From Cross-Validation to SURE: Asymptotic Risk of Tuned Regularized Estimators

Karun Adusumilli[*] Maximilian Kasy[†] Ashia Wilson[‡]

March 19, 2026

Abstract

We derive the asymptotic risk function of regularized empirical risk minimization (ERM) estimators tuned by n-fold cross-validation (CV). The out-of-sample prediction loss of such estimators converges in distribution to the squared-error loss (risk function) of shrinkage estimators in the normal means model, tuned by Stein's unbiased risk estimate (SURE). This risk function provides a more fine-grained picture of predictive performance than uniform bounds on worst-case regret, which are common in learning theory: it quantifies how risk varies with the true parameter.

As key intermediate steps, we show that (i) n-fold CV converges uniformly to SURE, and (ii) while SURE typically has multiple local minima, its global minimum is generically well separated. Well-separation ensures that uniform convergence of CV to SURE translates into convergence of the tuning parameter chosen by CV to that chosen by SURE.

[*] Department of Economics, University of Pennsylvania. akarun@sas.upenn.edu.

[†] Department of Economics, University of Oxford. maximilian.kasy@economics.ox.ac.uk. Maximilian Kasy was supported by the Alfred P. Sloan Foundation, under the grant "Social foundations for statistics and machine learning."

[‡] Department of Electrical Engineering and Computer Science, MIT.

1 Introduction

Background The goal of supervised learning is to produce good predictions for new observations.[1] An important class of estimators for supervised learning can be described as regularized empirical risk minimization (ERM) estimators that are tuned using cross-validation (CV).

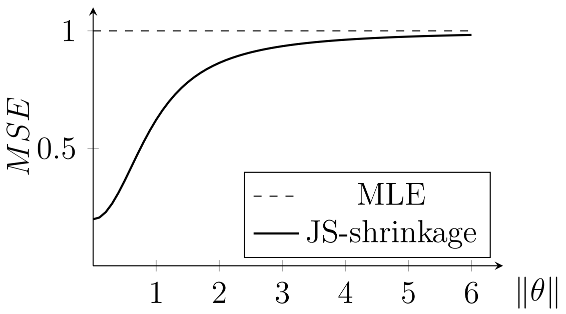
ERM estimators minimize in-sample average prediction loss (empirical risk), among a given class of predictors. Examples include ordinary least squares and maximum likelihood. ERM estimators are prone to overfitting if the class of predictors is large. Such estimators achieve low in-sample loss, but perform poorly for new observations. To counter overfitting, regularization is used. Regularization adds a penalty term to the ERM objective; common penalties include the L^2 norm of parameters (in Ridge regression) and the L^1 norm (in Lasso regression). Adding a penalty avoids overfitting, by reducing the estimator variance, at the cost of introducing some bias, which might result in underfitting.

To achieve good performance, avoiding both overfitting and underfitting, the amount of penalization needs to be carefully tuned. This can be done by choosing the weight on the penalty term as the minimizer of a cross-validation estimate of predictive loss. We focus on n-fold CV, where predictions are evaluated for one hold-out observation at a time, and predictive loss is estimated by averaging evaluations over each of the n observations.

Risk functions The present paper characterizes the behavior of estimators of this form by deriving an asymptotic approximation to their risk function. A large literature in learning theory characterizes such estimators by proving bounds on their worst-case regret—the supremum over data-generating processes (DGPs) of the difference between an estimator's risk and the expected loss of the best predictor in the given class. Such bounds provide strong robustness guarantees, but they might not be informative about the behavior of predictive algorithms for realistic DGPs. By focusing on the risk function,

[1] We would like to dedicate this paper to Gary Chamberlain, whose conversations provided the original inspiration for this project.

Figure 1: Risk function for JS-shrinkage, dimension 10



we obtain a more fine-grained characterization: the risk function tells us how expected predictive performance depends on the DGP, while worst-case regret bounds only characterize performance for the least favorable DGP.

Our asymptotic characterization relates the risk function of regularized ERM estimators tuned using CV to the risk function of the James-Stein (JS) shrinkage estimator James and Stein (1961), and generalizations thereof. JS shrinkage famously dominates maximum likelihood estimation (MLE) in the normal means setting: The risk function (mean squared error) of JS shrinkage, as characterized in Stein (1981), is lower than the risk of MLE, for any possible DGP. Figure 1 illustrates. Our result suggests that this same risk improvement carries over, asymptotically, to CV-tuned penalized estimation in general parametric models. Like CV, SURE provides an unbiased estimate of mean squared error; the JS estimator approximately minimizes SURE over the shrinkage intensity.

Main result Our main theorem states that the distribution of the out-of-sample prediction loss of CV-tuned regularized ERM estimators converges to the distribution of the squared-error loss of the corresponding SURE-tuned shrinkage estimator in the Gaussian limit experiment. In particular, the risk function (expected out-of-sample prediction loss, as a function of the true parameter) converges to the mean squared error of SURE-tuned penalized estimation in the normal means model. Two of our intermediate results are of independent interest: the uniform approximation of n-fold CV by SURE, and the generic well-separation of the global minimum of SURE.

Key steps Let us briefly outline the three main parts of our proof. First, we show that in large samples ERM estimators are approximately normally distributed, and that out-of-sample predictive loss is approximately equal to squared error loss. These are standard results, and we follow van der Vaart (2000) in proving this step. We furthermore need to show that this approximation carries over to penalized estimators, for fixed tuning parameters. Our asymptotic approximations are based on local-to-0 asymptotics: As sample size n increases, the parameter vector drifts to 0 (which is the minimizer of the penalty term) at a rate of $1/\sqrt{n}$. This rate is such that both bias and variance remain non-negligible in the limit. The drifting-parameter framework is natural for studying regularized estimators, because it is the regime in which the penalty has a first-order effect: if the true parameter were fixed away from 0, the penalty would become asymptotically irrelevant, while if it were exactly 0, no bias–variance tradeoff would arise.

Second, we need to show that n-fold CV, as a random function that maps tuning parameter values to estimates of predictive loss, converges uniformly to SURE. In this step of the proof, we build on the prior work of Wilson et al. (2020). This step involves an influence function approximation for leave-one-out estimators, and a second-order approximation to predictive risk. Uniformity of convergence in the tuning parameter is key to this step, and we need to carefully specify regularity conditions such that uniformity is guaranteed.

Third, we need to show that convergence of the CV criterion function for tuning is sufficient for convergence of its minimizer, the tuned parameter. This is non-trivial, because both CV and SURE typically have multiple local minima, and might have multiple global minima. We need to show that generically the global minimum is well-separated. We do this using separate arguments for L^1 and L^2 penalties, characterizing the shape and behavior of SURE in either case.

We should emphasize some limitations of our analysis: We do not consider penalties beyond L^1 and L^2, or k-fold CV with $k < n$; extending our results in these directions is left for future work. Our asymptotic approximations are furthermore not appropriate in the over-parametrized regime, when the number of parameters is of similar or larger magnitude than the number of observations.

