Implicit Regularization in Over-Parameterized Support Vector Machine

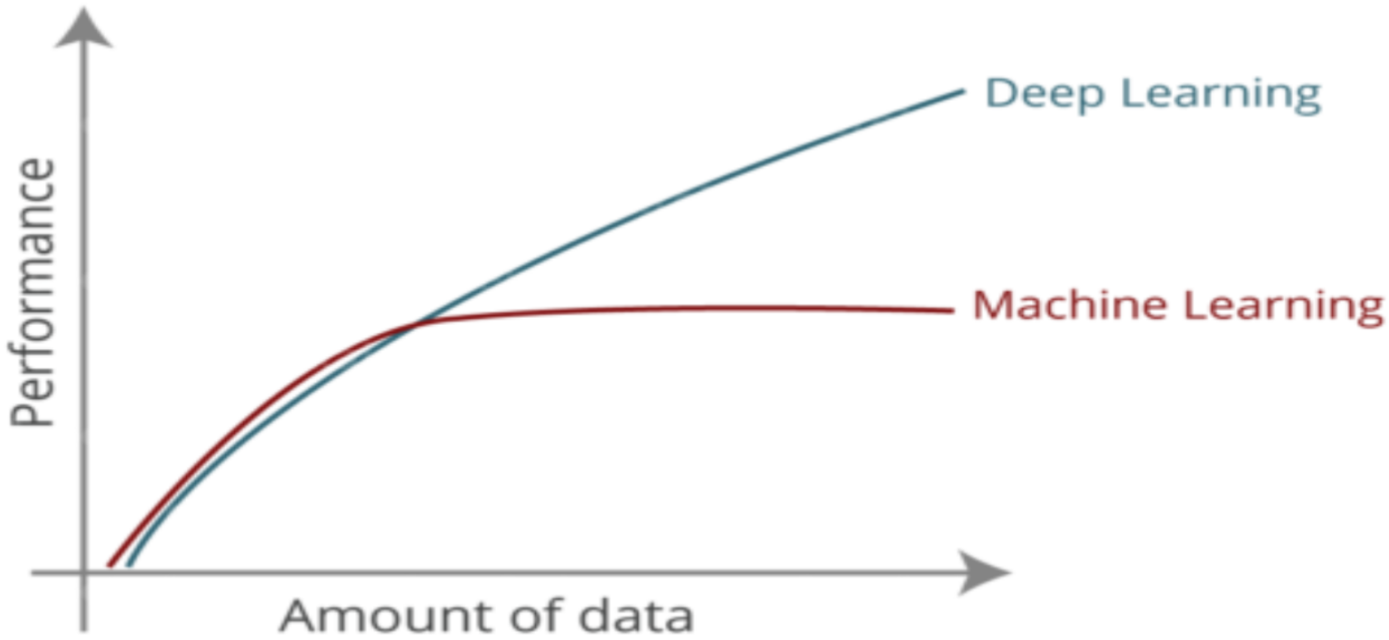
Yang Sui, Xin He and Yang Bai

Shanghai University of Finance and Economics

NeurIPS 2023

Motivation

- The performance of Deep Learning



- Applications of Deep Learning
  - Imaging classification
  - Natural language processing
  - Artificial intelligence

Motivation

- Applying deep learning to tasks such as regression and classification, the regression function or classifier is represented by a deep neural network
- Difficulties:
  - the loss function is nonconvex, with saddle points and local minima
  - the neural network is over-parameterized
- Simple algorithms such as gradient descent tend to find the global minimum of the loss function

Implicit Regularization

- Neyshabur et al. (2015), Zhang et al. (2016) show that the generalization stems from an **implicit regularization** of the optimization algorithm
- In over-parametrized models, although the optimization problems consist of bad local minima, the choice of optimization algorithm, usually gradient descent, guard the iterates from local minima
- Without any regularization in optimization objective, the implicit preference of the algorithm itself plays the role of regularization

Literature Review

- Matrix factorization: Gunasekar et al. (2017); Li et al. (2018)
  - $\min ||\mathbf{AX} - \mathbf{Y}||_2^2$, rewrite the parameter as $\mathbf{X} = \mathbf{U}\mathbf{U}^T$, estimate the true parameter by updating $\mathbf{U}$ via gradient descent
  - gradient descent biases towards the minimum nuclear norm solution

