Model Regularization

Overfitting, Bias-variance decomposition, L1 and L2 regularization, probabilistic interpretation

Machine Learning and Data Mining, 2021

Artem Maevskiy
National Research University Higher School of Economics

LAMBDA • HSE

September 22, 2021

Model Regularization PART 1

Overfitting, Bias-variance decomposition, L1 and L2 regularization, probabilistic interpretation

Machine Learning and Data Mining, 2021

Artem Maevskiy

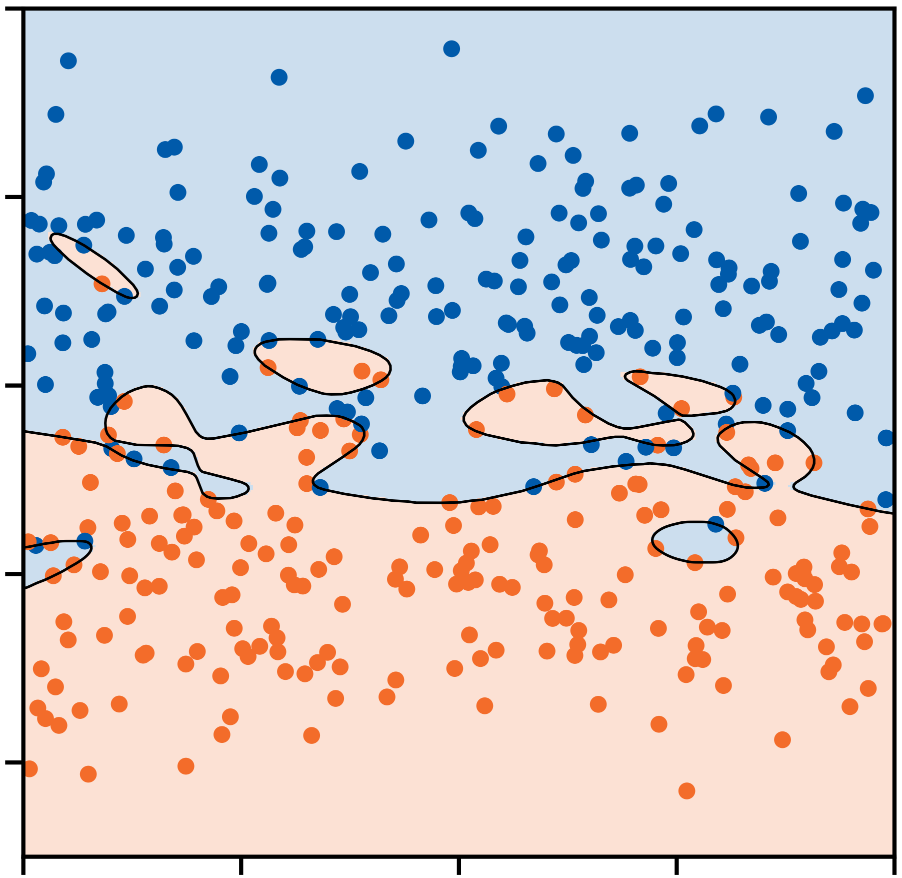
National Research University Higher School of Economics