Literature The analysis in this paper connects several lines of work in statistics, econometrics, and machine learning. Leave-one-out cross-validation was formalized by Stone (1974); its asymptotic optimality for model selection was established by Li (1987). A closely related family of risk estimators includes Mallows' C_p (Mallows, 1973), generalized cross-validation (Golub et al., 1979), and Stein's unbiased risk estimate (Stein, 1981). Efron (2004) provides a unifying perspective, showing that these criteria all take the form of in-sample error plus a covariance penalty. Arlot and Celisse (2010) provide a comprehensive survey of cross-validation procedures and their theoretical properties. A noteworthy contrast with our results is provided by Shao (1993), who shows that leave-one-out CV is asymptotically *inconsistent* for model selection—it selects overfitted models with positive probability—while k-fold CV with $k = o(n)$ is consistent. Our result is complementary in focus: rather than asking which of a finite list of models has the best predictive ability, we characterize the limiting distribution and risk function of the estimator selected by leave-one-out CV, showing it converges to that of the SURE-tuned normal-means estimator.

Shrinkage estimation originates with James and Stein (1961) and was analyzed in depth by Stein (1981). SURE-based tuning of shrinkage was extended to wavelet thresholding by Donoho and Johnstone (1995). The two penalty families we study—Ridge (Hoerl and Kennard, 1970) and Lasso (Tibshirani, 1996)—are the most widely used forms of regularization in practice, and both are commonly tuned by cross-validation. The close relationship between leave-one-out CV and covariance-penalty criteria such as SURE is further illuminated by Zou et al. (2007), who show—using SURE as the analytical tool—that the effective degrees of freedom of the Lasso equals the number of nonzero fitted coefficients.

Using local asymptotic frameworks to characterize decision problems was pioneered by Le Cam (1972); work using this approach is reviewed in Hirano and Porter (2020). The use of shrinkage asymptotics for parametric models in econometrics is discussed in Hansen (2016). We build directly on Wilson et al. (2020), who provide non-asymptotic deterministic guarantees for approximate cross-validation, showing that leave-one-out estimators can be well approximated by a single Newton step from the full-sample estimator.

Roadmap The remainder of this paper is structured as follows: In Section 2, we introduce our model and assumptions, and define all relevant notation. In Section 3, we first provide a heuristic outline of our proof, and then state a series of intermediate lemmas. Proofs of all lemmas are collected in the appendix. In Appendix A, we prove the lemmas corresponding to the first part of our argument, involving influence function approximations and asymptotic normality. In Appendix B, we prove the second part of the argument, namely the (uniform) approximation of n-fold CV by SURE. In Appendices C and D, we show that the global minimizer of SURE is generically well-separated for both L^2 and L^1 penalties, thereby proving the third part. In Appendix E, we conclude our derivation, proving the convergence of risk for tuned estimators.

2 Setup

In the following, we first set up our estimators, and their asymptotic counterparts, in a series of definitions. We then state the assumptions that will be invoked to justify our asymptotic approximations.

Throughout this paper, we consider the problem of estimating a parameter vector β_0 which is defined as the minimizer of expected loss $E[l(\beta, Z)]$. For prediction problems, typically $Z = (W, Y)$, for predictive features W and outcomes Y. Examples include linear OLS regression, where $l(\beta, Z) = (Y - W \cdot \beta)^2$, as well as the use of neural nets[2] or other parametric models for classification,

[2] With a small number of parameters relative to the sample size.

where $l(\beta, Z) = -\log(f(Y|W, \beta))$, and $f(Y|W, \beta)$ is the probability assigned to outcome Y by the model.

We use local-to-0 coordinates, $\theta = \sqrt{n} \cdot \beta$, for sample size n, and correspondingly $\theta_0 = \sqrt{n} \cdot \beta_0$. For each n, the random vectors Z_n^i are i.i.d. draws from the distribution μ_n, across i. Any finite-sample object will be denoted by a subscript n; objects without subscripts correspond to the limiting experiment.

2.1 Definitions

Definition 1 introduces notation for loss functions and for limiting loss functions. We evaluate estimates of θ in terms of their expected loss $\bar{L}_n(\theta, \theta_0)$. For supervised learning, $\bar{L}_n(\theta, \theta_0)$ is the *out-of-sample* expected prediction error.

Definition 1 (Loss, empirical loss, and expected loss)**.**
Given the loss function $l(\beta, z)$, define the following.

$$\begin{aligned}
l_n(\theta, z) &= l(\theta/\sqrt{n}, z) && \textit{Loss function in local parameter} \\
L_n(\theta) &= \sum_{i=1}^{n} l_n(\theta, Z_n^i) && \textit{Empirical loss} \\
\bar{L}_n(\theta, \theta_0) &= E\left[L_n(\theta) - L_n(\theta_0)\right] && \textit{Expected loss} \\
\bar{L}(\theta, \theta_0) &= \lim_{n \to \infty} \bar{L}_n(\theta, \theta_0). && \textit{Limiting expected loss}
\end{aligned}$$

We assume that the sequence of distributions μ_n is such that the limit in the last definition is well-defined.

Scaling Note that we could have equivalently defined

$$\bar{L}_n(\theta, \theta_0) = n \cdot E\left[l_n(\theta, Z_n^{n+1}) - l_n(\theta_0, Z_n^{n+1})\right],$$

that is, $\bar{L}_n(\theta, \theta_0)$ is the expected regret for out-of-sample predictions, multiplied by the sample size n. The multiplication by n is required because in Definition 1, we do *not* scale empirical loss $L_n(\theta)$ by a factor $\frac{1}{n}$, and therefore $L_n(\theta)$ diverges. In the definition of the local parameter vector θ we have however re-scaled the parameter vector β by a factor of $\frac{1}{\sqrt{n}}$. This implies that the Hessian (second derivative) of $L_n(\theta)$ with respect to θ, if it exists, is given by

$$\nabla^2_\theta L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla^2_\beta l(\theta/\sqrt{n}, Z_n^i).$$

We can thus expect that this Hessian converges, by a law of large numbers, under some additional regularity conditions.

Estimators of θ and their asymptotic counterparts We next specify a series of estimators, in Definition 2. We start with the standard ERM estimator $\hat{\theta}_n = \text{argmin}_\theta L_n(\theta)$, and its limiting counterpart $\hat{\theta}$, where the latter is normally distributed with mean θ_0 and variance Σ. We then consider the regularized versions of these estimators, that is, the penalized ERM estimator with penalty $\lambda \cdot \pi(\theta)$, and its limiting counterpart. Our choice of local parametrization ensures that a constant value of λ along the sequence indexed by n leads to a non-degenerate limit for the penalized ERM estimator, where neither variance nor bias of this estimator vanish.

In Definition 2, we furthermore introduce leave-one-out loss, and the corresponding leave-one-out ERM and penalized ERM estimator. These will serve as the building blocks of n-fold cross-validation.

