Bias/Variance Vs Approximation/Estimation

Definitions and intuitions..

Risk:

- Formal Definition:
  - $R(f) = \mathbb{E}_{x,y}[\ell(y, f(x))]$
  - —the expected loss of model f under the true data distribution P(x,y)
- Intuition:
  - Risk is how much error your model makes on average.

Excess Risk:

- Formal Definition:
  - $R(f) - R(y^*)$
  - where y∗ is the Bayes model (the function with minimal possible risk).
  - Approximation Error + Estimation Error
- Intuition:
  - Excess risk is how much worse your model is than the best possible one you could ever have.

Expected Risk:

- Formal Definition:
  - $\mathbb{E}_D[R(\hat{f})]$
  - Expected risk is the average performance of a trained model over all possible training sets.
  - Bias + Variance
- Intuition:
  - Think of training your model many times on different datasets — expected risk is the average test error you'd get across all those runs.

Bias and Variance

- Formal Definition:
  - $\text{Bias} = \mathbb{E}_x[(y^* - \mathbb{E}_D[\hat{f}(x)])^2]$
  - $\text{Variance} = \mathbb{E}_x[\mathbb{E}_D[(\hat{f}(x) - \mathbb{E}_D[\hat{f}(x)])^2]]$
- Intuition:
  - Learning errors come from three sources:
    - Bias: How far your model's average prediction is from the truth
    - Variance: How much your model's predictions jump around when trained on different datasets.
    - Noise: Irreducible randomness in the data.

Approximation, Estimation, and Optimization Errors

- Formal Definition:
  - $$R(\hat{f}) - R(y^*) = \underbrace{R(\hat{f}) - R(\hat{f}_{ERM})}_{\text{Optimization Error}} + \underbrace{R(\hat{f}_{ERM}) - R(f^*)}_{\text{Estimation Error}} + \underbrace{R(f^*) - R(y^*)}_{\text{Approximation Error}}.$$
  -
  - Approximation Error: Your model family is too limited (e.g., linear model can’t fit curves).
  - Estimation Error: You don’t have enough or diverse enough data.
  - Optimization Error: Your algorithm didn’t find the best model it could have.
- Intuition:
  - Learning errors come from three sources
    - Model limits
    - Not enough data
    - Imperfect training

Bregman Divergence

- Formal Definition:
  - The Bregman divergence between two points p and q (with respect to a strictly convex, differentiable function ϕ) is defined as:
  - $B_\phi(p,q) = \phi(p) - \phi(q) - \langle \nabla\phi(q), p-q \rangle$
- Intuition:
  - The Bregman divergence tells us how far one point is from another, based on how a convex function ϕ curves.
  - It compares the value of ϕ at a point to the value predicted by a straight-line (tangent) approximation at another point.

Bregman Left Centroid

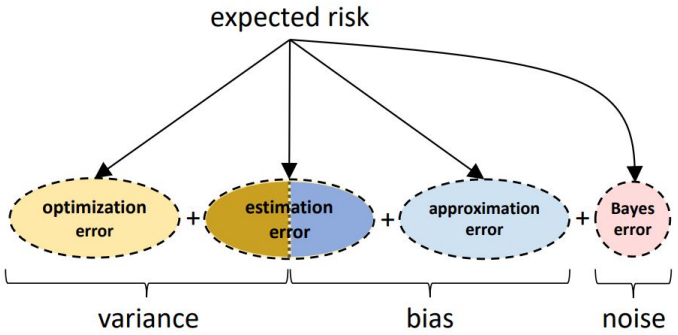
- Formal Definition:
  - $\circ f_{\phi}(x) = \operatorname{argmin}_{z} \mathbb{E}_{D}[B_{\phi}(z, \widehat{f}(x))] = (\nabla\phi)^{-1}(\mathbb{E}_{D}[\nabla\phi(\widehat{f}(x))]).$
  -
- Intuition:
  - The Bregman left centroid is the “average” prediction across different trained models — but measured in the geometry defined by the loss function (not always a simple mean).

