



Regularization via Early Stopping

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

Carlos Fernandez-Granda

Prerequisites

Ordinary least squares (OLS)

OLS coefficient analysis

OLS training and test error analysis

Ridge regression

Gradient descent

General procedure for minimizing cost function f

Intuition: At β move in the steepest direction $-\nabla f(\beta)$

Set the initial point $\beta^{(0)}$ to an arbitrary value

Update

$$\beta^{(k+1)} := \beta^{(k)} - \alpha_k \nabla f(\beta^{(k)})$$

where $\alpha_k > 0$ is the step size, until a stopping criterion is met

Least squares

Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{p \times n}$, $\beta \in \mathbb{R}^p$

The gradient of the least-squares cost function

$$f(\beta) := \frac{1}{2} \|y - X^T \beta\|_2^2 = \frac{1}{2} y^T y + \frac{1}{2} \beta^T X X^T \beta - y^T X^T \beta$$

equals

$$\begin{aligned} \nabla f(\beta) &= X X^T \beta - X y \\ &= X(X^T \beta - y) \end{aligned}$$

Gradient descent for least squares

Gradient descent updates are

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - \alpha_k \mathbf{X} \left(\mathbf{X}^T \beta^{(k)} - \mathbf{y} \right) \\ &= \beta^{(k)} + \alpha_k \sum_{i=1}^n \left(y_i - \langle \mathbf{x}_i, \beta^{(k)} \rangle \right) \mathbf{x}_i\end{aligned}$$

Initialization at origin with constant step size

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - \alpha X \left(X^T \beta^{(k)} - y \right) \\ &= \left(I - \alpha X X^T \right) \beta^{(k)} + \alpha X y\end{aligned}$$

$$\beta^{(0)} = 0$$

$$\beta^{(1)} = \alpha X y$$

$$\begin{aligned}\beta^{(2)} &= \left(I - \alpha X X^T \right) \beta^{(1)} + \alpha X y \\ &= \left(I - \alpha X X^T \right) \alpha X y + \alpha X y\end{aligned}$$

$$\begin{aligned}\beta^{(3)} &= \left(I - \alpha X X^T \right) \beta^{(2)} + \alpha X y \\ &= \left(I - \alpha X X^T \right)^2 \alpha X y + \left(I - \alpha X X^T \right) \alpha X y + \alpha X y\end{aligned}$$

$$\beta^{(k+1)} = \sum_{i=0}^k \left(I - \alpha X X^T \right)^i \alpha X y$$

Gradient descent iterates, starting at origin

$$\begin{aligned}\beta^{(k+1)} &= \sum_{i=0}^k \left(I - \alpha X X^T \right)^i \alpha X y \\ &= \alpha \sum_{i=0}^k \left(U U^T - \alpha U S^2 U^T \right)^i U S V^T y \\ &= \alpha \sum_{i=0}^k U \left(I - \alpha S^2 \right)^i U^T U S V^T y \\ &= \alpha U \operatorname{diag}_{j=1}^p \left(\sum_{i=0}^k \left(1 - \alpha s_j^2 \right)^i \right) S V^T y \\ &= \alpha U \operatorname{diag}_{j=1}^p \left(\frac{1 - \left(1 - \alpha s_j^2 \right)^{k+1}}{\alpha s_j} \right) V^T y\end{aligned}$$

Convergence

Condition for convergence? $|1 - \alpha s_j^2| < 1$

In that case

$$\begin{aligned}\lim_{k \rightarrow \infty} \beta^{(k)} &= \lim_{k \rightarrow \infty} U \operatorname{diag}_{j=1}^p \left(\frac{1 - (1 - \alpha s_j^2)^k}{s_j} \right) V^T y \\ &= US^{-1}V^T y = \beta_{\text{OLS}}\end{aligned}$$

