The bias variance tradeoff

The world we think of

$$
\begin{aligned}
&\text{Data-point :} && \{x, y\} \sim p(x, y) \\
&\text{Data-set :} && D = (X, Y) \sim p^n \\
&\text{Model class :} && \mathcal{A} \\
&\text{Model :} && h_D(x) \in \mathcal{A} \\
&\text{Expected model :} && \bar{h}(x) = \int_D h_D Pr(D)\,\mathrm{d}h_D = E_D[h_D] \\
&\text{Expected label :} && \bar{y}(x) = \int_y y Pr(y|x)\,\mathrm{d}y = E_{y|x}[y]
\end{aligned}
$$

The bias variance tradeoff is then encapsulated in

$$
\begin{aligned}
\underbrace{E_{x,y,D}\left[\left(h_D(x) - y\right)^2\right]}_{\text{Squared error}} &= \underbrace{E_{x,D}\left[\left(h_D(x) - \bar{h}(x)\right)^2\right]}_{\text{variance}} \\
&+ \underbrace{E_x\left[\left(\bar{h}(x) - \bar{y}(x)\right)^2\right]}_{\text{bias}^2} \\
&+ \underbrace{E_{x,y}\left[\left(\bar{y}(x) - y\right)^2\right]}_{\text{noise}}
\end{aligned}
$$

This is all explained in much more depth by the creators of CS4780 at Cornell http://www.cs.cornell.edu/courses/cs4780/2017sp/lectures/lecturenote11.html

So where's the tradeoff?

If we have $y = f(x) + e$ and consider measure related to the total error, the MSE, defined as

$$MSE(x) = \text{squared error} - \textit{noise}$$

We have a nice relationship

$$\begin{aligned} MSE &= \text{Var} + \text{bias}^2 && \Rightarrow \\ \sqrt{MSE}^2 &= \sigma^2 + \text{bias}^2 && \Rightarrow \\ RMSE^2 &= \sigma^2 + \text{bias}^2 \end{aligned}$$

It turns out that σ and bias are orthogonal vectors $\Rightarrow$ this is a right-angled triangle.

Imagine *RMSE* is minimized (no low hanging fruits to inependently lower either bias or variance). The relation between bias and σ is then similar to how two legs in a triangle can be scaled while preserving the length of the hypothenuse; clearly a tradeoff.