- Linear regression: Vaskevicius et al. (2019); Zhao et al. (2019)
  - re-parametrize the parameter using two vectors $\boldsymbol{\beta} = \mathbf{g} \odot \mathbf{h}$ or $\boldsymbol{\beta} = \mathbf{g} \odot \mathbf{g} - \mathbf{h} \odot \mathbf{h}$ via the Hadamard product
  - gradient descent yields an estimator with optimal statistical accuracy

Our Goal

- Study the implicit regularization of gradient descent for high-dimensional SVM
- Consider the non-differentiability of hinge loss
- Provide evidence of implicit regularization from theoretical and empirical perspective

Over-parameterization for ℓ_1-regularized SVM

Given a random sample $\mathcal{Z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, ℓ_1-regularized SVM that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \mathbf{x}_i^T \boldsymbol{\beta})_+ + \lambda \|\boldsymbol{\beta}\|_1,$$

Directly minimize the hinge loss and rewrite $\boldsymbol{\beta} = \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$,

$$\mathcal{E}_{\mathcal{Z}^n}(\mathbf{w}, \mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} \big(1 - y_i \mathbf{x}_i^T (\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v})\big)_+.$$

Update $\mathbf{w}_t$ and $\mathbf{v}_t$ via gradient descent and $\boldsymbol{\beta}_t = \mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t$

Over-parameterization for ℓ_1-regularized SVM

- The dimensionality of $\boldsymbol{\beta}$ is p, but $2p$-dimensional parameter is involved
- $\|\boldsymbol{\beta}\|_1 = \arg\min_{\boldsymbol{\beta}=\mathbf{a}\odot\mathbf{c}}(\|\mathbf{a}\|^2 + \|\mathbf{c}\|^2)/2$
- ℓ_1 regularization is to $\min_{\mathbf{a},\mathbf{c}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{a},\mathbf{c}) + \lambda\big(\|\mathbf{a}\|^2 + \|\mathbf{c}\|^2\big)/2$
- $\mathbf{w} = \frac{\mathbf{a}+\mathbf{c}}{2}$ and $\mathbf{v} = \frac{\mathbf{a}-\mathbf{c}}{2}$ and then $\boldsymbol{\beta} = \mathbf{w}\odot\mathbf{w} - \mathbf{v}\odot\mathbf{v}$
- Drop the explicit ℓ_2-regularized term and perform gradient descent to minimize $\mathcal{E}_{\mathcal{Z}^n}(\mathbf{w},\mathbf{v})$, following the neural network learning

Nesterov's smoothing

- Hinge loss function is not differentiable
- First-order methods such as sub-gradient and stochastic gradient methods converge slowly and are not suitable for large-scale problems
- Second-order methods like Newton and Quasi-Newton methods achieve better convergence rates, but the computational cost is expensive

Nesterov's smoothing:

$$\min_{\mathbf{w},\mathbf{v}} \mathcal{E}_{\mathcal{Z}^n}(\mathbf{w},\mathbf{v}) \equiv \min_{\mathbf{w},\mathbf{v}} \max_{\boldsymbol{\mu}\in\mathcal{P}_1} \frac{1}{n}\sum_{i=1}^{n}\big(1 - y_i\mathbf{x}_i^T(\mathbf{w}\odot\mathbf{w} - \mathbf{v}\odot\mathbf{v})\big)\mu_i,$$

where $\mathcal{P}_1 = \{\boldsymbol{\mu} \in \mathbb{R}^n : 0 \leq \mu_i \leq 1\}$.

Nesterov's smoothing

The above saddle point function can be smoothed by subtracting a prox-function $d_\gamma(\boldsymbol{\mu}) = \frac{\gamma}{2}\|\boldsymbol{\mu}\|^2$,

$$\mathcal{E}^*_{\mathcal{Z}^n,\gamma}(\mathbf{w},\mathbf{v}) \equiv \max_{\boldsymbol{\mu}\in\mathcal{P}_1} \Big\{ \frac{1}{n}\sum_{i=1}^{n} \big(1 - y_i \mathbf{x}_i^T(\mathbf{w}\odot\mathbf{w} - \mathbf{v}\odot\mathbf{v})\big)\mu_i - d_\gamma(\boldsymbol{\mu}) \Big\},$$

μ_i can be obtained directly,

$$\mu_i = \text{median}\Big(0, \frac{1 - y_i {\mathbf{x}_i}^T(\mathbf{w}\odot\mathbf{w} - \mathbf{v}\odot\mathbf{v})}{\gamma n}, 1\Big)$$

Larger γ yields larger approximation error

Implicit regularization via gradient descent

Algorithm 1: Gradient Descent for High-Dimensional Sparse SVM.

