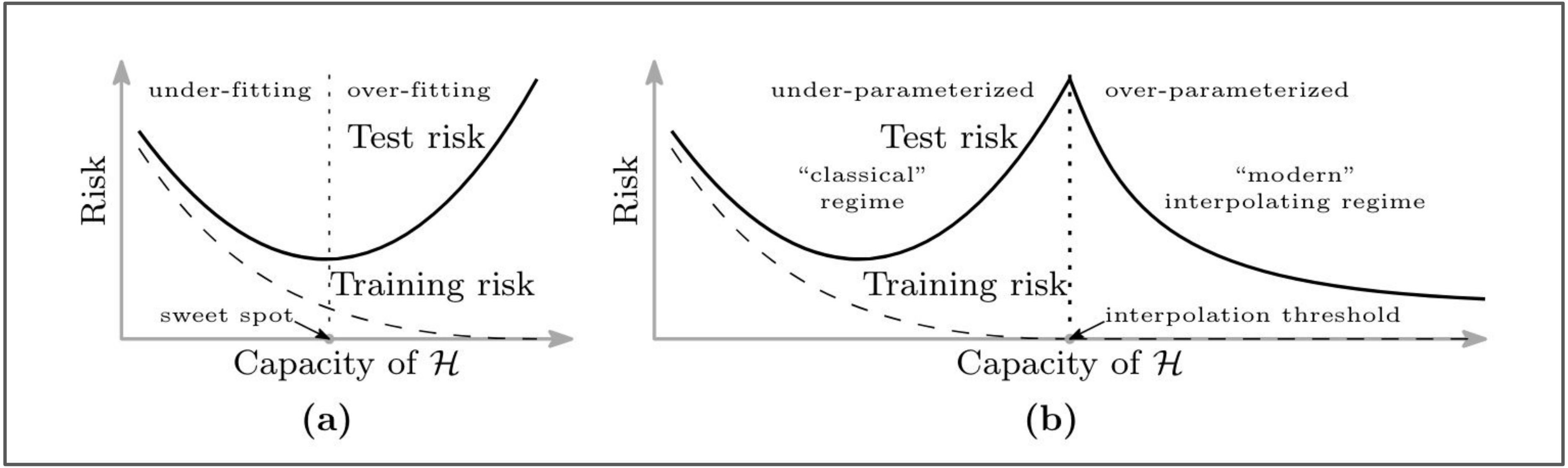
High-Dimensional Regression

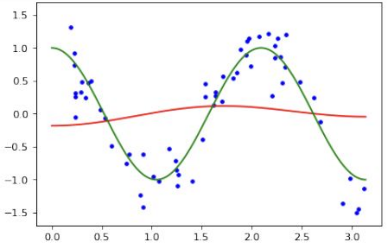
July 25, 2024
(The talk involved discussion using the whiteboard also)

Double Descent, Interpolation Regime

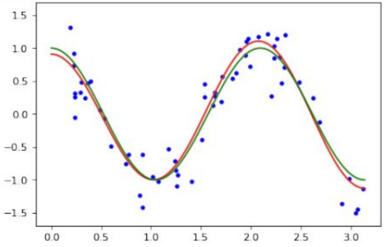


- Smallest risk can be in the overparameterized regime
- Overfitting is "benign" when highly overparameterized

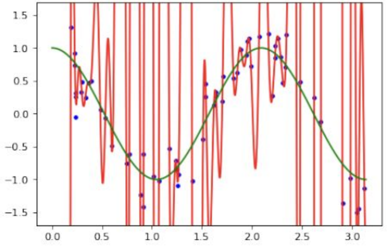
Benign Overfitting



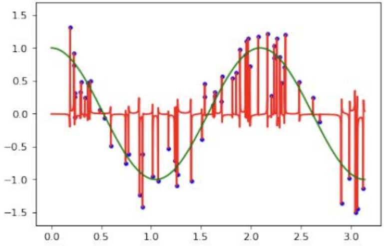
(a) Features $\{\cos(mx)\}_{m=1}^{2}$: underfitting. A linear combination of features cannot approximate the true dependence.



