Minimum L1-norm interpolators: Precise asymptotics and multiple descent

SeokHoon Park

January 17, 2022

Seoul National University

Table of Contents

Table of Contents

Over-parameterization

- In classic statistics, over-parameterization seems to hurt generalization

- However, an evolving an evolving line of works in machine learning observes empirical evidence that suggests, to the surprise of many statisticians, over-parameterization is not necessarily harmful.

Interpolation

- Example : Deep neural networks
- Many machine learning models such as DNN,Random forest are trained until the training error vanishes to zero meaning that they are able to perfectly interpolate the data while still generalizing well.

Minimum L-1 norm interpolator

- p : number of parameter , n : number of data
- generalization error : out-of-sample risk
- $Risk(\theta) = E((x_{new}\hat{\theta} - y_{new})^2)$

Minimum L-1 norm interpolator

- When it comes to the overparameterized regim where p>n , the system of equations $y_i =< x_i, \theta >$ is under determined.

- This implies the existence of multiple regression parameters θ that interpolate the training data perfectly.

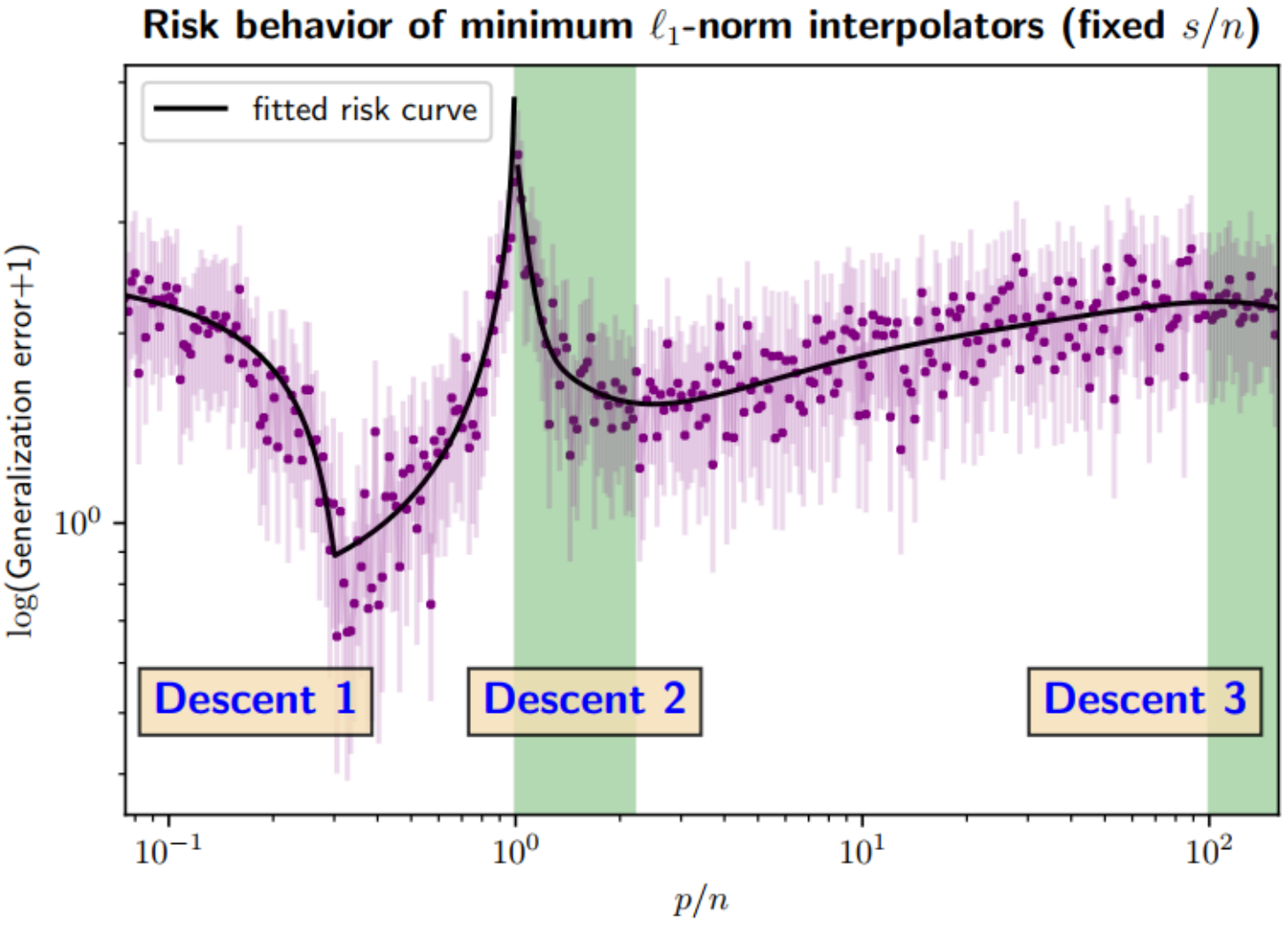
- Among all possible interpolates, the focal point of this paper is the minimum L-1 interpolater.