Definition 2 (Estimators of θ_0)**.** *We will consider the following estimators of θ_0, for finite n, and in the limit experiment:*

$$\begin{aligned}
\hat{\theta}_n &= \underset{\theta}{\text{argmin}}\ L_n(\theta) && \textit{ERM estimator}\\
\hat{\theta} &\sim N(\theta_0, \Sigma) && \textit{Limiting ERM estimator}\\
\hat{\theta}_n^\lambda &= \underset{\theta}{\text{argmin}}\ [L_n(\theta) + \lambda \cdot \pi(\theta)] && \textit{Penalized ERM estimator}\\
\hat{\theta}^\lambda &= \underset{\theta}{\text{argmin}}\ \left[\tfrac{1}{2}\|\theta - \hat{\theta}\|^2 + \lambda \cdot \pi(\theta)\right], && \textit{Limiting penalized ERM estimator}
\end{aligned}$$

where $\pi(\cdot)$ is convex and attains its minimum at 0.

We furthermore consider the following leave-one-out (LOO) loss and estima-

tors of θ_0:

$$
\begin{aligned}
L_n^{-i}(\theta) &= \sum_{j \neq i} l_n(\theta, Z_n^j) && \textit{LOO empirical loss}\\
\hat{\theta}_n^{-i} &= \underset{\theta}{\operatorname{argmin}}\ L_n^{-i}(\theta) && \textit{LOO ERM estimator}\\
\hat{\theta}_n^{\lambda,-i} &= \underset{\theta}{\operatorname{argmin}}\ \left[L_n^{-i}(\theta) + \lambda \cdot \pi(\theta)\right]. && \textit{LOO penalized ERM estimator}
\end{aligned}
$$

We can rewrite the limiting penalized ERM estimator as

$$\hat{\theta}^\lambda = \hat{\theta} + g^\lambda(\hat{\theta}),$$

where

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

Denote $\nabla g^\lambda(\theta)$ the derivative of $g^\lambda(\theta)$ where it exists, and define $\nabla g^\lambda(\theta) = 0$ at points where $g^\lambda(\theta)$ is not differentiable.[3]

Estimators of risk The preceding definition introduced penalized estimators for given, fixed values of the tuning parameter λ. We are interested in estimators which choose this tuning parameter in a data-dependent way, where λ minimizes an estimator of risk. For finite sample size n, we consider the n-fold crossvalidation (CV) criterion as an estimator of the risk of penalized ERM estimation. For the limit experiment, we consider Stein's Unbiased Risk Estimator (SURE) as an estimator of the risk of the penalized limiting ERM estimator.

Definition 3 (Estimators of risk)**.** *We consider the following estimators of*

[3] This convention is adopted for convenience; we will use it to handle Lasso (L^1) penalties in the proof of Lemma 4 below.

risk for penalized ERM estimators with fixed tuning parameter λ.

$$CV_n(\lambda) = \sum_i l_n(\hat{\theta}_n^{\lambda,-i}, Z_n^i), \qquad \textit{n-fold CV}$$

$$SURE(\lambda, \hat{\theta}, \Sigma) = \operatorname{trace}(\Sigma) + \|g^\lambda(\hat{\theta})\|^2 + 2\operatorname{trace}\left(\nabla g^\lambda(\hat{\theta}) \cdot \Sigma\right). \qquad \textit{SURE}$$

Tuned estimators We can now formally define our tuned estimators. $\hat{\theta}_n^*$ is the penalized ERM estimator using a tuning parameter λ_n^* which minimizes the n-fold CV estimator of risk. $\hat{\theta}^*$ is the penalized limiting ERM estimator using a tuning parameter λ^* which minimizes the SURE estimator of risk. The tuning parameter is chosen from a set $\Lambda \subset \mathbb{R}$. Later, we will consider $\Lambda = \mathbb{R}$ (for Ridge penalties), and Λ arbitrary but finite (for Lasso penalties).

Definition 4 (Tuned estimators of θ)**.**

$$\begin{aligned} \hat{\theta}_n^* &= \hat{\theta}_n^{\lambda_n^*}, && \textit{Penalized ERM tuned using CV} \\ \lambda_n^* &= \operatorname*{argmin}_{\lambda \in \Lambda} \; CV_n(\lambda) \\ \hat{\theta}^* &= \hat{\theta}^{\lambda^*}, && \textit{Limiting penalized ERM tuned using SURE} \\ \lambda^* &= \operatorname*{argmin}_{\lambda \in \Lambda} \; SURE(\lambda, \hat{\theta}, \Sigma). \end{aligned}$$

We evaluate estimators based on their expected loss for new data-points. For supervised learning, this corresponds to the out-of-sample expected prediction loss. In this paper, we do not consider global criteria such as worst-case risk (maximizing over θ_0) or Bayes risk (averaging over a prior distribution for θ_0). Instead, we are interested in the dependence of expected loss on the parameter θ_0, which is captured by the risk function. Our notation makes this dependence on θ_0 explicit. The risk function gives a more fine-grained picture of estimator performance, relative to global criteria such as worst-case risk, Bayes risk, or worst-case regret.

The parameter θ_0 enters the following expressions both directly, as an argument of $\bar{L}_n$ and $\bar{L}$, and implicitly, via the distribution of Z that the expectations are averaging over.

Definition 5 (Risk functions)**.**

$$\begin{aligned} R_n(\theta_0) &= E\left[\bar{L}_n(\hat{\theta}^*_n, \theta_0)\right] && \textit{Finite sample risk, tuned using CV} \\ R(\theta_0) &= E\left[\bar{L}(\hat{\theta}^*, \theta_0)\right] && \textit{Limiting risk, tuned using SURE.} \end{aligned}$$

2.2 Assumptions

Having defined our estimators and evaluation criteria, we next specify the assumptions invoked in our asymptotic analysis. Assumption 1 sets up a sequence of experiments, indexed by n. We assume that, for each n, the minimizer of expected loss $E[l(\beta, Z^i_n)]$ is given by $\theta_0/\sqrt{n}$. Put differently, the minimizer β of expected loss drifts towards 0. We furthermore assume that the variance Σ of the score $\nabla_\beta l(\theta_0/\sqrt{n}, Z^i_n)$ remains constant along our sequence.

Assumption 1 (Sequence of experiments)**.** *For each n, the random vectors Z^i_n are i.i.d. draws from the distribution μ_n, across i. The distributions μ_n are such that θ_0 and Σ do not depend on n, where*

$$\begin{aligned} \theta_0 &= \underset{\theta}{\operatorname{argmin}}\ E[l(\theta/\sqrt{n}, Z^i_n)], \\ \Sigma &= \operatorname{Var}\left(\nabla_\beta l(\theta_0/\sqrt{n}, Z^i_n)\right). \end{aligned}$$

The limiting Hessian $H = \nabla^2_\theta \bar{L}(\theta, \theta_0)$ is typically non-degenerate because of our scaling of $\bar{L}_n(\theta, \theta_0)$ and of θ. The following Assumption 2 is made for notational convenience. This assumption states that the Hessian H is equal to the identity I. This is a coordinate normalization that can be imposed without loss of generality.[4]

Assumption 2 (Normalized loss function)**.**

$$\nabla^2_\theta \bar{L}(\theta, \theta_0)|_{\theta=\theta_0} = I.$$

[4] By suitable choice of coordinates we can normalize *either* H, *or* the asymptotic variance Σ of Assumption 1, *or* the Hessian of the penalty function π (when the latter exists), but not more than one of these three matrices, in general. After normalizing the Hessian, we can however diagonalize one more matrix, without loss of generality.