Given: Training set $\mathcal{Z}^n$, initial value α, stepsize η, proxy parameter γ, maximum iteration number T_1, validation set $\widetilde{\mathcal{Z}}^n$;
Initialize: $\mathbf{w}_0 = \alpha\mathbf{1}$, $\mathbf{v}_0 = \alpha\mathbf{1}$, and set iteration index $t = 0$.
While $t < T_1$, **do**

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t + 2\eta\frac{1}{n}\sum_{i=1}^{n} y_i\mu_{t,i}\mathbf{x}_i \odot \mathbf{w}_t; \\
\mathbf{v}_{t+1} &= \mathbf{v}_t - 2\eta\frac{1}{n}\sum_{i=1}^{n} y_i\mu_{t,i}\mathbf{x}_i \odot \mathbf{v}_t; \\
\boldsymbol{\beta}_{t+1} &= \mathbf{w}_{t+1} \odot \mathbf{w}_{t+1} - \mathbf{v}_{t+1} \odot \mathbf{v}_{t+1}; \\
\mu_{t+1,i} &= \text{median}\left(0, \frac{1 - y_i{\mathbf{x}_i}^T\boldsymbol{\beta}_{t+1}}{n\gamma}, 1\right); \\
t &= t + 1;
\end{aligned}$$

End if $t > T_1$ or $\boldsymbol{\mu}_{t+1} = \mathbf{0}$.
Return Set $\widehat{\boldsymbol{\beta}}$ as $\boldsymbol{\beta}^t$.

Implicit regularization via gradient descent

Remarks:

- $\mathbf{w}_0 = \mathbf{v}_0 = \alpha \mathbf{1}_{p\times 1}$, where $\alpha > 0$ is a small constant. The zero component is initialized close to zero, while the non-zero component receives a non-zero initialization
- the stopping condition can be determined based on the value of $\boldsymbol{\mu}$
- Computational cost is the vector multiplication. Significant portion of the elements in $\boldsymbol{\mu}$ are either 0 or 1, and the proportion of these elements increases substantially as γ decreases

Theoretical Analysis

Assumpotions

- True parameters: $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}(1 - y\mathbf{x}^T\boldsymbol{\beta})_+$. $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is s-sparse signal
- Let $S \subset \{1, \ldots, p\}$ denote support of $\boldsymbol{\beta}^*$, and the size $|S|$ of S is s
- Within the s nonzero signal components of $\boldsymbol{\beta}^*$, define the index set of strong signals as $S_1 = \{i \in S : |\beta_i^*| \geq C_s \log p \sqrt{\log p/n}\}$ and weak signals as $S_2 = \{i \in S : |\beta_i^*| \leq C_w \sqrt{\log p/n}\}$. s_1 and s_2 are the cardinalities of S_1 and S_2
- $m = \min_{i \in S_1} |\beta_i^*|$ and κ is the condition number as the ratio between the largest absolute value of strong signal to the smallest one

Theoretical Analysis

Assumpotions

- The design matrix $\mathbf{X}/\sqrt{n}$ satisfies δ-incoherence with $0 < \delta \lesssim 1/(\kappa s \log p)$. In addition, every entry x of $\mathbf{X}$ is $i.i.d.$ zero-mean sub-Gaussian random variables with bounded sub-Gaussian norm σ
- The initialization size α satisfies $0 < \alpha \lesssim 1/p$, the parameter of prox-function γ satisfies $0 < \gamma \leq 1/n$, and the stepsize η satisfies $0 < \eta \lesssim 1/(\kappa \log p)$

Theoretical Analysis

Theory 1

Suppose that Assumptions hold, then there exist positive constants c_1, c_2, c_3 and c_4 such that there holds with probability at least $1 - c_1 n^{-1} - c_2 p^{-1}$ that, for every t with $c_3 \log(m/\alpha^2)/(\eta m) \leq t \leq c_4 \log(1/\alpha)/(\eta \log n)$, the solution of the t-th iteration in Algorithm, $\boldsymbol{\beta}_t = \mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t$, satisfies

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 \lesssim \frac{s \log p}{n}.$$

- the convergence rate in terms of the ℓ_2-norm is $\mathcal{O}(\sqrt{s \log p/n})$
- Such a convergence rate matches the near-oracle rate of sparse SVM and can be attained through explicit regularization like ℓ_1-norm penalty, as well as concave penalties