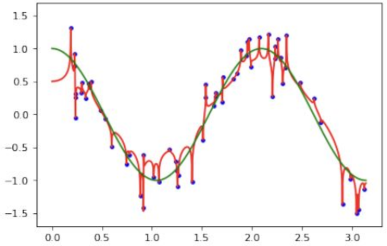
(b) Features $\{\cos(mx)\}_{m=1}^{3}$: the best fit. This is the minimum number of features that span the true dependence.



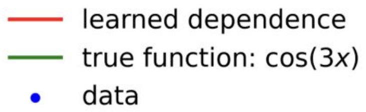
(c) Features $\{\cos(mx)\}_{m=1}^{50}$: overfitting. As the number of features approaches the number of data points, the effect of the noise becomes stronger.



(d) Features $\{\cos(mx)\}_{m=1}^{2000}$: isotropic overparameterization. As the number of cosine features grows above the interpolation threshold, the learned solution goes to zero out of sample.



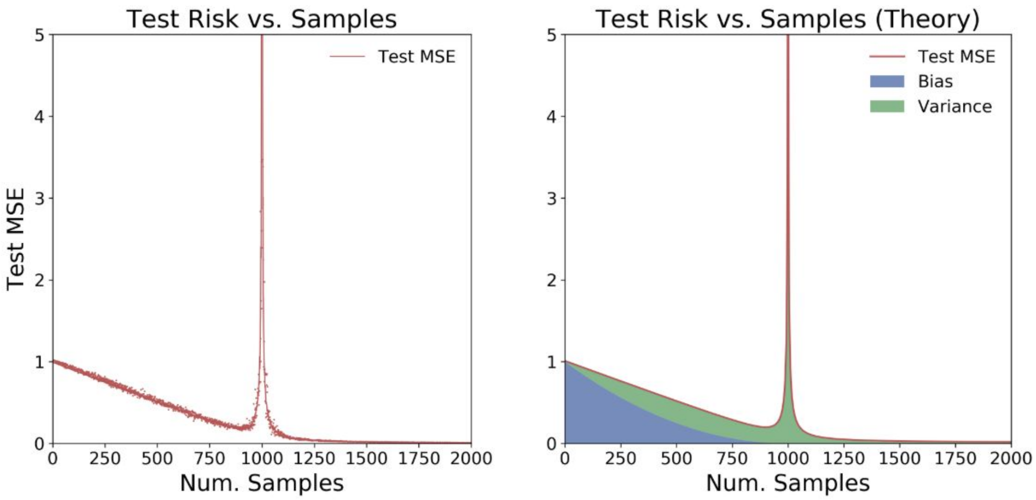
(e) Features $\{\cos(mx)/m\}_{m=1}^{2000}$: benign overfitting. Adding weights to cosine features results in interpolating the noise with high frequency features and learning the signal with low frequency features.



(f) Legend for all the plots.

Double Descent Sample-wise

- *n=p* is the interpolation threshold
- More data can hurt linear regr



(a) Test MSE for $d = 1000, \sigma = 0.1$. (b) Test MSE in theory for $d = 1000, \sigma = 0.1$

Figure 1: **Test MSE vs. Num. Train Samples for the min-norm ridgeless regression estimator in $d = 1000$ dimensions.** The distribution is a linear model with noise: covariates $x \sim \mathcal{N}(0, I_d)$ and response $y = \langle x, \beta \rangle + \mathcal{N}(0, \sigma^2)$, for $d = 1000, \sigma = 0.1$, and $||\beta||_2 = 1$. The estimator is $\hat{\beta} = X^\dagger y$. *Left:* Solid line shows mean over 50 trials, and individual points show a single trial. *Right:* Theoretical predictions for the bias, variance, and risk from Claims 1 and 2.

