Deep learning

2.3. Bias-variance dilemma

François Fleuret

https://fleuret.org/dlc/

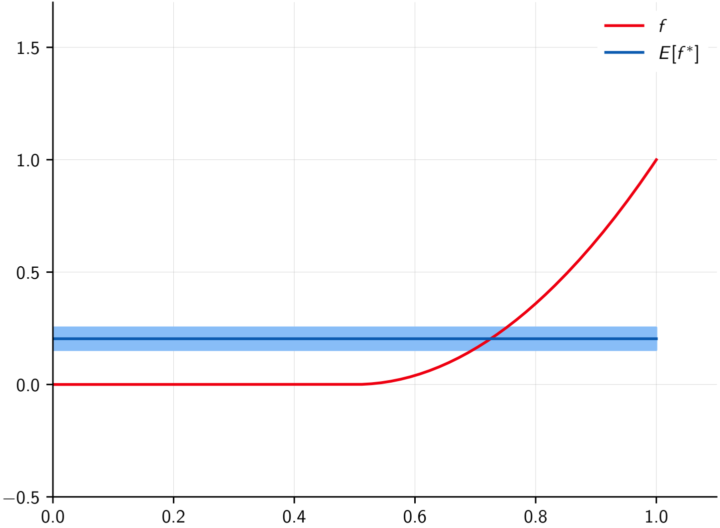
UNIVERSITÉ DE GENÈVE

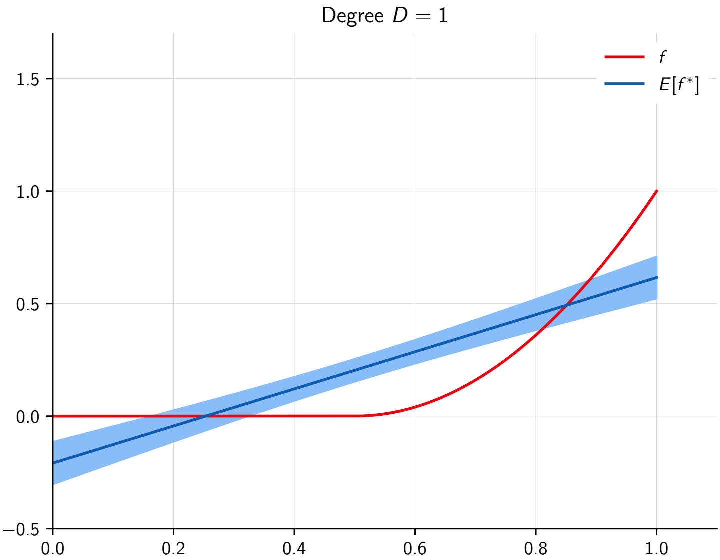
We can visualize over-fitting for our polynomial regression by generating multiple training sets $\mathscr{D}_1, \dots, \mathscr{D}_M$, training as many models $f_1, \dots, f_M$, and computing empirically the mean and standard deviation of the prediction at every point.

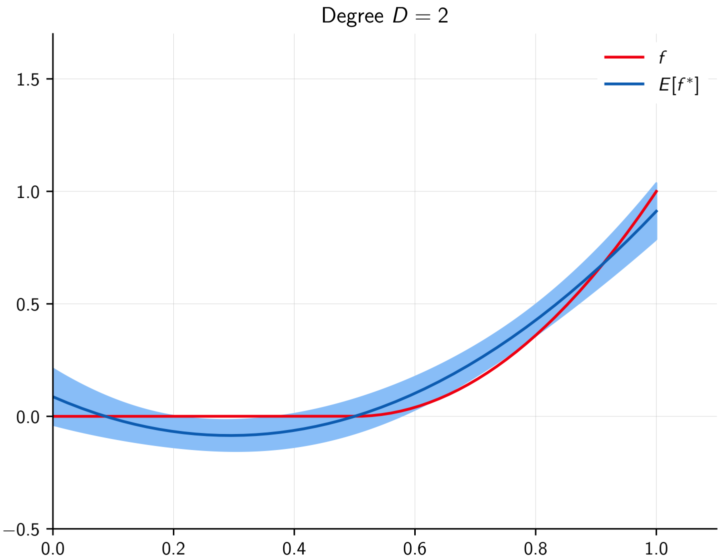
We can visualize over-fitting for our polynomial regression by generating multiple training sets $\mathscr{D}_1, \dots, \mathscr{D}_M$, training as many models $f_1, \dots, f_M$, and computing empirically the mean and standard deviation of the prediction at every point.

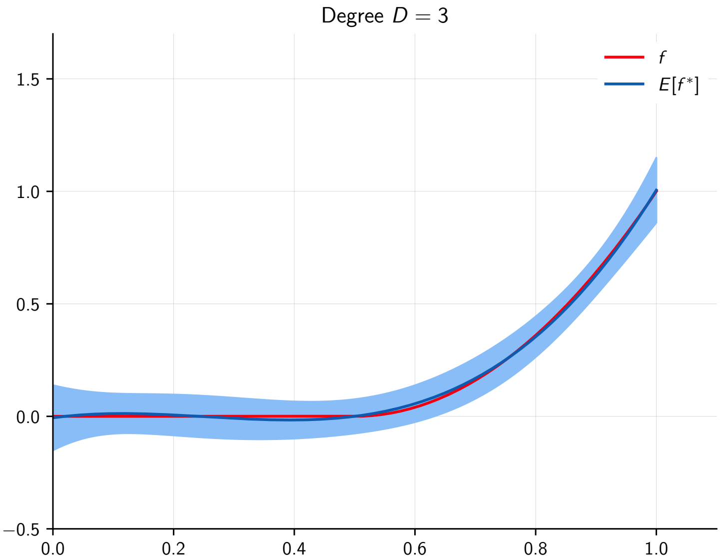
As we will see, when the capacity increases, or the regularization decreases, the mean of the predicted value gets right on target, but the prediction varies more across runs.

