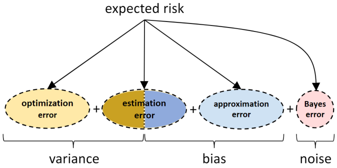
Introduction to Machine Learning

Advanced Risk Minimization
Bias-Variance 2:
Approximation and Estimation error



Learning goals

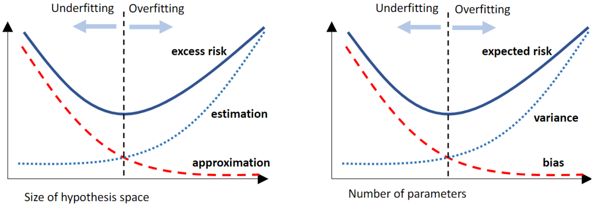
- Decomposing excess risk
- Into estimation, approx. and optim. error

APPROX./ESTIMATION ERROR ▸ Brown and Ali 2024

- BV decomp often confused with related (but different) decomp:

$$\underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{excess risk}} = \underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{approx. error}}$$

- Both commonly described using same figure and analogies

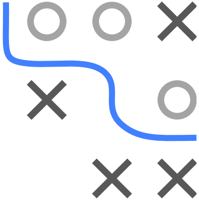


▸ Click for source

- BV decomp. only holds for certain losses, above is universal

APPROX./ESTIMATION ERROR

▸ Brown and Ali 2024



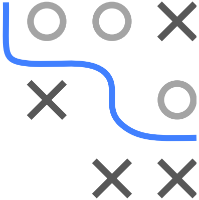
- Approx. error is a structural property of $\mathcal{H}$
- Estimation error is random due to dependence on data in $\hat{f}$
- Estimation error occurs as we choose $f \in \mathcal{H}$ with limited train data minimizing $\mathcal{R}_{\text{emp}}$ instead of $\mathcal{R}$
- Knowing $\hat{f}_{\mathcal{H}} \in \arg\inf_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$ assumes we found a global minimizer of $\mathcal{R}_{\text{emp}}$, which is often impossible (e.g. ANNs)
- In practice, optimizing $\mathcal{R}_{\text{emp}}$ gives us “best guess” $\tilde{f}_{\mathcal{H}} \in \mathcal{H}$ of $\hat{f}_{\mathcal{H}}$
- Can now decompose its excess risk finer as

$$\underbrace{\mathcal{R}(\tilde{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{excess risk}} = \underbrace{\mathcal{R}(\tilde{f}_{\mathcal{H}}) - \mathcal{R}(\hat{f}_{\mathcal{H}})}_{\text{optim. error}} + \underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{approx. error}}$$

- NB: Optim err. can be < 0, but $\mathcal{R}_{\text{emp}}(\tilde{f}_{\mathcal{H}}) \geq \mathcal{R}_{\text{emp}}(\hat{f}_{\mathcal{H}})$ always

APPROX./ESTIMATION ERROR

▸ Brown and Ali 2024



- We can further decompose estimation error more finely by defining the *centroid* model or “systematic” model part
- For $\hat{f}_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$ centroid model under L2 loss is mean prediction at each x over all $\mathcal{D}_n$, $f^{\circ}_{\mathcal{H}} := \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}[\hat{f}_{\mathcal{H}}]$
- With $f^{\circ}_{\mathcal{H}}$, can decompose expected estimation error as

$$\underbrace{\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}\left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^{*}_{\mathcal{H}})\right]}_{\text{expected estimation error}} = \underbrace{\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}\left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^{\circ}_{\mathcal{H}})\right]}_{\text{estimation variance}} + \underbrace{\mathcal{R}(f^{\circ}_{\mathcal{H}}) - \mathcal{R}(f^{*}_{\mathcal{H}})}_{\text{estimation bias}}$$

- Estimation bias measures distance of centroid model to risk minimizer over $\mathcal{H}$
- Estimation var. spread of ERM around centroid model induced by randomness due to $\mathcal{D}_n$

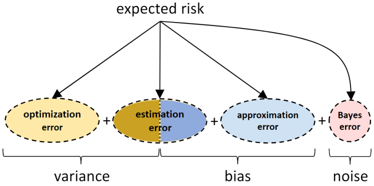
APPROX./ESTIMATION ERROR

▸ Brown and Ali 2024

- Can now connect derived quantities back to bias and variance
- Bias is not only approx. error and variance is not estimation error
- Many details skipped here, see paper!

$$\text{bias} = \text{approximation error} + \text{estimation bias}$$

$$\text{variance} = \text{optimization error} + \text{estimation variance}$$



▸ Click for source

- **NB**: For special case of LM and L2 loss, we have very small optim / numerical error and estimation bias; so both decomps agree