The last part of our proof requires showing that the global optimum of $SURE$ with respect to λ is generically unique and well-separated. We will prove this fact for both Ridge and Lasso penalties, using separate arguments for either case.

Assumption 3 (Penalty function and grid for tuning)**.**
The penalty $\pi(\theta)$ and the set Λ take one of the following two forms:

1. ***Ridge**: $\pi(\theta) = \frac{1}{2}\theta \cdot A^{-1} \cdot \theta$, where A is positive definite, and $\Lambda = \mathbb{R}^+$.*

2. ***Lasso**: $\pi(\theta) = \|A^{-1} \cdot \theta\|_1$, where A is an invertible matrix, and $\Lambda \subset \mathbb{R}^+$ is finite.*

The remaining assumptions state regularity conditions. The first item in Assumption 4 is a condition on the loss function which allows us to invoke results from empirical process theory. Similar assumptions are invoked in van der Vaart (2000) when deriving the properties of M-estimators. The second item in Assumption 4 is a weak high-level condition ruling out divergence of ERM estimators, which ensures the applicability of empirical process results.

Assumption 4 (Conditions for convergence of the ERM estimator)**.**

1. ***Lipschitz loss***
   The loss function $l(\beta, z)$ satisfies

$$|l(\beta_1, z) - l(\beta_2, z)| \leq m(z) \cdot \|\beta_1 - \beta_2\|,$$

   for all β_1, β_2 in a neighborhood of 0, where $\sup_n \mathrm{Var}(m(Z_n^i)) \leq \infty$. Furthermore, $l(\beta, Z_n^i)$ is differentiable w.r.t. β, for all β in a neighborhood of $\beta = 0$, with probability 1.

2. ***Stochastically bounded ERM estimator***
   The sequence $\hat{\theta}_n = \mathrm{argmin}_{\,\theta}\, L_n(\theta)$ is bounded in probability.

To prove the (uniform) convergence of CV_n to $SURE$, we finally impose the following additional regularity conditions.

Assumption 5 (Conditions for convergence of CV)**.**

1. ***Conditions on loss***
   There exist $\mu > 0, \nu < \infty$ independent of n such that $L_n(\theta)$ is μ-strongly convex and has ν-smooth Hessians with probability approaching 1 under μ_n.[5]

2. ***Conditions on scores***
   The function $\sqrt{n}\nabla_\theta l_n(\theta, Z_n^i)$ is Lipschitz continuous almost everywhere, i.e., there exists $B_n(Z_n^i)$ such that

$$\left\|\sqrt{n}\nabla_\theta l_n(\theta, Z_n^i) - \sqrt{n}\nabla_\theta l_n(\theta', Z_n^i)\right\| \leq B_n(Z_n^i)\left\|\theta - \theta'\right\| \;\; \forall \theta, \theta' \in \Theta$$

   and $E_{\mu_n}\left[\left\|B_n(Z_n^i)\right\|^2\right] < \infty$. Additionally, there exists $M < \infty$ independent of n such that

$$\mathbb{E}_{\mu_n}\left[\left\|\sqrt{n}\nabla_\theta l_n\left(\theta_0, Z_n^i\right)\right\|^4\right] \leq M.$$

3. ***Conditions on Hessians***
   The Hessian $\nabla_\theta^2 l_n(\theta, Z_n^i)$ is such that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|n\nabla_\theta^2 l_n\left(\theta_0, Z_n^i\right)\right\|^2 = O_{\mu_n}(1).$$

   Furthermore, there exists $C_n(Z_n^i)$ such that

$$\left\|n\nabla_\theta^2 l_n(\theta, Z_n^i) - n\nabla_\theta^2 l_n(\theta', Z_n^i)\right\| \leq C_n(Z_n^i)\left\|\theta - \theta'\right\| \;\; \forall \theta, \theta' \in \Theta$$

   and $\sup_n E_{\mu_n}\left[C_n(Z_n^i)^2\right] < \infty$.

4. ***Conditions on Fourth Derivatives***

[5] $L_n(\theta)$ is μ-strongly convex if $\nabla^2 L_n(\theta) - \mu I$ is positive semi-definite for all θ. A function is L-smooth if its gradients are Lipschitz continuous with Lipschitz constant L.

The fourth derivative tensor, $D_\theta^4 l_n(\theta, Z_n^i)$, of $l_n(\cdot, Z_n^i)$ is such that

$$\sup_n \mathbb{E}_{\mu_n}\left[\sup_{\theta\in\Theta}\left\|n^2 D_\theta^4 l_n\left(\theta, Z_n^i\right)\right\|^4\right] \leq M,$$

for some $M < \infty$.

3 Main result and intermediate lemmas

Our main goal in this paper is to prove the following result:

Theorem 1.

$$\bar{L}_n(\hat{\theta}_n^*, \theta_0) \to_d \tfrac{1}{2}\|\hat{\theta}^* - \theta_0\|^2.$$

In words, the distribution of loss of the penalized ERM estimator tuned using n-fold cross-validation converges to the distribution of squared error of the corresponding shrinkage estimator in the normal means model, tuned by Stein's Unbiased Risk Estimate. A special case of these limiting estimators are James-Stein shrinkage estimators, for which closed form characterizations of the risk function are known (Stein, 1981). An immediate corollary of Theorem 1 is the convergence of risk functions, subject to possible truncation of tail events.[6]

Corollary 1. *Let $M > 0$. Then*

$$E\left[\min\left(\bar{L}_n(\hat{\theta}_n^*, \theta_0), M\right)\right] \to E\left[\min\left(\|\hat{\theta}^* - \theta_0\|^2, M\right)\right].$$

3.1 Outline of proof

We will build up our argument that proves Theorem 1 in a series of lemmas. Before doing so, however, we first provide an intuitive sketch of our argument, while neglecting remainder terms. Subsequently, we will prove that these remainder terms are indeed asymptotically negligible.

[6] Truncation is necessary because, even for estimators such as linear OLS regression, risk is typically undefined, since some eigenvalues of the design matrix might be close to 0, so that the moments of $\hat{\theta}_n$ might not exist.

Influence function approximation We start by noting that empirical risk is asymptotically equivalent to quadratic error loss relative to the sample mean $\tilde{\theta}_n$,

$$L_n(\theta) \approx const. + \tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2, \qquad \tilde{\theta}_n = \theta_0 + \tfrac{1}{\sqrt{n}} \sum_i X_n^i,$$

where

$$X_n^i = -\nabla_\beta l(\theta_0/\sqrt{n}, Z_n^i)$$

is the influence function. Recall that we have normalized the Hessian in Assumption 2, which simplifies the expression for $\tilde{\theta}_n$. This approximation of empirical risk immediately implies an asymptotic linear approximation of the empirical risk minimization (ERM) estimator, $\hat{\theta}_n \approx \tilde{\theta}_n$. These are standard approximations that deliver asymptotic normality of ERM estimators; see for instance Theorem 5.21 in van der Vaart (2000).

We then get the corresponding approximation for the penalized ERM estimator, for fixed λ,

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

where we recall the definition $g^\lambda(\theta) = \text{argmin}_{\,g}\, \frac{1}{2}\|g\|^2 + \lambda \cdot \pi(\theta + g)$.