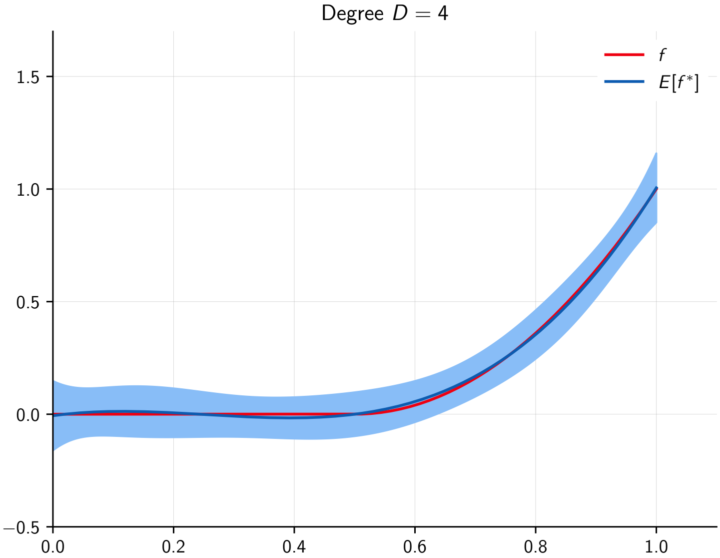
Degree $D = 0$

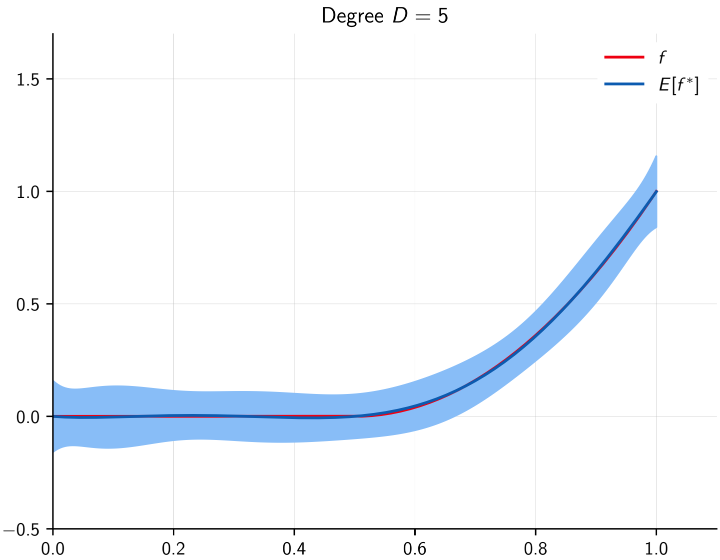


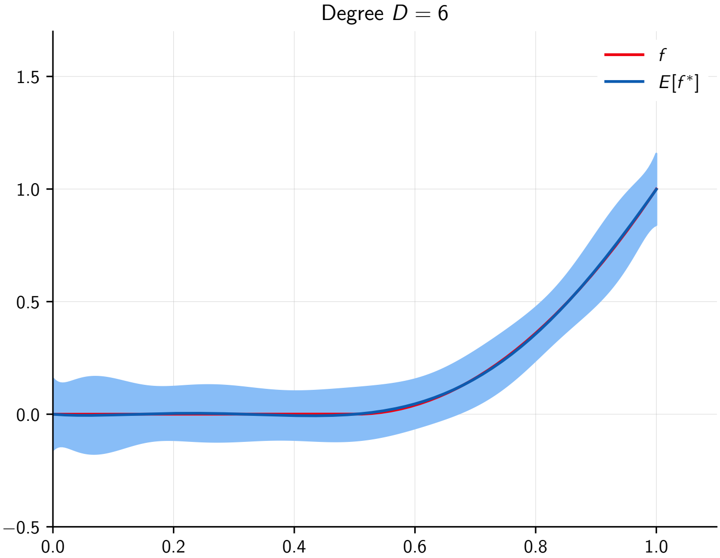

Degree $D = 1$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

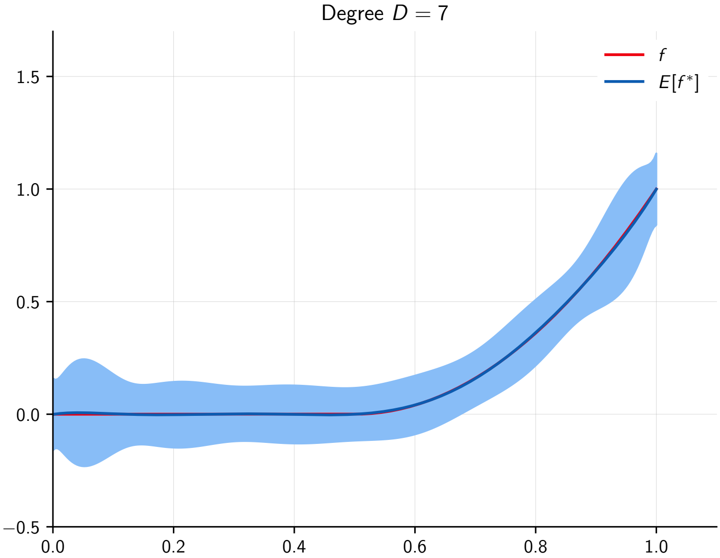

Degree $D = 2$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


Degree $D = 3$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

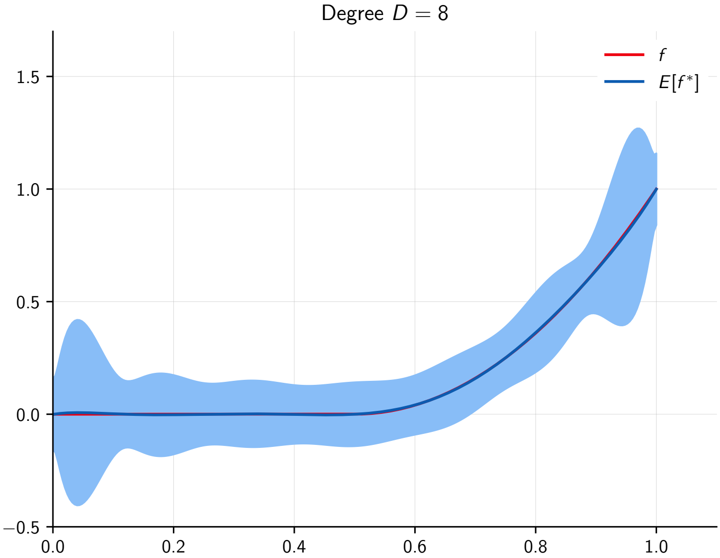

Degree $D = 4$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


