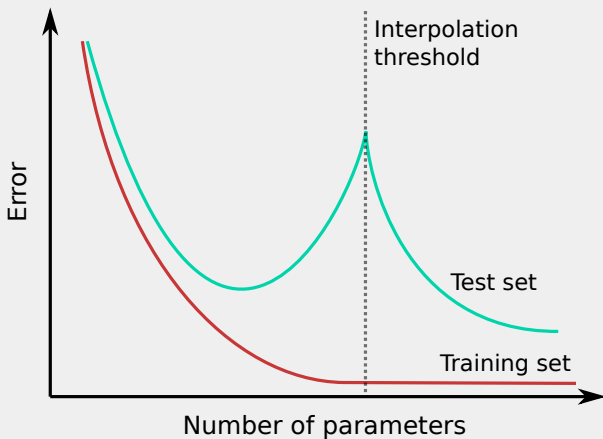


REGULARIZATION PATHS - $K=1$

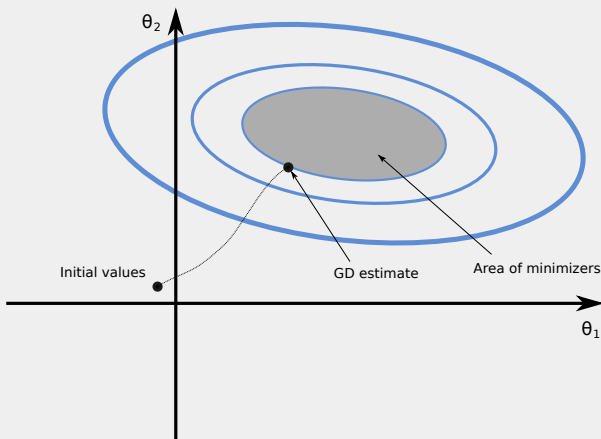


IMPLICIT REGULARIZATION AND DOUBLE DESCENT

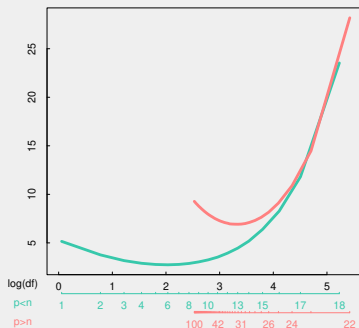
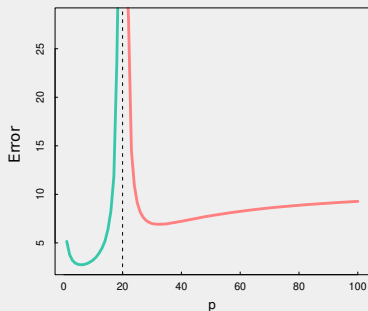
IMPLICIT REGULARIZATION - DOUBLE DESCENT



IMPLICIT REGULARIZATION - DOUBLE DESCENT



MINIMUM ℓ_2 -NORM ESTIMATE - DF



²Requires a more advanced definition of DF that treats X as random variable [Luan et al., 2021]

IMPLICIT REGULARIZATION

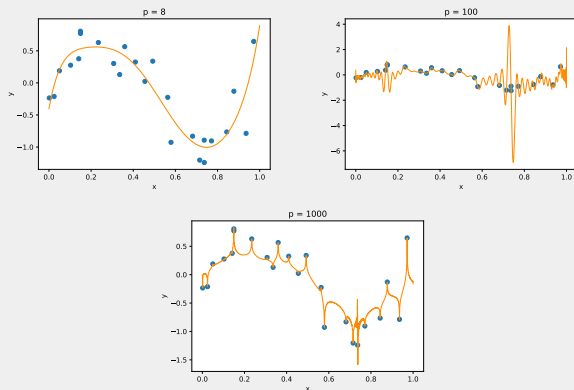


Figure: Fitting degree $d = p - 1$ Legendre polynomials. For $p > n$ the solution with the smallest ℓ_2 -norm is used.

²Legendre polynomials are quite useful, since their absolute value is bounded by one.

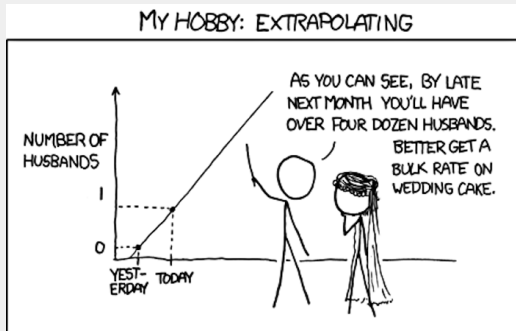
TAKE HOME MESSAGES

- Expected performance is the sum of training performance and model complexity
- Complex models require regularization to prevent overfitting
- The number of parameters does not correspond to the complexity of a model
- Increasing the number of features can reduce model complexity if a min- ℓ_2 -norm estimator is used
- If we have complex data and cannot make any assumptions on the generating process, we might be better off with an overparametrized model using regularization (success behind deep learning)

MORE REFERENCES

- Akaike information criterion (AIC)
[Akaike, 1974, Cavanaugh and Neath, 2019]
- Bayesian information criterion (BIC) [Schwarz, 1978]
- Deviance information criterion (DIC)
[Spiegelhalter et al., 2002]
- Fisher Information Approximation (FIA) [Rissanen, 1996, Grünwald, 2007, Cheema and Sugiyama, 2020]
- Degrees of freedom (DF)
[Tibshirani, 2015, Gao and Jojic, 2016, Luan et al., 2021]
- Implicit regularization and double descent
[Hastie et al., 2022, Luan et al., 2021, Derezhinski et al., 2020, Kobak et al., 2020]





OVERFITTING








- Sections 3.4, 7.3, 7.6, 7.7 and 7.9 [Hastie et al., 2009]