$\hat{\theta}^{Int} := argmin_{\theta \in R^p} ||\theta||_1$ subject to $y_i =< x_i, \theta >$, i=1,...,n

Example of overparameterization

- $y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$
- $y = X\theta$'s solution is $\theta_1 = 0, \theta_2 + \theta_3 = 1$

Multi Descent



Approach in the paper

- In this paper, they are trying to investigate the Risk curve of minimum L1-norm interpolator by using asymptotics of Risk.

Table of Contents

Setting

- In this paper, they have gathered n i.i.d noisy training data drawn from a linear model

$$y = X\theta^* + z \ , \ y = [y_i]_{i=1,..,n}, X = [x_1, ..., x_n]^t$$

where $\theta^* \in R^p, x_i \in R^p$: random design vector

- $x_i \sim^{iid} N(0, \frac{1}{n} I_p), i = 1, ..., n$

- $z \sim N(0, \sigma^2 I_n)$

Linear Sparsity

- $\theta_i^* \sim^{iid} \epsilon P_{M\sqrt{\delta}} + (1-\epsilon)P_0$

- M : some given quantity that determines the magnitude of a non-zero entry.

- scaler factor $\sqrt{\delta}$ is introduced soley for notational converience which ensures that the $signal - to - noise - ratio(SNR)$ obeys

$$SNR := \frac{E((x^T\theta^*)^2)}{\sigma^2} = \frac{\epsilon M^2}{\sigma^2}$$

Notation

- $y_{new} =< x_{new}, \theta^* > + z_{new}$,
  $x_{new} \sim N(0, \frac{1}{n} I_p), z_{new} \sim N(0, \sigma^2)$

- $Risk(\hat{\theta}) := E((x_{new}^t \hat{\theta} - y_{new})^2)$

- $Risk(\hat{\theta}; \delta) := \lim_{n/p=\delta \; n,p \to \infty} Risk(\hat{\theta})$

Table of Contents

Theorem 1

Suppose that $0 < \delta < 1$, $\delta = \frac{n}{p}$, setting (i.d.d gaussian desing and i.i.d gaussian nose) and linear sparsity Then the generalization of error of the minimum L-1 norm interpolator satisfies the following properites:

(1) There exists two constants $1 < \eta_1 < \eta_2 < \infty$ such that $Risk(\hat{\theta}^{Int}; \delta)$ decreases with p/n within the range p/n $\in (1, \eta_1) \cup (\eta_2, \infty)$

(2) $Risk(\hat{\theta}^{Int}; \delta)$ approaches the risk of the zero estimator(Risk(0)) as p/n tends to infinity.

Theorem 1

(3) For any fixed signal-to-noise ratio(SNR) , there exists a constant $\epsilon^* > 0$ such that if the sparsity ratio ϵ obeys $\epsilon < \epsilon^*$, then one can find a region within the range p/n $\in (\eta_1, \eta_2)$ such that $Risk(\hat{\theta}^{Int}; \delta)$ increases with p/n

(4) In addtion, for every given δ , there exists a threshold $\tilde{\epsilon}(\delta)$ such that $Risk(\hat{\theta}^{Int}; \delta)$ decreases with p/n at this particular point δ as long as the sparsity ratio ϵ satisfies $\epsilon \leq \tilde{\epsilon}(\delta)$

Theorem 1 insights

- (1) identifies two non-overlapping regions within the over-parameterizaed regime that exhibit risk descent.

- (2) the minimum L1 -norm interpolator is essentially no better than a trivial estimator (i.e., the zero estimator) when the over-parameterized ratio p/n is overly large.

- (3) indicates that the risk in between the above-mentioned two regions exhibits contrastingly different behavior depending on the sparsity ratio.

- (4) reveals that at any over-parameterized ratio, the generalization risk can be decreasing with p/n as long as the sparsity ratio is small enough.

Table of Contents

Theorem 2

Suppose that the setting is equal to before and the empirical distribtuion of θ^* converges weakly to a probability measure P_θ. Consider and given $0 < \delta < 1$. If $E(\theta^2) < \infty$ and $P(\theta \neq 0) > 0$, then the prediction risk of the minimum $L1$-norm interpolator obeys below.

$$\lim_{n/p=\delta\ n,p\to\infty} Risk(\hat{\theta}^{Int}) =^{a.s}\ r^{*2}$$

Here, r^*, α^* stands for the unique solution to the following system of equations

$$r^2 = \sigma^2 + \frac{1}{\delta} E((\eta(\theta + rZ; \alpha r) - \theta)^2)$$

$$\delta = P(|\theta + rZ| > \alpha r)$$

where $\theta \sim P_\theta$ and $Z \sim N(0, 1)$ and is independent of θ

Theorem 2 insights

- we can readily examine how r^* varies with δ
- By taking a close look at the solutions, they can analyze the shape of the risk curve and establish Theorem 1.

Table of Contents

Gaussian Experiments

- Data are generated from a linear model in setting
- The sparsity level is 0.05 and SNR=4
- The theoretical curve is computed by solving the equations in Thm2
- For each p/n, they generate a random instance, compute the minimum L-1 norm interpolator and its risk, and repeat this procedure for 30 times. They report the averasge risk and error bar over 30 independent runs.

Gaussian Experiments