Proof Sketch

- We can control the estimated strengths associated with the non-signal and weak signal components, denoted as $\|\mathbf{w}_t \odot \mathbf{1}_{S_1^c}\|_\infty$ and $\|\mathbf{v}_t \odot \mathbf{1}_{S_1^c}\|_\infty$ at the same order as the square root of the initial value α up to $\mathcal{O}(\log(1/\alpha)/(\eta \log n))$ steps. α governs the size of coordinates outside the signal support S_1
- Strong signal part, denoted as $\boldsymbol{\beta}_t \odot \mathbf{1}_{S_1}$, grows exponentially with an accuracy of approximately $\mathcal{O}(\log p/n)$ around the true parameter $\boldsymbol{\beta}^* \odot \mathbf{1}_{S_1}$ within approximately $\mathcal{O}(\log(m/\alpha^2)/(\eta m))$ steps

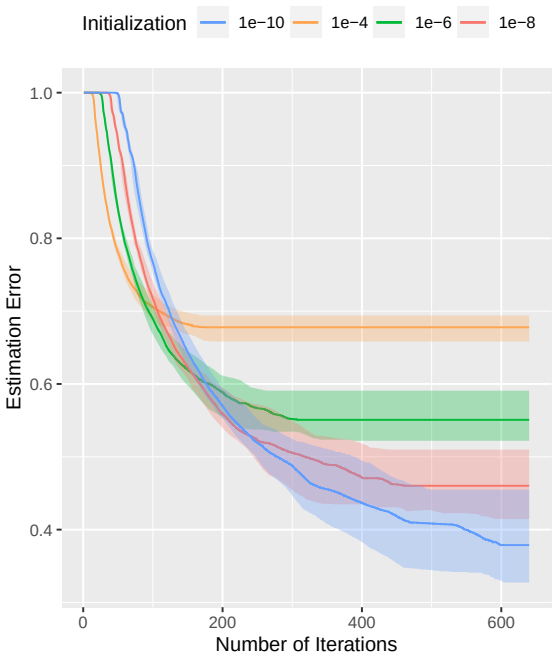
Numerical Study

Setup:

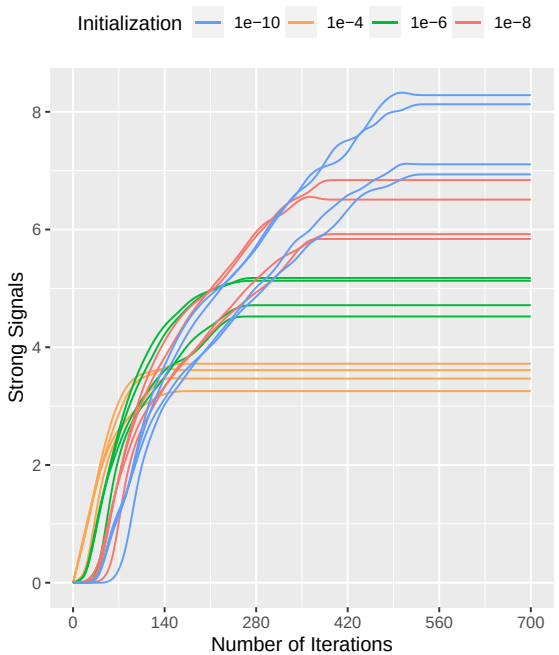
- Generate $3n$ independent observations, divided evenly into three parts: one for training, one for validation, and one for testing
- $\boldsymbol{\beta}^* = m\mathbf{1}_S$ for a constant m
- Entries of $\mathbf{X}$ are sampled as $i.i.d.$ zero-mean Gaussian r.v.s, and the labels y follow a binomial distribution
- True signal strength $m = 10$, number of signals $s = 4$, sample size $n = 200$, dimension $p = 400$, stepsize $\eta = 0.5$, prox-parameter $\gamma = 10^{-4}$, and initialization size $\alpha = 10^{-8}$
- Estimation error: $\|\boldsymbol{\beta}_t/\|\boldsymbol{\beta}_t\| - \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|\|$, prediction accuracy: $P(\hat{y} = y_{test})$
- Variable selection error "False positive" and "True negative"