LAMBDA • HSE

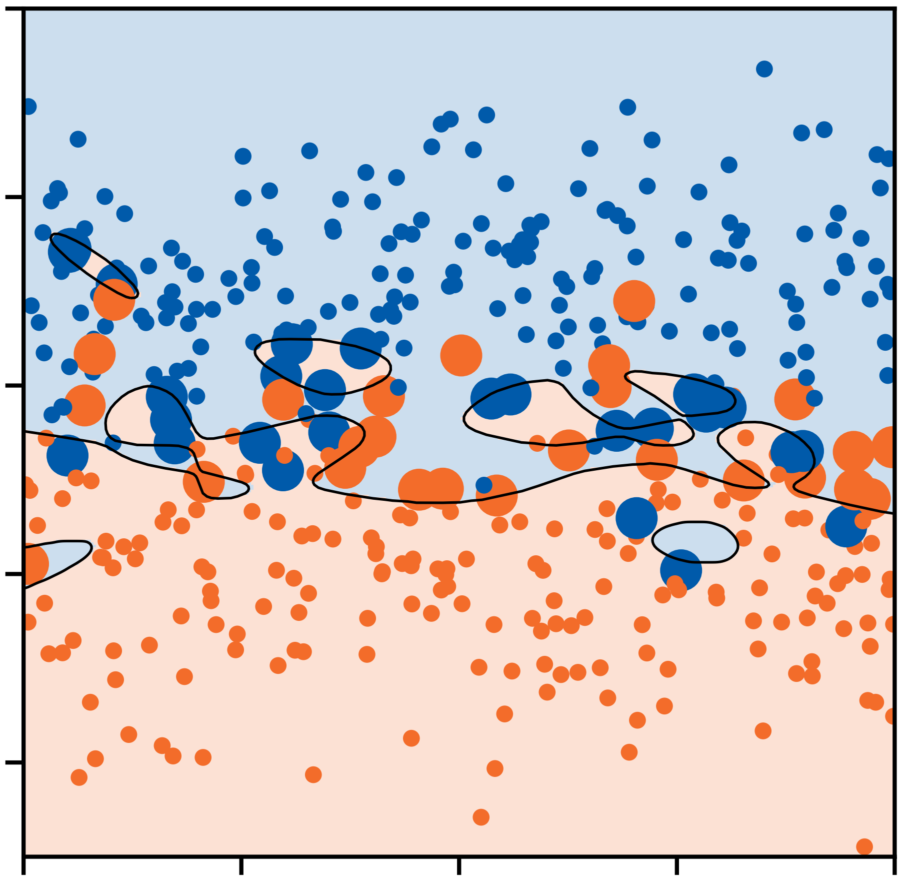
September 22, 2021

The problem of overfitting

Overfitting in classification



Training set

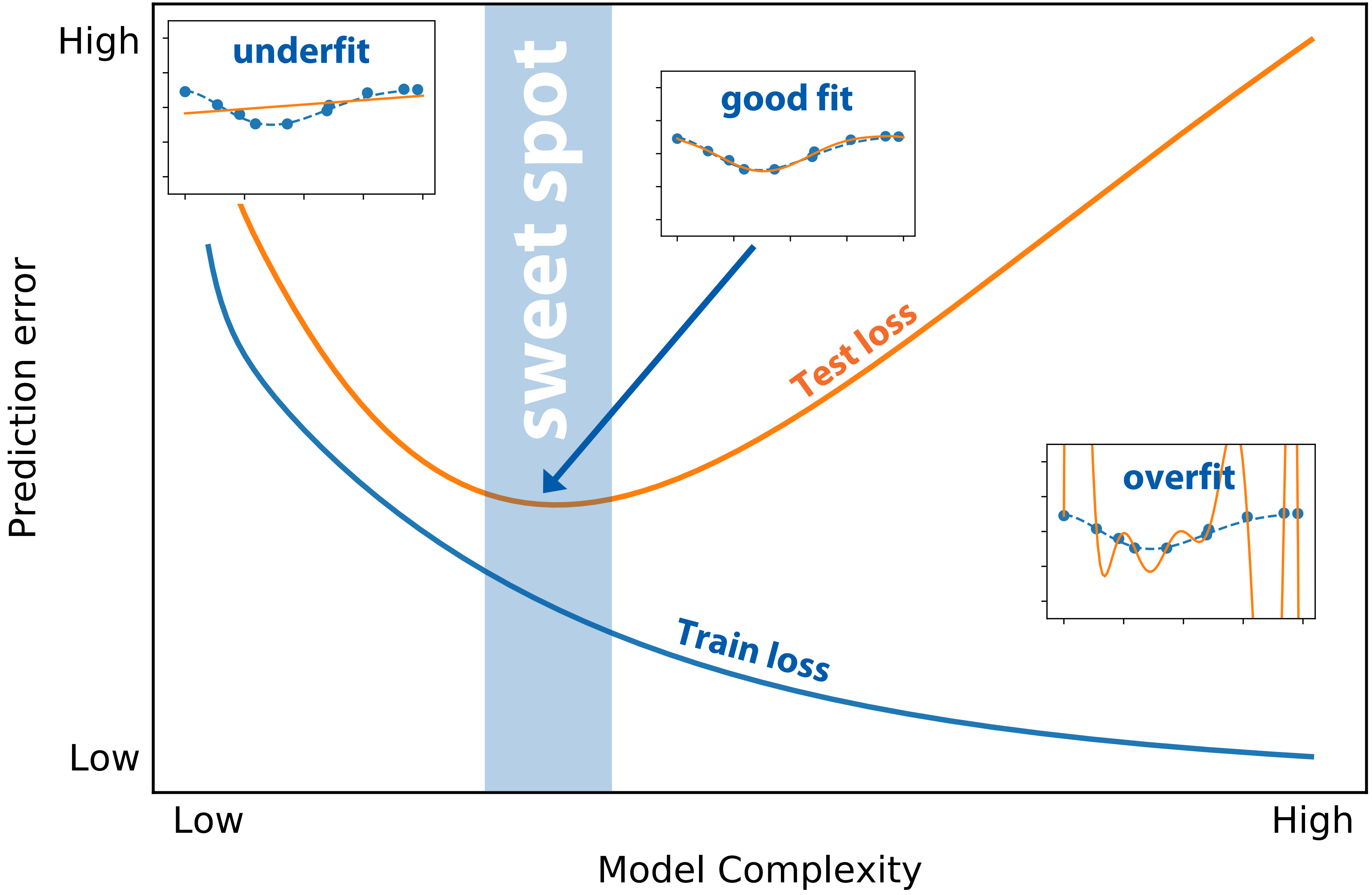


Test set

Large points =
classification error

Artem Maevskiy, NRU HSE

How to check whether a model is good?



Check the loss on the **test data** – i.e. data that the learning algorithm hasn't seen

The goal is to find the **right level of limitations** – not too strict, not too loose

Prediction error decomposition

Prediction error decomposition

Assume there's the following (unknown) **relation between the features and targets**:

$$y = f(x) + \varepsilon$$

where ε is some random noize:

$$\mathbb{E}[\varepsilon] = 0$$

$$\mathbb{D}[\varepsilon] = \sigma_\varepsilon^2$$

Prediction error decomposition

Assume there's the following (unknown) **relation between the features and targets**:

$$y = f(x) + \varepsilon$$

where ε is some random noize:

$$\mathbb{E}[\varepsilon] = 0$$

$$\mathbb{D}[\varepsilon] = \sigma_\varepsilon^2$$

Let's denote our training set as τ.

We want to study the **expected squared error** for the model $\hat{f}_\tau$ trained on it:

$$\mathrm{exp.sq.err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

Prediction error decomposition

$$\mathrm{exp.sq.err}(x) = \underset{\tau,y|x}{\mathbb{E}}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \underset{\tau,y|x}{\mathbb{E}}\left[\left(\hat{f}_\tau(x) \qquad\qquad\qquad\qquad\qquad\qquad - y\right)^2\right]$$

Prediction error decomposition

$$\text{exp. sq. err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right] + \mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right] - y\right)^2\right]$$

Prediction of the “expected model”

Prediction error decomposition

