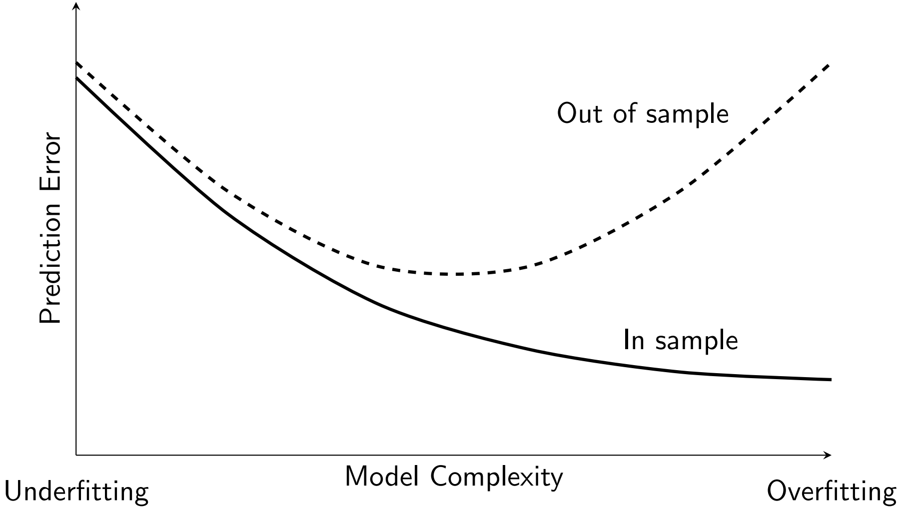
The risk function of regularized ERM estimators tuned using CV

Maximilian Kasy
with Karun Adusumilli and Ashia Wilson

February 2025

The basic tradeoff of supervised learning



Introduction

- Standard supervised learning:
  - How to control model complexity?
    A: penalization / regularization / shrinkage.
  - How to find the optimal amount of regularization?
    A: Minimize cross-validation estimate of out-of-sample loss.
  - How to evaluate estimators?
    A: Expected out of sample loss.
    - Standard theory: Worst-case regret.
    - This talk: Risk function - more fine-grained picture!
      Taking into account data-dependent tuning!
- Key idea in our paper: We can approximate risk by
  - the mean squared error (MSE)
  - of shrinkage estimators in the normal means model
  - tuned using Stein's unbiased risk estimate (SURE).

Review: Penalized ERM-estimation with tuning

- Penalized empirical risk minimization (ERM) estimator (in local coordinates):

$$\hat{\theta}_n^{\lambda} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} l(\theta/\sqrt{n}, Z_n^i) + \lambda \cdot \pi(\theta) \right].$$

- Cross-validated tuning parameter (definition uses leave-one-out estimator):

$$\lambda_n^* = \underset{\lambda}{\operatorname{argmin}} \sum_i l(\hat{\theta}_n^{\lambda,-i}/\sqrt{n}, Z_n^i).$$

- Risk function of the tuned estimator:

$$E\left[l(\hat{\theta}_n^{\lambda_n^*}/\sqrt{n}, Z) \right],$$

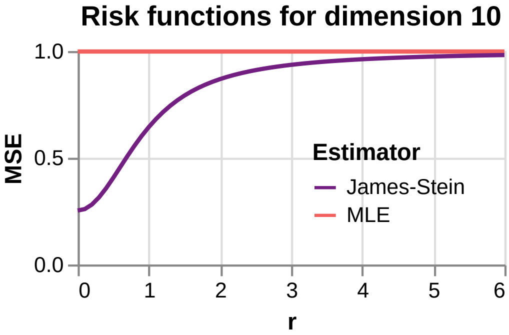
for Z an independent draw.

Review: James-Stein shrinkage

- Shrinkage estimator: For $\hat{\theta} \sim N(\theta, I_k)$,

$$\hat{\theta}^* = \left(1 - \frac{(k-2)}{\|X\|^2}\right) \cdot \hat{\theta}.$$

- Risk function (MSE): $1 - \frac{1}{k} E\left[\frac{(k-2)^2}{\|\hat{\theta}\|^2}\right]$.
  Denoting $r = \|\theta_0\|$:



Review: Stein's unbiased risk estimate

- Suppose $\hat{\theta} \sim N(\theta_0, \Sigma)$. Let

$$\hat{\theta}^\lambda = \hat{\theta} + g^\lambda(\hat{\theta}),$$
$$g^\lambda(\theta) = \operatorname*{argmin}_g \ \tfrac{1}{2}\|g\|^2 + \lambda \cdot \pi(\theta + g).$$

- Then

$$SURE(\lambda, \hat{\theta}, \Sigma) = \text{trace}(\Sigma) + \underbrace{\|g^\lambda(\hat{\theta})\|^2}_{\text{In-sample loss}} + 2 \cdot \underbrace{\text{trace}\left(\nabla g^\lambda(\hat{\theta}) \cdot \Sigma\right)}_{\text{Overfitting penalty}}.$$

is an unbiased estimator of the MSE of $\hat{\theta}^\lambda$.