Effects of Small Initialization

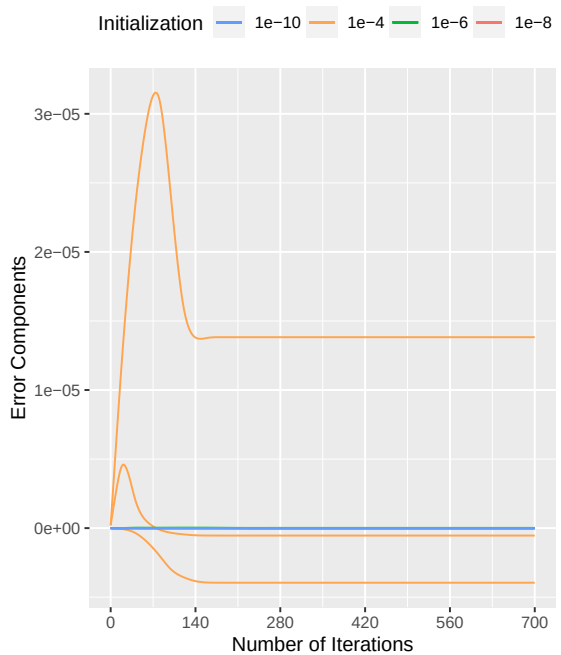
Initialization size: $\alpha = \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}\}$



Effects of Initialization



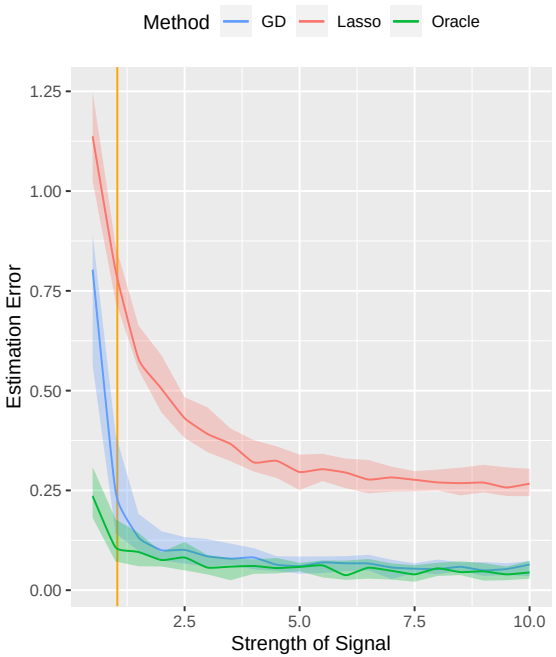
Paths of Signals



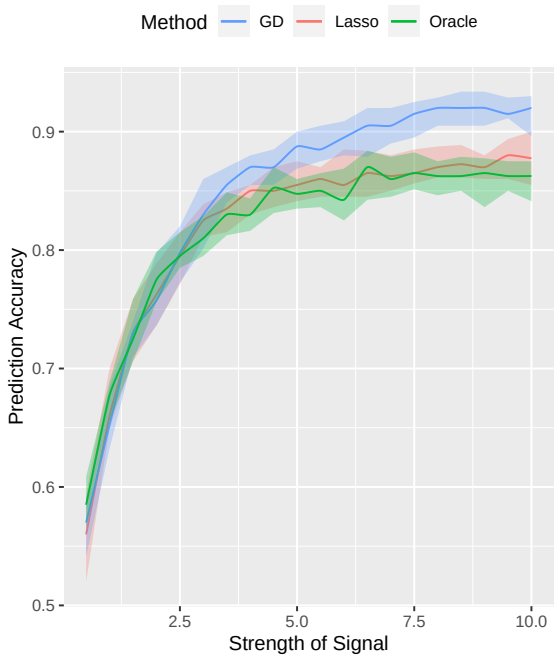
Paths of Errors

Effects of Signal Strength and Sample Size

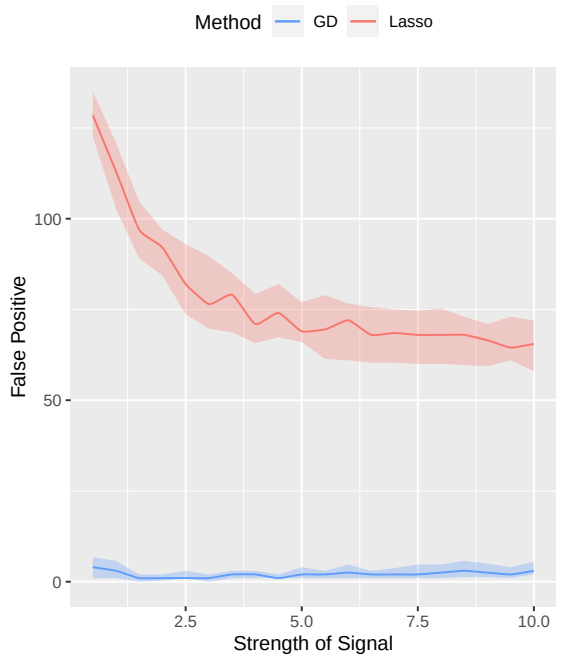
True signal strength: $m = 0.5 * k, k = 1, \ldots, 20$