$$\text{exp.sq.err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right] + \mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right] - f(x) + f(x) - y\right)^2\right]$$

Ground truth (without the noise)

Prediction error decomposition

$$\text{exp. sq. err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right]\right) + \left(\mathop{\mathbb{E}}_{\tau'}\left[\hat{f}_{\tau'}(x)\right] - f(x)\right) + (f(x) - y)\right)^2\right]$$

(grouping the terms, then expanding the square)

Prediction error decomposition

$$\text{exp. sq. err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right) + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right) + (f(x) - y)\right)^2\right]$$

(easy to show that all the cross term expectations are 0)

$$= \mathop{\mathbb{E}}_{\tau}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right)^2\right] + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right)^2 + \mathop{\mathbb{E}}_{y|x}[(f(x) - y)^2]$$

Variance of the model

i.e. how "unstable" the model is wrt the noise in the training data

Prediction error decomposition

$$\text{exp. sq. err}(x) = \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau,y|x}\left[\left(\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right) + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right) + (f(x) - y)\right)^2\right]$$

(easy to show that all the cross term expectations are 0)

$$= \mathop{\mathbb{E}}_{\tau}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right)^2\right] + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right)^2 + \mathop{\mathbb{E}}_{y|x}[(f(x) - y)^2]$$

how much the “expected model” differs from the ground truth

Squared bias

Prediction error decomposition

$$\text{exp. sq. err}(x) = \mathop{\mathbb{E}}_{\tau, y|x}\left[\left(\hat{f}_\tau(x) - y\right)^2\right]$$