Degree $D = 5$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

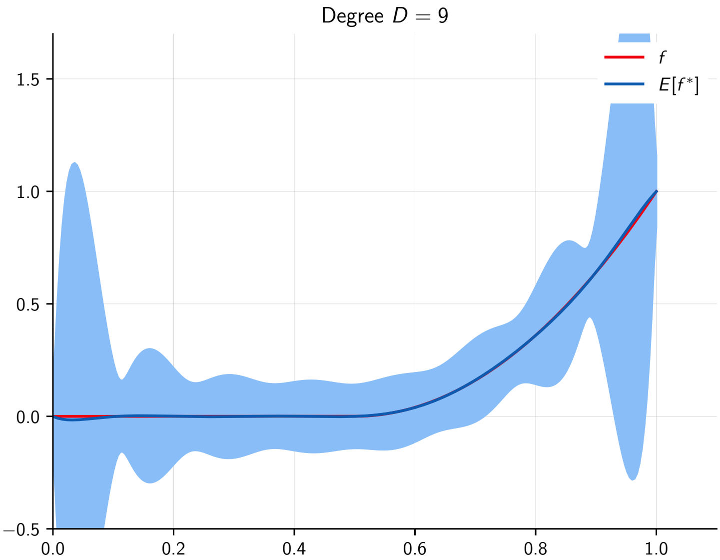

Degree $D = 6$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

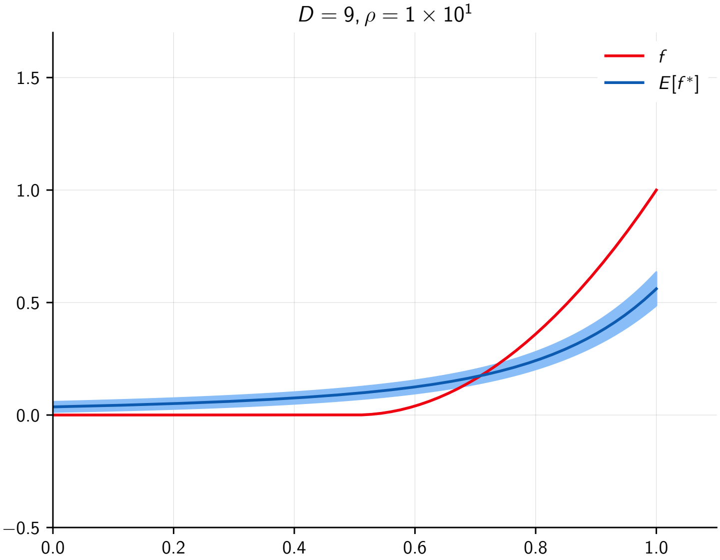


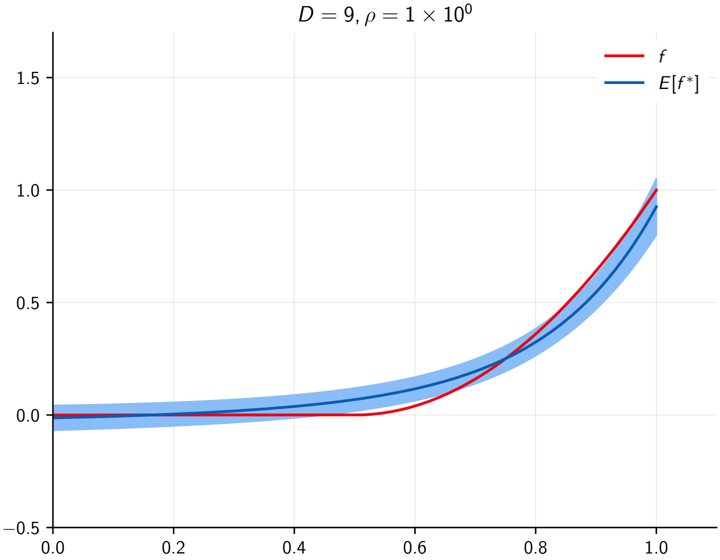
Degree $D = 7$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0



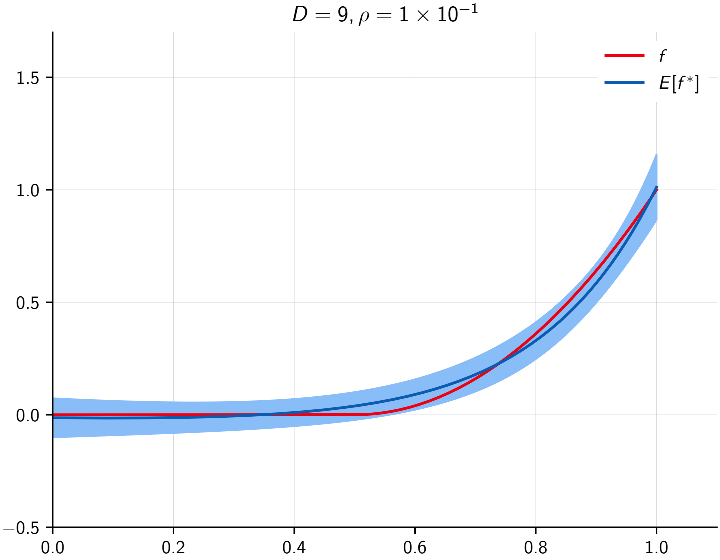
Degree $D = 8$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