- Special cases of SURE:
    - Mallows's C_p (for homoskedastic Gaussian linear regression).
    - Akaike information criterion (for correctly specified unregularized parametric models).
- JS shrinkage can be obtained from Ridge, tuned by minimizing SURE.
  (Up to a small degrees of freedom correction.)

$$\hat{\theta}^* = \hat{\theta}^{\lambda^*},$$
$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\ SURE(\lambda, \hat{\theta}, \Sigma).$$

Asymptotic approximations

- Consider large n, fixed k, θ_0 local to 0.
- Then:
  1. Penalized ERM $\approx$ shrinkage in the normal means model.
  2. Tuning using n-fold CV $\approx$ tuning using SURE.
  3. Out of sample predictive error $\approx$ MSE.
- **Our main result**:
  The risk function of tuned penalized ERM
  converges to
  the MSE of (generalized) JS shrinkage.
- Formally: Under suitable assumptions,

$$E\left[l(\hat{\theta}_n^{\lambda_n^*}/\sqrt{n}, Z)\right] \to \frac{1}{2}E[\|\hat{\theta}^* - \theta_0\|^2].$$

Key steps

1. Influence function approximations of
   - ERM ($\Rightarrow$ asymptotically normal),
   - penalized ERM,
   - leave-one-out estimators.
2. Taylor-approximation of CV
   $\Rightarrow$ CV $\approx$ SURE.
3. $\Rightarrow$ minimizer of CV $\approx$ minimizer of SURE
4. Approximation of average out-of-sample loss by squared error.

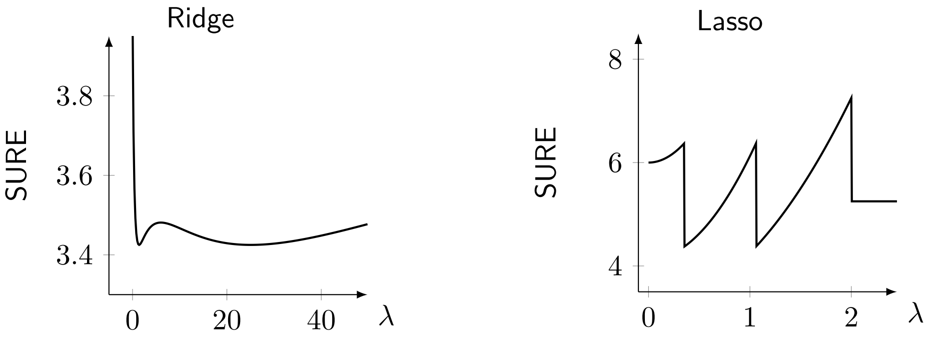
Challenges

1. *Convergence of penalized ERM estimators*:
   Standard empirical process results, plus arguments from convex analysis.

2. *Uniformity of convergence of CV to SURE*:
   Need to deal with points of non-differentiability.
   For Lasso: Restrict attention to a grid.

3. *Multimodality of SURE / CV in λ*:
   Non-standard arguments, separately for Ridge and Lasso.
   Show that the problem arises with sufficiently small probability.

Examples of multi-modality of $SURE$



- Examples for fixed (handpicked) values of $\hat{\theta}$ and Σ.
- L^2 penalty (Ridge) and L^1 penalty (Lasso).

Influence function approximations

- Loss:

$$\sum_{i=1}^{n} l(\theta/\sqrt{n}, Z_n^i) \approx const. + \tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2,$$

  where

$$\tilde{\theta}_n = \theta_0 + \tfrac{1}{\sqrt{n}} \sum_i X_n^i, \qquad X_n^i = -\nabla_\beta l(\theta_0/\sqrt{n}, Z_n^i).$$

- Asymptotic normality of ERM:

$$\hat{\theta}_n \to^d \hat{\theta} \sim N(\theta_0, \Sigma).$$

- Penalized ERM:

$$\hat{\theta}_n^\lambda \approx \tilde{\theta}_n^\lambda = \tilde{\theta}_n + g^\lambda(\tilde{\theta}_n).$$