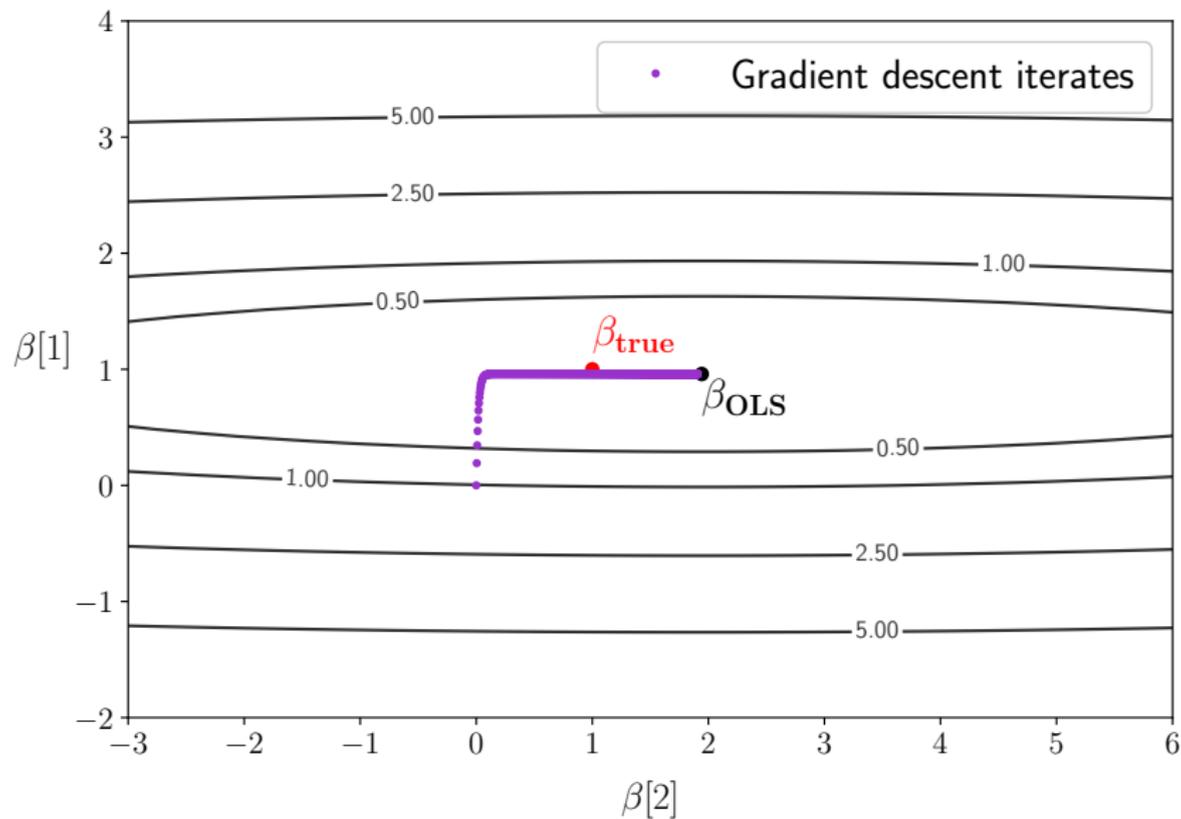
Guaranteed by $\alpha \leq 2/s_1^2$

Convergence rate

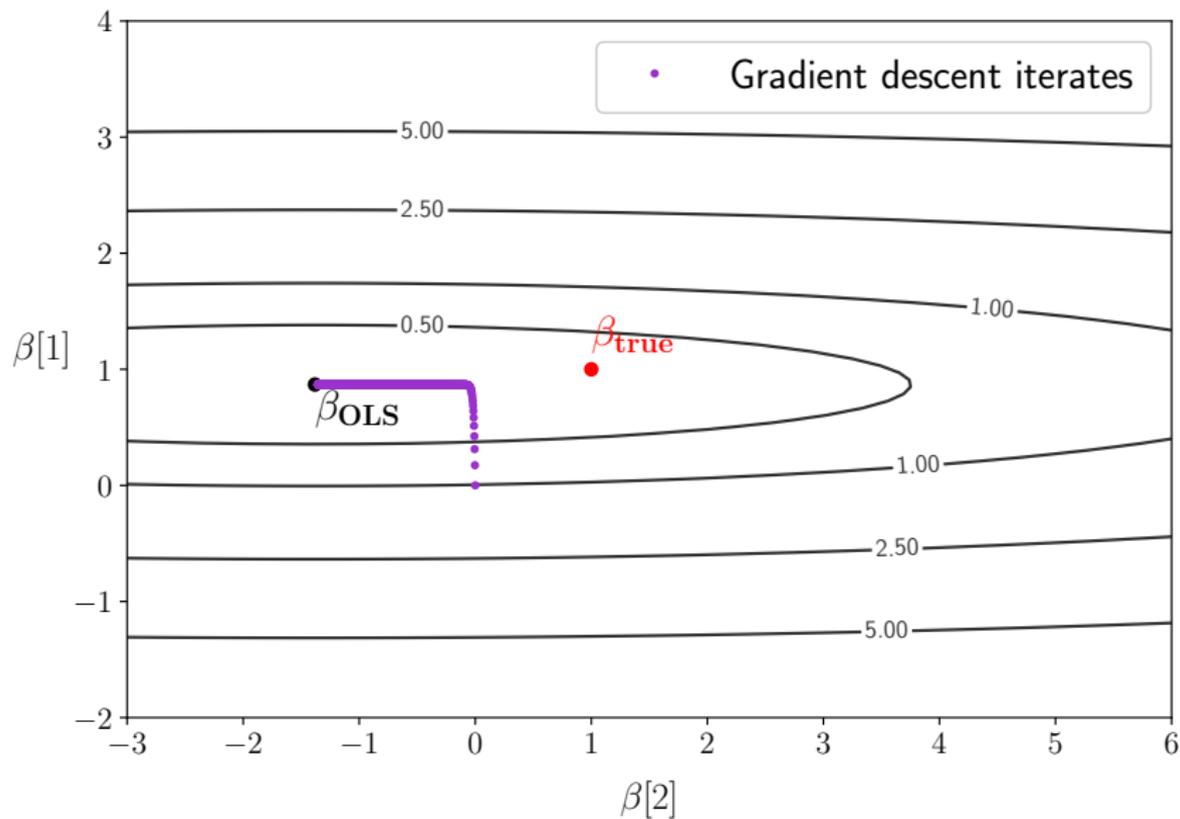
$$\beta^{(k+1)} = \alpha U \operatorname{diag}_{j=1}^p \left(\frac{1 - (1 - \alpha s_j^2)^{k+1}}{\alpha s_j} \right) V^T y$$

If $\alpha \approx 1/s_1^2$ convergence of each component governed by s_j/s_1

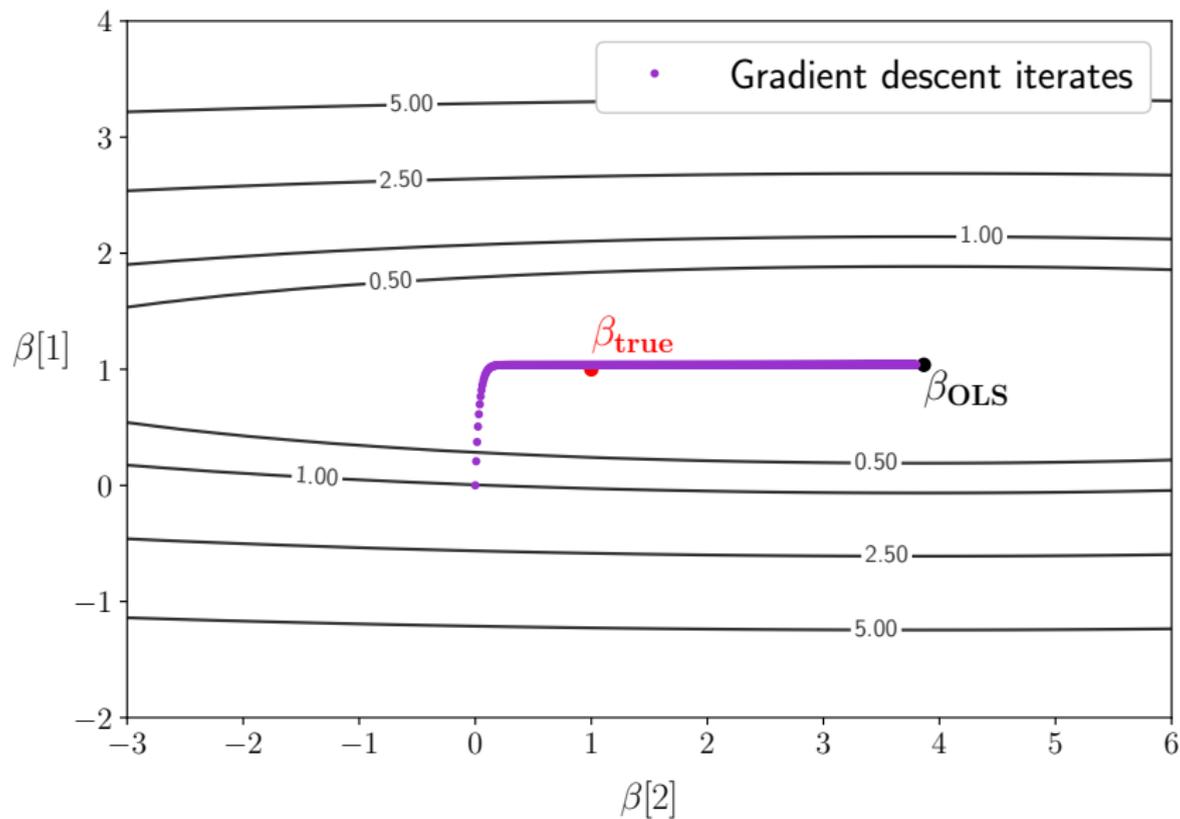
Additive model ($s_1 = 1, s_2 = 0.1$)



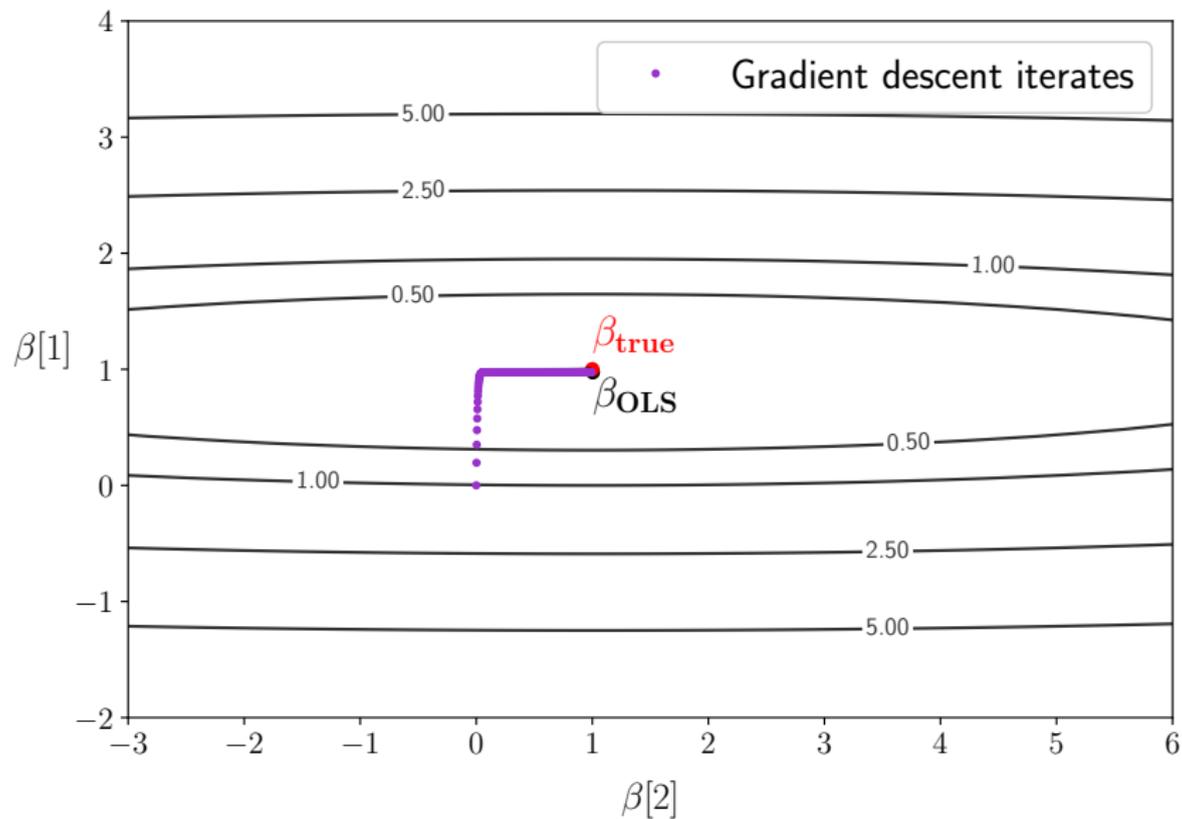
Additive model ($s_1 = 1, s_2 = 0.1$)