Estimation Result



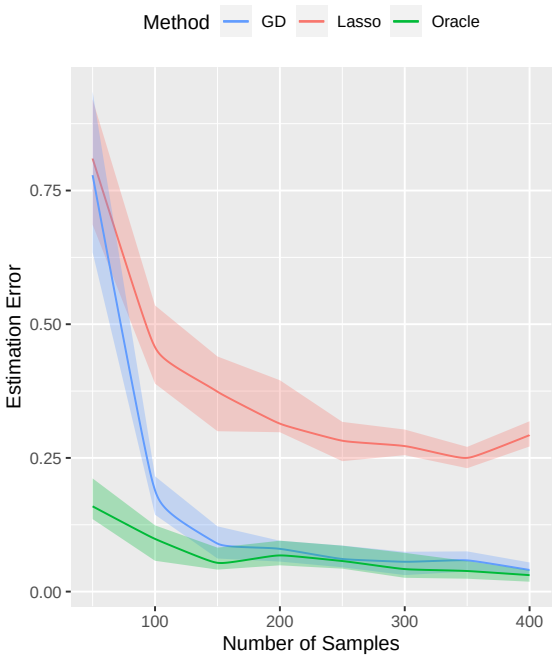
Prediction Result



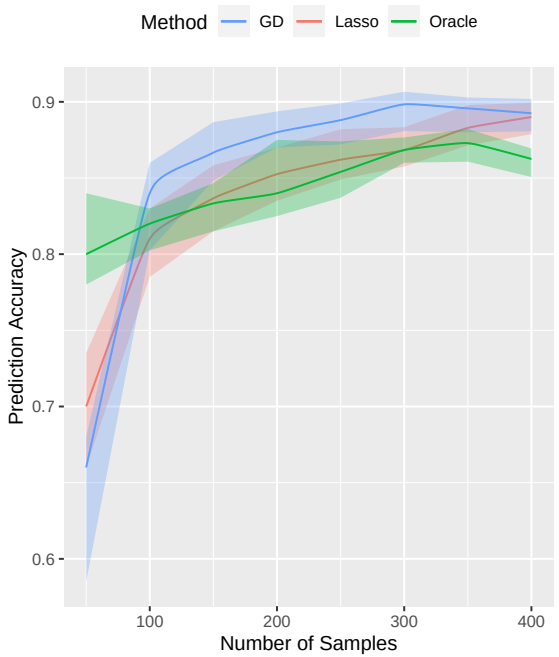
Variable Selection Result

Effects of Signal Strength and Sample Size

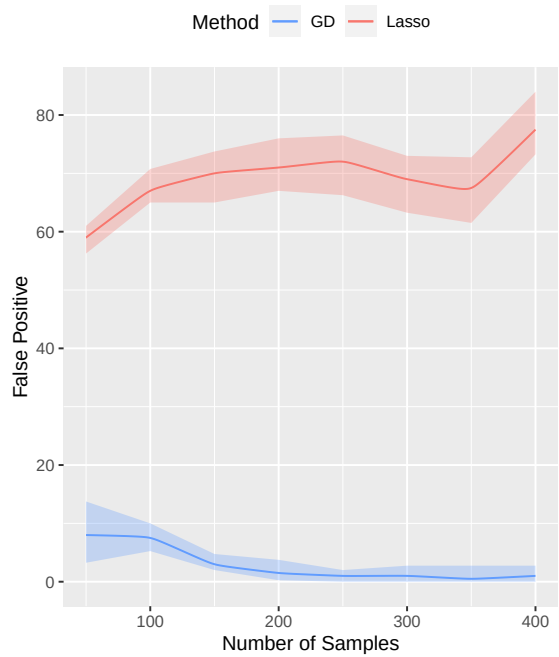
Sample size: $n = 50 * k$ for $k = 1, \ldots, 8$