Influence function approximations for leave-one-out

- Leave-one-out (LOO) loss:

$$\sum_{j\neq i} l(\theta/\sqrt{n}, Z_j^n) \approx const. + \tfrac{1}{2}\|\theta - \tilde{\theta}_n^{-i}\|^2,$$

where

$$\tilde{\theta}_n^{-i} = \tilde{\theta}_n - \tfrac{1}{\sqrt{n}} X_n^i.$$

- Penalized LOO estimator:

$$\begin{aligned}\hat{\theta}_n^{\lambda,-i} &\approx \tilde{\theta}_n^{-i} + g^\lambda(\tilde{\theta}_n^{-i}) \\ &\approx \tilde{\theta}_n^\lambda - \tfrac{1}{\sqrt{n}}(I + \nabla g^\lambda(\tilde{\theta}_n)) \cdot X_n^i.\end{aligned}$$

Local linear approximation of g^λ: have to be careful for Lasso, which has kinks.

Taylor expansion of CV

$$
\begin{aligned}
CV_n(\lambda) &= \sum_i l(\hat{\theta}_n^{\lambda,-i}/\sqrt{n}, Z_n^i) \\
&\approx const. + \frac{1}{n}\sum_i \|\hat{\theta}_n^{\lambda,-i} - \theta_0 - \sqrt{n}X_n^i\|^2 \\
&\approx const. + \frac{1}{n}\sum_i \|\underbrace{\tilde{\theta}_n + g^\lambda(\tilde{\theta}_n) - \frac{1}{\sqrt{n}}(I + \nabla g^\lambda(\tilde{\theta}_n)) \cdot X_n^i}_{\approx \hat{\theta}_n^{\lambda,-i}} - \theta_0 - \sqrt{n}X_n^i\|^2 \\
&\approx const. + \frac{1}{n}\sum_i \|g^\lambda(\tilde{\theta}_n)\|^2 + \frac{2}{n}\sum_i \langle \nabla g^\lambda(\tilde{\theta}_n) \cdot X_n^i, X_n^i \rangle \\
&\approx const. + \|g^\lambda(\hat{\theta}_n)\|^2 + 2 \cdot \mathrm{trace}(\nabla g^\lambda(\hat{\theta}_n) \cdot \hat{\Sigma}_n) \\
&= const. + SURE(\lambda, \hat{\theta}, \hat{\Sigma}_n),
\end{aligned}
$$

Uniform convergence of CV

- Suppose that

$$\lim_{\|\delta\|\to 0} \sup_{\lambda\in\Lambda} \frac{\left\|R^\lambda(\delta;\theta)\right\|}{\|\delta\|} = 0$$

for (Lebesgue) almost every θ, where

$$R^\lambda(\delta;\theta) = g^\lambda(\theta+\delta) - g^\lambda(\theta) - \nabla g^\lambda(\theta)\cdot\delta.$$

- Then (assuming regularity conditions) the n-fold crossvalidation criterion satisfies

$$\sup_{\lambda\in\Lambda}\left|CV_n(\lambda) - SURE(\lambda,\hat\theta_n,\Sigma)\right| \to^p 0.$$

Dealing with multi-modality: Ridge

- Penalty $\pi(\theta) = \frac{1}{2}\theta \cdot A^{-1} \cdot \theta$, where A is positive definite. $\Rightarrow$

$$g^\lambda(\theta) = C_\lambda \cdot \theta, \qquad \nabla g^\lambda(\theta) = C_\lambda, \qquad C_\lambda = (\tfrac{1}{\lambda}A + I)^{-1}.$$

- Thus

$$SURE(\lambda, \theta, \Sigma) = \mathrm{trace}(\Sigma) + \|C_\lambda \cdot \theta\|^2 + 2\,\mathrm{trace}\left(C_\lambda \cdot \Sigma\right).$$

- Change of coordinates, with slight abuse of notation:
  For $R = \|\hat{\theta}\|$ and $\nu = \hat{\theta}/R$,

$$SURE(\lambda, R, \nu) := SURE(\lambda, \hat{\theta}, \Sigma).$$

SURE for Ridge

1. For every θ,

$$\lim_{\theta' \to \theta} \sup_{\lambda} |SURE(\lambda, \theta', \Sigma) - SURE(\lambda, \theta, \Sigma)| = 0$$