Estimation Bias/Variance

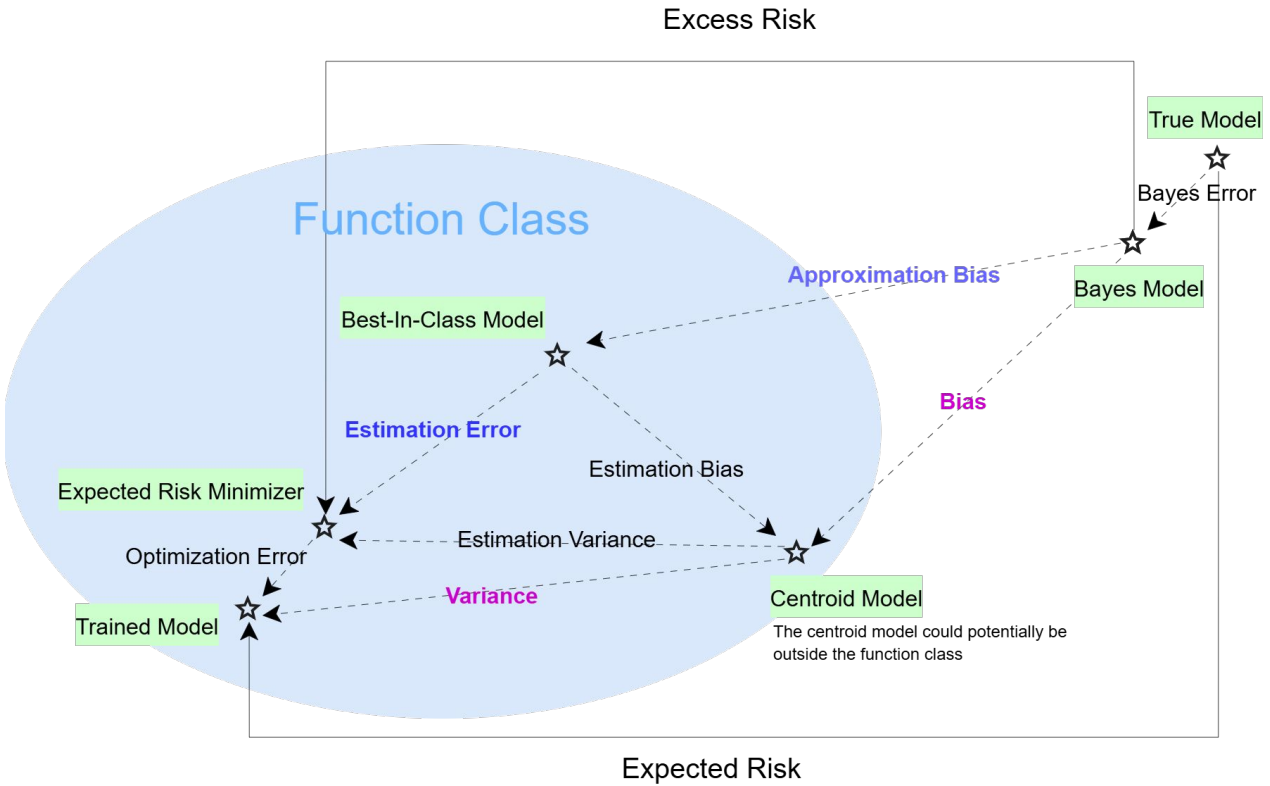
- Formal Definition:

$$\underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(f^*)\right]}_{\text{expected estimation error}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f}_\phi)\right]}_{\text{estimation variance}} + \underbrace{R(\mathring{f}_\phi) - R(f^*)}_{\text{estimation bias}}.$$

- Intuition:
  - Bias exists in estimation error as well and comes from the specific choice of sample
  - Bias = bias from choice of function class + bias from choice of data sample



Relationship between concepts



*Not Euclidean Distance

Discussions & Indications

- Linear least squares
  - Linear least squares (LLS) are **unbiased** estimators
    - The unbiased estimate = best-in-class models (= ordinary least square (OLS))
  - Centroid of LLS: expectation over all estimators trained on different samples
    - Centroid of LLS = expected value of the estimator = unbiased estimate = OLS
  - In LLS, the centroid model is the best-in-class-model
    - Estimation bias = 0
    - Closed form solution: optimization error = 0
    - Bias/Variance = Approximation/Estimation

Discussions & Indications

- Flawed proxy for model capacity
  - Model capacity is ultimately measured by the approximation error
  - Centroid model not necessarily inside the function class
    - Sufficient condition: dual-convex hypothesis (model) class
    - Not necessarily **R(**centroid model**)** > **R(**best-in-class model**)**
    - There could be: estimation bias < 0
  - A decrease in bias does not indicate a decrease in approximation error

