STATISTICS AND DATA SCIENCE 365 / 665
INTERMEDIATE MACHINE LEARNING

A note on the double descent phenomenon

How can we understand the surprising resistance to overfitting seen in many neural network models? One view comes from the "double descent phenomenon" that has been recently uncovered. As neurons are added to the network, the model becomes overparameterized, interpolating the data, with no training error. A distinguished interpolating model is the minimum norm solution to a linear system. As the level of overparameterization decreases, the variance decreases while the bias remains constant, leading to a double descent behavior in the risk curve.

1. The random features model: A computational "fruit fly"

The "random features model" is a simple framework that can be used as a tool to gain understanding for more complex neural network models — similar to how the fruit fly is studied to understand more complex species. In its most basic form, the random features model is a two-layer neural network

$$f_\beta(x) = \beta^T h(x) = \beta^T \varphi(Wx + b)$$

where the first layer parameters W (with intercepts b) are random and fixed. If the input x is d dimensional then $W \in \mathbb{R}^{p\times d}$ is a $p \times d$ matrix.

This is nothing more than a linear model in the random feature space $h(x) = \varphi(Wx + b) \in \mathbb{R}^p$. One reason this is a reasonable approximation to what's going on with large neural networks is that as $p \to \infty$, when trained with stochastic gradient descent, the first layer parameters W will not change much from their random initializations. Moreover, if β is initialized at zero, then stochastic gradient descent will tend to converge to the minimum norm solution, which we introduce next.

2. The minimum norm estimator

Consider the linear model $f(x) = \beta^T h(x)$ in the feature space $x \to h(x) \in \mathbb{R}^p$, with training data $(X_1, y_1), \ldots, (X_n, y_n)$ where $X_i = h(x_i) \in \mathbb{R}^p$ is a feature vector. If $p < n$ then we can compute the ordinary least squares solution

$$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

where $\mathbb{X}$ is the design matrix with rows $\mathbb{X}_i = X_i^T$ and $Y = (Y_1, \ldots, Y_n)^T$ is the vector of responses. The fitted values are then $\widehat{Y} = \mathbb{X}\widehat{\beta}$.

If $n < p$, however, the least squares solution is underdetermined. A distinguished solution to the linear system $\mathbb{X}\beta = Y$ is the minimum norm solution

$$\widehat{\beta}_{\text{mn}} = \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y$$

where we make the (weak) assumption that $\mathbb{X}\mathbb{X}^T \in \mathbb{R}^{n\times n}$ is nonsingular in the overparameterized regime $n < p$. This clearly satisfies $\widehat{Y} = \mathbb{X}\widehat{\beta}_{\text{mn}} = Y$. This is described as "interpolating the data" since the fitted values are equal to the data points.

Why is it the minimum norm solution? That is, why does it minimize $\|\beta\|$ among all solutions to $\mathbb{X}\beta = Y$? Note that for any such solution

$$\begin{aligned}(\beta - \widehat{\beta}_{\text{mn}})^T\widehat{\beta}_{\text{mn}} &= (\beta - \widehat{\beta}_{\text{mn}})^T\mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}Y \\ &= \left(\mathbb{X}(\beta - \widehat{\beta}_{\text{mn}})\right)^T(\mathbb{X}\mathbb{X}^T)^{-1}Y \\ &= 0.\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\|\beta\|^2 &= \|\beta - \widehat{\beta}_{\text{mn}} + \widehat{\beta}_{\text{mn}}\|^2 \\ &= \|\widehat{\beta}_{\text{mn}}\|^2 + \|\beta - \widehat{\beta}_{\text{mn}}\|^2 \\ &\geq \|\widehat{\beta}_{\text{mn}}\|^2.\end{aligned}$$

Geometrically, we can describe $\widehat{\beta}_{\text{mn}}$ is the orthogonal projection of the zero vector in $\mathbb{R}^p$ onto the n-dimensional hyperplane $\{\beta \ : \ \mathbb{X}\beta = Y\}$.