2. $SURE(\lambda, R, \nu)$ is strictly **supermodular** in λ and R. This implies:
   2.1 $\lambda(R, \nu) = \text{argmin}_{\lambda \in \mathbb{R}^+} SURE(\lambda, R, \nu)$ is **monotonically decreasing** in R, given ν.
   2.2 $\lambda(R, \nu)$ has at most **countably many discontinuities**, as a function of R, given ν.
3. Fix ν and R such that $\lambda(\cdot)$ is continuous in R at (R, ν), and let $\bar{\lambda} = \lambda(R, \nu)$.
   Then supermodularity implies that **the minimum of $SURE$ is well separated**:
   For any $\epsilon > 0$,

$$\inf_{\lambda \in \mathbb{R}^+ \setminus [\bar{\lambda} - \epsilon, \bar{\lambda} + \epsilon]} SURE(\lambda, R, \nu) - SURE(\bar{\lambda}, R, \nu) > 0.$$

Almost everywhere continuity for Ridge

- Define

$$w = (\theta, \Delta), \qquad \qquad \|w\| = \|\theta\| + \sup_{\lambda} |\Delta(\lambda)|$$

- For **almost every** θ, the mapping from w to

$$g^{\tilde{\lambda}(\theta,\Delta)}(\theta) = \left(\frac{1}{\tilde{\lambda}(\theta,\Delta)} A + I\right)^{-1} \cdot \theta, \text{ where}$$

$$\tilde{\lambda}(\theta, \Delta) = \min \left(\operatorname*{argmin}_{\lambda} \; [SURE(\lambda, \theta, \Sigma) + \Delta(\lambda)] \right),$$

is **continuous at** $w = (\theta, 0)$ with respect to the norm $\|w\|$.

- Continuous mapping theorem, uniform integrability
$\Rightarrow$ convergence of risk.

Dealing with multi-modality: Lasso

- Penalty $\pi(\theta) = \|A^{-1} \cdot \theta\|_1$, where A is an invertible matrix.
- Denote $h^\lambda(\theta) = A^{-1}(\theta + g^\lambda(\theta)) \Rightarrow$

$$h^\lambda(\theta) = \underset{h}{\text{argmin}} \ \frac{1}{2}\|A \cdot h - \theta\|^2 + \lambda \cdot \|h\|_1.$$

- Solution:

$$h_J^\lambda(\theta) = (A_J' A_J)^{-1} \cdot [A_J'\theta - \lambda \eta_J],$$

where

  - $\eta_j = sign(h_j^\lambda)$,
  - $J = \{j : \eta_j \neq 0\}$.

SURE for Lasso

1. As a function of λ, the graph of $SURE(\lambda, R, \nu)$ consists of **at most** 3^k **segments** on which η and $J = \{j : \eta_j \neq 0\}$ are constant, and

$$SURE(\lambda, R, \nu) = const. + \lambda^2 \cdot \eta_J'(A_J' A_J)^{-1} \eta_J.$$

2. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ $(m \leq 3^k)$ be the local minimizers of $SURE(\lambda, 1, \nu)$. Then $R \cdot \lambda_1, R \cdot \lambda_2, \ldots, R \cdot \lambda_m$ are the **local minimizers** of $SURE(\lambda, R, \nu)$.

3. Let $\lambda(R, \nu) = \text{argmin}_{\lambda \in \mathbb{R}^+} SURE(\lambda, R, \nu)$. Then
   - $\lambda(R, \nu) = R \cdot \lambda_{j(R)}$,
   - where $j(R) \in \{1, 2, \ldots, m\}$ is a **monotonically decreasing**.

4. Fix ν and R such that $\lambda(\cdot)$ is continuous in R at (R, ν), and let $\bar{\lambda} = \lambda(R, \nu)$ be such that $\eta \neq 0$. Then **the minimum of $SURE$ is well separated**: For any $\epsilon > 0$,

$$\inf_{\lambda \in \mathbb{R}^+ \setminus [\bar{\lambda} - \epsilon, \bar{\lambda} + \epsilon]} SURE(\lambda, R, \nu) - SURE(\bar{\lambda}, R, \nu) > 0.$$

Conclusion

- Majority of supervised learning methods:
  - Empirical risk minimizers
  - with regularization
  - tuned using cross-validation.
- We show:
  - Such methods behave approximately like (generalized) James-Stein shrinkage:
  - Uniform dominance relative to un-regularized estimators.
  - Largest gains for: Large k, small $\|\theta\|$.

Open issues and limitations

- Our asymptotic result for Lasso holds for a fixed grid Λ.
  - Extension to sequence of grids?
  - More refined argument to cover the case $\Lambda = \mathbb{R}$?
- Our approximations hold for fixed k, large n.
  - $\Rightarrow$ We can leverage asymptotic normality.
  - But what about the over-parametrized case $k > n$?
    Important in deep learning!
- Risk $\approx$ average loss for point prediction.
  - Conformal inference:
    Turns point prediction into predictive intervals with guaranteed coverage.
  - Can we map our risk results into results about average size of predictive sets?

Thank you!