

Lecture 7: Bias-Variance Tradeoff

Lecturer: Yiqiao Zhong

Last update: 19:34:01, 2024-10-02

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

Our goal is to introduce a fundamental concept underpinning many statistical theories and applications.

1 A simple example revisited

In the previous lecture, we have learned that natural decision rules may not be optimal. In a simple case where $X \sim N(\mu, 1)$, the “trivial” estimator $T_2(X) \equiv 0$ has zero variance but a nonzero bias unless $\mu = 0$. On the other hand, the natural estimator $T_2(X) = X$ incurs no bias but there is a constant variance 1. Which estimator is better depends on μ .

A more refined example. (Example 1.3.4. in [3])

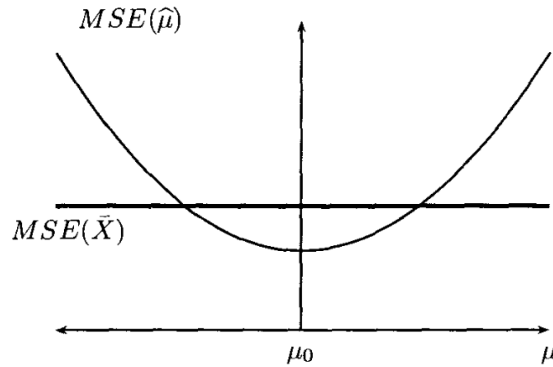


Figure 1: Risk function with varying unknown “true” parameter μ . Figure from [3]

Suppose that $X_1, X_2, \dots, X_n$ are independent measurements (random variables) of household income in a certain area. We may combine sample average $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ with a fixed value (from U.S. census for example) μ_0 , so we construct the estimator $\hat{\mu}$:

$$\hat{\mu}(X) = (0.2)\mu_0 + (0.8)\bar{X}.$$

We can derive the bias, variance, and risk (using MSE).

$$\text{Bias}(\hat{\mu}) = 0.2(\mu_0 - \mu)$$

$$\text{Var}(\hat{\mu}) = 0.64 \frac{\sigma^2}{n}$$

$$R(\mu, \hat{\mu}) = 0.04(\mu_0 - \mu)^2 + 0.64 \frac{\sigma^2}{n}.$$

This estimator $\hat{\mu}$ is better than the sample mean estimator if μ is sufficiently close to $\hat{\mu}_0$.

2 Ridge regression

Suppose that we have n observations (x_i, y_i) where $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$, and they satisfy the linear model

$$y_i = x_i^\top \beta^* + \varepsilon_i$$

where we want to estimate the unknown parameters $\beta^* \in \mathbb{R}^p$. By convention, denote $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}, y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$.

The ridge regression finds the estimator by solving the following minimization problem.

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_2^2 \quad (1)$$

where $\lambda \geq 0$ is a ridge tuning parameter (or hyperparameter). The solution is given by (exercise)

$$\hat{\beta}_\lambda = \left(\frac{1}{n} X^\top X + \lambda I_p \right)^{-1} \frac{1}{n} X^\top y$$

In particular, if $\lambda = 0$, then $\hat{\beta}_0$ is simply the least-square estimator.

To evaluate ridge regression, let us assume that X is a nonrandom and nondegenerate matrix, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. random variables with zero mean and finite second moment. Consider the MSE of the estimator

$$\text{MSE} = \mathbb{E}[\|\hat{\beta}_\lambda - \beta^*\|_2^2] = \underbrace{\mathbb{E}[\|\hat{\beta}_\lambda - \mathbb{E}\hat{\beta}_\lambda\|_2^2]}_{\text{Variance}} + \underbrace{\|\mathbb{E}\hat{\beta}_\lambda - \beta^*\|_2^2}_{\text{Bias}^2}.$$

For $\lambda = 0$, the variance is $\sigma^2 \text{Tr}((X^\top X)^{-1})$. In general, the Variance decreases in λ and Bias² increases in λ .

Is there a tradeoff? Suppose that we apply a small ridge regularization. Some calculation shows that (exercise)

$$\text{Bias} = O(\lambda) \quad \sigma^2 \text{Tr}((X^\top X)^{-1}) - \text{Variance} = O(\lambda).$$

This means that at small λ , the benefit of reduced variance outweighs the incurred bias.

3 A general bias-variance tradeoff

The bias-variance tradeoff in the linear model is a quite general feature of statistics.