This can also be seen as "ridgeless" regression. In ridge regression we minimize

$$\|\mathbb{X}\beta - Y\|^2 + \lambda\|\beta\|^2$$

which has the closed form solution $\widehat{\beta}_\lambda = \left(\mathbb{X}^T X + \lambda I\right)^{-1}\mathbb{X}^T Y$. In the overparameterized regime $p > n$ it can be shown that, as $\lambda \to 0$,

$$\left(\mathbb{X}^T\mathbb{X} + \lambda I\right)^{-1}\mathbb{X}^T \longrightarrow \mathbb{X}^T(\mathbb{X}\mathbb{X}^T)^{-1}$$

and that $\widehat{\beta}_\lambda \longrightarrow \widehat{\beta}_{\text{mn}}$. To see this, we use the identity

$$(X^TX + \lambda I_p)^{-1}X^T = X^T(XX^T + \lambda I_n)^{-1} \tag{1}$$

which is a consequence of the Woodbury formula

$$(I + UV^T)^{-1} = I - U(I + V^TU)^{-1}V^T. \tag{2}$$

Defining $\gamma = \frac{p}{n}$, we can summarize this as follows:

As the regularization parameter λ decreases to zero, the ridge regression estimate:

- Converges to the least squares solution in the "classical regime" $\gamma < 1$
- Converges to the least norm solution in the "overparameterized regime" $\gamma > 1$

3. Double descent

The double descent phenomenon was only recently recognized as an important ingredient in understanding the behavior of large neural networks. The bias of both the minimum norm solution and the least squares estimator are each small relative to the best linear model. The surprise lies in the behavior of the variance of the minimum norm estimator as γ gets large. In the plots of the following figure, we see that this variance decreases as γ gets large. This can be analyzed using random matrix theory for the matrix $\mathbb{X}\mathbb{X}^T$.

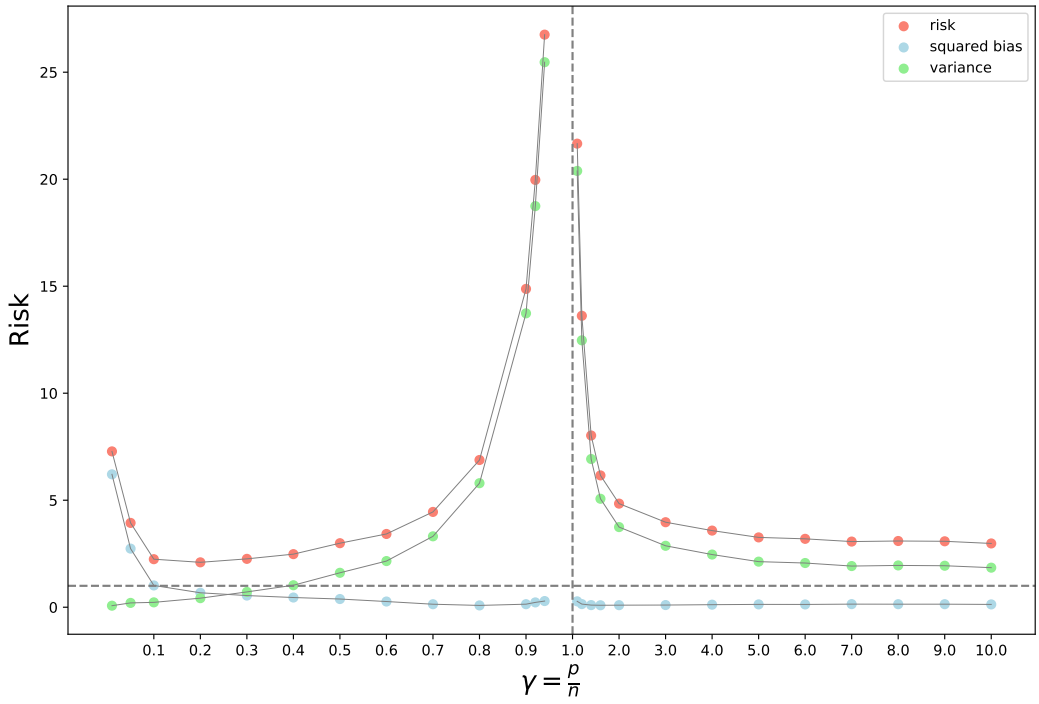

FIG 1. *The bias-variance behavior underlying the double descent curve for a random features model. In the classical regime where $\gamma < 1$, the bias decreases while the variance increases, creating the U-shaped risk curve. In the overparameterized regime the bias is constant while the variance decreases. Note that the scale of γ in the plot is different in the two regimes.*

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

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.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.