Bias-Variance Tradeoff

Matthieu R. Bloch

We have formalized the problem of supervised learning as finding a function (or hypothesis) h in a given set $\mathcal{H}$ that minimizes the true risk $R(h)$. In the context of classification we hope to approximate the optimal Bayes classifier while in the context of regression we hope to approximate the true underlying function. We have already seen that the choice of $\mathcal{H}$ must strike a delicate tradeoff between two desirable characteristics:

- a more complex $\mathcal{H}$ leads to better chance of *approximating* ideal classifier/function;
- a less complex $\mathcal{H}$ leads to better chance of *generalizing* to unseen data.

Regularization plays a similar role by biasing answer away from complex functions. This is particularly crucial for regression in which the complexity must be carefully limited to avoid *overfitting*.

In the context of classification, we have already seen that the tradeoff can be precisely quantified in terms of the *VC generalization bound*, which takes the form

$$R(h) \leqslant \widehat{R}_N(h) + \epsilon(\mathcal{H}, N) \text{ with high probability.}$$

We now develop an alternative method to quantify the tradeoff called the *bias-variance decomposition* which takes the form

$$R(h) \approx \text{bias}^2 + \text{variance}.$$

Thererin, the *bias* captures how well $\mathcal{H}$ can approximate the true h^*, while the *variance* captures how likely we are to pick a good $h \in \mathcal{H}$. This approach generalizes more easily to regression than the VC dimension approach developed for classification.

1 Setup for bias-variance decomposition analysis

We formalize the bias-variance tradeoff assuming the following:

- $f : \mathbb{R}^d \to \mathbb{R}$ is the unknown target function that we are trying to learn;
- $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is the dataset, where $(\mathbf{x}_i, y_i)$ are independent and identically distributed (i.i.d.); specifically, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i = f(\mathbf{x}_i) + \varepsilon_i \in \mathbb{R}$, where ε_i is a zero-mean noise random variable independent of $\mathbf{x}_i$ with variance σ_ε^2 (for instance $\epsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$);
- $\hat{h}_\mathcal{D} : \mathbb{R}^d \to \mathbb{R}$ is our choice of function in $\mathcal{H}$, selected using $\mathcal{D}$;
- The performance of $\hat{h}_\mathcal{D}$ is measured in terms of the mean squared error $R(\hat{h}_\mathcal{D}) = \mathbb{E}_{XY}\Big((\hat{h}_\mathcal{D}(X) - Y)^2\Big)$;

Note that the random variables (X, Y) denote the data at *testing* and should not be confused with the random variable $\mathcal{D}$ representing the *training* data.

Lemma 1.1 (Bias-variance decomposition).

$$\mathbb{E}_{\mathcal{D}}\Big(R(\hat{h}_{\mathcal{D}})\Big) = \sigma_{\varepsilon}^2 + \mathbb{E}_X\Big(\text{Var}\Big(\hat{h}_{\mathcal{D}}(X)\Big)|X\Big) + \mathbb{E}_X\Big(\text{Bias}(\hat{h}_{\mathcal{D}}(X))^2|X\Big)$$

with

$$\text{Var}\Big(\hat{h}_{\mathcal{D}}(X)\Big) \triangleq \mathbb{E}_{\mathcal{D}}\bigg(\Big(\hat{h}_{\mathcal{D}}(X) - \mathbb{E}_{\mathcal{D}}\Big(\hat{h}_{\mathcal{D}}(X)\Big)\Big)^2\bigg)$$

$$\text{Bias}(\hat{h}_{\mathcal{D}}(X)) \triangleq \mathbb{E}_{\mathcal{D}}\Big(\hat{h}_{\mathcal{D}}(X)\Big) - f(X)$$

Proof. For clarity, set $\bar{h}(X) \triangleq \mathbb{E}_{\mathcal{D}}\Big(\hat{h}_{\mathcal{D}}(X)\Big)$. Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\Big(R(\hat{h}_{\mathcal{D}})\Big) &= \mathbb{E}_{\mathcal{D}}\Big(\mathbb{E}_{XY}\Big((\hat{h}_{\mathcal{D}}(X) - Y)^2\Big)\Big) && (1)\\
&= \mathbb{E}_{\mathcal{D}}\Big(\mathbb{E}_{X\varepsilon}\Big((\hat{h}_{\mathcal{D}}(X) - f(X) - \varepsilon)^2\Big)\Big) && (2)\\
&= \mathbb{E}_{\mathcal{D}}\Big(\mathbb{E}_{X\varepsilon}\Big((\hat{h}_{\mathcal{D}}(X) - \bar{h}(X) + \bar{h}(X) - f(X) - \varepsilon)^2\Big)\Big) && (3)\\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))^2 + (\bar{h}(X) - f(X))^2 + \varepsilon^2 \\
&\qquad +2(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - f(X)) - 2(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))\varepsilon \\
&\qquad\qquad -2(\bar{h}(X) - f(X))\varepsilon\Big] && (4)
\end{aligned}$$