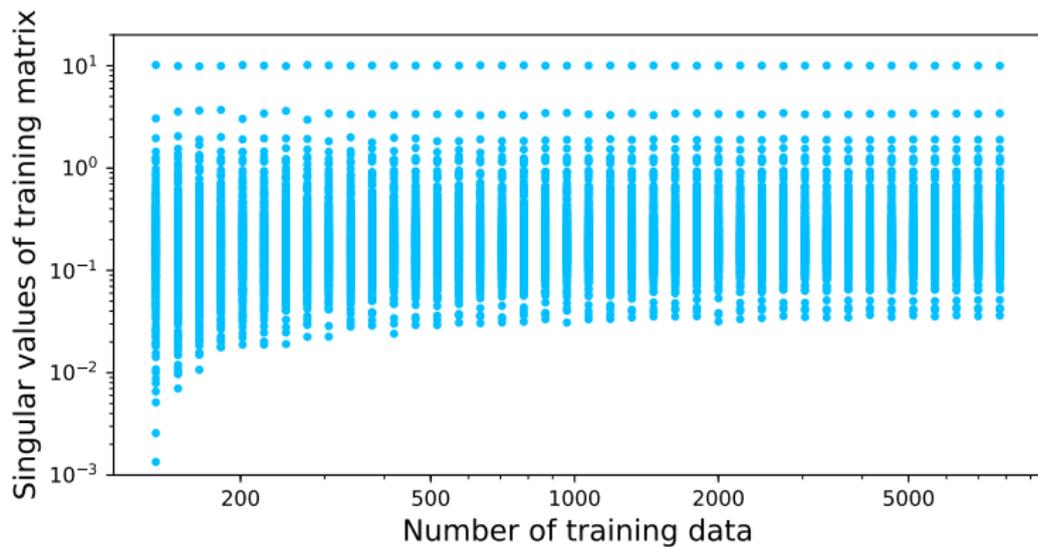
Additive model ($s_1 = 1, s_2 = 0.1$)



Additive model ($s_1 = 1, s_2 = 0.1$)



Temperature prediction via linear regression



Gradient descent for linear regression

Bad news: Convergence very slow

But, *do we want to converge?*

Additive model

Assume additive model for regression problem

$$y_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

Estimate coefficients via gradient descent up to iteration k

Gradient descent iterates

$$\tau_j := 1 - \alpha s_j^2$$

$$\begin{aligned}\tilde{\beta}^{(k)} &= U \operatorname{diag}_{j=1}^p \left(\frac{1 - \tau_j^k}{s_j} \right) V^T (X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}) \\ &= U \operatorname{diag}_{j=1}^p \left(\frac{1 - \tau_j^k}{s_j} \right) V^T (VSU^T \beta_{\text{true}} + \tilde{z}_{\text{train}}) \\ &= U \operatorname{diag}_{j=1}^p (1 - \tau_j^k) U^T \beta_{\text{true}} + U \operatorname{diag}_{j=1}^p \left(\frac{1 - \tau_j^k}{s_j} \right) V^T \tilde{z}_{\text{train}}\end{aligned}$$

Gradient descent coefficient estimate

$$\tilde{\beta}_{\text{GD}} = U \text{diag}_{j=1}^p (1 - \tau_j^k) U^T \beta_{\text{true}} + U \text{diag}_{j=1}^p \left(\frac{1 - \tau_j^k}{s_j} \right) V^T \tilde{z}_{\text{train}}$$

Distribution? Gaussian with mean

$$\beta_{\text{bias}} := \sum_{j=1}^p \left(1 - (1 - \alpha s_j^2)^k \right) \langle u_j, \beta_{\text{true}} \rangle u_j$$

and covariance matrix

$$\Sigma_{\text{GD}} := \sigma^2 U \text{diag}_{j=1}^p \left(\frac{(1 - (1 - \alpha s_j^2)^k)^2}{s_j^2} \right) U^T$$

Bias

Like ridge regression, early stopping produces systematic error

$$\mathbb{E}(\beta_{\text{true}} - \tilde{\beta}_{\text{GD}}) = \sum_{j=1}^p (1 - \alpha s_j^2)^k \langle u_j, \beta_{\text{true}} \rangle u_j$$

Bias decreases with k

Variance

Variance in direction of u_j equals $\frac{\sigma^2(1-(1-\alpha s_j^2)^k)^2}{s_j^2}$