Consider a nonparametric regression by k -nearest-neighbor (KNN). Suppose that the statistical model

$$Y = f(X) + \varepsilon, \quad \text{where } \mathbb{E}\varepsilon = 0, \text{ Var}(\varepsilon) = \sigma^2.$$

where $X \in \mathbb{R}^p$ and f is an unknown function. We collect data $(x_i, y_i)_{i=1}^n$ that follow the above model. For simplicity, we assume that $x_1, \dots, x_n$ are nonrandom, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. We define our KNN estimator in the following way: given a pre-specified K and test input x_0 , we take

$$\hat{f}_K(x_0) = \frac{1}{K} \sum_{k=1}^K y_{x_{(k)}}$$

where the $x_{(1)}, \dots, x_{(K)}$ are K nearest points (under the Euclidean distance) in the data to x_0 . In other words, our prediction is based on the K nearest points in the observations. We derive the MSE at the given test input x_0 .

$$\mathbb{E}[(Y - \hat{f}_K(x_0))^2 | X = x_0] = \underbrace{\sigma^2 + \left[f(x_0) - \frac{1}{K} \sum_{k=1}^K f(x_{(k)}) \right]^2}_{\text{Bias}^2} + \underbrace{\frac{\sigma^2}{K}}_{\text{Variance}}$$

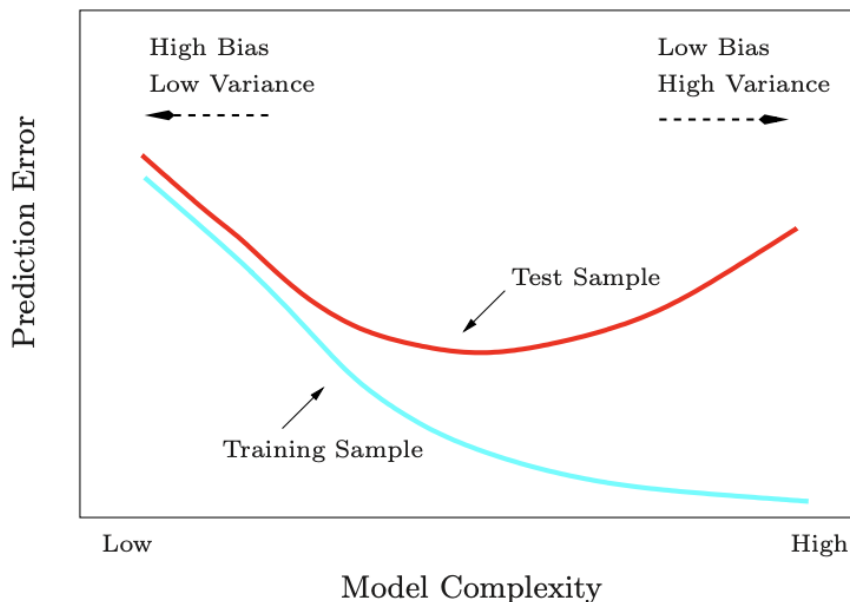


Figure 2: The risk function exhibits a U-shaped curve with varying model complexity. A larger model generally induces less bias at the expense of increased variance. Figure is taken from [4].

The first term σ^2 is the irreducible error due to our observation noise term, and the variance decreases in K as the average reduces variance.

The bias term measures how representative the K neighbors are to the test point. If the regression function f is smooth, we expect this bias to be small for small K . When K is large, the prediction relies on potentially irrelevant observations, so we cannot control the bias.

Overfitting happens when the function estimator fits excessively to data, causing the estimator to depend heavily on noise in the observation. It is associated with a high model complexity: larger p or small λ in ridge regression, and small K in KNN. In this case, the fitted function has a high degree of freedom. In the extreme case, when $p = n$, $\lambda = 0$ or $K = 1$, the fitted function interpolates exactly the observation—this is also known as **memorization**.

Figure 2 shows the general picture of prediction errors plotted against model complexity. There is a sweet spot in model complexity for achieving optimal error.

Regularization. In classical statistics, there are multiple ways of constraining the complexity of the models. For prediction problems, *model selection* picks the model with the smallest predicted error, for instance, with cross-validation. Adding a regularizer also reduces the **effective degrees of freedom**. For example,

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_q^q \quad (2)$$

adds regularization terms known as ℓ_q regularizer (or penalty), which includes ridge regression as a special case. We will return to regularized models later in this course.