Estimation Result



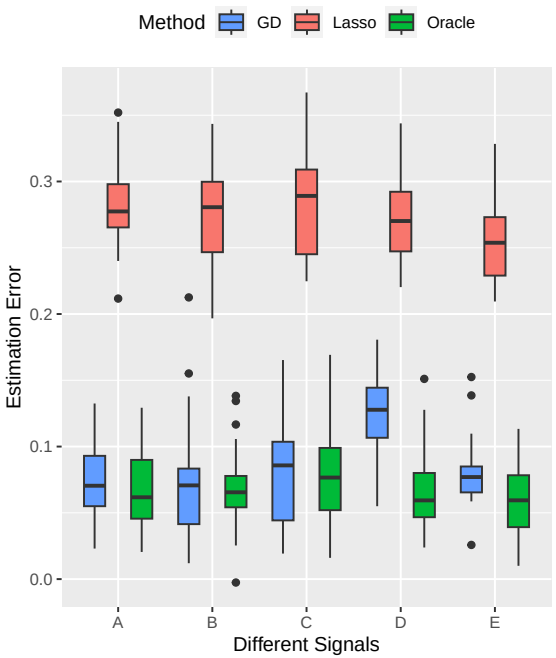
Prediction Result



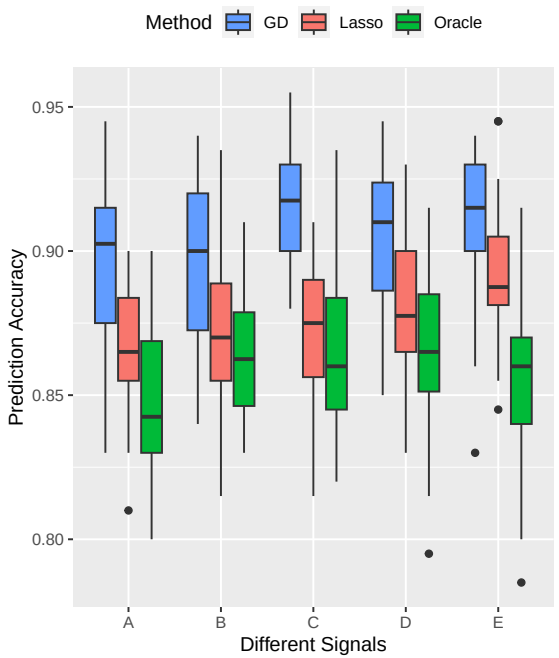
Variable Selection Result

Performance on Complex Signal Structure

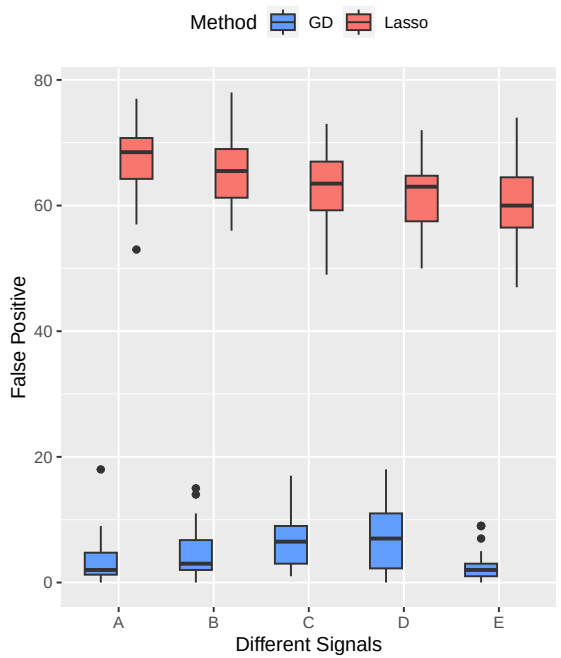
Five signal structures: $\mathbf{A} - (5, 6, 7, 8)$, $\mathbf{B} - (4, 6, 8, 9)$, $\mathbf{C} - (3, 6, 9, 10)$, $\mathbf{D} - (2, 6, 10, 11)$ and $\mathbf{E} - (1, 6, 11, 12)$