Degree $D = 9$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

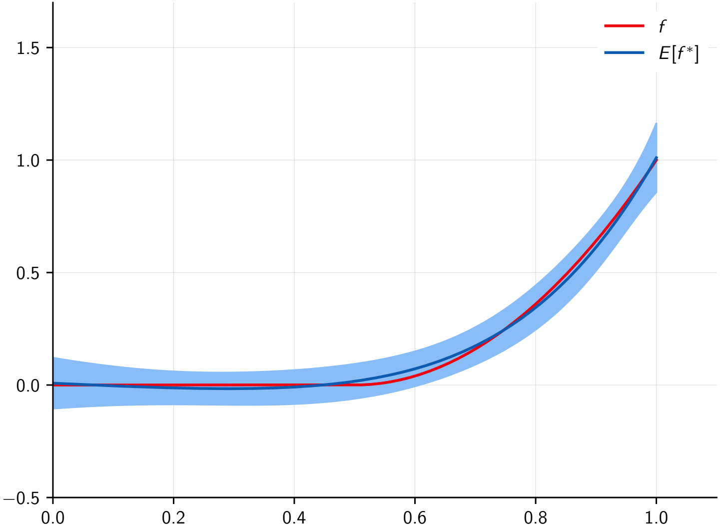

$D = 9, \rho = 1 \times 10^1$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0



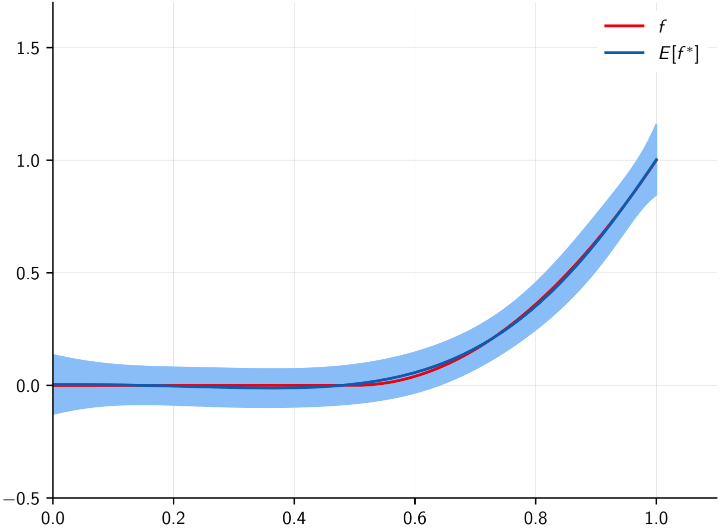
$D = 9, \rho = 1 \times 10^0$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

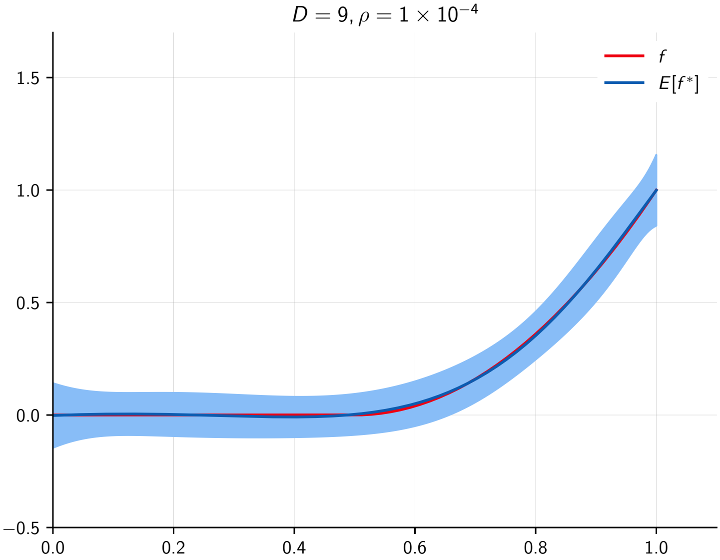

$D = 9, \rho = 1 \times 10^{-1}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