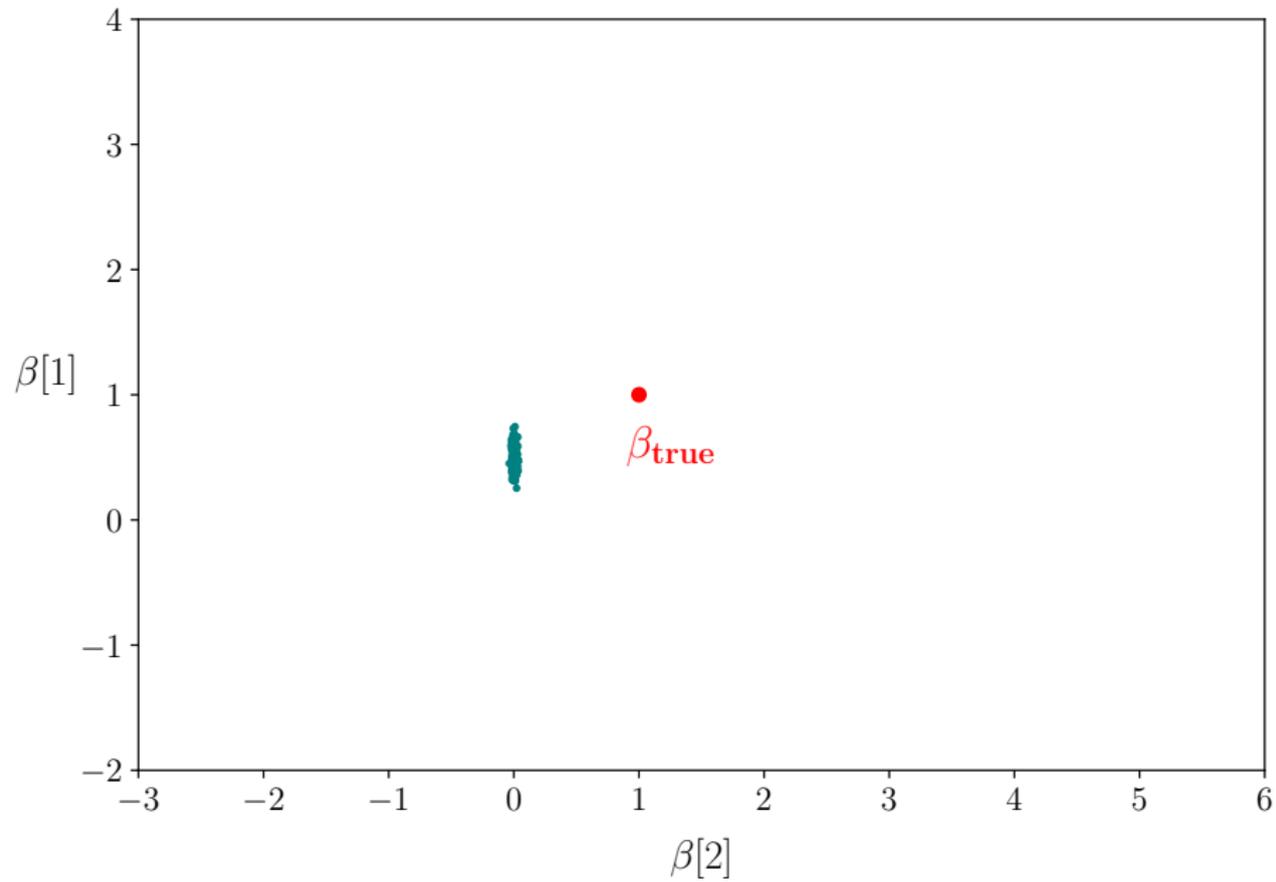
Small s_j blow up variance of OLS

For small k and αs_j , $(1 - \alpha s_j^2)^k \approx 1 - k\alpha s_j^2$

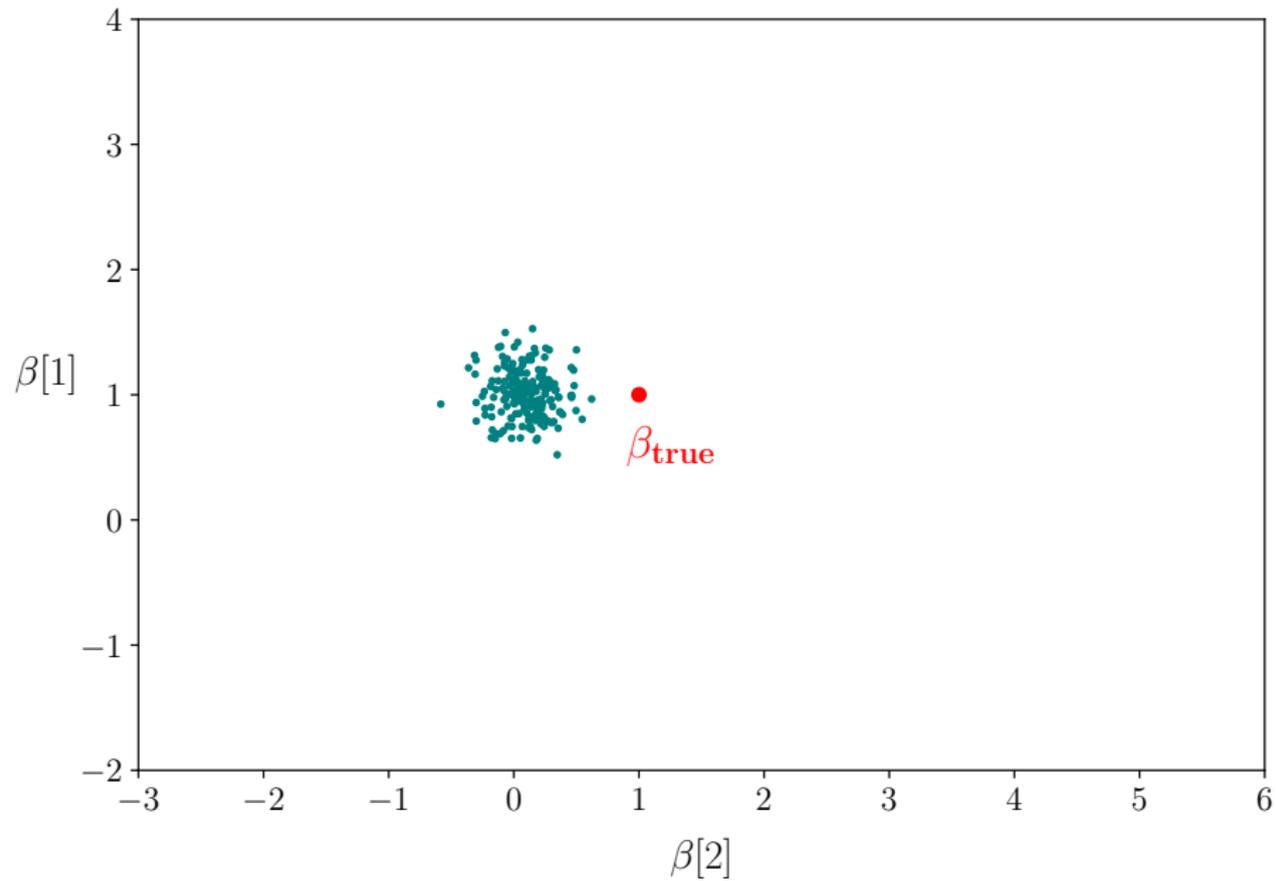
The corresponding variance component $\approx \sigma^2 k^2 \alpha^2 s_j^2$

