1 Bias Variance Decomposition

Throughout, the training set $\mathcal{D}$ has N examples, the rows of $\mathbf{X}_\mathcal{D}$ are the inputs $\mathbf{x}_n^\top$, and expectations over $\mathcal{D}$ average over both the training inputs and the noise in the training targets. We write $\bar{t}(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$ and $\bar{y}(\mathbf{x}) = \mathbb{E}_\mathcal{D}[y(\mathbf{x};\mathcal{D})]$. The question's formulas use $(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}$, so we assume $\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D}$ is invertible for the training sets considered and that the required expectations are finite. The test example is independent of the training set, and the Gaussian noise is independent of the inputs. We retain the question's term "bias" for $\mathbb{E}_\mathbf{x}[(\bar{y}(\mathbf{x}) - \bar{t}(\mathbf{x}))^2]$; this nonnegative quantity is conventionally called the *squared bias*.

Problem 1.1. The noise ϵ has mean zero and does not depend on $\mathbf{x}$, so

$$\bar{t}(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \mathbb{E}[\mathbf{w}_*^\top\mathbf{x} + \epsilon \,|\, \mathbf{x}] = \mathbf{w}_*^\top\mathbf{x} + \mathbb{E}[\epsilon] = \mathbf{w}_*^\top\mathbf{x} = \mathbf{x}^\top\mathbf{w}_*,$$

where the last step uses $\mathbf{w}_*^\top\mathbf{x} = \mathbf{x}^\top\mathbf{w}_*$ (both are the same scalar). Equivalently, $p(t|\mathbf{x}) = \mathcal{N}(t|\mathbf{w}_*^\top\mathbf{x}, 1)$, and $\bar{t}(\mathbf{x})$ is the mean of this Gaussian.

Problem 1.2. Each training target follows the same model, $t_n = \mathbf{x}_n^\top\mathbf{w}_* + \epsilon_n$, with independent noise $\epsilon_n \sim \mathcal{N}(0,1)$. Stacking the N equations row by row gives $\mathbf{t}_\mathcal{D} = \mathbf{X}_\mathcal{D}\mathbf{w}_* + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$. The components of $\boldsymbol{\epsilon}$ are independent with mean 0 and variance 1, so $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Problem 1.3. Substitute $\mathbf{t}_\mathcal{D}$ from part 2 into $\mathbf{w}_\mathcal{D}$:

$$\mathbf{w}_\mathcal{D} = (\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top(\mathbf{X}_\mathcal{D}\mathbf{w}_* + \boldsymbol{\epsilon}) = \mathbf{w}_* + (\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top\boldsymbol{\epsilon},$$

so

$$y(\mathbf{x};\mathcal{D}) = \mathbf{x}^\top\mathbf{w}_* + \mathbf{x}^\top(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top\boldsymbol{\epsilon}.$$

The first term does not depend on $\mathcal{D}$. For the second term, first average over the noise with the training inputs held fixed. The noise is independent of the inputs and has mean zero, so $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}_\mathcal{D}] = \mathbf{0}$ and

$$\mathbb{E}\big[\mathbf{x}^\top(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top\boldsymbol{\epsilon} \,\big|\, \mathbf{X}_\mathcal{D}\big] = \mathbf{x}^\top(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^\top\,\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}_\mathcal{D}] = 0.$$

Averaging this zero over the training inputs (law of total expectation) still gives zero. Therefore

$$\boxed{\bar{y}(\mathbf{x}) = \mathbf{x}^\top\mathbf{w}_*.}$$

Problem 1.4. By parts 1 and 3, $\bar{y}(\mathbf{x}) = \bar{t}(\mathbf{x})$ for every $\mathbf{x}$, so

$$\text{bias} = \mathbb{E}_\mathbf{x}\big[(\bar{y}(\mathbf{x}) - \bar{t}(\mathbf{x}))^2\big] = \mathbb{E}_\mathbf{x}[0] = 0.$$

Averaged over training sets, the least-squares prediction equals the true regression function. This relies on the assumption: the true $\bar{t}(\mathbf{x})$ is linear in $\mathbf{x}$, so it can be represented by the model. The expected test error is still positive, because of the variance and noise terms.

Problem 2. Use $\mathbf{I} = \mathbf{Q}\mathbf{Q}^{-1}$ and factor out $\mathbf{Q}$ on the left and $\mathbf{Q}^{-1}$ on the right:

$$\begin{aligned}\mathbf{A} + \lambda\mathbf{I} &= \mathbf{Q}\mathbf{S}\mathbf{Q}^{-1} + \lambda\mathbf{Q}\mathbf{Q}^{-1} \\ &= \mathbf{Q}\mathbf{S}\mathbf{Q}^{-1} + \mathbf{Q}(\lambda\mathbf{I})\mathbf{Q}^{-1} \\ &= \mathbf{Q}(\mathbf{S} + \lambda\mathbf{I})\mathbf{Q}^{-1} \\ &= \mathbf{Q}\mathbf{S}'\mathbf{Q}^{-1},\end{aligned}$$

since $\mathbf{S} + \lambda\mathbf{I} = \mathrm{diag}(s_1 + \lambda, \ldots, s_n + \lambda) = \mathbf{S}'$. Adding $\lambda\mathbf{I}$ keeps the eigenvectors and shifts every eigenvalue by λ.