  - Insight to bias/variance trade-off
    - Estimation error > 0
    - If the estimation bias < 0, the estimation variance would need to increase

Discussions & Indications

- A more general bias-variance decomposition
  - Extending to losses beyond the Bregman divergences

Proposition 1 (Bias/Variance Effects, in terms of Approximation-Estimation) *For any loss ℓ, assuming a centroid model exists, we have the following decomposition of the bias-effect and variance-effect.*

$$\underbrace{R(\mathring{f}) - R(\boldsymbol{y}^*)}_{\textbf{bias-effect}} = \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}} + \underbrace{R(\mathring{f}) - R(f^*)}_{\textbf{estimation bias}}, \tag{24}$$

$$\underbrace{\mathbb{E}_D[R(\hat{f}) - R(\mathring{f})]}_{\textbf{variance-effect}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}) - R(\hat{f}_{erm})\right]}_{\textbf{optimisation error}} + \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f})\right]}_{\textbf{estimation variance}}. \tag{25}$$

Numerical Example: Polynomial Ridge Regression

Let our true data generating process be given by:

$$y = x^3 + \epsilon, \quad \epsilon \sim N(0, 0.3^2)$$

Goal: Fit a polynomial regressor of degree 1 across a range of ridge strengths λ to examine the effect on Bias/Variance and Approximation/Estimation.

For each λ, we randomly sample **60 data sets** and **train 60 independent ridge models** on these datasets using ridge regression.

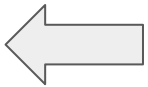
The outcomes of these models on a test set of inputs are averaged to estimate bias, variance, approximation and estimation error.

Numerical Example: Polynomial Ridge Regression

Bayes model: $y^* = x^3$

Here, the learning algorithm is analytical - it minimizes the mean squared error by computing the ridge regression solution:

$$\hat{\theta}_\lambda = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$



OLS for λ=0

$$\hat{f}_\lambda(x) = \hat{\theta}_{\lambda,0} + \hat{\theta}_{\lambda,1} x + \ldots + \hat{\theta}_{\lambda,d} x^d$$

Question: Is the optimisation error for this algorithm zero?

Numerical Example: Polynomial Ridge Regression

It's complicated…

It depends on how you treat the regularisation: is it a change in the definition of empirical risk, or is it a change in the learning procedure?

Here, we treat the regularisation as a change in the learning procedure and keep the definition of risk fixed from the OLS case for the function class:

$$R_{\text{emp}}^{(\text{OLS})}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

$$\hat{f}_{\text{erm}} \in \arg\inf_{f\in\mathcal{F}} R_{\text{emp}}^{(\text{OLS})}(f)$$