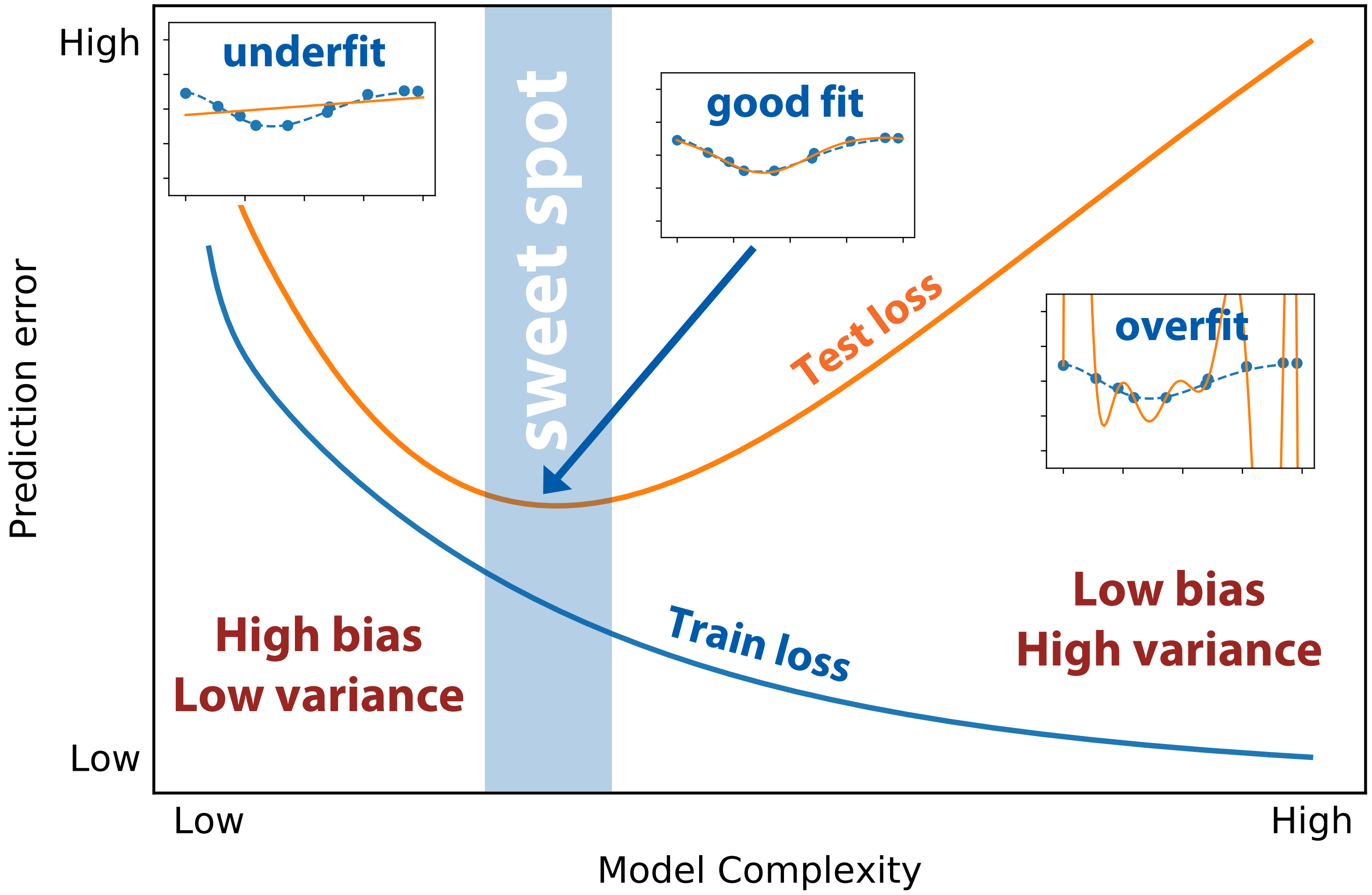
$$= \mathop{\mathbb{E}}_{\tau, y|x}\left[\left(\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right) + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right) + (f(x) - y)\right)^2\right]$$

(easy to show that all the cross term expectations are 0)

$$= \mathop{\mathbb{E}}_{\tau}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right)^2\right] + \left(\mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)] - f(x)\right)^2 + \mathop{\mathbb{E}}_{y|x}[(f(x) - y)^2]$$

Irreducible error
$(= \mathbb{E}[\varepsilon^2] = \sigma_\varepsilon^2)$

Bias-variance tradeoff



Typically there's a **tradeoff** between the two sources of error

Example: bias and variance of a linear model

Bias and variance error components can be calculated analytically for linear models