Problem 3. Write $\mathbf{A}_\mathcal{D} = \mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D}$. As in Problem 1.3, substitute $\mathbf{t}_\mathcal{D} = \mathbf{X}_\mathcal{D}\mathbf{w}_* + \boldsymbol{\epsilon}$:

$$\mathbf{w}_\mathcal{D} = (\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{A}_\mathcal{D}\,\mathbf{w}_* + (\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{X}_\mathcal{D}^\top\boldsymbol{\epsilon}.$$

The noise term has expectation zero for the same reason as in Problem 1.3, so

$$\bar{y}(\mathbf{x}) = \mathbf{x}^\top\,\mathbb{E}_\mathcal{D}\big[(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{A}_\mathcal{D}\big]\,\mathbf{w}_*.$$

Simplify with the eigendecomposition. $\mathbf{A}_\mathcal{D}$ is symmetric positive semidefinite, so $\mathbf{A}_\mathcal{D} = \mathbf{Q}\mathbf{S}\mathbf{Q}^{-1}$ with eigenvalues $s_i \geq 0$ (here $\mathbf{Q}$ and $\mathbf{S}$ depend on $\mathcal{D}$). By Problem 2, $\mathbf{A}_\mathcal{D} + \lambda\mathbf{I} = \mathbf{Q}\mathbf{S}'\mathbf{Q}^{-1}$, and for $\lambda > 0$ all its eigenvalues $s_i + \lambda$ are positive, so $(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1} = \mathbf{Q}\mathbf{S}'^{-1}\mathbf{Q}^{-1}$. Then

$$(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{A}_\mathcal{D} = \mathbf{Q}\mathbf{S}'^{-1}\mathbf{Q}^{-1}\,\mathbf{Q}\mathbf{S}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{S}'^{-1}\mathbf{S}\,\mathbf{Q}^{-1} = \mathbf{Q}\,\mathrm{diag}\Big(\frac{s_i}{s_i + \lambda}\Big)\mathbf{Q}^{-1},$$

$$(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{A}_\mathcal{D} - \mathbf{I} = \mathbf{Q}\,\mathrm{diag}\Big(\frac{s_i}{s_i + \lambda} - 1\Big)\mathbf{Q}^{-1} = -\mathbf{Q}\,\mathrm{diag}\Big(\frac{\lambda}{s_i + \lambda}\Big)\mathbf{Q}^{-1}.$$

Since $\bar{t}(\mathbf{x}) = \mathbf{x}^\top\mathbf{w}_* = \mathbf{x}^\top\mathbf{I}\mathbf{w}_*$,

$$\bar{y}(\mathbf{x}) - \bar{t}(\mathbf{x}) = -\mathbf{x}^\top\mathbf{B}_\lambda\mathbf{w}_*, \qquad \mathbf{B}_\lambda = \mathbb{E}_\mathcal{D}\Big[\mathbf{Q}\,\mathrm{diag}\Big(\frac{\lambda}{s_i + \lambda}\Big)\mathbf{Q}^{-1}\Big] = \lambda\,\mathbb{E}_\mathcal{D}\big[(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\big].$$

The bias is therefore

$$\boxed{\text{bias} = \mathbb{E}_\mathbf{x}\Big[\big(\mathbf{x}^\top\mathbf{B}_\lambda\mathbf{w}_*\big)^2\Big] = \lambda^2\,\mathbb{E}_\mathbf{x}\Big[\big(\mathbf{x}^\top\mathbb{E}_\mathcal{D}[(\mathbf{X}_\mathcal{D}^\top\mathbf{X}_\mathcal{D} + \lambda\mathbf{I})^{-1}]\,\mathbf{w}_*\big)^2\Big].}$$

Effect of λ. Along each eigenvector of $\mathbf{A}_\mathcal{D}$, ridge regression shrinks the fitted weights by the factor $s_i/(s_i + \lambda)$, so a fraction $\lambda/(s_i + \lambda)$ of the true weight in that direction is lost.

- $\lambda = 0$ (with $\mathbf{A}_\mathcal{D}$ invertible, so every $s_i > 0$): every fraction is 0, $\mathbf{B}_0 = \mathbf{0}$, and the bias is 0, as in Problem 1.
- $\lambda > 0$: the bias is positive whenever $\mathbf{w}_* \neq \mathbf{0}$ and $\mathbb{E}_\mathbf{x}[\mathbf{x}\mathbf{x}^\top]$ is positive definite. Indeed, $\mathbf{B}_\lambda$ is an average of positive definite matrices, so $\mathbf{v} = \mathbf{B}_\lambda\mathbf{w}_* \neq \mathbf{0}$ and the bias equals $\mathbf{v}^\top\mathbb{E}_\mathbf{x}[\mathbf{x}\mathbf{x}^\top]\mathbf{v} > 0$.
- Each fraction $\lambda/(s_i + \lambda)$ increases with λ and tends to 1 as $\lambda \to \infty$. Then $\mathbf{B}_\lambda \to \mathbf{I}$, $\bar{y}(\mathbf{x}) \to 0$, and the bias tends to $\mathbb{E}_\mathbf{x}[(\mathbf{x}^\top\mathbf{w}_*)^2]$: in the limit, the model predicts 0 everywhere.