Numerical Example: Polynomial Ridge Regression

$$R_{\text{Ridge}} = R_{\text{OLS}} + \lambda \|\hat{\theta}_\lambda\|^2$$

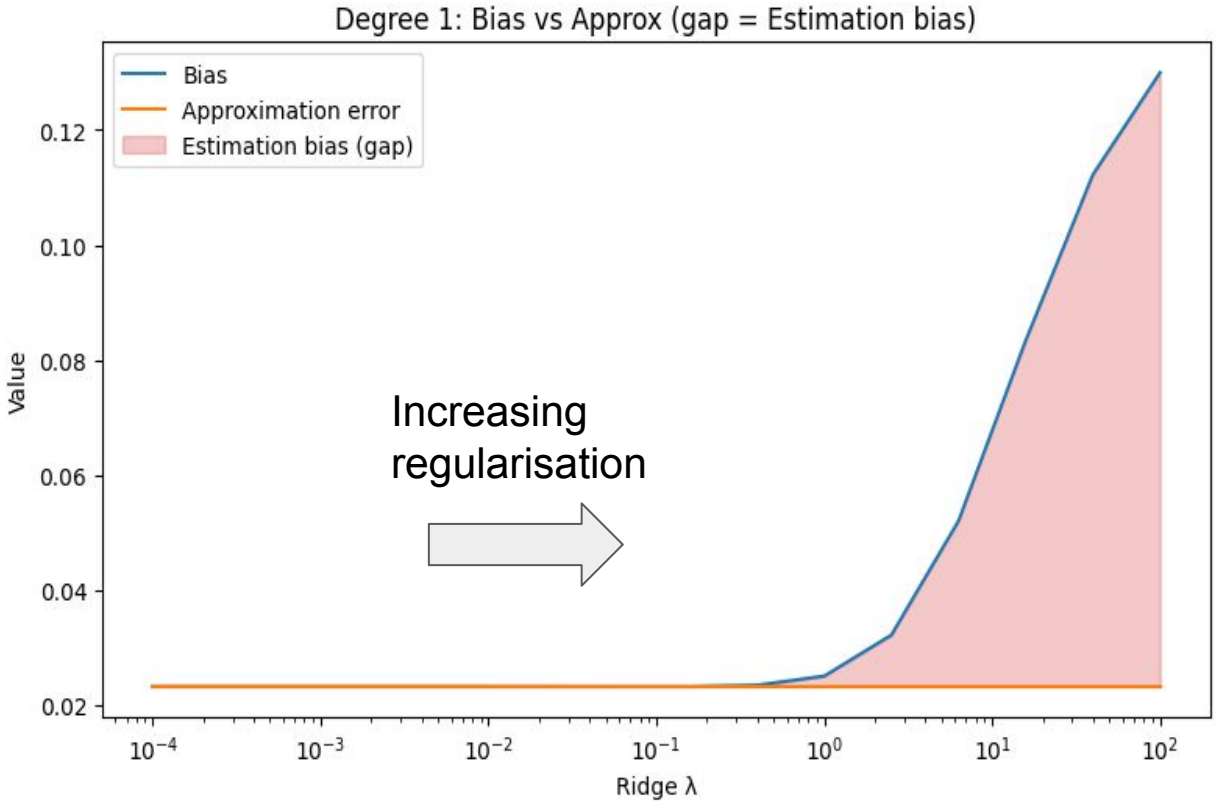
$$R(\hat{f}_\lambda) = R(\hat{f}_{\text{Ridge, ERM}}) \neq R(\hat{f}_{\text{OLS, ERM}}) \Rightarrow \textbf{optimisation error} \neq 0$$

The ERM reached by ridge regression is not necessarily the same as the ERM for the function class because it's minimising the risk for ridge regularisation, as opposed to that of OLS.

Numerical Example: Polynomial Ridge Regression

We obtain an estimate for the best in class model for polynomial degree d by fitting a d-degree polynomial with zero ridge to a large dataset (2000 data points). The labels of this dataset are given by the Bayes model.

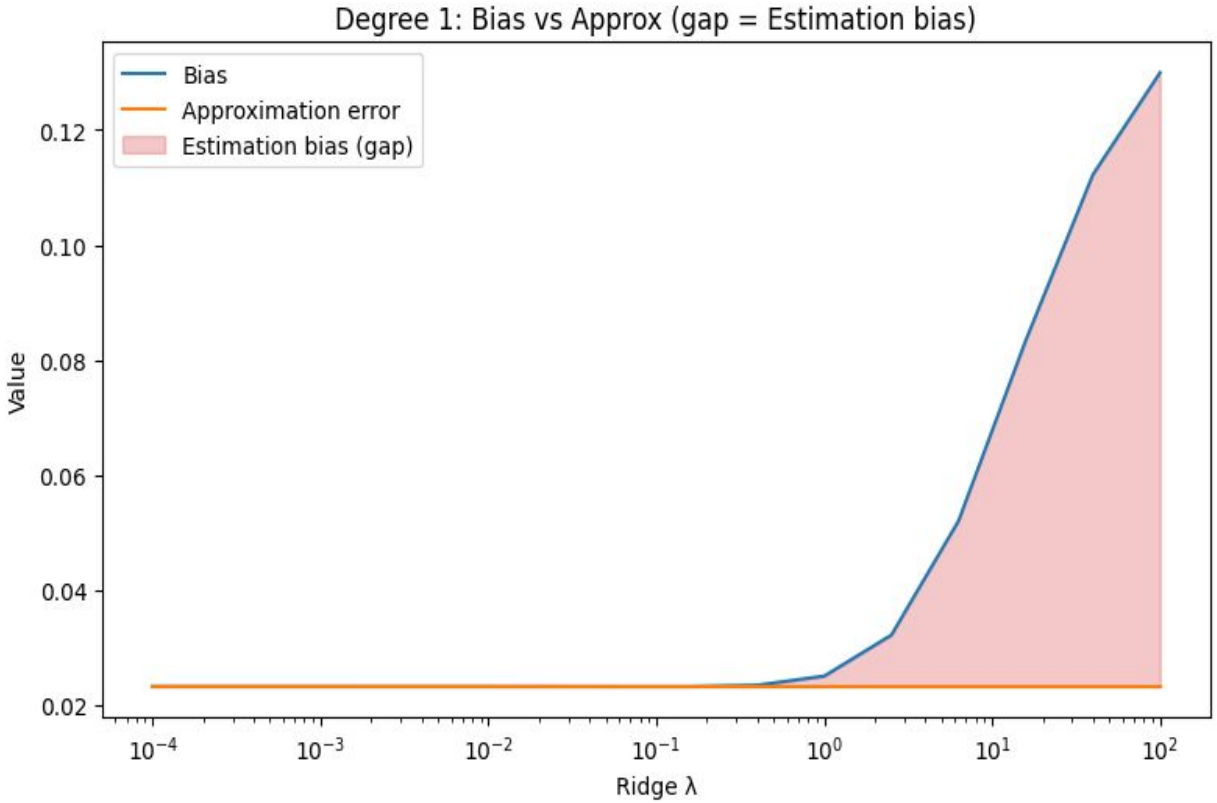
Linear Polynomial Fit: Bias Vs Approximation