Convergence of CV to SURE An analogous approximation holds for leave-one-out (LOO) loss. To obtain the LOO sample mean, the influence function $\frac{1}{\sqrt{n}} X_n^i$ is subtracted from the sample mean $\tilde{\theta}_n$, which gives

$$L_n^{-i}(\theta) \approx const. + \tfrac{1}{2}\|\theta - \tilde{\theta}_n^{-i}\|^2, \qquad \text{where } \tilde{\theta}_n^{-i} = \tilde{\theta}_n - \tfrac{1}{\sqrt{n}} X_n^i.$$

The penalized LOO estimator is then approximately given by

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

In the last step we have replaced g^λ by its first-order Taylor approximation around $\tilde{\theta}_n$, at points $\tilde{\theta}_n$ where g^λ is differentiable. (This approximation won't hold at kink-points of g^λ, which exist for Lasso penalties, in particular.)

The n-fold cross-validation estimator of the risk of $\hat{\theta}_n^\lambda$ can be approximated by

$$CV_n(\lambda) = \sum_i l_n(\hat{\theta}_n^{\lambda,-i}, Z_n^i) \approx const. + \tfrac{1}{n} \sum_i \|\hat{\theta}_n^{\lambda,-i} - \theta_0 - \sqrt{n} X_n^i\|^2.$$

When we take this expression, plug in the approximate form of $\hat{\theta}_n^{\lambda,-i}$, multiply out the inner products, and omit terms which do not depend on λ, we obtain

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

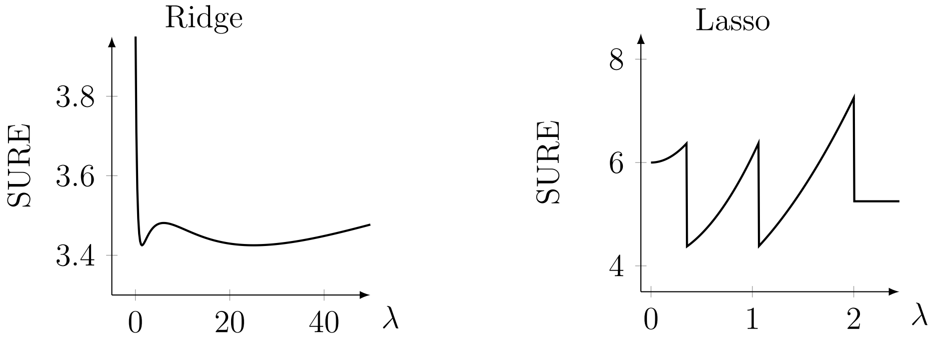
where $\hat{\Sigma}_n$ is the sample second moment of X_n. In the second line, *const.* subsumes any terms that do not depend on λ, while the approximation omits terms that depend on λ but are of order $1/\sqrt{n}$. This approximation to CV_n has the form of Stein's Unbiased Risk Estimate. The first term in this approximation to $CV_n(\lambda)$ corresponds to the average in-sample error, the second term has the form of a covariance penalty (Efron, 2004).

Convergence of tuning parameter and tuned estimators We will need to show that this approximation is uniformly valid in λ. We will furthermore need to show that uniform proximity of CV_n to $SURE$ is enough to guarantee proximity of the corresponding optimized tuning parameters,

$$\underset{\lambda}{\operatorname{argmin}}\; CV_n(\lambda) \approx \underset{\lambda}{\operatorname{argmin}}\; SURE(\lambda, \hat{\theta}_n, \hat{\Sigma}_n).$$

This latter step is non-trivial, because the criterion function $SURE(\lambda, \hat{\theta}_n, \hat{\Sigma}_n)$

Figure 2: Examples of multi-modality of $SURE$


Ridge

Lasso

Notes: These plots show examples of multi-modality for SURE, for the case of L^2 penalties (Ridge) and L^1 penalties (Lasso). Both examples are reproduced from Wilson et al. (2020).

typically has multiple local minima. For certain values of $\hat{\theta}_n$ this function furthermore has multiple *global* minima. When $SURE$ has multiple global (near-)minima, uniform closeness of CV to $SURE$ is not enough to ensure closeness of the minimizer of CV to the minimizer of $SURE$. The plots in Figure 2, which are reproduced from Wilson et al. (2020),[7] illustrate two numerical examples (realizations of $\hat{\theta}_n$ and of $\hat{\Sigma}_n$) for which $SURE$ indeed has multiple global minima.

Using separate arguments for Ridge (Appendix C) and Lasso (Appendix D), we will prove, however, that the global minimum of $SURE$ with respect to λ is unique and well separated almost everywhere, in a suitable sense. Put differently, cases such as those represented in Figure 2 are non-generic, such that they do not lead to a breakdown of convergence for the optimized tuning parameter. The arguments proving that multiple global minima only occur on a set of measure 0 might be the most non-standard part of our proof

Well-separation ensures that the argmin functional is continuous at almost every realization of $\hat{\theta}$. From these results we thus conclude that the mapping from $\hat{\theta}_n$ and CV_n to the tuned estimate $\hat{\theta}^*_n$ is almost everywhere continuous.

[7] The numerical values corresponding to these examples are as follows: (a) SURE for Ridge: $\hat{\theta} = (1.3893, 1.5)$, $L(\theta) = (\theta - \hat{\theta})\,\mathrm{diag}(1, 40)(\theta - \hat{\theta})$, $\pi(\theta) = \|\theta\|^2$, (b) SURE for Lasso: $\hat{\theta} = \frac{1}{\sqrt{n}}(\sqrt{1/8}, \sqrt{9/8}, 2)$, $\pi(\theta) = \sum |\theta_j|$.

This allows us to invoke the continuous mapping and dominated convergence theorems, and to conclude the proof of Theorem 1.

3.2 Intermediate lemmas

Let us now turn to a more formal exposition of our argument. We will prove Theorem 1 in a series of Lemmas. The lemmas are stated in this section, their proofs in the appendices. All results impose the assumptions stated in Section 2.

Lemma 1 (Lipschitz g^λ)**.** *For any* $\lambda \geq 0$*, if* $\pi(\cdot)$ *is convex then* $g^\lambda(\theta) = \operatorname{argmin}_g \frac{1}{2}\|g\|^2 + \lambda \cdot \pi(\theta + g)$ *is Lipschitz with Lipschitz constant 1.*

Lemma 2 (Influence function approximation)**.**

$$L_n(\theta) - L_n(\theta_0) = \tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \epsilon_n(\theta), \tag{1}$$

where $\sup_{\theta:\|\theta\|\leq C} \epsilon_n(\theta) = o_{\mu_n}(1)$ *and* $\sup_{\theta:\|\theta\|\leq C} \nabla\epsilon_n(\theta) = o_{\mu_n}(1)$ *for any* $C < \infty$*, and*

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

The ERM and penalized ERM estimators

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmin}}\ L_n(\theta), \qquad \hat{\theta}_n^\lambda = \underset{\theta}{\operatorname{argmin}}\ [L_n(\theta) + \lambda \cdot \pi(\theta)]$$

satisfy

$$\hat{\theta}_n = \tilde{\theta}_n + o_{\mu_n}(1), \qquad \sup_\lambda \|\hat{\theta}_n^\lambda - \tilde{\theta}_n - g^\lambda(\tilde{\theta}_n)\| = o_{\mu_n}(1).$$

Lemma 3 (Limiting squared error loss)**.** *The limiting expected loss is well defined and given by*

$$\bar{L}(\theta, \theta_0) = \tfrac{1}{2}\|\theta - \theta_0\|^2.$$

Convergence of $\bar{L}_n(\theta, \theta_0)$ to this limit is uniform in any bounded neighborhood of θ_0: $\sup_{\theta:\|\theta-\theta_0\|\leq C} |\bar{L}_n(\theta, \theta_0) - \bar{L}(\theta, \theta_0)| \to 0$ for all $C < \infty$.

