Fundamental Limits of Ridge-Regularized Empirical Risk Minimization in High-Dimensions

Hossein Taheri

Joint with Ramtin Pedarsani and Christos Thrampoulidis

University of California, Santa Barbara.

March 2021

Motivation

Empirical Risk Minimization is a method widely used for inference from data:

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(y_i, \mathbf{a}_i^T\mathbf{x}\right) + \lambda\|\mathbf{x}\|_2^2.$$

Despite being used in every-day applications, a theory for explaining their **generalization properties** and specifying **optimal choices of loss function and regularization** is lacking; especially in modern high-dimensional models.

Problem Statement

Feature Vectors and Labels

Feature Vectors : $\mathbf{a}_i \in \mathbb{R}^n$, labels $y_i \in \mathbb{R}$, $1 \leq i \leq m$

Problem Statement

Feature Vectors and Labels

Feature Vectors : $\mathbf{a}_i \in \mathbb{R}^n$, labels $y_i \in \mathbb{R}$, $1 \leq i \leq m$

1. Binary model with a link function :
   $y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$

Problem Statement

Feature Vectors and Labels

Feature Vectors : $\mathbf{a}_i \in \mathbb{R}^n$, labels $y_i \in \mathbb{R}$, $1 \leq i \leq m$

1. Binary model with a link function :
   $y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$

   Random function $f : \mathbb{R} \to \{\pm 1\}$

   - **Noisy signed**: $f(t) = \begin{cases} \text{sign}(t) & \text{, w.p. } 1-\varepsilon, \\ -\text{sign}(t) & \text{, w.p. } \varepsilon. \end{cases}$ $\varepsilon \in [0, 1/2]$

Problem Statement

Feature Vectors and Labels

Feature Vectors : $\mathbf{a}_i \in \mathbb{R}^n$, labels $y_i \in \mathbb{R}$, $1 \leq i \leq m$

1. Binary model with a link function :
   $y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$

   Random function $f : \mathbb{R} \to \{\pm 1\}$

   - **Noisy signed**: $f(t) = \begin{cases} \text{sign}(t) & \text{, w.p. } 1-\varepsilon, \\ -\text{sign}(t) & \text{, w.p. } \varepsilon. \end{cases} \quad \varepsilon \in [0, 1/2]$
   - **Logistic**: $f(t) = \begin{cases} +1 & \text{, w.p. } \frac{1}{1+e^{-t}}, \\ -1 & \text{, w.p. } 1 - \frac{1}{1+e^{-t}}. \end{cases}$

Problem Statement

Feature Vectors and Labels

Feature Vectors : $\mathbf{a}_i \in \mathbb{R}^n$, labels $y_i \in \mathbb{R}$, $1 \leq i \leq m$

1. Binary model with a link function :
   $y_i = f(\mathbf{a}_i^T \mathbf{x}_0), \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$

   Random function $f : \mathbb{R} \to \{\pm 1\}$

   - **Noisy signed**: $f(t) = \begin{cases} \text{sign}(t) & \text{, w.p. } 1-\varepsilon, \\ -\text{sign}(t) & \text{, w.p. } \varepsilon. \end{cases} \quad \varepsilon \in [0, 1/2]$
   - **Logistic**: $f(t) = \begin{cases} +1 & \text{, w.p. } \frac{1}{1+e^{-t}}, \\ -1 & \text{, w.p. } 1 - \frac{1}{1+e^{-t}}. \end{cases}$

2. Linear Models: $y_i = \mathbf{a}_i^T \mathbf{x}_0 + z_i$, where $z_i \overset{\text{iid}}{\sim} P_Z$

Performance Measure

- Linear Models

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(y_i - \mathbf{a}_i^T\mathbf{x}\right) + \lambda\|\mathbf{x}\|_2^2.$$

Performance Measure: $\alpha_{\mathcal{L},\lambda} := \|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2$

Performance Measure

- Linear Models

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(y_i - \mathbf{a}_i^T\mathbf{x}\right) + \lambda\|\mathbf{x}\|_2^2.$$

  Performance Measure: $\alpha_{\mathcal{L},\lambda} := \|\widehat{\mathbf{x}}_{\mathcal{L},\lambda} - \mathbf{x}_0\|_2$

- Binary Linear Models

$$\widehat{\mathbf{x}}_{\mathcal{L},\lambda} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(1 - y_i\mathbf{a}_i^T\mathbf{x}\right) + \lambda\|\mathbf{x}\|_2^2.$$

  Performance Measure: $\mathcal{E}_{\mathcal{L},\lambda} := \mathbb{P}_{\mathbf{a},y}\left(y \neq \operatorname{sign}\left(\mathbf{a}^T\widehat{\mathbf{x}}_{\mathcal{L},\lambda}\right)\right)$

Theoretical Results: Fundamental Limits

For a r.v. H, we define the Fisher information of H as $\mathcal{I}(H) := \mathbb{E}[(p'(H)/p(H))^2]$

Main Theorem(Binary Models)

For $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ define the random variable $W_s := s\,G + S\,f(S)$ parameterized by $s \in \mathbb{R}$. Fix any $\delta := m/n > 0$ and define $\sigma_\star$ as

$$\min_{0 \le x < 1/\delta} \left[s > 0 : \frac{1 - s^2(1 - s^2\mathcal{I}(W_s))}{\delta s^2(s^2\mathcal{I}(W_s) + \mathcal{I}(W_s) - 1)} - 2x + x^2\delta\left(1 + \frac{1}{s^2}\right) = 1 \right].$$

Then, for any convex loss function and $\lambda \ge 0$, it holds that $\sigma_\star \le \sigma_{\mathcal{L},\lambda}$ and in particular

$$\mathcal{E}_\star := \mathbb{P}(\sigma_\star G + Sf(S) < 0) \le \mathcal{E}_{\mathcal{L},\lambda}$$

Theoretical Results: Optimal Loss and Regularization

For given $\delta > 0$ and binary link function f, let $\sigma_\star > 0, x_\star \in [0, 1/\delta)$ be the optimal solution in the minimization in the theorem. Denote $\lambda_\star = x_\star$ and define $W_\star := \sigma_\star G + Sf(S)$. Consider the loss function

$$\mathcal{L}_\star(x) := -\mathcal{M}_{\frac{\lambda_\star \delta - 1}{\delta(\eta - \mathcal{I}(W_\star))}(\eta Q + \log P_{W_\star})}(x, 1),$$

where $\eta := 1 - \mathcal{I}(W_\star) \cdot (\sigma_\star^2 - \sigma_\star^2 \lambda_\star \delta - \lambda_\star \delta) - \lambda_\star \delta$ and $Q(w) := w^2/2$.

Then $\mathcal{L}_\star$ and $\lambda_\star$ are the optimal pair of loss function and ridge-regularization parameter.

Final Remarks

- The optimal choices of loss function and ridge-regularization parameter, depend on $\delta := \frac{\textit{sample size}}{\textit{dimension}}$, SNR and data model.

- In particular, ridge-regularized least-squares can be optimal for low SNR logistic models or when δ is very small. As δ grows or as SNR increases, ridge-regularized least-squares becomes sub-optimal.