Simplification:

for each expectation term $\mathop{\mathbb{E}}_{\tau}$ let's consider **the features fixed**, i.e. $X_\tau \equiv X$ (the design matrix is constant), and only the **target vector y_τ is random**)

Example: bias and variance of a linear model

Bias and variance error components can be calculated analytically for linear models

Simplification:

for each expectation term $\mathop{\mathbb{E}}_{\tau}$ let's consider **the features fixed**, i.e. $X_\tau \equiv X$ (the design matrix is constant), and only the **target vector y_τ is random**)

Recall the solution for the linear regression model with the MSE loss:

$$\widehat{f}_\tau(x) = \theta_\tau^{\mathrm{T}} x = x^{\mathrm{T}} \theta_\tau$$

$$\theta_\tau = \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} y_\tau$$

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\mathrm{bias}(x) = \mathbb{E}_{\tau}\left[\widehat{f}_{\tau}(x)\right] - f(x)$$

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\text{bias}(x) = \mathop{\mathbb{E}}_{\tau}\left[\widehat{f}_{\tau}(x)\right] - f(x) = \mathop{\mathbb{E}}_{\tau}\left[x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_{\tau}\right] - x^{\mathrm{T}}\theta_{\text{true}}$$

We'll also assume that the **true dependence is linear** indeed

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\mathrm{bias}(x) = \mathop{\mathbb{E}}_{\tau}\left[\widehat{f}_{\tau}(x)\right] - f(x) = \mathop{\mathbb{E}}_{\tau}\left[x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_{\tau}\right] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\mathop{\mathbb{E}}_{\tau}[y_{\tau}] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\text{bias}(x) = \mathop{\mathbb{E}}_{\tau}\left[\widehat{f}_{\tau}(x)\right] - f(x) = \mathop{\mathbb{E}}_{\tau}\left[x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_{\tau}\right] - x^{\mathrm{T}}\theta_{\text{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\mathop{\mathbb{E}}_{\tau}[y_{\tau}] - x^{\mathrm{T}}\theta_{\text{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}X\theta_{\text{true}} - x^{\mathrm{T}}\theta_{\text{true}}$$

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\mathrm{bias}(x) = \mathop{\mathbb{E}}_{\tau}\left[\hat{f}_\tau(x)\right] - f(x) = \mathop{\mathbb{E}}_{\tau}\left[x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_\tau\right] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\mathop{\mathbb{E}}_{\tau}[y_\tau] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}X\theta_{\mathrm{true}} - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

Example: bias and variance of a linear model

Let's look at the **bias term** from the error decomposition:

$$\mathrm{bias}(x) = \mathop{\mathbb{E}}_{\tau}\left[\widehat{f}_{\tau}(x)\right] - f(x) = \mathop{\mathbb{E}}_{\tau}\left[x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_{\tau}\right] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\mathop{\mathbb{E}}_{\tau}[y_{\tau}] - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}X\theta_{\mathrm{true}} - x^{\mathrm{T}}\theta_{\mathrm{true}}$$

$$= x^{\mathrm{T}}\theta_{\mathrm{true}} - x^{\mathrm{T}}\theta_{\mathrm{true}} = 0$$

I.e. linear regression model is **unbiased**

as long as the true dependence is linear

Example: bias and variance of a linear model

Now let's look at the **variance term**:

$$\text{variance}(x) = \mathbb{E}_{\tau}\left[\left(\hat{f}_{\tau}(x) - \mathbb{E}_{\tau'}[\hat{f}_{\tau'}(x)]\right)^2\right]$$

It can then be shown that:

$$\text{variance}(x) = \sigma_{\varepsilon}^2 x^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x$$

So the variance error component is a **quadratic form**, defined by the $\left(X^{\mathrm{T}} X\right)^{-1}$ matrix.

