Machine Learning in Bioinformatics

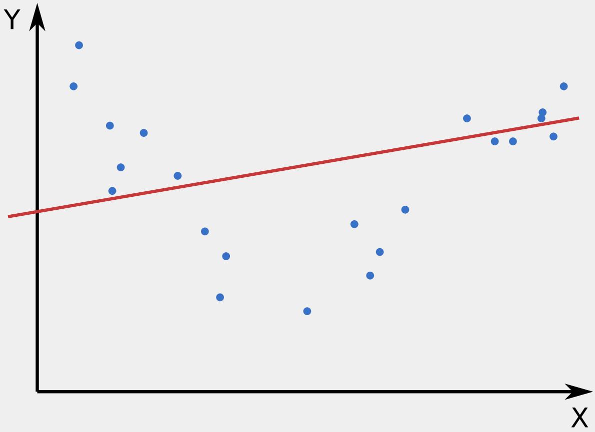
Model Selection and Regularization

Philipp Benner
philipp.benner@bam.de

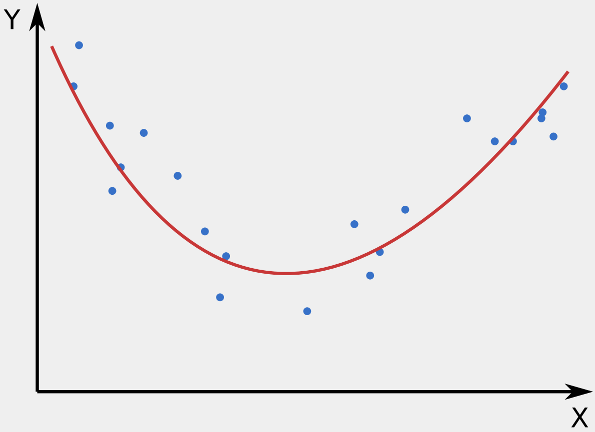
S.3 - eScience
Federal Institute for Materials Research and Testing (BAM)

June 25, 2023

Model selection problem

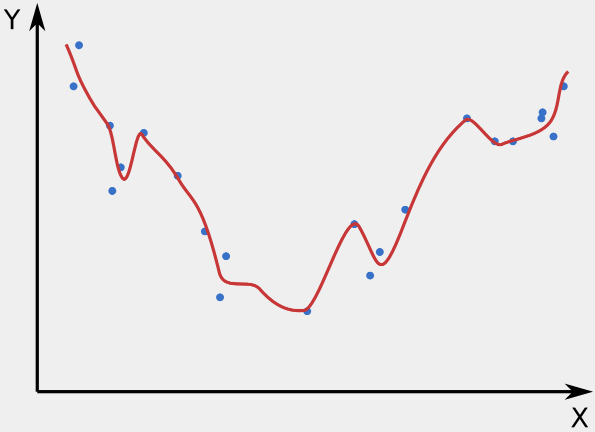


Linear model class



Quadratic model class

Model selection



Polynomial model class

Bias-variance decomposition and tradeoff

Bias-Variance decomposition

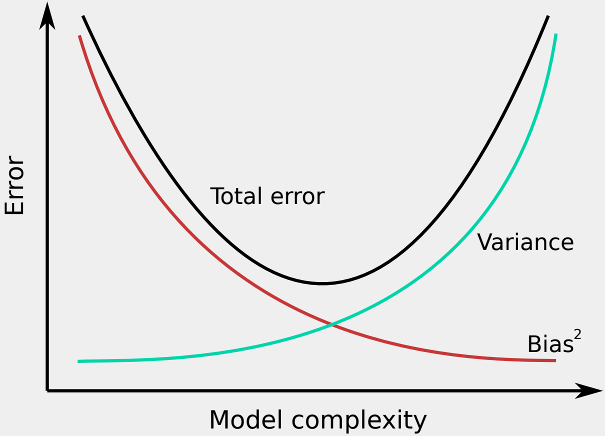
- Let $\mathbf{Y}, \mathbf{X}$ and ϵ be random variables such that $\mathbf{Y} = f(\mathbf{X}) + \epsilon$, with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{var}[\epsilon] = \sigma^2$
- Assume that $\hat{f}_D$ has been estimated on some training data $D = (X, y)$, where X is a matrix of n observations from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