"All models are wrong, but some are useful."
[Moody, 1991]





REFERENCES I

-  AKAIKE, H. (1974).
A NEW LOOK AT THE STATISTICAL MODEL IDENTIFICATION.
IEEE transactions on automatic control, 19(6):716–723.
-  BERGER, J. O. (2013).
STATISTICAL DECISION THEORY AND BAYESIAN ANALYSIS.
Springer Science & Business Media.
-  CAVANAUGH, J. E. AND NEATH, A. A. (2019).
THE AKAIKE INFORMATION CRITERION: BACKGROUND, DERIVATION, PROPERTIES, APPLICATION, INTERPRETATION, AND REFINEMENTS.
Wiley Interdisciplinary Reviews: Computational Statistics, 11(3):e1460.
-  CHEEMA, P. AND SUGIYAMA, M. (2020).
DOUBLE DESCENT RISK AND VOLUME SATURATION EFFECTS: A GEOMETRIC PERSPECTIVE.
arXiv preprint arXiv:2006.04366.






REFERENCES II

-  COOK, R. D. (1977).
DETECTION OF INFLUENTIAL OBSERVATION IN LINEAR REGRESSION.
Technometrics, 19(1):15–18.
-  DEREZINSKI, M., LIANG, F. T., AND MAHONEY, M. W. (2020).
**EXACT EXPRESSIONS FOR DOUBLE DESCENT AND IMPLICIT
REGULARIZATION VIA SURROGATE RANDOM DESIGN.**
Advances in neural information processing systems, 33:5152–5164.
-  EFRON, B. (1986).
HOW BIASED IS THE APPARENT ERROR RATE OF A PREDICTION RULE?
Journal of the American statistical Association, 81(394):461–470.
-  EFRON, B. AND MORRIS, C. (1977).
STEIN'S PARADOX IN STATISTICS.
Scientific American, 236(5):119–127.
-  GAO, T. AND JOJIC, V. (2016).
DEGREES OF FREEDOM IN DEEP NEURAL NETWORKS.
arXiv preprint arXiv:1603.09260.

REFERENCES III

-  GRÜNWALD, P. D. (2007).
THE MINIMUM DESCRIPTION LENGTH PRINCIPLE.
MIT press.
-  HASTIE, T., MONTANARI, A., ROSSET, S., AND TIBSHIRANI, R. J. (2022).
SURPRISES IN HIGH-DIMENSIONAL RIDGELESS LEAST SQUARES INTERPOLATION.
The Annals of Statistics, 50(2):949–986.
-  HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. (2009).
THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION.
Springer Science & Business Media.
-  JAMES, W., STEIN, C., ET AL. (1961).
ESTIMATION WITH QUADRATIC LOSS.
In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379. University of California Press.

REFERENCES IV

-  JANSON, L., FITHIAN, W., AND HASTIE, T. J. (2015).
EFFECTIVE DEGREES OF FREEDOM: A FLAWED METAPHOR.
Biometrika, 102(2):479–485.
-  KASS, R. E. AND RAFTERY, A. E. (1995).
BAYES FACTORS.
Journal of the american statistical association, 90(430):773–795.
-  KOBAK, D., LOMOND, J., AND SANCHEZ, B. (2020).
THE OPTIMAL RIDGE PENALTY FOR REAL-WORLD HIGH-DIMENSIONAL DATA CAN BE ZERO OR NEGATIVE DUE TO THE IMPLICIT RIDGE REGULARIZATION.
J. Mach. Learn. Res., 21:169–1.
-  LEHMANN, E. L. AND CASELLA, G. (1998).
THEORY OF POINT ESTIMATION.
Springer.
-  LUAN, B., LEE, Y., AND ZHU, Y. (2021).
PREDICTIVE MODEL DEGREES OF FREEDOM IN LINEAR REGRESSION.
arXiv preprint arXiv:2106.15682.

REFERENCES V

 LY, A., MARSMAN, M., VERHAGEN, J., GRASMAN, R. P., AND WAGENMAKERS, E.-J. (2017).


A TUTORIAL ON FISHER INFORMATION.

Journal of Mathematical Psychology, 80:40–55.

 MALLOWS, C. L. (2000).


SOME COMMENTS ON CP.

Technometrics, 42(1):87–94.

 MOODY, J. (1991).

**THE EFFECTIVE NUMBER OF PARAMETERS: AN ANALYSIS OF
GENERALIZATION AND REGULARIZATION IN NONLINEAR LEARNING
SYSTEMS.**





Advances in neural information processing systems, 4.

 RISSANEN, J. J. (1996).

FISHER INFORMATION AND STOCHASTIC COMPLEXITY.

IEEE transactions on information theory, 42(1):40–47.

REFERENCES VI

-  SCHWARZ, G. (1978).
ESTIMATING THE DIMENSION OF A MODEL.
The annals of statistics, pages 461–464.
-  SHALEV-SHWARTZ, S. AND BEN-DAVID, S. (2014).
UNDERSTANDING MACHINE LEARNING: FROM THEORY TO ALGORITHMS.
Cambridge university press.
-  SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., AND VAN DER LINDE, A. (2002).
BAYESIAN MEASURES OF MODEL COMPLEXITY AND FIT.
Journal of the royal statistical society: Series b (statistical methodology), 64(4):583–639.
-  TIBSHIRANI, R. J. (2015).
DEGREES OF FREEDOM AND MODEL SEARCH.
Statistica Sinica, pages 1265–1296.

REFERENCES VII



WOLPERT, D. H. AND MACREADY, W. G. (1997).

NO FREE LUNCH THEOREMS FOR OPTIMIZATION.

IEEE transactions on evolutionary computation, 1(1):67–82.