- The fraction is largest in directions with small $s_i = \sum_n (\mathbf{q}_i^\top \mathbf{x}_n)^2$ (taking unit-length eigenvectors $\mathbf{q}_i$), the directions in which the training inputs have small components.

Variance. For fixed training inputs (only the noise varies between training sets), regularisation reduces the variance: the noise term $(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{X}_\mathcal{D}^\top\boldsymbol{\epsilon}$ has covariance $(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}\mathbf{A}_\mathcal{D}(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1} = \mathbf{Q}\,\mathrm{diag}\big(s_i/(s_i + \lambda)^2\big)\mathbf{Q}^{-1}$, whose eigenvalues decrease as λ grows. Prediction variance therefore decreases at every fixed test input. When the training inputs also vary, the total variance includes variation of the conditional mean, so this argument alone does not establish monotonicity of the total variance.

Clarification about the tradeoff. Regularisation introduces shrinkage bias and reduces the conditional variance above, but the prediction bias need not increase monotonically with λ for an arbitrary test-input distribution. Contributions from different directions can cancel. For example, with fixed $\mathbf{A}_\mathcal{D} = \mathrm{diag}(1, 4)$, $\mathbf{w}_* = (1, 1)^\top$ and test input $\mathbf{x} = (1, -5/2)^\top$, the squared bias is

$$\left(\frac{\lambda}{1+\lambda} - \frac{5\lambda}{2(4+\lambda)}\right)^2.$$

It equals $1/324$ at $\lambda = 1/2$ and 0 at $\lambda = 1$. The increasing-bias conclusion does hold in the eigenbasis approximation below. In practice λ is chosen by validation.

Bonus: an explicit formula. To see the effect of λ in a formula, approximate $\mathbb{E}_\mathcal{D}[(\mathbf{A}_\mathcal{D} + \lambda\mathbf{I})^{-1}]$ by $(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}$, where $\boldsymbol{\Sigma} = \mathbb{E}_\mathcal{D}[\mathbf{A}_\mathcal{D}]$. This is exact when every training set has the same inputs and only the noise varies; otherwise it is an approximation, which is close when $\mathbf{A}_\mathcal{D}$ varies little around its mean (for example, when N is much larger than the input dimension). If the N training inputs and the test input come from the same distribution, $\boldsymbol{\Sigma} = \sum_n \mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top] = N\,\mathbb{E}_\mathbf{x}[\mathbf{x}\mathbf{x}^\top]$. Then

$$\text{bias} \approx \lambda^2\,\mathbb{E}_\mathbf{x}\Big[\mathbf{w}_*^\top(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{x}\mathbf{x}^\top(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{w}_*\Big] = \frac{\lambda^2}{N}\,\mathbf{w}_*^\top(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{w}_*.$$

$\boldsymbol{\Sigma}$ is symmetric positive semidefinite, so it has orthonormal eigenvectors: $\boldsymbol{\Sigma} = \mathbf{V}\,\mathrm{diag}(r_i)\mathbf{V}^\top$ with $\mathbf{V}^{-1} = \mathbf{V}^\top$ and $r_i \geq 0$. Let $\mathbf{u} = \mathbf{V}^\top\mathbf{w}_*$ be the true weights in this basis. All three matrices share the eigenvectors, so

$$\boxed{\text{bias} \approx \sum_i \underbrace{\frac{r_i u_i^2}{N}}_{\text{signal in direction } i} \Big(\frac{\lambda}{r_i + \lambda}\Big)^2.}$$

The mean square of the true function is $\mathbb{E}_\mathbf{x}[(\mathbf{x}^\top\mathbf{w}_*)^2] = \mathbf{w}_*^\top\frac{\boldsymbol{\Sigma}}{N}\mathbf{w}_* = \sum_i r_i u_i^2/N$, so $r_i u_i^2/N$ is the part of it that comes from direction i. The bias is the fraction $\big(\lambda/(r_i + \lambda)\big)^2$ of each part:

- As $\lambda \to 0$, the bias tends to 0. As $\lambda \to \infty$, it tends to $\mathbb{E}_\mathbf{x}[(\mathbf{x}^\top\mathbf{w}_*)^2]$. In between it increases with λ, since each fraction does.

- Directions with $r_i \ll \lambda$ lose almost all of their part, but that part is itself small when r_i is small; directions with $r_i \gg \lambda$ lose little. For a fixed u_i^2, the contribution of direction i is largest when $r_i = \lambda$.