[derivation]

Now let's look at the **variance term**:

$$\text{variance}(x) = \mathop{\mathbb{E}}_{\tau}\left[\left(\hat{f}_\tau(x) - \mathop{\mathbb{E}}_{\tau'}[\hat{f}_{\tau'}(x)]\right)^2\right]$$

Note that $\widehat{f_\tau}(x)$ can be thought of as a **linear transformation** to the training targets vector y_τ:

$$\widehat{f_\tau}(x) = x^{\mathrm{T}}\theta_\tau = x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y_\tau = h^{\mathrm{T}}(x)y_\tau$$

$$h^{\mathrm{T}}(x) = x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$$

[derivation]

$$\text{variance}(x) = \mathop{\mathbb{E}}_{\tau}\left[\left(h^{\mathrm{T}}(x)y_\tau - \mathop{\mathbb{E}}_{\tau'}[h^{\mathrm{T}}(x)y_{\tau'}]\right)^2\right] = \mathop{\mathbb{E}}_{\tau}\left[\left(h^{\mathrm{T}}(x)\left(y_\tau - \mathop{\mathbb{E}}_{\tau'}[y_{\tau'}]\right)\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\tau}\left[h^{\mathrm{T}}(x)\left(y_\tau - \mathop{\mathbb{E}}_{\tau'}[y_{\tau'}]\right)\left(y_\tau - \mathop{\mathbb{E}}_{\tau'}[y_{\tau'}]\right)^{\mathrm{T}} h(x)\right]$$

$$= h^{\mathrm{T}}(x)\mathop{\mathbb{E}}_{\tau}\left[\left(y_\tau - \mathop{\mathbb{E}}_{\tau'}[y_{\tau'}]\right)\left(y_\tau - \mathop{\mathbb{E}}_{\tau'}[y_{\tau'}]\right)^{\mathrm{T}}\right] h(x)$$

$$= h^{\mathrm{T}}(x)\mathop{\mathrm{cov}}_{\tau}[y_\tau, y_\tau]\, h(x) = \sigma_\varepsilon^2 h^{\mathrm{T}}(x)h(x)$$

[derivation]

$$\text{variance}(x) = \sigma_\varepsilon^2 h^{\mathrm{T}}(x) h(x)$$

$$= \sigma_\varepsilon^2 x^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} X \left(X^{\mathrm{T}} X\right)^{-1} x$$

$h^{\mathrm{T}}(x) = x^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}}$

$$= \sigma_\varepsilon^2 x^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x$$

So the variance error component is a **quadratic form**, defined by the $\left(X^{\mathrm{T}} X\right)^{-1}$ matrix.

Example: bias and variance of a linear model

We can diagonalize $X^{\mathrm{T}}X$:

$$\text{variance}(x) = \sigma_\varepsilon^2 x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}x = \sigma_\varepsilon^2 \tilde{x}^{\mathrm{T}}\Lambda^{-1}\tilde{x}$$

where $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_d\}$ is the matrix of eigenvalues of $X^{\mathrm{T}}X$.

Example: bias and variance of a linear model

We can diagonalize $X^{\mathrm{T}}X$:

$$\text{variance}(x) = \sigma_\varepsilon^2 x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}x = \sigma_\varepsilon^2 \tilde{x}^{\mathrm{T}}\Lambda^{-1}\tilde{x}$$

where $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_d\}$ is the matrix of eigenvalues of $X^{\mathrm{T}}X$.

This means that **small eigenvalues amplify the model variance**.

Example: bias and variance of a linear model

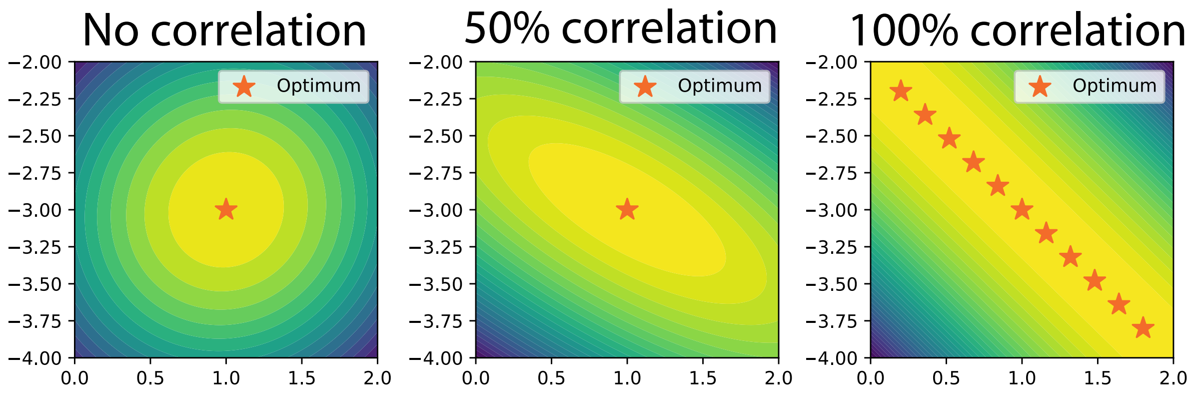
We can diagonalize $X^{\mathrm{T}}X$:

