Bias–Variance Tradeoff

Statistics · Module 10 — Model Selection Learning

Node 10.1

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.1.

The bias–variance tradeoff is the foundational decomposition underlying all model selection procedures. It quantifies the tension between underfitting (high bias) and overfitting (high variance), establishing that no estimator can simultaneously minimize both for finite sample size. Every criterion in this module—AIC, BIC, cross-validation, regularization—is a procedure for navigating this tradeoff.

Formal Definition

Let $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ be an estimator of a regression function f_0, fitted from data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ where $Y_i = f_0(X_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and $\varepsilon_i \perp\!\!\!\perp \mathcal{D}_n$.

The **prediction risk** (conditional on x) is

$$R(\hat{f}_n, x) = \mathbb{E}_{\mathcal{D}_n, \varepsilon}\big[(Y - \hat{f}_n(x))^2\big].$$

Bias–Variance Decomposition. The prediction risk decomposes as

$$R(\hat{f}_n, x) = \underbrace{\big(\mathbb{E}_{\mathcal{D}_n}[\hat{f}_n(x)] - f_0(x)\big)^2}_{\mathrm{Bias}^2(\hat{f}_n(x))} + \underbrace{\mathrm{Var}_{\mathcal{D}_n}(\hat{f}_n(x))}_{\mathrm{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible noise}}.$$

The integrated prediction risk is $R(\hat{f}_n) = \mathbb{E}_X[R(\hat{f}_n, X)]$.

Intuition

Every estimator makes two kinds of errors. Bias is systematic: the estimator consistently misses the truth because the model class is too restrictive. Variance is random: the estimator fluctuates across different training samples because the model class is too flexible. The irreducible noise σ^2 is the floor—no estimator can beat it.

Simple models (e.g., linear regression with few features) have high bias and low variance. Complex models (e.g., high-degree polynomials, deep trees) have low bias and high variance. The optimal model complexity is the one that minimizes their sum, and this optimum shifts rightward as n grows.

Structural Role

The bias–variance tradeoff is the fundamental organizing principle of model selection. It determines the optimal model complexity as a function of sample size n: as $n \to \infty$, variance decreases at rate $O(1/n)$, allowing progressively richer models.

Dependencies: Relies on the mean squared error framework (Node 3.1) and the asymptotic theory of estimators (Node 7.3). **Enables:** All subsequent model selection criteria—AIC (Node 10.2), cross-validation (Node 10.3), regularization (Node 10.4)—are procedures for navigating this decomposition.

Mathematical Development

Full Proof of the Decomposition

Write $Y = f_0(x) + \varepsilon$ and let $\bar{f}(x) = \mathbb{E}_{\mathcal{D}_n}[\hat{f}_n(x)]$. Then

$$\mathbb{E}[(Y - \hat{f}_n(x))^2] = \mathbb{E}[(f_0(x) + \varepsilon - \hat{f}_n(x))^2].$$

Add and subtract $\bar{f}(x)$:

$$= \mathbb{E}\big[(\varepsilon + f_0(x) - \bar{f}(x) + \bar{f}(x) - \hat{f}_n(x))^2\big].$$

Let $A = \varepsilon$, $B = f_0(x) - \bar{f}(x)$, $C = \bar{f}(x) - \hat{f}_n(x)$. Then

$$\mathbb{E}[(A + B + C)^2] = \mathbb{E}[A^2] + B^2 + \mathbb{E}[C^2] + 2B\,\mathbb{E}[C] + 2\,\mathbb{E}[AC] + 2B\,\mathbb{E}[A].$$

Now: $\mathbb{E}[A] = \mathbb{E}[\varepsilon] = 0$, so the last term vanishes. $\mathbb{E}[C] = \bar{f}(x) - \mathbb{E}[\hat{f}_n(x)] = 0$, so the $2B\,\mathbb{E}[C]$ term vanishes. $\mathbb{E}[AC] = \mathbb{E}[\varepsilon(\bar{f}(x) - \hat{f}_n(x))] = 0$ by independence of ε from $\mathcal{D}_n$. Therefore

$$R(\hat{f}_n, x) = \sigma^2 + (f_0(x) - \bar{f}(x))^2 + \mathbb{E}[(\hat{f}_n(x) - \bar{f}(x))^2] = \sigma^2 + \text{Bias}^2 + \text{Variance}.$$

Parametric MSE Decomposition

For a parametric estimator $\hat{\theta}_n$ of $\theta_0 \in \mathbb{R}^k$,

$$\text{MSE}(\hat{\theta}_n) = \|\text{Bias}(\hat{\theta}_n)\|^2 + \text{Tr}\,\text{Cov}(\hat{\theta}_n).$$

The Cramer–Rao bound (Node 3.4) gives $\text{Tr}\,\text{Cov}(\hat{\theta}_n) \geq \text{Tr}\, I(\theta_0)^{-1}$ for unbiased estimators. A biased estimator may achieve lower total MSE when the bias term is small relative to the variance reduction. Ridge regression (Node 10.4) exploits this: it introduces bias $O(\lambda)$ while reducing variance by $O(\lambda^2)$ near the optimal λ.