When λ is small, we are approximately in the OLS regime → Bias ≅ Approx

As λ increases, the regularization introduces estimation bias and the two quantities diverge. The gap between them is the estimation bias.

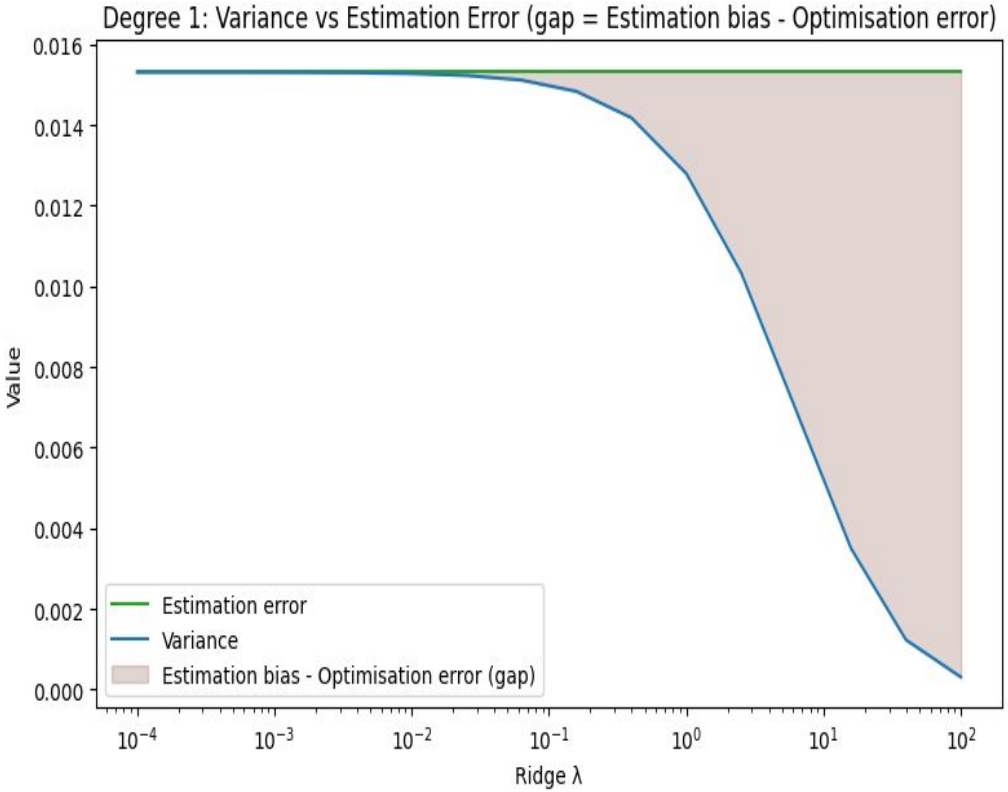
Linear Polynomial Fit: Bias Vs Approximation



Note that the **approximation error remains fixed** regardless of λ, because ridge regularisation only affects the learning procedure (and hence bias).