Today we will discuss:

Journal of Machine Learning Research 21 (2020) 1-16 Submitted 10/19; Revised 6/20; Published 7/20

The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization

Dmitry Kobak DMITRY.KOBAK@UNI-TUEBINGEN.DE
Institute for Ophthalmic Research
University of Tübingen
Otfried-Müller-Straße 25, 72076 Tübingen

Jonathan Lomond JONATHAN.LOMOND@GMAIL.COM
Toronto, Canada

Benoit Sanchez BEN.SAN3@GMAIL.COM
Paris, France

Editor: Ambuj Tewari

Empirically observed optimal ridge penalty for a “real-world” dataset

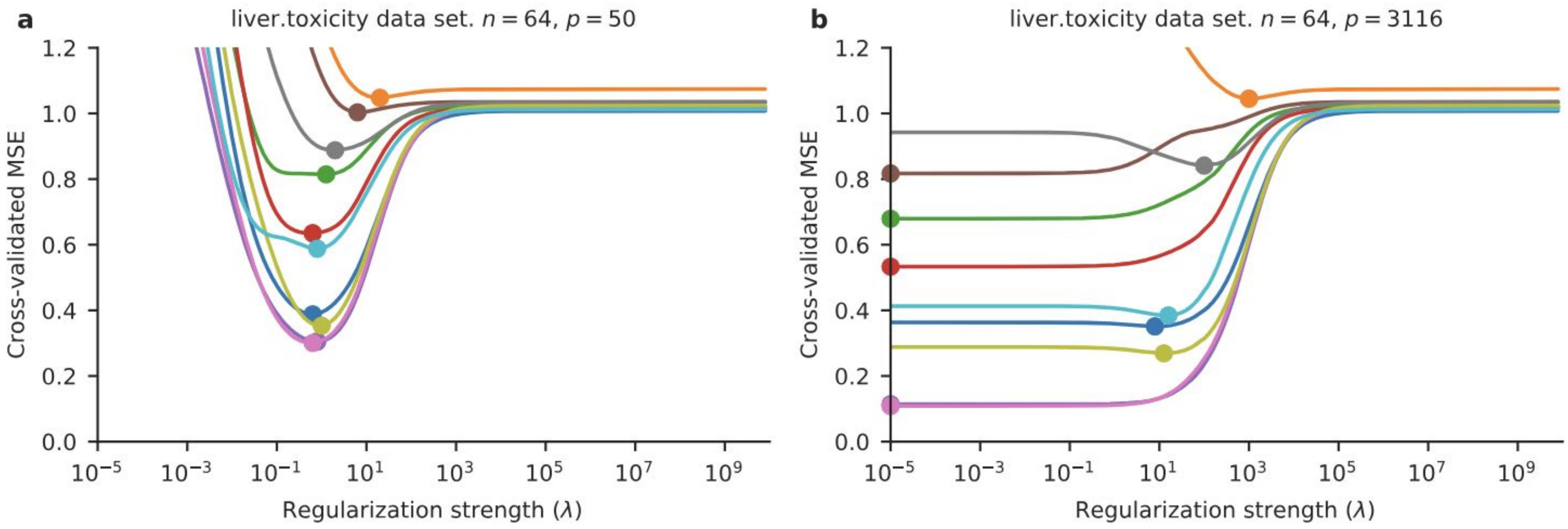
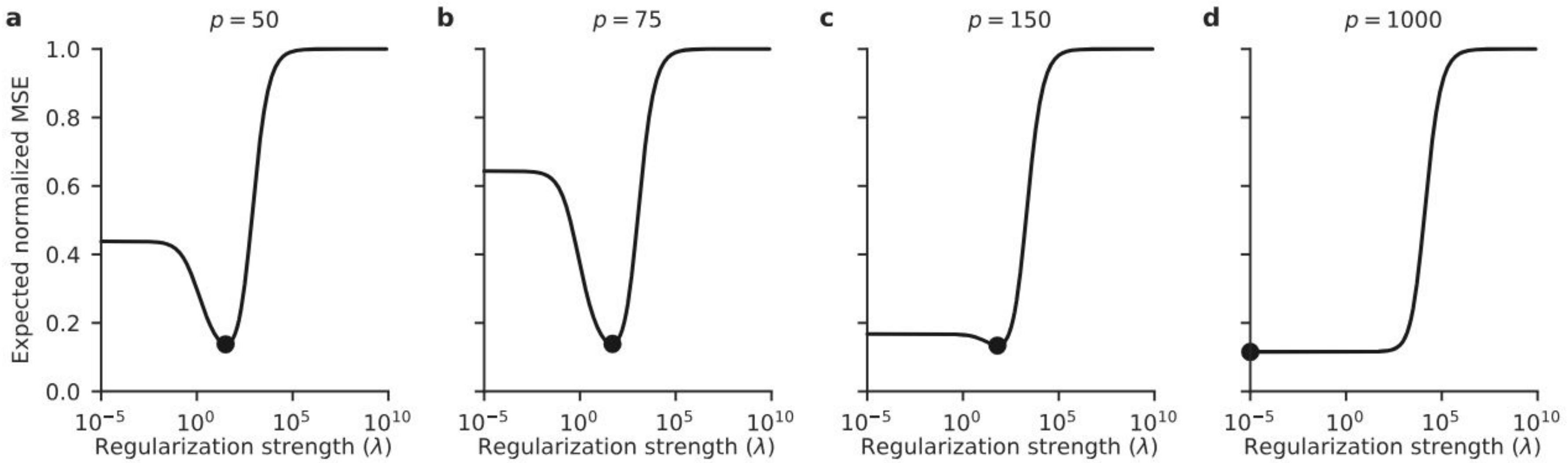