Ideal λ achieves **bias-variance tradeoff**

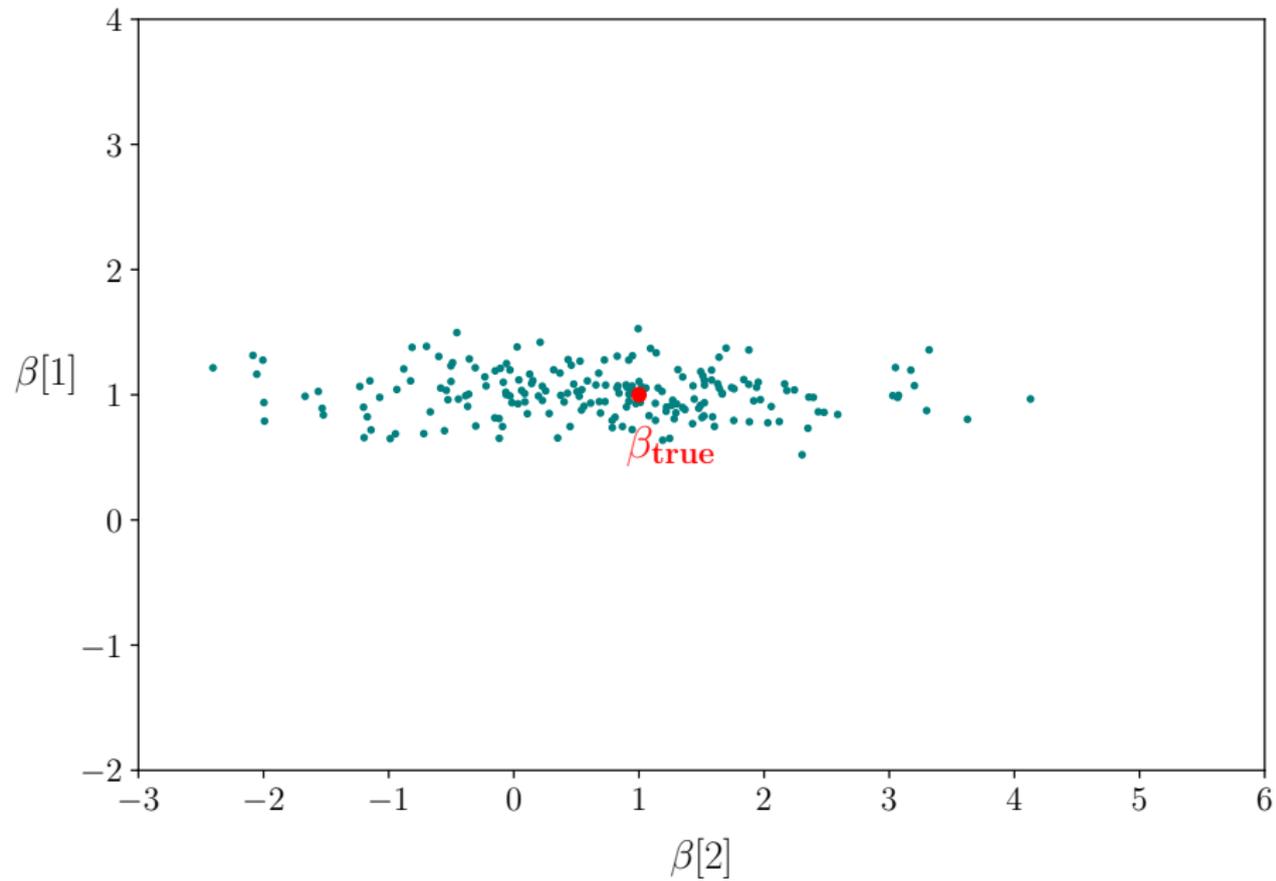
$k = 3$



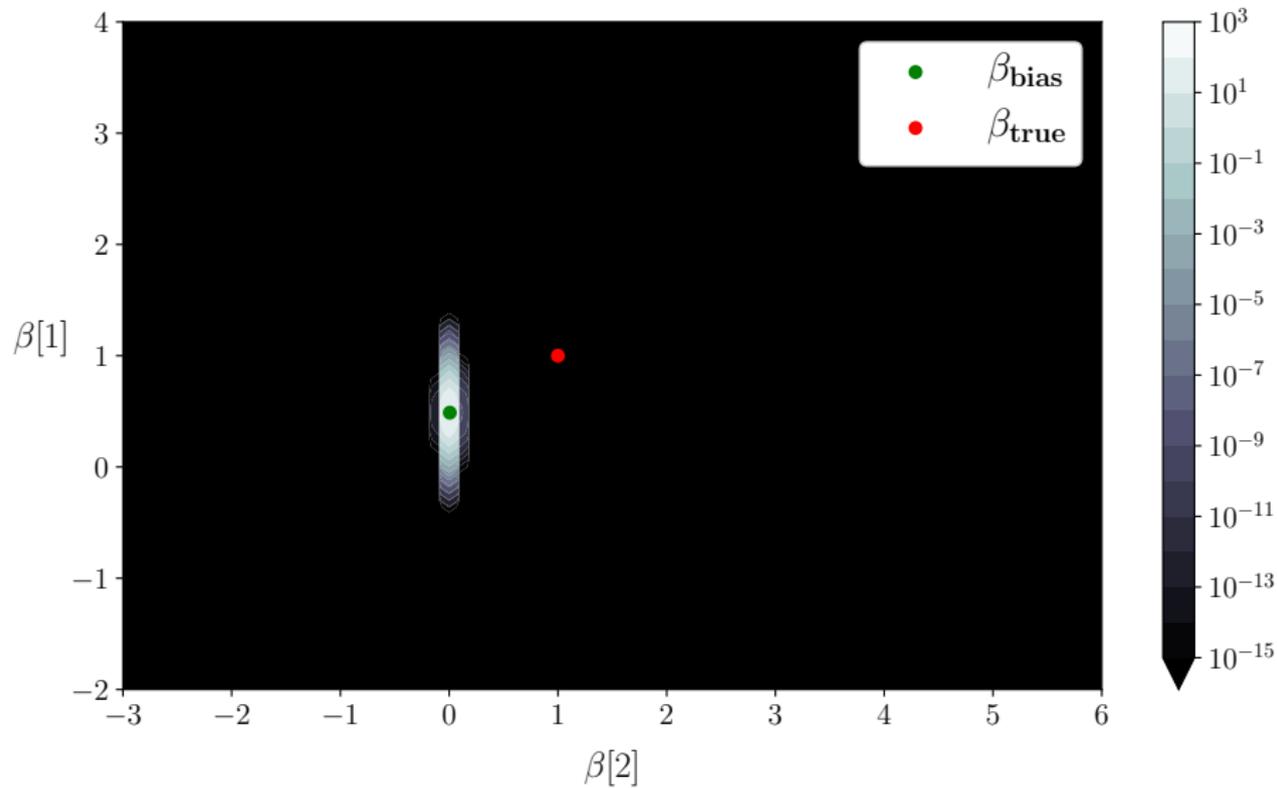
$k = 50$



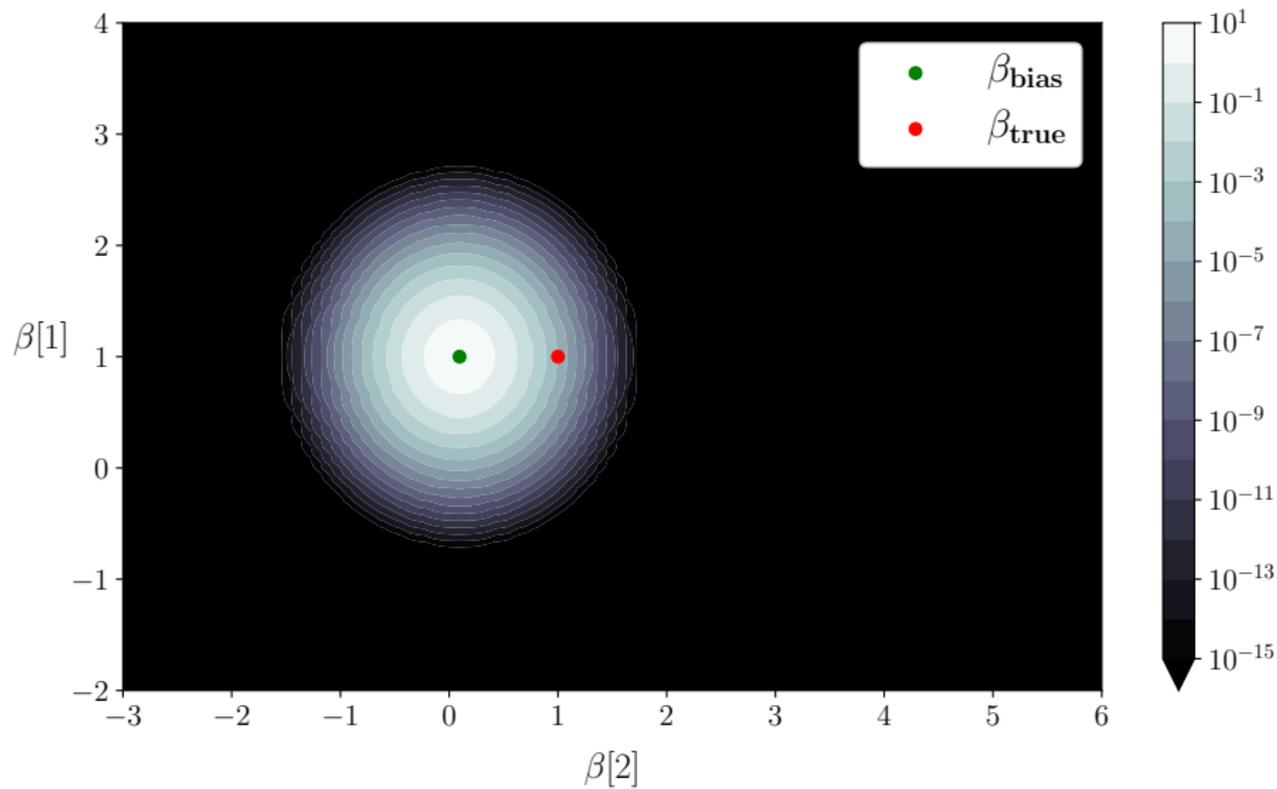
$k = 500$