The function class remains the same, and thus so does approximation error.

Linear Polynomial Fit: Variance Vs Estimation

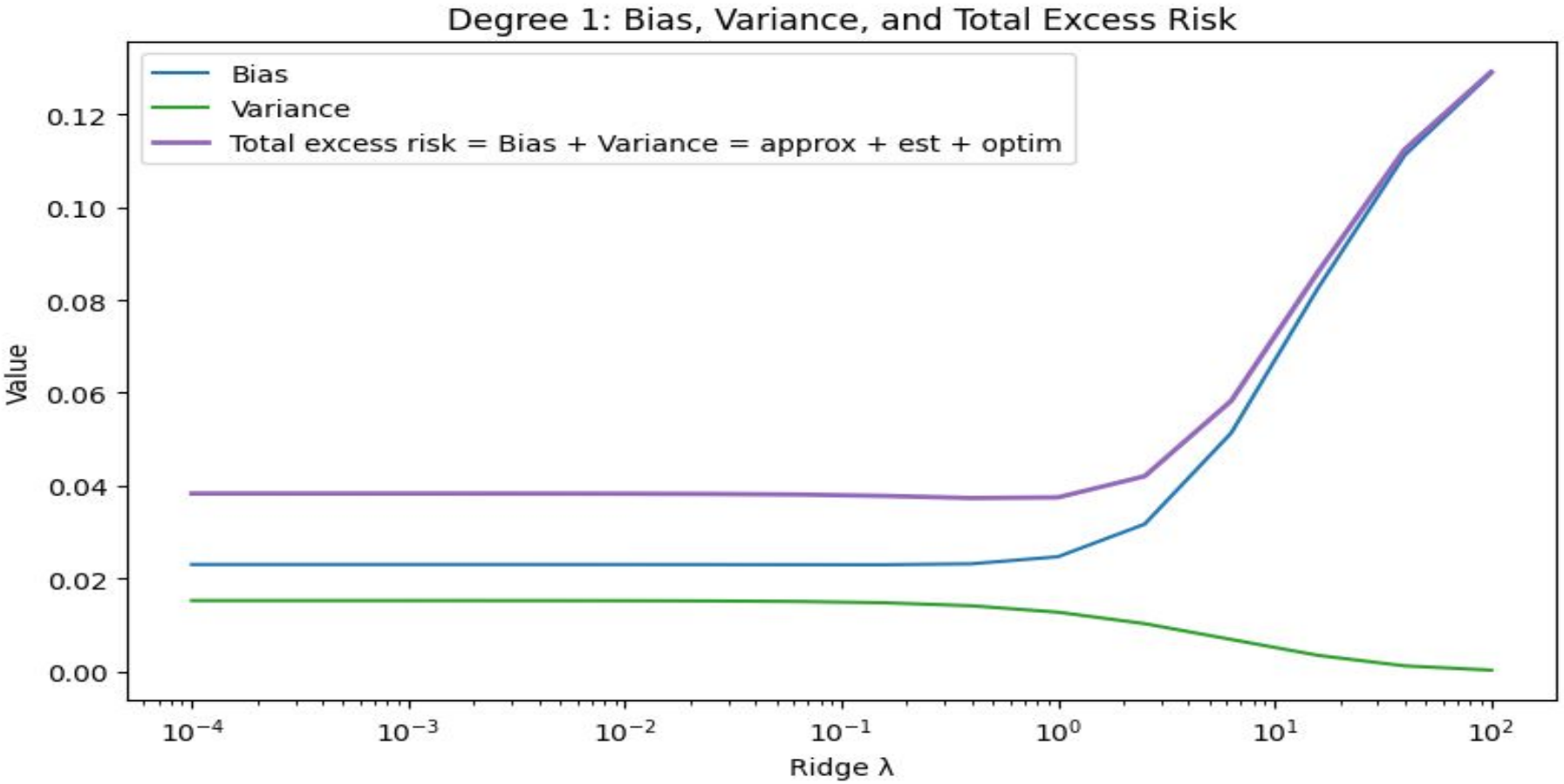


Again, for low regularisation,
variance = estimation variance
= estimation error
because estimation bias = 0 and optimisation error = 0 for small λ.