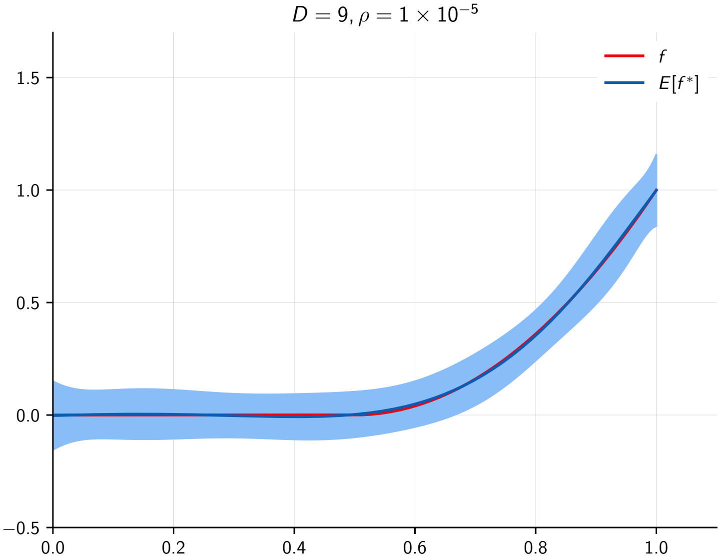
$D = 9, \rho = 1 \times 10^{-2}$



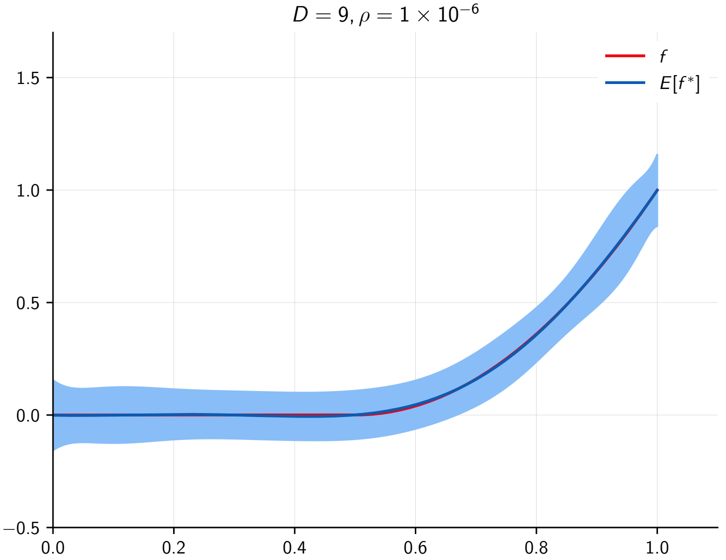
$D = 9, \rho = 1 \times 10^{-3}$



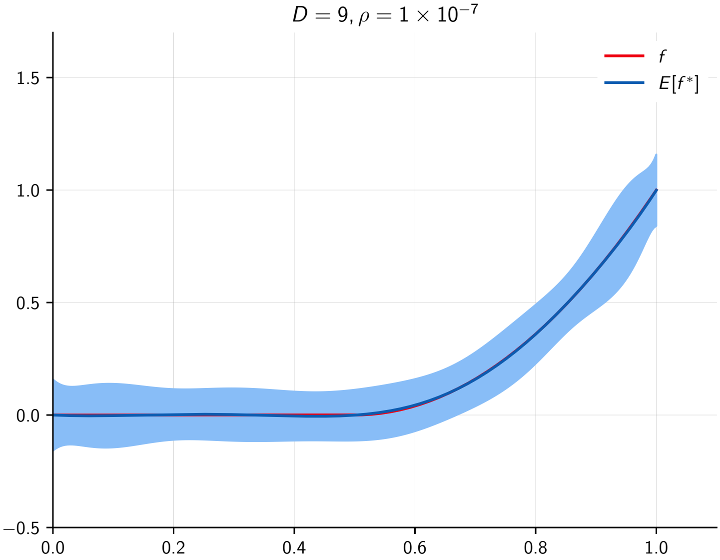

$D = 9, \rho = 1 \times 10^{-4}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0



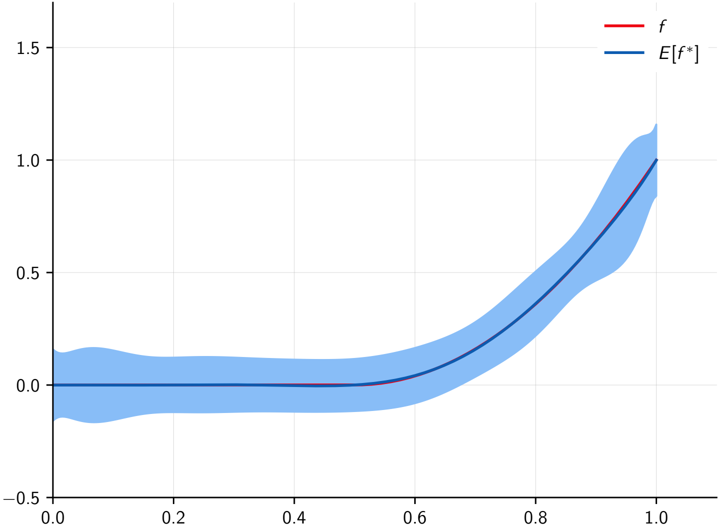
$D = 9, \rho = 1 \times 10^{-5}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

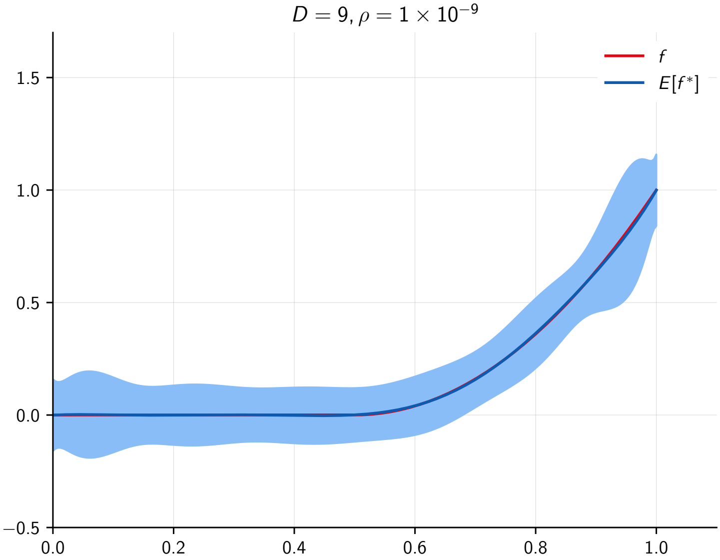


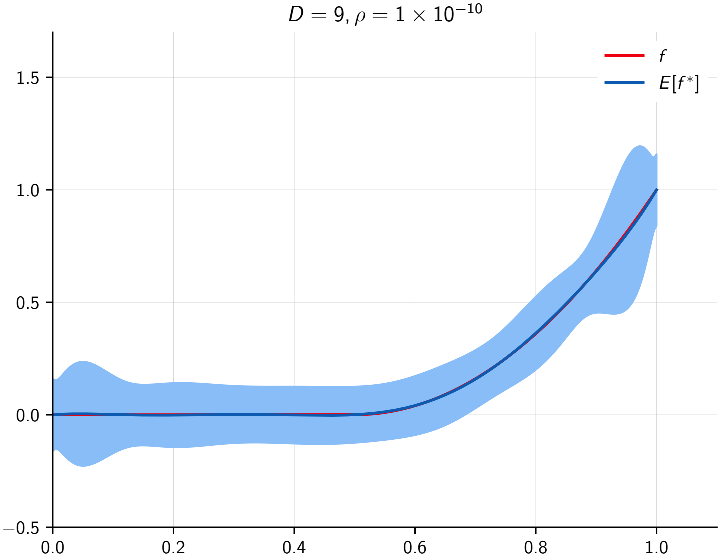
$D = 9, \rho = 1 \times 10^{-6}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

