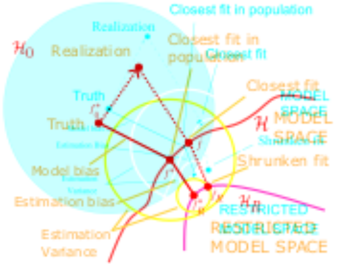
Introduction to Machine Learning

Regularization
Bias-variance Tradeoff



Learning goals

- Understand the bias-variance trade-off
- Know the definition of model bias, estimation bias, and estimation variance

BIAS-VARIANCE TRADEOFF

In this slide set, we will visualize the bias-variance trade-off.

We consider a DGP $\mathbb{P}_{xy}$ with $\mathcal{Y} \subset \mathbb{R}$ and the L2 loss L. We measure the distance between models $f : \mathcal{X} \to \mathbb{R}^g$ via

$$d(f, f') = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\mathbf{x}} \left[L(f(\mathbf{x}), f'(\mathbf{x})) \right].$$

We define f_0^* as the risk minimizer such that

$$f_0^* \in \arg\min_{f \in \mathcal{H}_0} \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}_{xy}} \left[L(y, f(\mathbf{x})) \right]$$

where $\mathcal{H}_0 = \{ f : \mathcal{X} \to \mathbb{R} |\ d(\underline{0}, f) < \infty \}$ and $\underline{0} : \mathcal{X} \to \{0\}$.

... numerical encoding of $\mathcal{Y}$. We measure the distance between models $f : \mathcal{X} \to \mathbb{R}^g$ via

$$d(f, f') = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\mathbf{x}} \left[L(f(\mathbf{x}), f'(\mathbf{x})) \right].$$

... d becomes a metric, e.g., L1-loss, L2-loss, etc.

We define f_{true} as the risk minimizer such that

$$f_{\text{true}} \in \arg\min_{f \in \mathcal{H}_0} \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}_{xy}} \left[L(y, f(\mathbf{x})) \right]$$

where $\mathcal{H}_0 = \{ f : \mathcal{X} \to \mathbb{R}^g |\ d(\underline{0}, f) < \infty \}$ and $\underline{0} : \mathcal{X} \to \{0\}$.

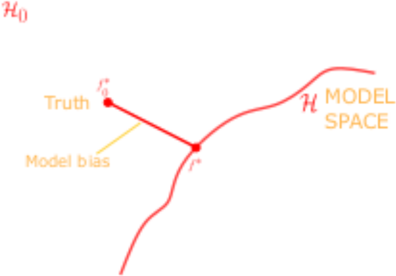
BIAS-VARIANCE TRADEOFF / 2

- Our model space $\mathcal{H}$ usually is a proper subset of $\mathcal{H}_0$ and in general $f_0^* \notin \mathcal{H}$.

We define f^* as the risk minimizer in $\mathcal{H}$, i.e.,

$$f^* \in \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{xy}} \left[L(f(\mathbf{x}, y)) \right].$$

- $f^* \in \mathcal{H}$ is closest to f_0^*, and we call $d(f_0^*, f^*)$ the model bias.



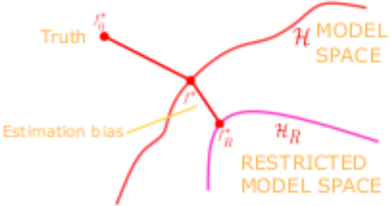
BIAS-VARIANCE TRADEOFF / 3

By regularizing our model, we further restrict the model space so that $\mathcal{H}_R$ is a proper subset of $\mathcal{H}$. We define f_R^* as the risk minimizer in $\mathcal{H}_R$, i.e.,

$$f_R^* \in \arg\min_{f \in \mathcal{H}_R} \mathbb{E}_{(\mathbf{x},y)\sim \mathbb{P}_{xy}} \left[L(f(\mathbf{x}, y)) \right].$$

$f_R^* \in \mathcal{H}_R$ is closest to f_{true}, and we call $d(f_R^*, f^*)$ the estimation bias.

$\mathcal{H}_0$



BIAS-VARIANCE TRADEOFF / 4

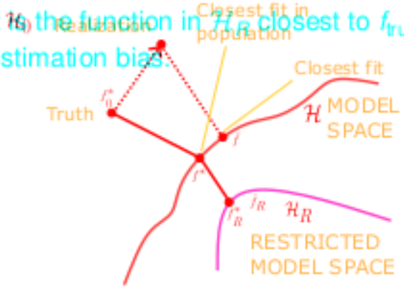
We sample a finite dataset $\mathcal{D} = \left(\mathbf{x}^{(i)}, y^{(i)}\right)_{i=1}^{n} \in (\mathbb{P}_{xy})^n$ and find via ERM

$$\hat{f} \in \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f(\mathbf{x}^{(i)})\right).$$