Note that in (4) we have used the fact that $\mathcal{D}$, X, and ε are independent. Notice that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))^2\Big] &\triangleq \mathbb{E}_X\Big(\text{Var}\Big(\hat{h}_{\mathcal{D}}(X)|X\Big)\Big) && (5)\\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[(\bar{h}(X) - f(X))^2\Big] &\triangleq \mathbb{E}_X\Big(\text{Bias}(\hat{h}_{\mathcal{D}}(X))^2\Big) && (6)\\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[\varepsilon^2\Big] &\triangleq \sigma_{\varepsilon}^2. && (7)
\end{aligned}$$

The last three terms turn out to be zero since

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\Big[(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - f(X))\Big] &= \mathbb{E}_X\Big[(\mathbb{E}_{\mathcal{D}}\Big(\hat{h}_{\mathcal{D}}(X)\Big) - \bar{h}(X))(\bar{h}(X) - f(X))\Big] \\
&= 0 && (8)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X))\varepsilon\Big] &= \mathbb{E}_X\Big(\mathbb{E}_{\mathcal{D}}\Big(\hat{h}_{\mathcal{D}}(X) - \bar{h}(X)\Big)\Big)\mathbb{E}_{\varepsilon}(\varepsilon) = 0 && (9)\\
\mathbb{E}_{\mathcal{D}}\mathbb{E}_X\mathbb{E}_{\varepsilon}\Big[(\bar{h}(X) - f(X))\varepsilon\Big] &= \mathbb{E}_X\big(\bar{h}(X) - f(X)\big)\mathbb{E}_{\varepsilon}(\varepsilon) = 0. && (10)
\end{aligned}$$

■

2 Intuition behind the bias-variance tradeoff

The intuition behind the bias-variance tradeoff is illustrated in Fig. 1. The gray area around the true function f represents the variance of the perception of the true resulting from the noisy samples that we obtain. The model space represent, for instance, all the linear models while the regularized model represents the regularized models. The blue area represents the variance of the model while the orange area represents the variance of the regularized model. The regularized model offers a smaller variance at the expense of an increased bias.

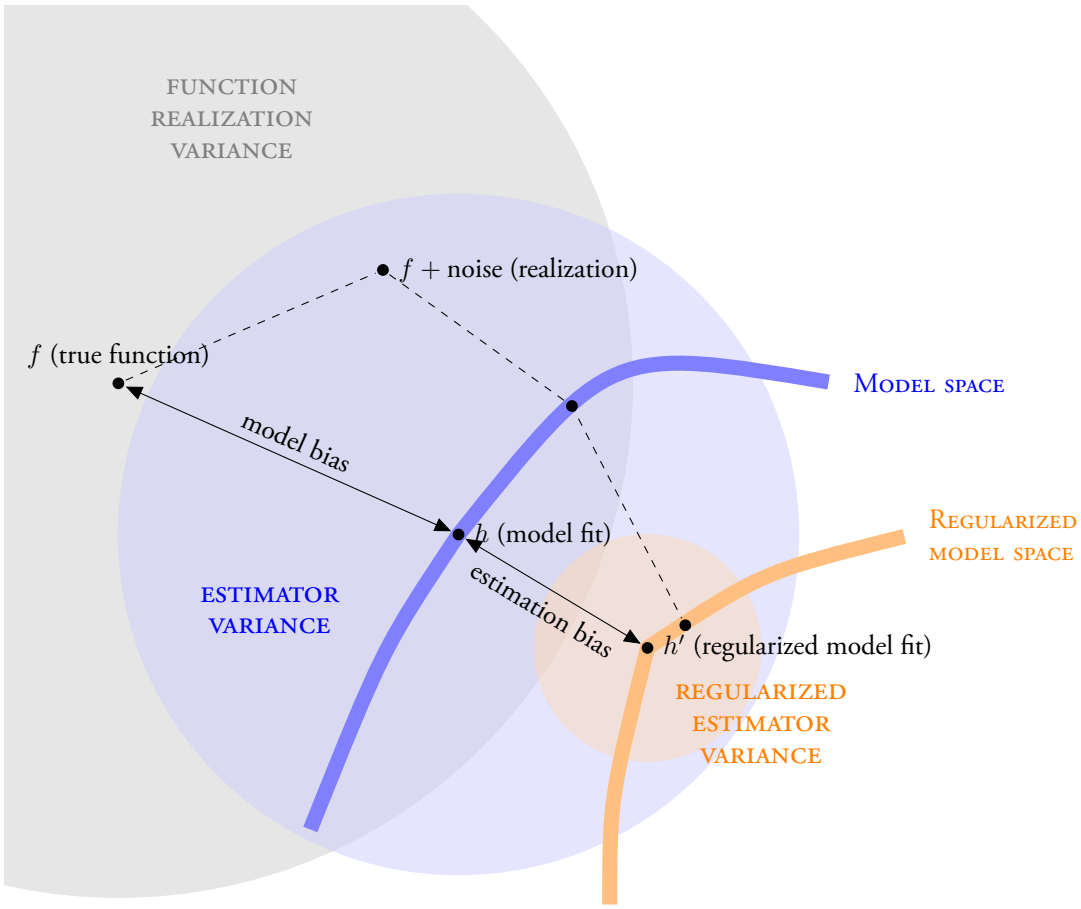

Figure 1: Illustration of bias-variance tradeoff adapted from [1, Figure 7.2]

References

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer series in statistics. Springer, 2009.