4 Overparametrization

The variance of fitted linear function $\hat{f}(x) = x^\top \hat{\beta}_0$ is, when $X^\top X$ is invertible,

$$\text{Var}(\hat{f}(x)) = \sigma^2 x^\top (X^\top X)^{-1} x.$$

If x_0 has zero mean and identity covariance matrix and denote $\sigma^2 = \text{Var}(\varepsilon_i)$, then

$$\text{Var}(\hat{f}(x_0)) = \sigma^2 \mathbb{E} \left[x_0^\top (X^\top X)^{-1} x_0 \right] = \sigma^2 \mathbb{E} \left[\text{Tr}((X^\top X)^{-1} x_0 x_0^\top) \right] = \sigma^2 \text{Tr}((X^\top X)^{-1}). \quad (3)$$

The variance generally increases in dimension p —in the case that $X^\top X = I_p$, the variance is p/n ; if the smallest eigenvalue of $X^\top X$ is bounded away from 0, then the variance is $O(p/n)$.

Suppose $p > n$ in (1). Since $\text{rank}(X^\top X) \leq n < p$, the classical least square estimator $\hat{\beta}_0$ is not well defined (non-invertible matrix). In fact, the linear equations

$$x_i^\top \beta = y_i, \quad i = 1, 2, \dots, n$$

has infinitely many solutions: denoting $\ker(X) = \{z : Xz = 0\}$, then any $\beta = \beta^* + \ker(X)$ solves the linear equations, which means perfect fit (or interpolation) of our observations.

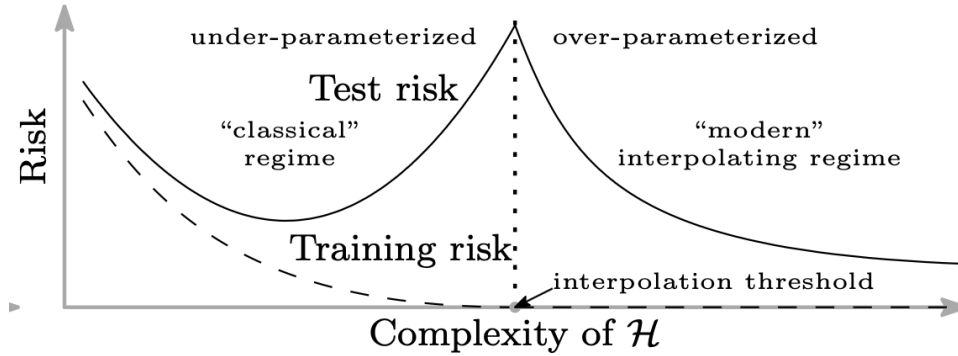


Figure 3: The risk function exhibits a second descent when the model complexity passes a threshold (the interpolation threshold). Figure is from [2]. This general picture applies to deep neural networks [5].

Double descent and min-norm solution. The above analysis suggests that the least square estimator suffers from singularity when p is close to n , and we need regularization to prevent this, especially for larger p . An interesting phenomenon is that we actually need *little regularization* when $p \gg n$.

When $p > n$, the min-norm interpolating estimator solves

$$\begin{aligned} \min_{\beta} \quad & \|\beta\|_2^2 \\ \text{subject to} \quad & X\beta = y \end{aligned} \quad (4)$$

It can be shown that the above optimization problem has a unique solution and is in fact $\lim_{\lambda \rightarrow 0+} \beta_\lambda$ (exercise). For this reason, it is also called the **ridgeless estimator**.

Theorem 1. Suppose that $p > n$ and that $XX^\top \in \mathbb{R}^{n \times n}$ has full rank n . Then, the optimization problem (4) has a unique solution, denoted by $\hat{\beta}_{0+}$ which is given by $\hat{\beta}_{0+} = \lim_{\lambda \rightarrow 0+} \beta_\lambda$. Moreover,

$$\hat{\beta}_{0+} = X^\top (XX^\top)^{-1} y. \quad (5)$$

To prove the formula in (5), we can use a fact in linear algebra

$$X^\top \left(\frac{1}{n} X X^\top + \lambda I_n \right)^{-1} y = \left(\frac{1}{n} X^\top X + \lambda I_p \right)^{-1} X^\top y, \quad \text{for all } \lambda > 0$$

and then taking $\lambda \rightarrow 0+$. The above identity can be proved by using singular value decomposition (exercise).