Let $\mathcal{H}_R$ be a restricted model space such that $\mathcal{H}_R$ is a proper subset of $\mathcal{H}$. We define f_R^* as the risk minimizer in $\mathcal{H}_R$, i.e.,

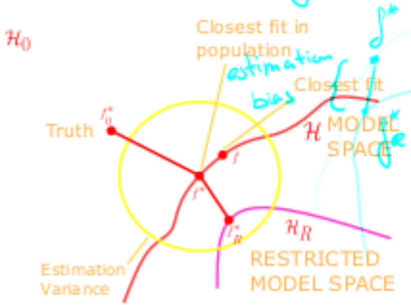
$$f_R^* \in \arg\min_{f \in \mathcal{H}_R} \mathbb{E}_{(\mathbf{x},y)\sim \mathbb{P}_{xy}}\left[L(f(\mathbf{x}), y)\right].$$

It is the function in $\mathcal{H}_R$ closest to f_{true}, and we call $d(f_R^*, f^*)$ the estimation bias.



Note that the realization is only shown in the visualization for didactic purposes but is not an element of $\mathcal{H}_0$.

Let's assume that $\hat{f}$ is an unbiased estimate of f^* (e.g., valid for linear regression), and we repeat the sampling process of $\hat{f}$.

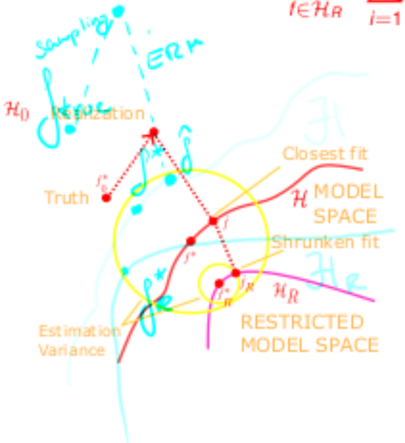


- We can measure the spread of sampled $\hat{f}$ around f^* via $\delta = \text{Var}_{\mathcal{D}}\left[d(f^*, \hat{f})\right]$ which we call the estimation variance.
- We visualize this as a circle around f^* with radius δ.

BIAS-VARIANCE TRADEOFF / 6

We repeat the previous construction in the restricted model space $\mathcal{H}_R$ and sample $\hat{f}_R$ such that

$$\hat{f}_R \in \arg\min_{f \in \mathcal{H}_R} \sum_{i=1}^{n} L\left(y^{(i)}, \hat{f}(\mathbf{x}^{(i)})\right).$$



- We can measure the spread of sampled $\hat{f}_R$ around f_R^* via $\delta = \text{Var}_{\mathcal{D}}\left[d(f_R^*, \hat{f}_R)\right]$ which we also call estimation variance.
- We observe that the increased bias results in a smaller estimation variance in $\mathcal{H}_R$ compared to $\mathcal{H}$.

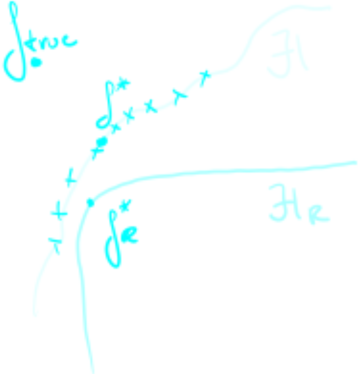
$$\hat{f} \in \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, \hat{f}(\mathbf{x}^{(i)})\right).$$

Note:

- $L : \mathcal{Y} \times \mathbb{R}^g \to \mathbb{R}$ is overloaded.
- The samples are only shown in the visualization for didactic purposes but are not an element of $\mathcal{H}$.

BIAS-VARIANCE TRADEOFF / 7

Let's assume that $\hat{f}$ is an unbiased estimate of f^* (e.g., valid for linear regression), and we repeat the sampling process of $\hat{f}$.



- We can measure the spread of sampled $\hat{f}$ around f^* via $\delta = \text{Var}_{\mathcal{D}}\left[d(f^*, \hat{f})\right]$ which we call the estimation variance.
- We visualize this as a circle around f^* with radius δ.

We repeat the previous construction in the restricted model space $\mathcal{H}_R$ and sample $\hat{f}_R$ such that

$$\hat{f}_R \in \underset{f \in \mathcal{H}_R}{\arg\min} \sum_{i=1}^{n} L\left(y^{(i)}, \hat{f}(\mathbf{x}^{(i)})\right).$$

- We can measure the spread of sampled $\hat{f}_R$ around f_R^* via $\delta = \text{Var}_{\mathcal{D}}\left[d(f^*, \hat{f}_R)\right]$ which we also call estimation variance.
- We observe that the increased bias results in a smaller estimation variance in $\mathcal{H}_R$ compared to $\mathcal{H}$.