Approximation vs. Estimation Error

The total excess risk decomposes as

$$R(\hat{f}_n) - R(f_0) = \underbrace{R(f^*_{\mathcal{F}}) - R(f_0)}_{\text{Approximation error}} + \underbrace{R(\hat{f}_n) - R(f^*_{\mathcal{F}})}_{\text{Estimation error}},$$

where $f^*_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} R(f)$. Approximation error decreases as $\mathcal{F}$ grows (richer model classes); estimation error increases with the complexity of $\mathcal{F}$ and decreases with n. For a model class with VC dimension d, classical bounds give estimation error $O(\sqrt{d/n})$.

Double Descent and Interpolation

In overparameterized models ($p > n$), the classical U-shaped risk curve breaks down. The **double descent** phenomenon: as p increases past n, the minimum-norm interpolating estimator (which achieves zero training error) can have decreasing risk. The bias–variance decomposition still holds, but the variance term behaves non-monotonically near $p = n$ (the interpolation threshold) where the condition number of $X^\top X$ diverges. For $p \gg n$, implicit regularization from the minimum-norm constraint reduces variance, producing a second descent in risk.

This does not violate the bias–variance tradeoff; rather, it reveals that the effective complexity of an estimator is not simply the number of parameters but depends on the geometry of the estimator class and the implicit constraints of the optimization procedure.

Optimism and Degrees of Freedom

For a linear smoother $\hat{Y} = HY$ with hat matrix H, the **optimism** (gap between in-sample and out-of-sample risk) is

$$\text{Optimism} = \frac{2\sigma^2}{n}\text{Tr}(H) = \frac{2\sigma^2}{n}\,\text{df},$$

where $\text{df} = \text{Tr}(H)$ is the **effective degrees of freedom**. This connects the bias–variance tradeoff to model selection: AIC (Node 10.2) corrects for exactly this optimism, and Mallows' C_p uses df to penalize complex models.

For OLS, $\text{df} = p$ (number of parameters). For ridge regression, $\text{df}(\lambda) = \sum_j d_j^2/(d_j^2 + \lambda) < p$. For k-NN, $\text{df} = n/k$.

Rates for Specific Estimators

k-Nearest Neighbors. For k-NN regression in $\mathbb{R}^d$ with Lipschitz f_0:

$$\text{Bias}^2 = O(k^{-2/d} \cdot n^{-2/d}), \quad \text{Variance} = O(\sigma^2/k).$$

Optimizing over k gives $k^* \sim n^{2/(2+d)}$ and risk $R - \sigma^2 = O(n^{-2/(2+d)})$, exhibiting the curse of dimensionality.

Polynomial Regression of Degree p. For a degree-p polynomial fit to data from a smooth function on $[0, 1]$ with m continuous derivatives:

$$\text{Bias}^2 = O(n^{-2m/(2m+1)}) \text{ (when } p \sim n^{1/(2m+1)}), \quad \text{Variance} = O(p/n).$$

The optimal degree balances the two at the minimax rate $n^{-2m/(2m+1)}$.

Worked Examples

Example 1: Polynomial Regression

Consider $Y_i = \sin(X_i) + \varepsilon_i$ with $X_i \sim \text{Uniform}(0, 2\pi)$, $\varepsilon_i \sim \mathcal{N}(0, 0.25)$, and $n = 50$. Fit polynomials of degree $p = 1, 3, 10, 25$.

Degree 1 (linear): $\bar{f}(x) \approx a + bx$. Since $\sin(x)$ is nonlinear, Bias^2 is large. Two parameters yield very low variance. The fit systematically misses the curvature.

Degree 3: Captures the dominant shape of $\sin(x)$ well. Bias is moderate (the Taylor expansion $\sin(x) \approx x - x^3/6$ is a degree-3 polynomial). Variance with 4 parameters on 50 points is still small.

Degree 10: Bias is near zero (degree-10 polynomial can approximate sin on a compact interval to high accuracy). Variance increases: 11 parameters on 50 points. Moderate overfitting at the boundaries.

Degree 25: Bias is essentially zero. Variance is very large: 26 parameters on 50 points means the fit interpolates noise. Test MSE is much worse than degree 3.

Computing $\widehat{\text{MSE}}$ on an independent test set of size 1000: degree 1 gives ≈ 0.35, degree 3 gives ≈ 0.27, degree 10 gives ≈ 0.30, degree 25 gives ≈ 1.2. The minimum is at an intermediate complexity.

Example 2: k-NN in Dimension $d = 1$

Data: $Y_i = X_i^2 + \varepsilon_i$, $X_i \sim \text{Uniform}(0, 1)$, $\sigma^2 = 0.1$, $n = 200$. Evaluate the prediction at $x = 0.5$.

$k = 1$**:** The predictor is Y_{j^*} where $j^* = \arg\min_j |X_j - 0.5|$. Bias ≈ 0 (nearest neighbor is close to x). Variance $= \sigma^2 = 0.1$.