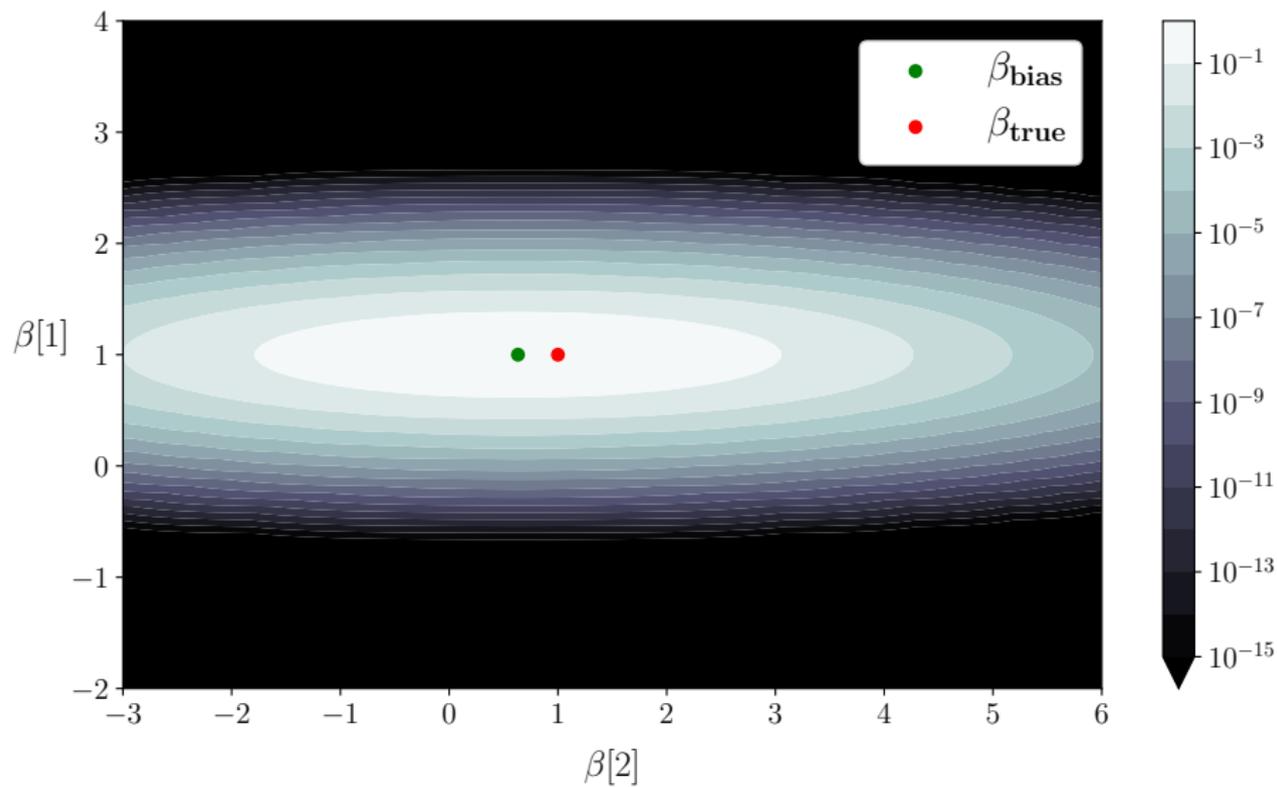
$k = 3$



$k = 50$



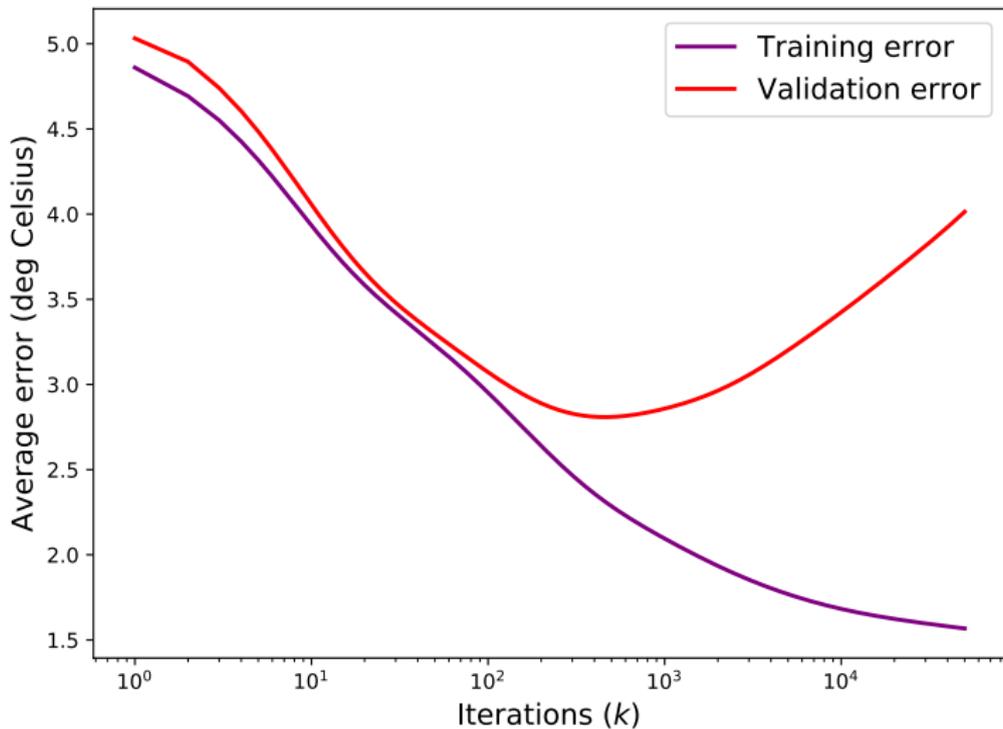
$k = 500$



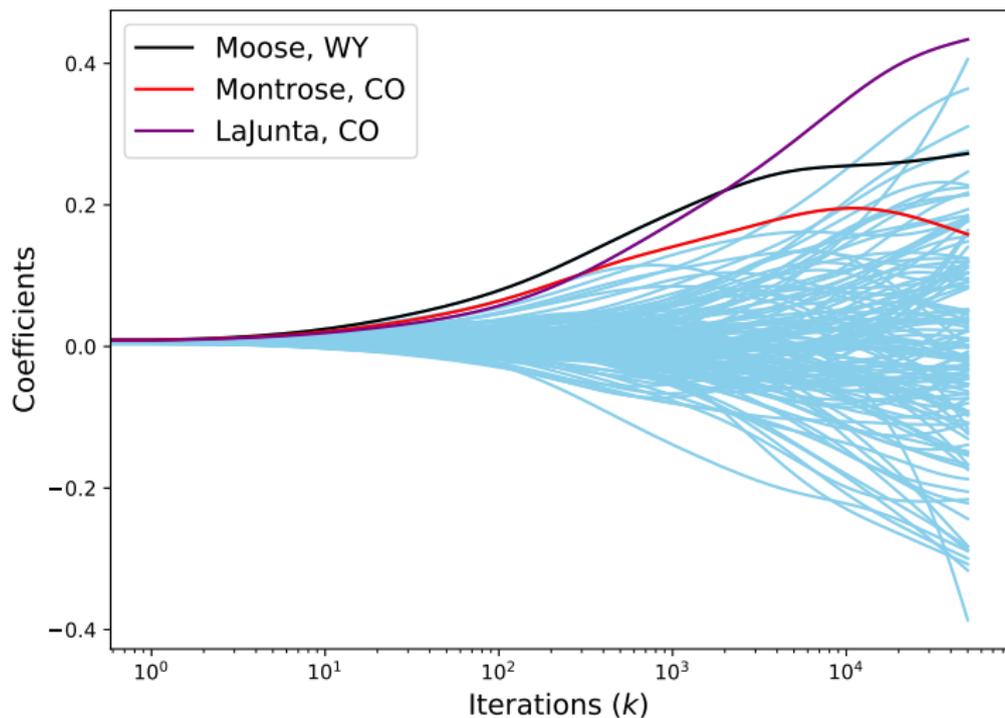
Temperature prediction via linear regression