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

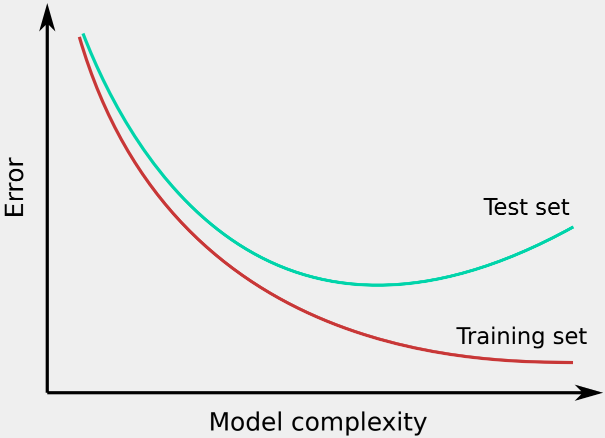
- bias: Is there a bias towards a particular kind of solution (e.g. linear model)? (inductive bias)
- variance: How much does the estimated model change if you train on a different data set? (overfitting)

Bias-Variance decomposition

Bias-Variance decomposition



[0] Note that here we average over multiple data sets. On a single data set we might observe bumps when increasing model complexity



[0] Note that here we average over multiple data sets. On a single data set we might observe bumps when increasing model complexity

Bias-Variance decomposition - Lessions 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]

[1] The *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.

Complexity measures

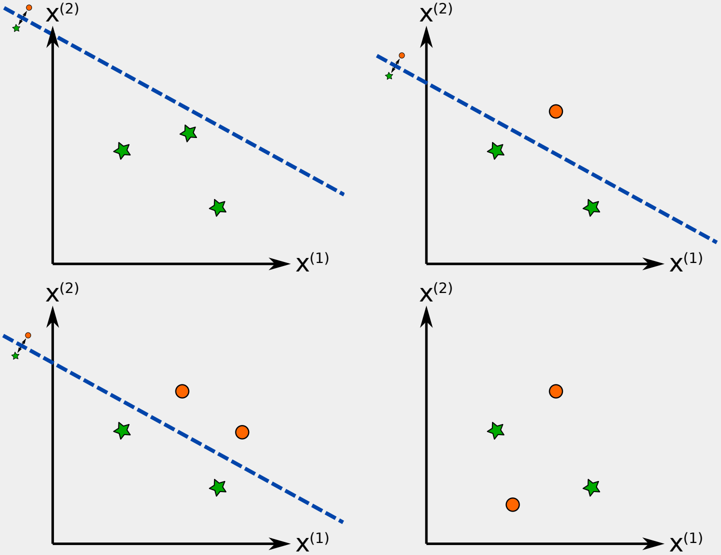
Complexity of classifiers - VC dimension

VC-Dimension (Vapnik Chervonenkis)

Let $\mathbb{F}_p$ be a set of classifiers on an n-dimensional input space. The VC-dimension $\mathrm{VC}(\mathbb{F}_p)$ 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 $\mathbb{R}^p$: $\mathrm{VC} = p + 1$
  - SVM with RBF kernel: $\mathrm{VC} = \infty$
  - Neural network with n_e edges, n_v nodes and sigmoid activation function: $\Omega(n_e^2) < \mathrm{VC} < \mathcal{O}(n_e^2 n_v^2)$ [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

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

where

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

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

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

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

- Degrees of freedom for the OLS estimate:

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

- $\mathrm{df}(\hat{y}) = p$, i.e. the number of parameters, assuming independent feature vectors (i.e. columns of X)

- This result holds for $p < n$

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

- Ridge regression is defined as

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

for some regularization strength $\lambda \geq 0$

- The ridge estimator has

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

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 samples[2]
- Traditional statistics: Combine measure of accuracy (in-sample-error) with a penalty for complexity

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

- Leave-one-out Cross-Validation (LOO-CV) at iteration $i = 1, 2, \ldots, n$:
  - 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):

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

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]

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

- The matrix

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

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

- 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?