$$\text{variance}(x) = \sigma_\varepsilon^2 x^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}x = \sigma_\varepsilon^2 \tilde{x}^{\mathrm{T}}\Lambda^{-1}\tilde{x}$$

where $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_d\}$ is the matrix of eigenvalues of $X^{\mathrm{T}}X$.

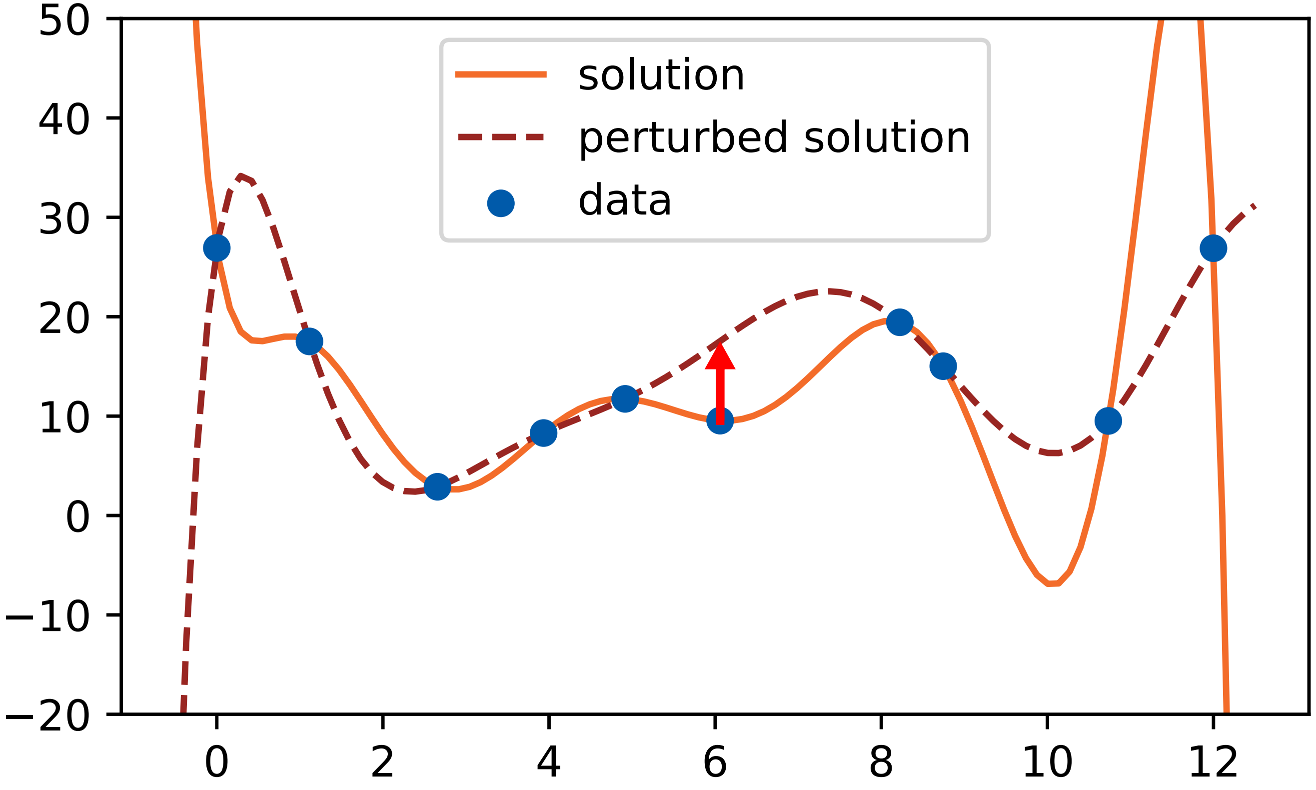
This means that **small eigenvalues amplify the model variance**.

This happens when $X^{\mathrm{T}}X$ is ill-defined e.g. when the features are correlated



MSE loss values
as a function
of model parameters

High-variance model



Small perturbation in data

⇩

Large change in prediction

To be continued…