$k = 50$**:** The predictor averages 50 neighbors. If these span roughly $[0.35, 0.65]$, then $\bar{f}(0.5) \approx \mathbb{E}[X^2 \mid X \in [0.35, 0.65]] \approx 0.253$ versus $f_0(0.5) = 0.25$. Bias ≈ 0.003. Variance $\approx \sigma^2/50 = 0.002$. MSE ≈ 0.002.

$k = 200$**:** All data used. $\bar{f}(0.5) = \mathbb{E}[X^2] = 1/3 \approx 0.333$ versus $f_0(0.5) = 0.25$. Bias$^2 \approx 0.007$. Variance $\approx \sigma^2/200 = 0.0005$. MSE ≈ 0.007.

The optimal k is intermediate: large enough to average away noise, small enough to avoid smoothing bias. Theory gives $k^* \sim n^{2/3} \approx 34$ for $d = 1$.

Cross-Links

AI. The bias–variance tradeoff motivates regularization (weight decay, dropout, early stopping) and architecture search in deep learning. The double descent phenomenon in overparameterized models challenges the classical U-shaped risk curve but can be explained through implicit regularization effects.

Economics. In forecasting and structural estimation, the tradeoff between model parsimony (few parameters, interpretable) and fit (many parameters, flexible) governs model specification. Shrinkage estimators in empirical Bayes (James–Stein) dominate the MLE in dimension ≥ 3 precisely because they exploit this tradeoff.

Linear Algebra. The singular value decomposition of the design matrix $X = U\Sigma V^\top$ provides the geometric view: ridge regression shrinks the components along small singular values (where variance is large) proportionally more, and the bias–variance tradeoff is controlled by the spectrum of $X^\top X$.

Quantum Mechanics. The uncertainty principle $\Delta x \cdot \Delta p \geq \hbar/2$ is structurally analogous: simultaneous localization in two conjugate domains is impossible. In quantum state tomography, the bias–variance tradeoff governs the choice of measurement basis and estimator complexity.

Structural Compression

Invariant

MSE $=$ Bias^2 + Variance $+ \sigma^2$: the three components are orthogonal in the L^2 decomposition of prediction error.

Mechanism

Second-moment expansion of the squared error; all cross terms vanish by independence of the noise ε from the training-data-dependent estimator $\hat{f}_n$ and by the definition of the bias as $\mathbb{E}[\hat{f}_n(x)] - f_0(x)$.

Cross-System Echo

The bias–variance tradeoff is the statistical instance of the approximation-generalization tradeoff in statistical learning theory (Vapnik). In AI, it motivates regularization, dropout, and early stopping as variance-reduction techniques. In economics, the tradeoff between model parsimony and fit appears in forecasting and structural estimation. In physics, the uncertainty principle is the conjugate-variable analogue: precision in one domain costs precision in the other.

Exercises

1. Prove the bias–variance decomposition in the multivariate case: for $\hat{\theta}_n \in \mathbb{R}^k$, show that $\mathbb{E}[\|\hat{\theta}_n - \theta_0\|^2] = \|\mathrm{Bias}(\hat{\theta}_n)\|^2 + \mathrm{Tr}\,\mathrm{Cov}(\hat{\theta}_n)$.
2. For the k-NN estimator in dimension d, derive the rates $\mathrm{Bias}^2 = O(k/n)^{2/d}$ and $\mathrm{Variance} = O(\sigma^2/k)$ assuming f_0 is Lipschitz and X has a bounded density.
3. Consider the James–Stein estimator $\hat{\theta}^{\mathrm{JS}} = (1-(k-2)\sigma^2/\|\bar{X}\|^2)\bar{X}$ for $\theta \in \mathbb{R}^k$, $k \geq 3$. Show that its MSE is strictly less than the MLE's MSE $k\sigma^2/n$ for every θ, and explain this in terms of the bias–variance tradeoff.
4. For a linear smoother $\hat{Y} = HY$, show that the in-sample optimism is $\mathrm{Optimism} = (2\sigma^2/n)\,\mathrm{Tr}(H)$, where $\mathrm{Tr}(H)$ is the effective degrees of freedom.
5. Suppose you observe $n = 100$ points from $Y = \sin(2\pi X) + \varepsilon$ with $\sigma = 0.5$. Compute the bias, variance, and MSE at $x = 0.5$ for the constant estimator $\hat{f}(x) = \bar{Y}$ and for the local average over the 10 nearest neighbors.
6. In ridge regression with design matrix X having singular values $d_1 \geq \cdots \geq d_p$, express both the bias and variance in terms of $\{d_j\}$, λ, and the true coefficients β_0. Show that the variance is $\sigma^2 \sum_j d_j^2/(d_j^2+\lambda)^2$.
7. Argue, structurally, why the bias–variance decomposition must hold for any loss function admitting a second-moment expansion, and identify conditions under which it fails (e.g., non-squared losses, dependent noise).

Statistics · Module 10 — Model Selection Learning · Node 10.1
Concepts: bias-variance-tradeoff, expectation-value, variance-and-covariance
Prerequisites: 3.1, 7.3
Enables: 10.2, 10.3, 10.4

hbar.university — own.your.cognition