- 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]

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

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

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

- 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)

- 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

$$\mathrm{pr}(m_i \mid x) = \frac{\mathrm{pr}(x \mid m_i)\mathrm{pr}(m_i)}{\sum_j \mathrm{pr}(x \mid m_j)\mathrm{pr}(m_j)} = \frac{\mathrm{pr}(x \mid m_i)\mathrm{pr}(m_i)}{\mathrm{pr}(x)}$$

- We compare two models m_i and m_j using

$$\frac{\mathrm{pr}(m_i \mid x)}{\mathrm{pr}(m_j \mid x)} = \frac{\frac{\mathrm{pr}(x \mid m_i)\mathrm{pr}(m_i)}{\mathrm{pr}(x)}}{\frac{\mathrm{pr}(x \mid m_j)\mathrm{pr}(m_j)}{\mathrm{pr}(x)}} = \frac{\mathrm{pr}(x \mid m_i)\mathrm{pr}(m_i)}{\mathrm{pr}(x \mid m_j)\mathrm{pr}(m_j)}$$

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

Model selection approaches - Bayes factor

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

$$\frac{\mathrm{pr}(x \mid m_i)}{\mathrm{pr}(x \mid m_j)}$$

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

$$\mathrm{pr}(x \mid m) = \int_\theta \mathrm{pr}(x \mid \theta, m)\mathrm{pr}(\theta \mid m)\mathrm{d}\theta$$

where θ are the model parameters

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

- 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

$$\mathrm{pr}(x \mid m) \approx \exp\left\{-\frac{1}{2}\mathrm{BIC}(x; m)\right\}$$

$$\mathrm{BIC}(x; m) = -2\log\mathrm{pr}(x \mid \hat{\theta}, m) + p\log(n)$$

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

- 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 $\epsilon \sim \mathrm{Normal}(0, \sigma^2)$ the BIC is related to the mean squared error with complexity penalty

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

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]:

$$\mathrm{pr}(x \mid m) \approx \exp\{-\mathrm{FIA}(x; m)\}$$

$$\mathrm{FIA}(x; m) = \underbrace{-\log \mathrm{pr}(x \mid \hat{\theta}, m) + \frac{p}{2}\log\left(\frac{n}{2\pi}\right)}_{\text{BIC like term}} + \log C_m$$

$$C_m = \underbrace{\int_\theta \sqrt{\det \mathcal{I}_m(\theta)}\mathrm{d}\theta}_{\text{Geometric complexity}}$$

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

$$\omega(\theta) = -\log \mathrm{pr}_\theta(y) \quad \text{(maximum likelihood), or}$$
$$\omega(\theta) = \|y - X\theta\|_2^2 \quad \text{(linear regression)}$$

Regularized estimate with ℓ_k-norm penalty

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

where

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

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

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 Lagriangian

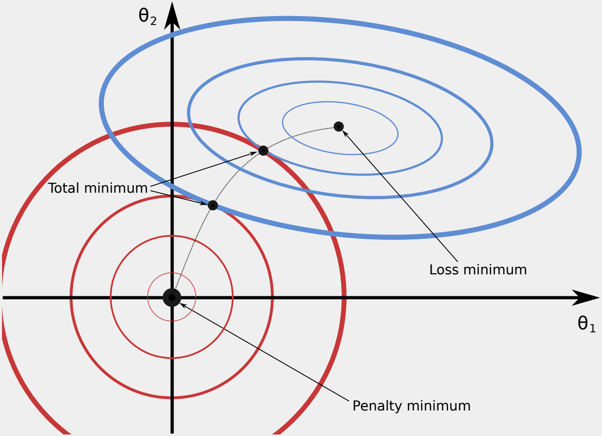
$$\hat{\theta}(\lambda) = \arg\min_{\theta} \omega(\theta) + \lambda \|\theta\|_k^k$$