Corollary 2 (Asymptotic distribution for fixed tuning parameter)**.** *The ERM and penalized ERM estimators satisfy*

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

Lemma 4 (Convergence of CV to SURE)**.**
The n-fold crossvalidation criterion satisfies

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

Lemma 5 (Joint convergence of tuning parameter and tuned estimators)**.**

$$(\lambda_n^*, \hat{\theta}_n) \to^d (\lambda^*, \hat{\theta}).$$

and

$$\hat{\theta}_n^* \to^d \hat{\theta}^*.$$

From 5, we then show Theorem 1. The appendices prove each of these Lemmas in turn.

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Hirano, K. and Porter, J. R. (2020). Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of Econometrics*, pages 283–354. Elsevier BV.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Le Cam, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 245–261. University of California Press.

Li, K.-C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975.

Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Rudin, W. (1991). *Principles of mathematical analysis*. McGraw-Hill.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS)*, 108.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

A Proofs: Influence function approximations

A.1 Proof of Lemma 1 (Lipschitz g^λ)

Fix $0 \leq \lambda < \infty$. Recall that $g^\lambda(\theta) = \text{argmin}_g \frac{1}{2}\|g\|^2 + \lambda \cdot \pi(\theta + g)$. If $\pi(\cdot)$ is convex, then so is the objective function on the right. This implies that there exists a sub-gradient $\nabla\pi$ of π such that the first order condition

$$g + \lambda \cdot \nabla\pi(\theta + g) = 0$$

holds for $g = g^\lambda(\theta)$. Consider two values θ_1, θ_2 of θ, and the corresponding solutions g_1, g_2 and sub-gradients $\nabla\pi_1, \nabla\pi_2$, as well as the differences $\Delta\theta = \theta_2 - \theta_1$ and $\Delta g = g_2 - g_1$. Taking the difference of the first order condition across the two values yields.

$$\Delta g + \lambda \cdot [\nabla\pi_2 - \nabla\pi_1] = 0.$$

Convexity of π implies

$$\langle \nabla\pi_2 - \nabla\pi_1, \Delta\theta + \Delta g \rangle \geq 0.$$

Combining the last two equations yields

$$\langle \Delta g, \Delta\theta + \Delta g \rangle = -\lambda \cdot \langle \nabla\pi_2 - \nabla\pi_1, \Delta\theta + \Delta g \rangle \leq 0,$$

and thus (using Cauchy-Schwartz to get the second inequality),

$$\|\Delta g\|^2 \leq \langle \Delta g, -\Delta\theta \rangle \leq \|\Delta g\| \cdot \|\Delta\theta\|,$$

so that

$$\|\Delta g\| \leq \|\Delta\theta\|.$$

This proves that $g^\lambda(\theta)$ is Lipschitz with Lipschitz-constant 1. □

A.2 Proof of Lemma 2 (Influence function approximation)

Assume first that Equation (1) holds, so that $L_n(\theta) - L_n(\theta_0) = \frac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \epsilon_n(\theta)$. We show, under this assumption, that $\sup_\lambda \|\hat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\|^2 = o_{\mu_n}(1)$. Leveraging convexity of π and Lipschitz continuity of g^λ, we first bound the corresponding difference in penalized squared error, which allows us to bound the difference in squared error, and finally the difference between the estimators themselves.

Bounding the difference in penalized squared error loss Define

$$\tilde{\theta}_n^\lambda = \underset{\theta}{\operatorname{argmin}} \left[\tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\theta)\right] = \tilde{\theta}_n + g^\lambda(\tilde{\theta}_n).$$

By definition,

$$\hat{\theta}_n^\lambda = \underset{\theta}{\operatorname{argmin}} \left[L_n(\theta) + \lambda \cdot \pi(\theta)\right]$$

and thus

$$L_n(\hat{\theta}_n^\lambda) + \lambda \cdot \pi(\hat{\theta}_n^\lambda) \leq L_n(\tilde{\theta}_n^\lambda) + \lambda \cdot \pi(\tilde{\theta}_n^\lambda).$$

Substituting for $L_n(\cdot)$ on both sides of this inequality, using Equation (1) applied to both $\theta = \hat{\theta}_n^\lambda$ and $\theta = \tilde{\theta}_n^\lambda$, and rearranging yields

$$\left[\tfrac{1}{2}\|\hat{\theta}_n^\lambda - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\hat{\theta}_n^\lambda)\right] - \left[\tfrac{1}{2}\|\tilde{\theta}_n^\lambda - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\tilde{\theta}_n^\lambda)\right] \leq \epsilon_n(\tilde{\theta}_n^\lambda) - \epsilon_n(\hat{\theta}_n^\lambda). \quad (2)$$

Bounding the difference in squared error loss We next prove the following claim: Convexity of π, and the definition $\tilde{\theta}_n^\lambda = \operatorname{argmin}_\theta \left[\tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\theta)\right]$, imply that, for any θ,

$$\tfrac{1}{2}\|\theta - \tilde{\theta}_n^\lambda\|^2 \leq \left[\tfrac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\theta)\right] - \left[\tfrac{1}{2}\|\tilde{\theta}_n^\lambda - \tilde{\theta}_n\|^2 + \lambda \cdot \pi(\tilde{\theta}_n^\lambda)\right]. \quad (3)$$

To show (3), denote $a(\theta) = \frac{1}{2}\|\theta - \tilde{\theta}_n\|^2$ and $b(\theta) = \lambda \cdot \pi(\theta)$. We can write

$$\tfrac{1}{2}\|\theta - \tilde{\theta}_n^\lambda\|^2 = a(\theta) - a(\tilde{\theta}_n^\lambda) - \nabla a(\tilde{\theta}_n^\lambda) \cdot (\theta - \tilde{\theta}_n^\lambda).$$

By convexity of π and optimality of $\tilde{\theta}_n^\lambda$, there exists a subgradient ∇b of b such that

$$\nabla a(\tilde{\theta}_n^\lambda) + \nabla b(\tilde{\theta}_n^\lambda) = 0.$$

Eliminating the common terms $a(\theta) - a(\tilde{\theta}_n^\lambda)$ on the left and right hand side, we can now rewrite (3) as

$$-\nabla a(\tilde{\theta}_n^\lambda) \cdot (\theta - \tilde{\theta}_n^\lambda) \leq b(\theta) - b(\tilde{\theta}_n^\lambda).$$

But since $-\nabla a(\tilde{\theta}_n^\lambda) = \nabla b(\tilde{\theta}_n^\lambda)$, this inequality holds by convexity of b and the definition of a subgradient, and the claim follows.