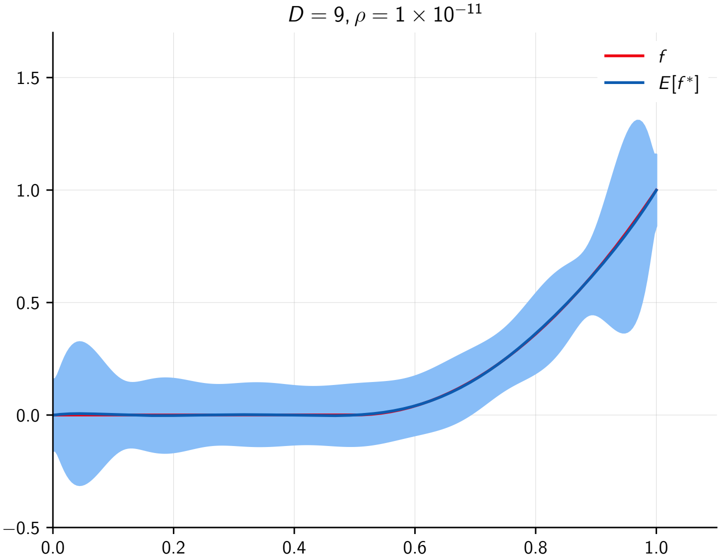

$D = 9, \rho = 1 \times 10^{-7}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

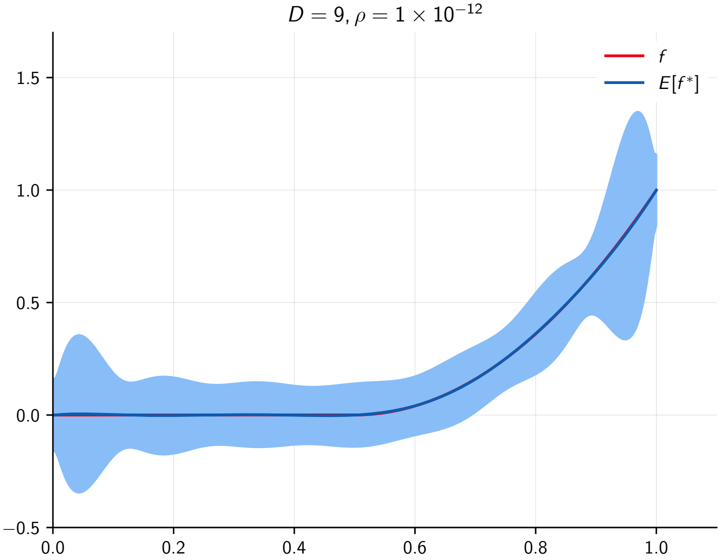
$D = 9, \rho = 1 \times 10^{-8}$

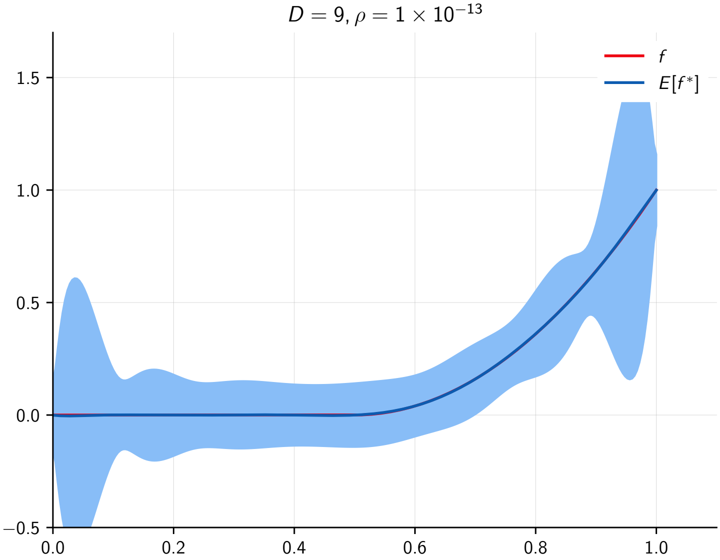


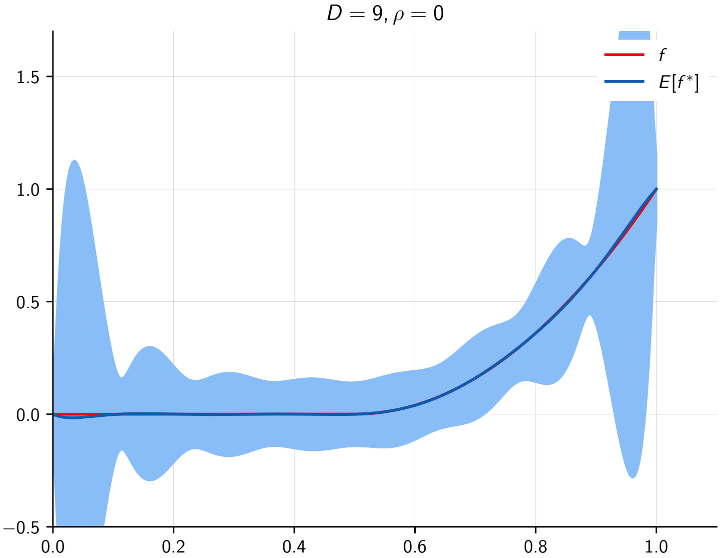

$D = 9, \rho = 1 \times 10^{-9}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


$D = 9, \rho = 1 \times 10^{-10}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


$D = 9, \rho = 1 \times 10^{-11}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


$D = 9, \rho = 1 \times 10^{-12}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


$D = 9, \rho = 1 \times 10^{-13}$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0



$D = 9, \rho = 0$
f
$E[f^*]$
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0

We can formalize these observations as follows:

Let x be fixed, y the "true" value associated to it, f^* the predictor we learned from the data-set $\mathscr{D}$, and $Y = f^*(x)$ be the value we predict at x.