At the optimum we must have

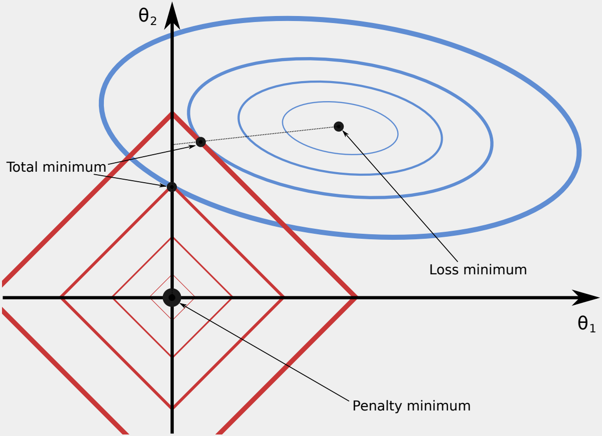
$$\nabla_\theta\, \omega(\theta) + \lambda \nabla_\theta \|\theta\|_k^k = 0$$

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

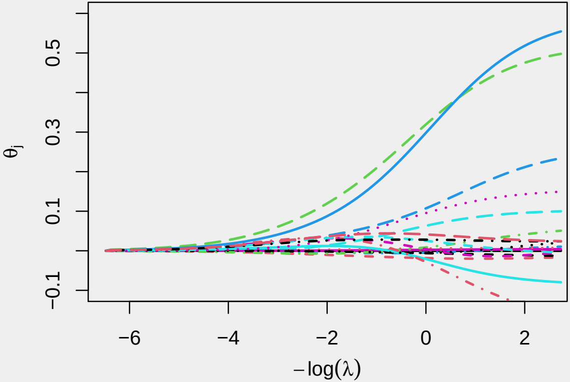
REGULARIZATION - K=2



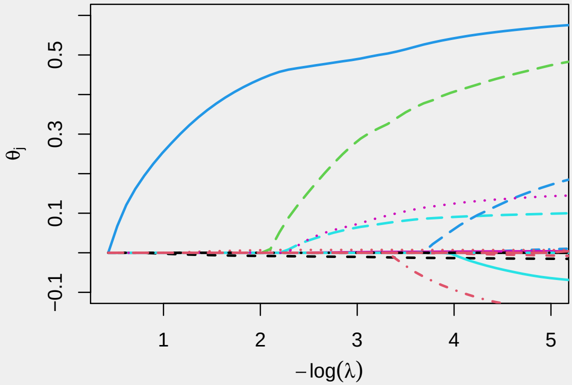
Regularization - k=1



Regularization paths - K=2

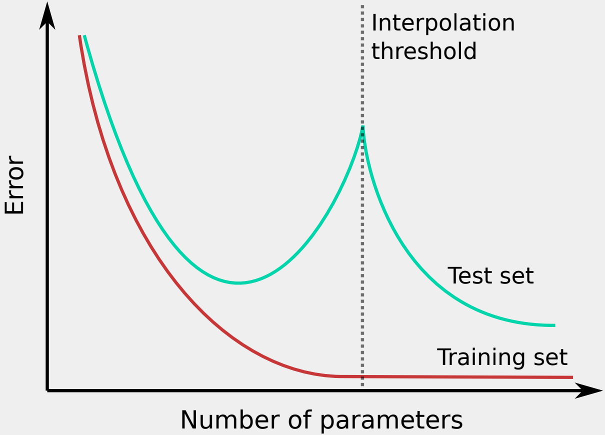


Regularization paths - k=1

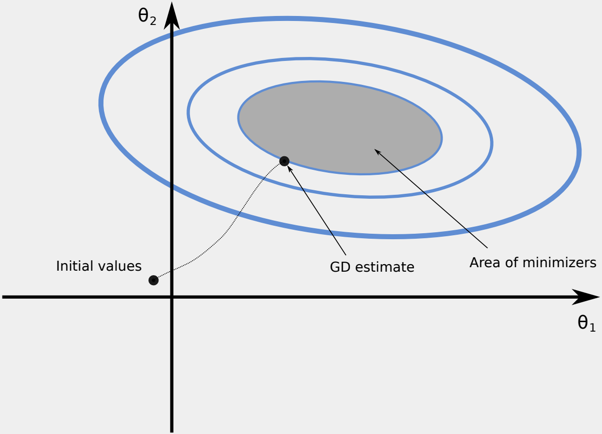


Implicit Regularization and Double Descent

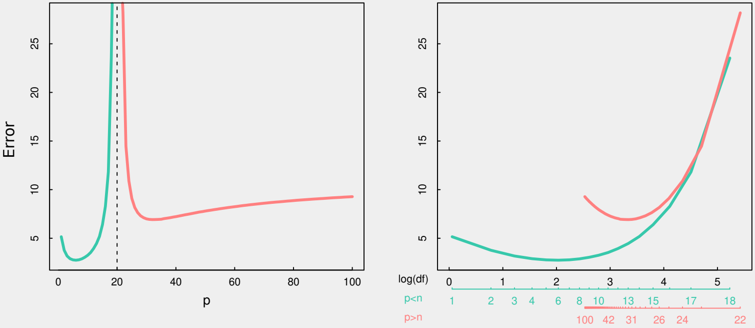
Implicit Regularization - Double descent



Implicit Regularization - Double descent



Minimum ℓ_2-norm estimate - DF



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

Implicit regularization

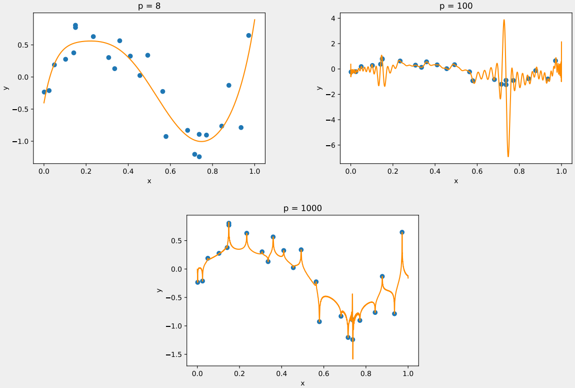


Figure: Fitting degree $d = p - 1$ Legendre polynomials. For $p > n$ the solution with the smallest ℓ_2-norm is used.

[2] Legendre 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 correspont 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, Derezinski et al., 2020, Kobak et al., 2020]

Reading

- 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.

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.

Cook, R. D. (1977).
Detection of influential observation in linear regression.
Technometrics, 19(1):15–18.

References II

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.

Gao, T. and Jojic, V. (2016).
Degrees of freedom in deep neural networks.
arXiv preprint arXiv:1603.09260.

Grünwald, P. D. (2007).
The minimum description length principle.
MIT press.

References III

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.

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.

References IV

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.

Luan, B., Lee, Y., and Zhu, Y. (2021).
Predictive model degrees of freedom in linear regression.
arXiv preprint arXiv:2106.15682.

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.

References V

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.

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.

References VI

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.

Wolpert, D. H. and Macready, W. G. (1997).
No free lunch theorems for optimization.
IEEE transactions on evolutionary computation, 1(1):67–82.