Bounding the distance between estimators Combining two inequalities (2) and (3) yields

$$\tfrac{1}{2}\|\hat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\|^2 \leq \epsilon_n(\tilde{\theta}_n^\lambda) - \epsilon_n(\hat{\theta}_n^\lambda).$$

It follows from Assumption 4 item 2, which states that $\hat{\theta}_n^\lambda$ is bounded in probability, and the Lipschitzness of g^λ (Lemma 1), which implies $\|\tilde{\theta}_n^\lambda\| \leq 2\|\tilde{\theta}_n\|$, that both $\hat{\theta}_n^\lambda$ and $\tilde{\theta}_n^\lambda$ are bounded in probability. Combined with equation (1), we consequently obtain that with probability approaching 1 under μ_n, there exists $C < \infty$ such that

$$\sup_\lambda \tfrac{1}{2}\|\hat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\|^2 \leq \left(2 \sup_{\|\theta\| \leq C} \epsilon_n(\theta)\right) = o(1).$$

The statement for the ERM estimator follows as a special case, where $\lambda = 0$.

Proving Equation (1)**, using empirical process theory** It remains to show that that (1) holds, where $\sup_{\|\theta\| \leq C} \epsilon_n(\theta) = o_{\mu_n}(1)$. This claim follows from a straightforward generalization of the proof of Lemma 19.31 in van der Vaart (2000) to the case of drifting distributions.

Applicability of arguments of Lemma 19.31 in van der Vaart (2000) is guaranteed by the conditions in Assumption 4, item 1. In particular, pointwise convergence, for fixed θ, follows from almost sure differentiability of l, by dominated convergence, given the uniform bound on the variance of $m(Z_n^i)$ in Assumption 4. To get uniform convergence across values of θ in any ball of radius δ around 0, tightness needs to be shown. Tightness follows from a bound on the bracketing number of the class of functions $\{\sqrt{n}(l_n(\theta, \cdot) - l_n(0, \cdot)) : \|\theta\| \leq \delta\}$. The bound in the proof of Lemma 19.31 in van der Vaart (2000) applies verbatim, with a constant C that does not depend on n, based on the uniform bound on the variance of $m(Z_n^i)$ in Assumption 4. The claim follows.

The claim that $\sup_{\theta:\|\theta\|\leq C} \nabla \epsilon_n(\theta) = o_{\mu_n}(1)$ follows from the same argument, applied to $\nabla_\theta l_n(\theta, Z_n^i)$, using the condition on scores in item 2 of Assumption 5. □

A.3 Proof of Lemma 3 (Limiting squared error loss)

By Definition 1, $\bar{L}_n(\theta, \theta_0) = E\left[L_n(\theta) - L_n(\theta_0)\right]$, and $\bar{L}_n(\theta, \theta_0)$ is minimized at $\theta = \theta_0$. By a second order Taylor expansion around $\theta = \theta_0$,

$$\bar{L}_n(\theta, \theta_0) = \tfrac{1}{2}(\theta - \theta_0) \cdot \nabla_\theta^2 \bar{L}_n(\tilde{\theta}, \theta_0) \cdot (\theta - \theta_0).$$

for some $\tilde{\theta}$ between θ and θ_0. By Assumption 2, $\nabla_\theta^2 \bar{L}(\theta, \theta_0)|_{\theta=\theta_0} = I$. By definition,

$$\nabla_\theta^2 \bar{L}_n(\theta, \theta_0) = \nabla_\beta^2 E\left[l(\beta, Z_n^i)\right]|_{\beta=\theta/\sqrt{n}}.$$

The claim of the lemma then follows from continuity of the Hessian of $\nabla_\beta^2 E\left[l(\beta, Z_n^i)\right]$ at $\beta = 0$. Continuity of the Hessian follows from item 3 of Assumption 5:

$$\begin{aligned}
&\left\|\nabla_\beta^2 E\left[l(\beta, Z_n^i)\right]|_{\beta=\theta/\sqrt{n}} - \nabla_\beta^2 E\left[l(\beta, Z_n^i)\right]|_{\beta=0}\right\| \\
\leq & E\left[\left\|\nabla_\beta^2 l(\beta, Z_n^i)\big|_{\beta=\theta/\sqrt{n}} - \nabla_\beta^2 l(\beta, Z_n^i)\big|_{\beta=0}\right\|\right] \\
\leq & E\left[C_n(Z_n^i)\right] \cdot \frac{\|\theta\|}{\sqrt{n}},
\end{aligned}$$

where $\sup_n E\left[C_n(Z_n^i)\right] < \infty$; this follows from $\sup_n E[C_n(Z_n^i)^2] < \infty$ (Assumption 5.3) via Jensen's inequality.

□

A.4 Proof of Corollary 2 (Asymptotic distribution for fixed tuning parameter)

Recall that $\mathrm{Var}(X_n^1) = \Sigma$ is constant in n, by assumption. Note furthermore that Assumption 4 (item 1) implies the Lindeberg condition

$$E\left[\|X_n^1\|^2 \cdot \mathbf{1}(\|X_n^1\| > \sqrt{n}M)\right] \to 0$$

for all $M > 0$, since $\|X_n^1\| \leq m(Z_n^1)$: $X_n^i = -\nabla_\beta l(\theta_0/\sqrt{n}, Z_n^i)$, and the Lipschitz condition in Assumption 4.1 together with a.e. differentiability implies $\|\nabla_\beta l(\beta, Z_n^i)\| \leq m(Z_n^i)$ at all points of differentiability, and the variance of the latter is uniformly bounded. The Lindeberg-Feller central limit theorem (Proposition 2.27 in van der Vaart 2000), applied to the triangular array (X_n^i), therefore implies $\tilde{\theta}_n \to^d N(\theta_0, \Sigma)$.

The claims of Corollary 2 then follow from Lemma 2, and the continuous mapping theorem, where continuity of g^λ follows from convexity of π, by Lemma 1.