Note that the roles of p and n are flipped in the matrix X when $p > n$. For that reason, $\hat{\beta}_{0+}$ generally has small variance if $p \gg n$ at the expense of bias. Comparing with (3), now we have

$$\text{Var}(\hat{f}(x_0)) = \mathbb{E} \left[x_0^\top X^\top (X X^\top)^{-2} X x_0 \right] = \mathbb{E} \left[(X X^\top)^{-1} X X^\top x_0 x_0^\top \right] = \sigma^2 \text{Tr}((X X^\top)^{-1}).$$

If the smallest eigenvalue of $X^\top X$ is bounded away from 0, then the variance is $O(n/p)$.

5 The implicit regularization of gradient descent

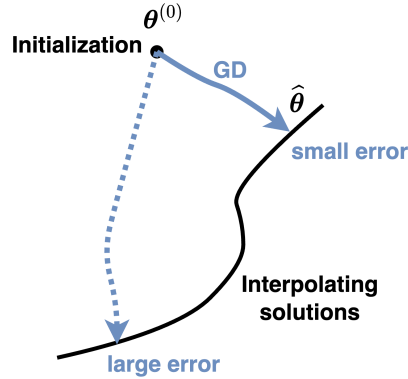


Figure 4: An illustration that shows implicit regularization of gradient descent.

There are many ways to numerically calculate the least square estimator. Here we consider the generic gradient descent (GD) algorithm. Denote the square loss by $L(\beta) = \|y - X\beta\|_2^2$. We start from initialization $\beta^{(0)} = 0$ and then iteratively computes

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla L(\beta)$$

where $\eta > 0$ is the step size or learning rate. We state the following deterministic result (not assuming random variables).

Theorem 2. Suppose that $p > n$ and that $X X^\top$ has full rank n . Then, if $\eta < [2\lambda_{\max}(X X^\top)]^{-1}$, then $\beta^{(t)}$ converges, and

$$\lim_{t \rightarrow \infty} \beta^{(t)} = \beta_{0+}.$$

This theorem says that, even without explicit regularization, GD prefers a solution—in the space of infinite interpolating solutions—with the minimum ℓ_2 norm.

The convergence is a standard result in convex optimization [], and the limit is because

$$\nabla L(\beta) \in \text{span}(X^\top) \implies \beta^{(t)} \in \text{span}(X^\top) \implies \lim_{t \rightarrow \infty} \beta^{(t)} \in \text{span}(X^\top)$$

We note that β_{0+} is the unique solution in the space $\text{span}(X^\top)$ among interpolating solutions $\beta^* + \ker(X)$, since $\text{span}(X^\top) \perp \ker(X)$. This proves that $\lim_{t \rightarrow \infty} \beta^{(t)} = \beta_{0+}$.

Theorem (Soudry, et al. 2018). Consider the logistic loss and any linearly separable data. From any initializer $\beta^{(0)} \in \mathbb{R}^d$, the gradient descent iterate $\beta^{(k+1)} = \beta^{(k)} - \eta \nabla L_n(\beta^{(k)})$ satisfies

$$\beta^{(k)} = \hat{\beta} \log k + \Delta^{(k)}$$

where residual $\|\Delta^{(k)}\|_2 = O(\log \log k)$ and $\hat{\beta}$ is the max-margin solution

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\beta\|_2^2 \quad \text{s.t. } \beta^\top x_i \geq 1, \text{ for all } i = 1, 2, \dots, n.$$

Figure 5: GD converges to the max-margin classifier in the overparametrized setting (fitting all labels correctly).

Min-norm solution in classification. So far, our analysis focuses on regression. The general idea applies to classification problems as well. It is known that GD applied to logistic regression loss induces a min-norm interpolation solution [6].

6 Further reading

Bias-variance tradeoff is a classical topic in statistics. See [4], especially Chapter 7. Our example in Section 3 is a reformulation therein.

There have been recent interests in overparametrized models since the success of deep learning. See [1] for a survey of recent theoretical results.

References

- [1] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.
- [2] Mikhail Belkin, Daniel J. Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849 – 15854, 2018.
- [3] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.
- [4] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [5] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.
- [6] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.