Figure 1: Cross-validation estimate of ridge regression performance for the `liver.toxicity` dataset. **a.** Using $p = 50$ randomly chosen predictors. **b.** Using all $p = 3116$ predictors. Lines correspond to 10 dependent variables. Dots show minimum values.

Simulation using spiked covariance model

$$\boldsymbol{\Sigma} = \mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top$$

- *n=64*
- *ρ=0.1*



Definition of the ridge estimator with negative λ?

- $\lambda \geq 0$

$$\hat{\beta}_\lambda := \arg\min_\beta \|y - X\beta\|^2 + \lambda\|\beta\|^2$$
$$= (X^\top X + \lambda I)^{-1} X y$$
$$= V \frac{S}{S^2+\lambda} U^\top y$$
$$(\text{where } X = USV^\top)$$

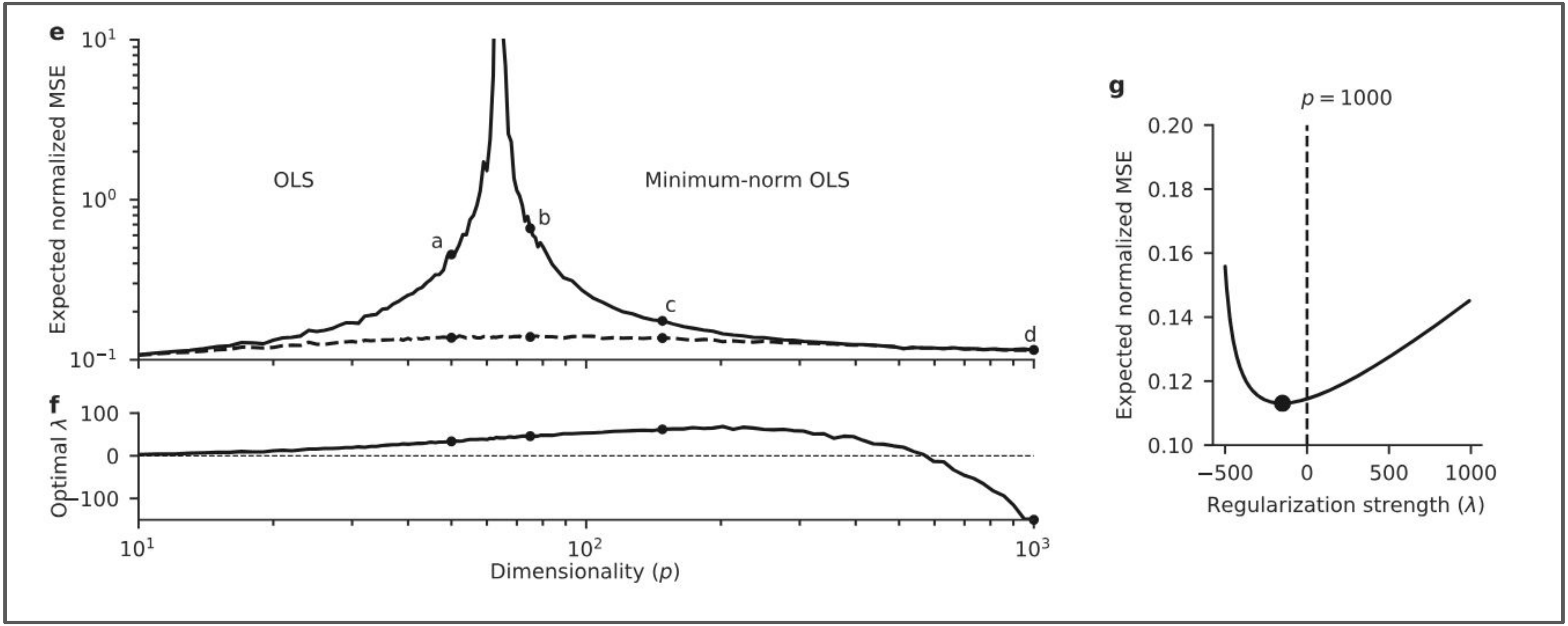
- $\lambda < 0$
  - Can't define using *argmin* problem, because its solution is not defined (since $||\beta|| \to \infty$)

$$\hat{\beta}_\lambda := V \frac{S}{S^2 + \lambda} U^\top y$$

computed $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{V}\frac{\mathbf{S}}{\mathbf{S}^2+\lambda}\mathbf{U}^\top\mathbf{y}$ for various values of λ and then found MSE (risk) of $\hat{\boldsymbol{\beta}}_\lambda$ using the formula

$$R(\hat{\boldsymbol{\beta}}_\lambda) = \mathbb{E}_{\mathbf{x},\varepsilon}\left[\left((\mathbf{x}^\top\boldsymbol{\beta} + \varepsilon) - \mathbf{x}^\top\hat{\boldsymbol{\beta}}_\lambda\right)^2\right] = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) + \sigma^2. \quad (8)$$

Simulation using spiked covariance model (contd)



Simulation using spiked covariance model (contd)