- ▶ Dataset of hourly temperatures measured at weather stations all over the US
- ▶ Goal: Predict temperature in Yosemite from other temperatures
- ▶ Response: Temperature in Yosemite
- ▶ Features: Temperatures in 133 other stations ($p = 133$) in 2015
- ▶ Test set: 10^3 measurements
- ▶ Additional test set: All measurements from 2016

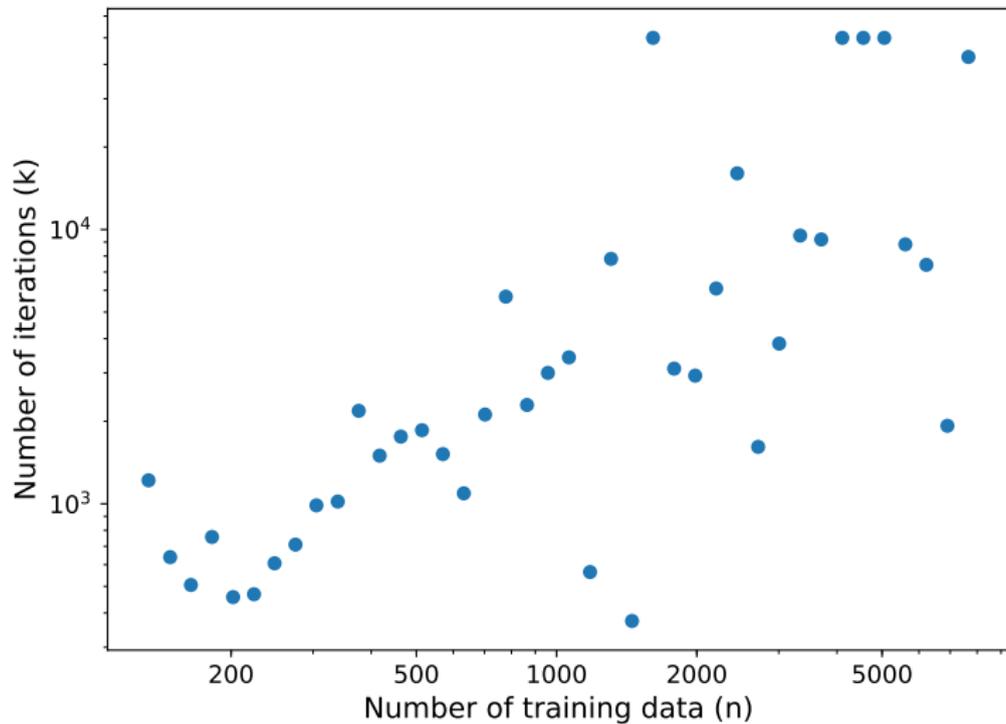
Gradient-descent estimator ($n = 200$)



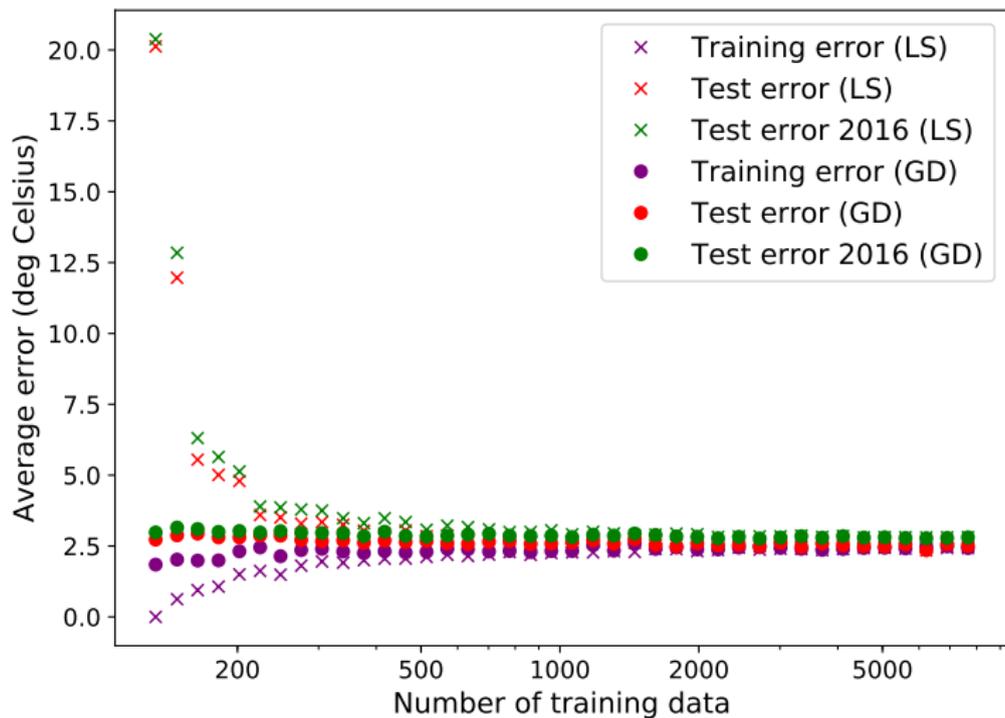
Gradient-descent estimator ($n = 200$)