If we consider that the training set $\mathscr{D}$ is a random quantity, then f^* is random, and consequently Y is.

We have

$$\mathbb{E}_{\mathscr{D}}\left((Y - y)^2\right)$$

We have

$$\mathbb{E}_{\mathscr{D}}\left((Y-y)^2\right) = \mathbb{E}_{\mathscr{D}}\left(Y^2 - 2Yy + y^2\right)$$

We have

$$\begin{aligned}\mathbb{E}_{\mathscr{D}}\left((Y-y)^2\right) &= \mathbb{E}_{\mathscr{D}}\left(Y^2 - 2Yy + y^2\right) \\ &= \mathbb{E}_{\mathscr{D}}\left(Y^2\right) - 2\mathbb{E}_{\mathscr{D}}\left(Y\right) y + y^2\end{aligned}$$

We have

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}}\left((Y-y)^2\right) &= \mathbb{E}_{\mathscr{D}}\left(Y^2 - 2Yy + y^2\right) \\
&= \mathbb{E}_{\mathscr{D}}\left(Y^2\right) - 2\mathbb{E}_{\mathscr{D}}\left(Y\right) y + y^2 \\
&= \underbrace{\mathbb{E}_{\mathscr{D}}\left(Y^2\right) - \mathbb{E}_{\mathscr{D}}\left(Y\right)^2}_{\mathbb{V}_{\mathscr{D}}(Y)} + \underbrace{\mathbb{E}_{\mathscr{D}}\left(Y\right)^2 - 2\mathbb{E}_{\mathscr{D}}\left(Y\right) y + y^2}_{(\mathbb{E}_{\mathscr{D}}(Y)-y)^2}
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}}\left((Y-y)^2\right) &= \mathbb{E}_{\mathscr{D}}\left(Y^2 - 2Yy + y^2\right) \\
&= \mathbb{E}_{\mathscr{D}}\left(Y^2\right) - 2\mathbb{E}_{\mathscr{D}}(Y)\, y + y^2 \\
&= \underbrace{\mathbb{E}_{\mathscr{D}}\left(Y^2\right) - \mathbb{E}_{\mathscr{D}}(Y)^2}_{\mathbb{V}_{\mathscr{D}}(Y)} + \underbrace{\mathbb{E}_{\mathscr{D}}(Y)^2 - 2\mathbb{E}_{\mathscr{D}}(Y)\, y + y^2}_{(\mathbb{E}_{\mathscr{D}}(Y)-y)^2} \\
&= \underbrace{\left(\mathbb{E}_{\mathscr{D}}(Y) - y\right)^2}_{\text{Bias}} + \underbrace{\mathbb{V}_{\mathscr{D}}(Y)}_{\text{Variance}}
\end{aligned}
$$

We have

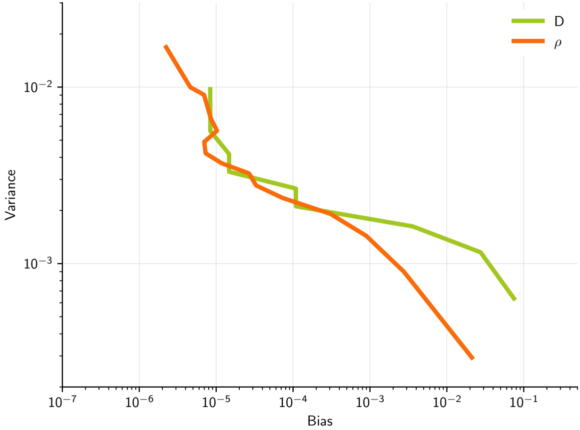
$$
\begin{aligned}
\mathbb{E}_{\mathscr{D}}\left((Y-y)^2\right) &= \mathbb{E}_{\mathscr{D}}\left(Y^2 - 2Yy + y^2\right) \\
&= \mathbb{E}_{\mathscr{D}}\left(Y^2\right) - 2\mathbb{E}_{\mathscr{D}}\left(Y\right) y + y^2 \\
&= \underbrace{\mathbb{E}_{\mathscr{D}}\left(Y^2\right) - \mathbb{E}_{\mathscr{D}}\left(Y\right)^2}_{\mathbb{V}_{\mathscr{D}}(Y)} + \underbrace{\mathbb{E}_{\mathscr{D}}\left(Y\right)^2 - 2\mathbb{E}_{\mathscr{D}}\left(Y\right) y + y^2}_{(\mathbb{E}_{\mathscr{D}}(Y)-y)^2} \\
&= \underbrace{\left(\mathbb{E}_{\mathscr{D}}\left(Y\right) - y\right)^2}_{\text{Bias}} + \underbrace{\mathbb{V}_{\mathscr{D}}\left(Y\right)}_{\text{Variance}}
\end{aligned}
$$

This is the **bias-variance decomposition:**

- the bias term quantifies how much the model fits the data on average,
- the variance term quantifies how much the model changes across data-sets.

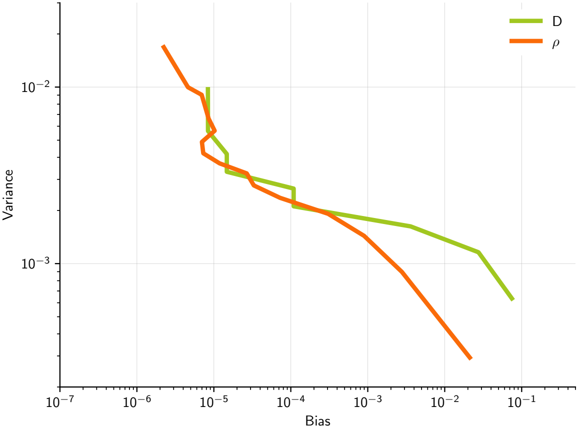
(Geman and Bienenstock, 1992)

From this comes the **bias-variance tradeoff:**



Reducing the capacity makes f^* fit the data less on average, which increases the bias term.