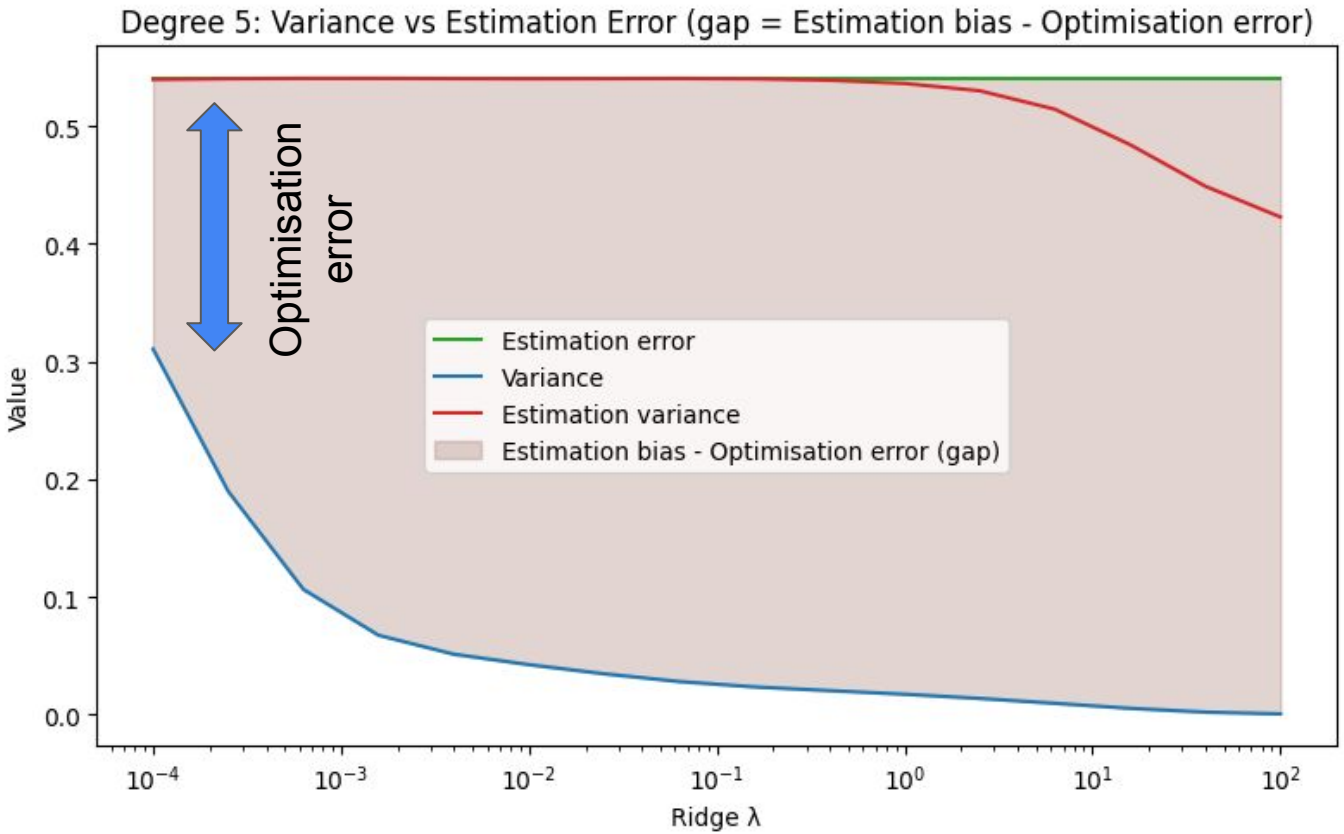
As λ increases, and estimation bias + optimisation error is introduced, the two quantities diverge.

Estimation error is constant, because only the learning procedure is changed by regularisation.

Linear Polynomial Fit: Total Expected Excess Risk



Bonus: Negative optimisation error?!



Summary

- Bias/Variance examines the performance of the learner, whereas Approximation/Estimation examines the performance of the function class
  - These two decompositions are only equivalent in the case of MSE loss
- Bias/Variance decomposition of expected risk and their relation to approximation, estimation and optimisation error only holds for certain losses e.g. those expressible as a Bregman divergence
- Approximation/Estimation decomposition of excess risk always holds