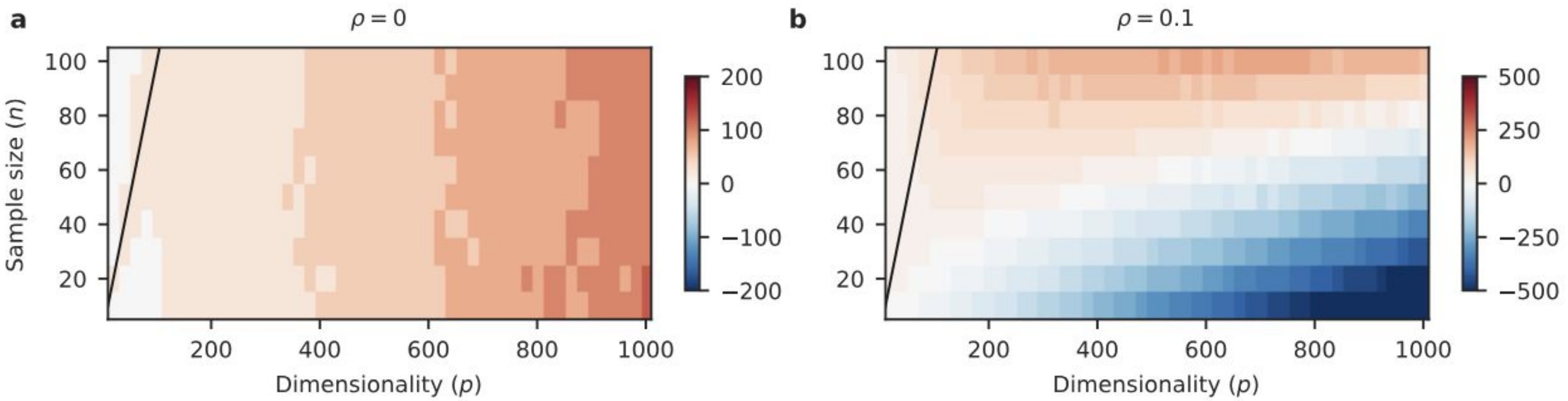


Figure 3: a. The optimal regularization parameter λ_{opt} as a function of sample size (n) and dimensionality (p) in the model with uncorrelated predictors ($\rho = 0$). In this case $\lambda_{\text{opt}} = p\sigma^2/\|\boldsymbol{\beta}\| = p/\alpha$. Black line corresponds to $n = p$. **b.** The optimal regularization parameter λ_{opt} in the model with correlated predictors ($\rho = 0.1$).

Analysis for the spiked covariance model

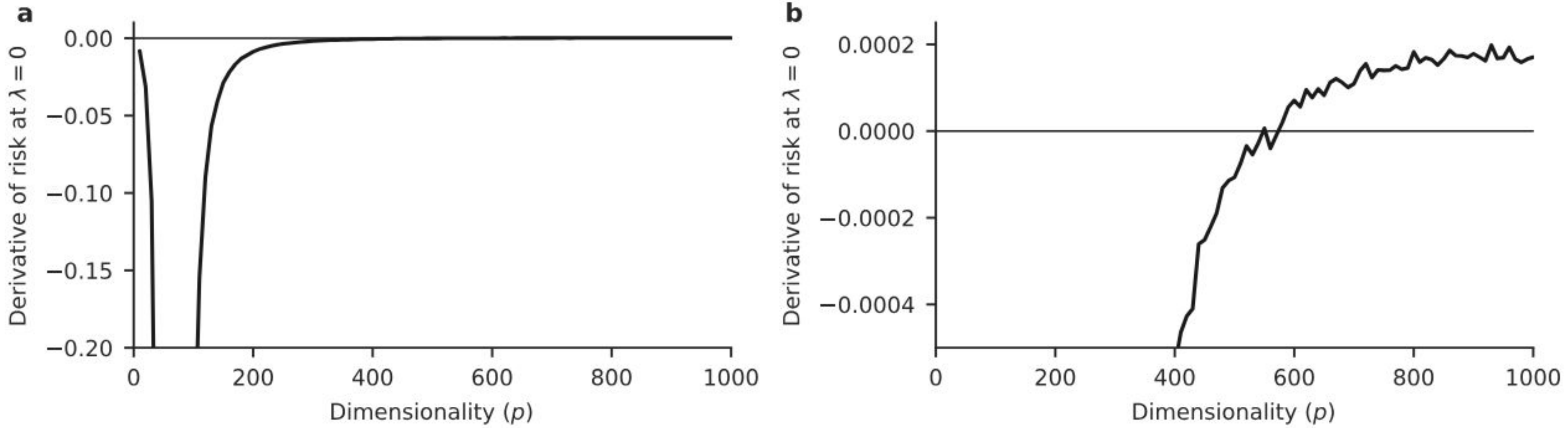


Figure 5: a. The derivative of the expected risk as a function of ridge penalty λ at $\lambda = 0$, in the model with p weakly correlated predictors (Eq. 24). Sample size $n = 64$. **b.** Zoom-in into panel (a). The derivative becomes positive for $p \gtrsim 600$, implying that $\lambda_{\mathrm{opt}} \leq 0$.

Discussion