Estimation Result



Prediction Result

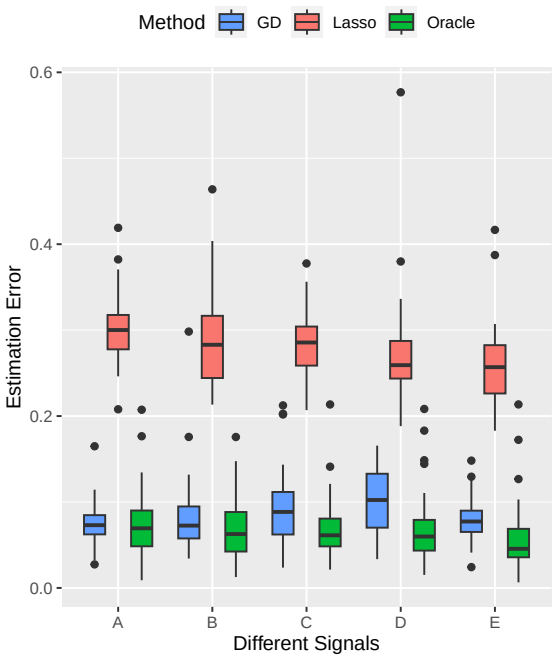


Variable Selection Result

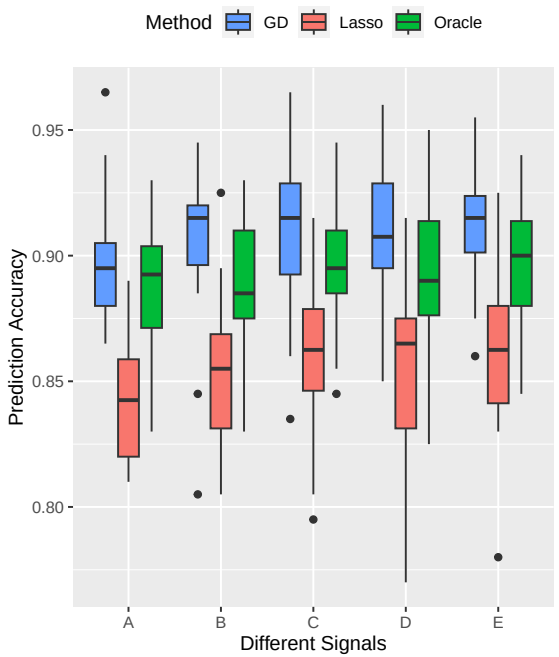
Performance on Heavy-tailed Distribution

Five signal structures: $\mathbf{A} - (5, 6, 7, 8)$, $\mathbf{B} - (4, 6, 8, 9)$, $\mathbf{C} - (3, 6, 9, 10)$, $\mathbf{D} - (2, 6, 10, 11)$ and $\mathbf{E} - (1, 6, 11, 12)$

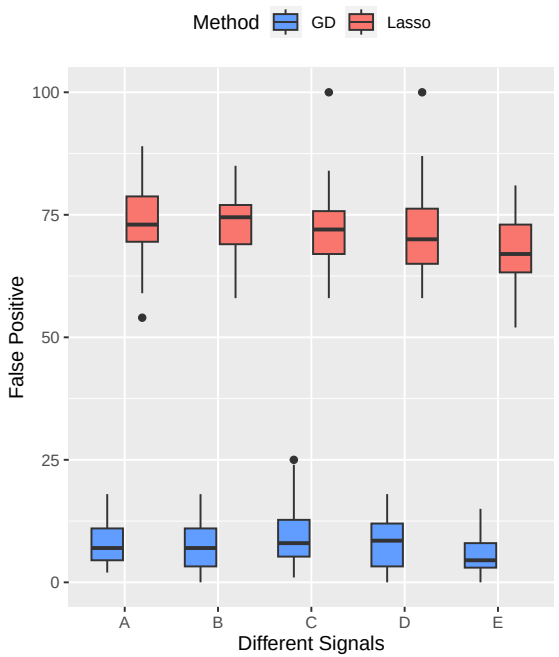
Sample $\mathbf{X}$ from $t(3)$ distribution



Estimation Result



Prediction Result



Variable Selection Result

Summary and Follow-up Work

- **Summary**:
  - leverage over-parameterization to design unregularized gradient-based algorithm for SVM
  - provide theoretical guarantees for implicit regularization
  - Nesterov's method is employed to smooth the re-parameterized hinge loss function
- **Follow-up work**:
  - whether our results are still valid without the incoherence
  - explore the deeper depths of re-parameterization in classification
  - consider non-linear SVM

References

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *International Conference on Learning Representations*, 2015.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2016.

Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2(4):8, 2019.