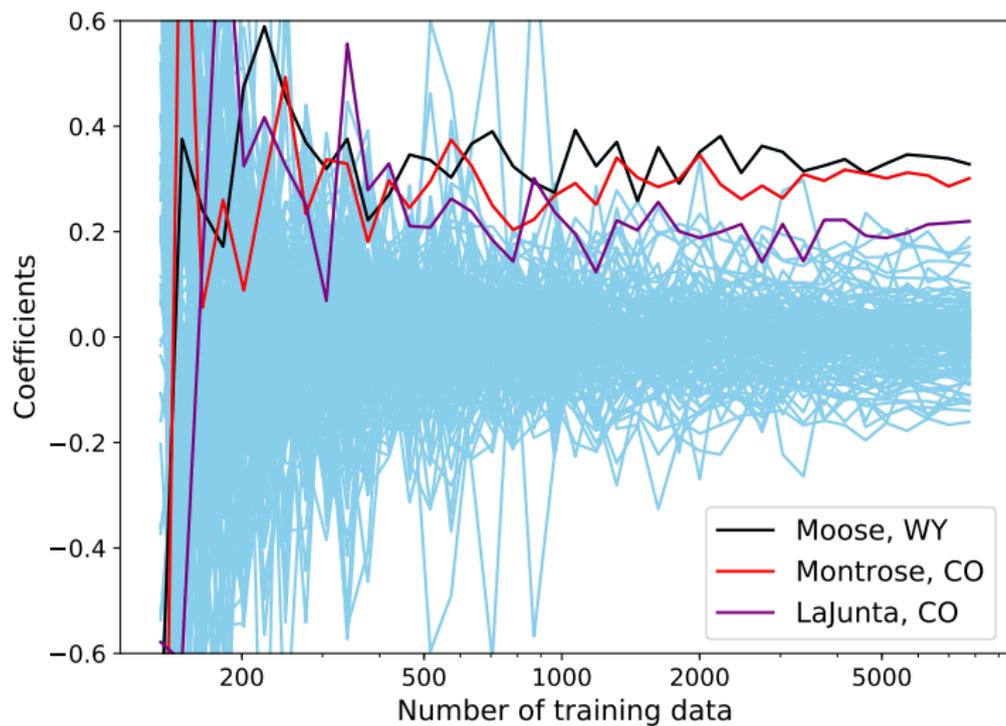
Selected number of iterations



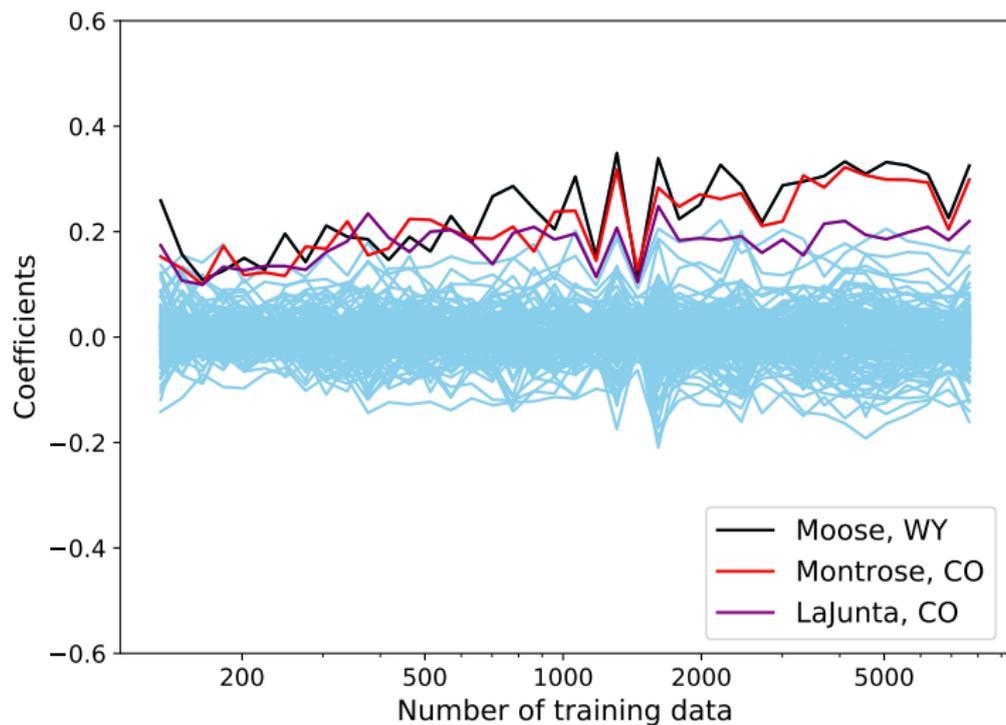
Comparison to least squares



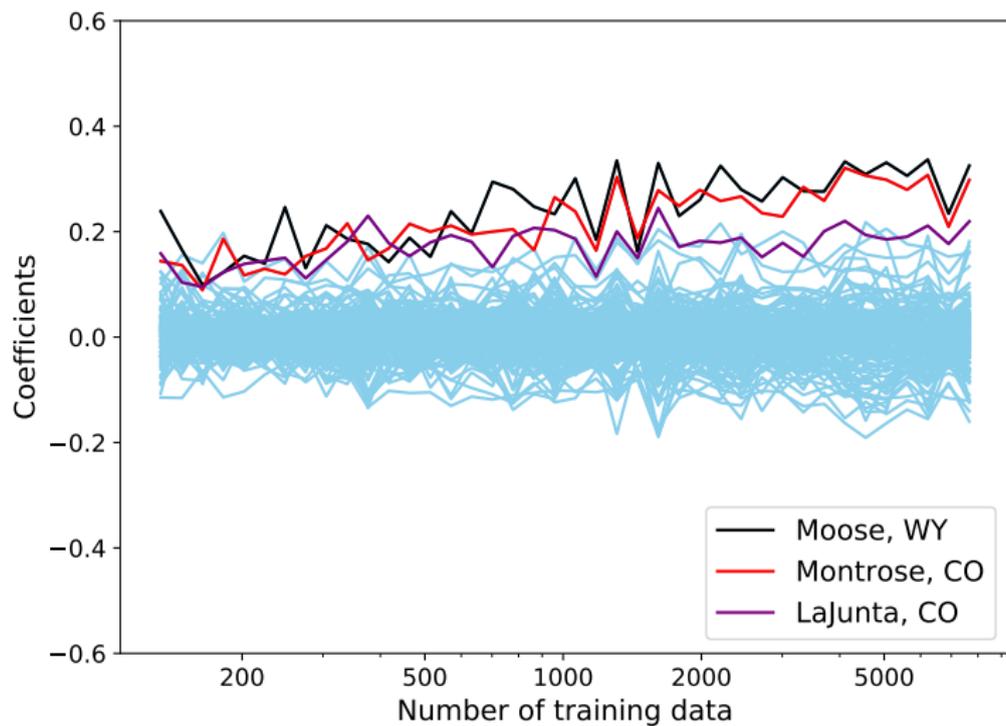
Least-squares coefficients



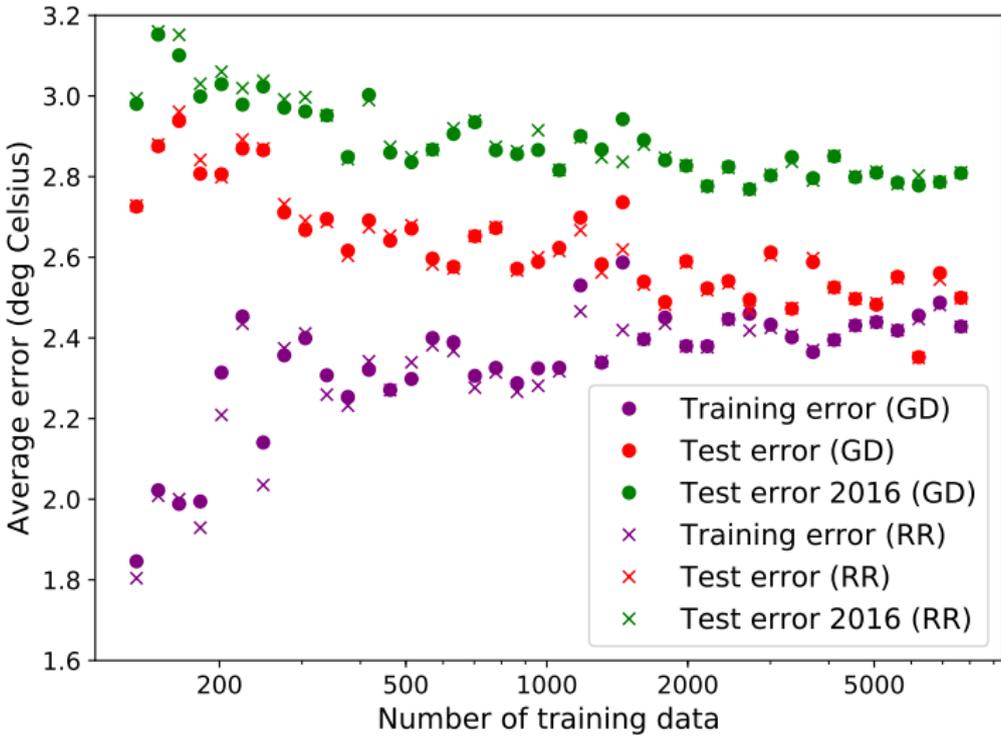
Gradient-descent coefficients



Ridge-regression coefficients



Comparison to ridge regression



What have we learned

- ▶ Gradient descent converges OLS estimate
- ▶ On the way it produces a biased estimate (under linear data model with additive noise)
- ▶ The estimate balances bias and variance from small singular values of feature matrix, just like ridge regression!