From this comes the **bias-variance tradeoff:**



Reducing the capacity makes f^* fit the data less on average, which increases the bias term. Increasing the capacity makes f^* vary a lot with the training data, which increases the variance term.

Is all this probabilistic?

Conceptually model-fitting and regularization can be interpreted as Bayesian inference.

Conceptually model-fitting and regularization can be interpreted as Bayesian inference.

This approach consists of **modeling the parameters A of the model themselves as random quantities following a prior distribution μ_A.**

Conceptually model-fitting and regularization can be interpreted as Bayesian inference.

This approach consists of **modeling the parameters A of the model themselves as random quantities following a prior distribution μ_A.**

By looking at the data $\mathscr{D}$, we can estimate a posterior distribution for the said parameters,

$$\mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \propto \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha),$$

and from that their most likely values.

Conceptually model-fitting and regularization can be interpreted as Bayesian inference.

This approach consists of **modeling the parameters A of the model themselves as random quantities following a prior distribution μ_A.**

By looking at the data $\mathscr{D}$, we can estimate a posterior distribution for the said parameters,

$$\mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \propto \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha),$$

and from that their most likely values.

So instead of a penalty term, we define a prior distribution, which is usually more intellectually satisfying.

For instance, consider a polynomial model with Gaussian prior, that is

$$\forall n,\ Y_n = \sum_{d=0}^{D} A_d\, X_n^d + \Delta_n,$$

where

$$\forall d,\ A_d \sim \mathcal{N}(0, \xi),\ \forall n,\ X_n \sim \mu_X,\ \Delta_n \sim \mathcal{N}(0, \sigma)$$

all independent.

For instance, consider a polynomial model with Gaussian prior, that is

$$\forall n,\ Y_n = \sum_{d=0}^{D} A_d\, X_n^d + \Delta_n,$$

where

$$\forall d,\ A_d \sim \mathcal{N}(0, \xi),\ \forall n,\ X_n \sim \mu_X,\ \Delta_n \sim \mathcal{N}(0, \sigma)$$

all independent.

For clarity, let $A = (A_0, \dots, A_D)$ and $\alpha = (\alpha_0, \dots, \alpha_D)$.

Remember that $\mathscr{D} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$ is the (random) training set and $\mathbf{d} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ is a realization.

$$\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d})$$

$$
\begin{aligned}
\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&= \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})}
\end{aligned}
$$

$$\begin{aligned}
\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&= \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&= \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z
\end{aligned}$$

$$
\begin{aligned}
&\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&\qquad = \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&\qquad = \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\qquad = \log \prod_n \mu(x_n, y_n \mid A = \alpha) + \log \mu_A(\alpha) - \log Z
\end{aligned}
$$

$$
\begin{aligned}
&\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&\qquad = \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&\qquad = \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\qquad = \log \prod_n \mu(x_n, y_n \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\qquad = \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha)\ \underbrace{\mu(x_n \mid A = \alpha)}_{= \mu(x_n)} + \log \mu_A(\alpha) - \log Z
\end{aligned}
$$

$$
\begin{aligned}
&\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&\quad = \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&\quad = \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(x_n, y_n \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha)\ \underbrace{\mu(x_n \mid A = \alpha)}_{= \mu(x_n)} + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha) + \log \mu_A(\alpha) - \log Z'
\end{aligned}
$$

$$
\begin{aligned}
\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d})
&= \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&= \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&= \log \prod_n \mu(x_n, y_n \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&= \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha)\, \underbrace{\mu(x_n \mid A = \alpha)}_{= \mu(x_n)} + \log \mu_A(\alpha) - \log Z \\
&= \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha) + \log \mu_A(\alpha) - \log Z' \\
&= \underbrace{-\frac{1}{2\sigma^2} \sum_n \left(y_n - \sum_d \alpha_d x_n^d \right)^2}_{\text{Gaussian noise on } Y} - \underbrace{\frac{1}{2\xi^2} \sum_d \alpha_d^2}_{\text{Gaussian prior on } A} - \log Z''.
\end{aligned}
$$

Taking $\rho = \sigma^2 / \xi^2$ gives the penalty term of the previous slides.

$$
\begin{aligned}
&\log \mu_A(\alpha \mid \mathscr{D} = \mathbf{d}) \\
&\quad = \log \frac{\mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha)\, \mu_A(\alpha)}{\mu_{\mathscr{D}}(\mathbf{d})} \\
&\quad = \log \mu_{\mathscr{D}}(\mathbf{d} \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(x_n, y_n \mid A = \alpha) + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha)\, \underbrace{\mu(x_n \mid A = \alpha)}_{= \mu(x_n)} + \log \mu_A(\alpha) - \log Z \\
&\quad = \log \prod_n \mu(y_n \mid X_n = x_n, A = \alpha) + \log \mu_A(\alpha) - \log Z' \\
&\quad = \underbrace{- \frac{1}{2\sigma^2} \sum_n \left(y_n - \sum_d \alpha_d x_n^d \right)^2}_{\text{Gaussian noise on } Y} - \underbrace{\frac{1}{2\xi^2} \sum_d \alpha_d^2}_{\text{Gaussian prior on } A} - \log Z''.
\end{aligned}
$$

Taking $\rho = \sigma^2 / \xi^2$ gives the penalty term of the previous slides.

Regularization seen through that prism is intuitive: The stronger the prior, the more evidence you need to deviate from it.

The end

References

S. Geman and E. Bienenstock. **Neural networks and the bias/variance dilemma**. Neural Computation, 4:1–58, 1992.