□

B Proof of Lemma 4

Step 0 (Preliminary observations):

We start by stating some useful results for the proof. First, note that by Lemma 1 in Wilson et al. (2020) and Assumption 5(i), it follows

$$\sup_{\lambda} \left\| \hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^{\lambda} \right\| = O_{\mu_n}\left(\frac{1}{\mu} \left\| \nabla_\theta l_n(\hat{\theta}_n, Z_n^i) \right\| \right) = O_{\mu_n}(n^{-1/2}). \tag{4}$$

An analogous argument implies

$$\sup_{\lambda} \left\| \tilde{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^{\lambda} \right\| = O_{\mu_n}(n^{-1/2}), \tag{5}$$

where

$$\tilde{\theta}^{\lambda,-i} := \operatorname*{argmin}_{\theta} \left[\tfrac{1}{2} \| \theta - \tilde{\theta}_n^{-i} \|^2 + \lambda \cdot \pi(\theta) \right]. \tag{6}$$

Second, recall from Lemma 2 that

$$\sup_{\lambda} \left\| \hat{\theta}_n^{\lambda} - \tilde{\theta}_n^{\lambda} \right\| = o_{\mu_n}(1). \tag{7}$$

The next set of results concern the properties of $g^\lambda(\cdot)$. Since $g^\lambda(\cdot)$ is Lipschitz continuous by Lemma 1, it is differentiable almost everywhere (by Rademacher's theorem). In particular, there exists an $R^\lambda(\cdot;\theta)$ such that

$$g^\lambda(\theta + \delta) = g^\lambda(\theta) + \nabla g^\lambda(\theta)^\intercal \delta + R^\lambda(\delta;\theta), \tag{8}$$

and

$$\lim_{\|\delta\| \to 0} \frac{\left\| R^\lambda(\delta;\theta) \right\|}{\|\delta\|} = 0 \text{ for each } \lambda \text{ and (Lebesgue) almost every } \theta. \tag{9}$$

In fact, under Assumption 3, we can strengthen (9) to:

$$\lim_{\|\delta\|\to 0} \sup_{\lambda\in\Lambda} \frac{\left\|R^\lambda(\delta;\theta)\right\|}{\|\delta\|} = 0 \text{ for (Lebesgue) almost every } \theta. \tag{10}$$

For Ridge, (10) is immediate, since $R^\lambda(\delta,\theta) = 0$. For Lasso, it follows from Lemma 8, which implies $R^\lambda(\delta,\theta) = 0$ for δ small enough, except on a set of θ values with Lebesgue measure 0.

For values of θ where $\nabla g^\lambda(\theta)$ does not exist, we somewhat arbitrarily set $\nabla g^\lambda(\theta) = 0$ and define $R^\lambda(\delta;\theta) = g^\lambda(\theta+\delta) - g^\lambda(\theta)$ for these values; that way (8) always holds. Observe that due to Lemma 1, $\left\|g^\lambda(\theta+\delta) - g^\lambda(\theta)\right\| \leq \|\delta\|$ and $\left\|\nabla g^\lambda(\theta)\right\| \leq 1$ (whenever the gradient exists), so

$$\sup_{\lambda,\theta} \left\|R^\lambda(\delta;\theta)\right\| \leq 2\,\|\delta\|\,. \tag{11}$$

Step 1:

We first show that

$$\mathrm{CV}_n(\lambda) = \sum_{i=1}^n l_n\left(\hat{\theta}_n^\lambda, Z_n^i\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle \tilde{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^\lambda, \sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^\lambda, Z_n^i\right)\right\rangle + o_{\mu_n}(1), \tag{12}$$

uniformly over λ. By a first order Taylor expansion,

$$\begin{aligned}
\mathrm{CV}_n(\lambda) &= \sum_{i=1}^n l_n\left(\hat{\theta}_n^{\lambda,-i}, Z_n^i\right) \\
&= \sum_{i=1}^n l_n\left(\hat{\theta}_n^\lambda, Z_n^i\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle \hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^\lambda, \sqrt{n}\nabla_\theta l_n\left(\hat{\theta}_n^\lambda, Z_n^i\right)\right\rangle \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{i=1}^n r_n\left(\hat{\theta}_n^{\lambda,-i}, \hat{\theta}_n^\lambda, Z_n^i\right),
\end{aligned}$$

where

$$\left|r_n\left(\hat{\theta}_n^{\lambda,-i}, \hat{\theta}_n^\lambda, Z_n^i\right)\right| \leq B_n(Z_n^i) \cdot \left\|\hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^\lambda\right\|^2$$

by Assumption 5. Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_n\left(\hat{\theta}_n^{\lambda,-i}, \hat{\theta}_n^{\lambda}, Z_n^i\right)\right| \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}\left|B_n(Z_n^i)\right|^2\right)^{1/2}\left(\sum_{i=1}^{n}\left\|\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda}\right\|^4\right)^{1/2}.
\end{aligned}$$

The first term on the right hand side of the above expression is $O_{\mu_n}(1)$ by Assumption 5, while the second term is $O_{\mu_n}(n^{-1/2})$ by (4) and the two requirements of Assumption 5 since

$$\begin{aligned}
&\sup_{\lambda}\sum_{i=1}^{n}\left\|\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda}\right\|^4 \\
&\leq \frac{1}{n^2\mu^4}\sum_{i=1}^{n}\left\|\sqrt{n}\nabla_\theta l_n(\hat{\theta}_n, Z_n^i)\right\|^4 \\
&\leq \frac{8}{n^2\mu^4}\sum_{i=1}^{n}\left\|\sqrt{n}\nabla_\theta l_n(\hat{\theta}_n, Z_n^i)-\sqrt{n}\nabla_\theta l_n(\theta_0, Z_n^i)\right\|^4+\frac{8}{n^2\mu^4}\sum_{i=1}^{n}\left\|\sqrt{n}\nabla_\theta l_n(\theta_0, Z_n^i)\right\|^4 \\
&= O_{\mu_n}(n^{-1}), \qquad (13)
\end{aligned}$$

so the expression overall is $O_{\mu_n}(n^{-1/2})$.

We now show that, uniformly over λ, one can approximate

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda}, \sqrt{n}\nabla_\theta l_n\left(\hat{\theta}_n^{\lambda}, Z_n^i\right)\right\rangle. \qquad (14)$$

with

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle\tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, \sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda}, Z_n^i\right)\right\rangle.$$

To this end, we first argue that (14) can be approximated with

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda}, \sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda}, Z_n^i\right)\right\rangle.$$

By the Cauchy-Schwarz inequality, the approximation error is bounded by

$$\sup_{\lambda\in\Lambda}\left(\sum_{i=1}^{n}\left\|\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda}\right\|^2\right)^{1/2}\cdot\sup_{\lambda\in\Lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\sqrt{n}\nabla_\theta l_n\left(\hat{\theta}_n^{\lambda},Z_n^i\right)-\sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda},Z_n^i\right)\right\|^2\right)^{1/2}. \tag{15}$$

The first term in (15) is $O_{\mu_n}(1)$ by (4) and Assumption 5 (the argument is analogous to 13). The second term in (15) is $o_{\mu_n}(1)$ under (7) and Assumption 5(ii).

It then remains to show

$$\sup_{\lambda\in\Lambda}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle\hat{\theta}_n^{\lambda,-i}-\hat{\theta}_n^{\lambda},\sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda},Z_n^i\right)\right\rangle\right.$$
$$\left.-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle\tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda},\sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda},Z_n^i\right)\right\rangle\right|=o_{\mu_n}(1).$$

By the Cauchy-Schwarz inequality, the expression on the left is bounded by

$$\sup_{\lambda\in\Lambda}\left(\sum_{i=1}^{n}\left\|\left(\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right)-\left(\hat{\theta}_n^{\lambda}-\tilde{\theta}_n^{\lambda}\right)\right\|^2\right)^{1/2}\cdot\sup_{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda},Z_n^i\right)\right\|^2\right)^{1/2}.$$

The second term in the above expression is $O_{\mu_n}(1)$ by Assumption 5(ii) (4th moment bound on scores) and the Lipschitz condition, since $\tilde{\theta}_n^{\lambda}$ is bounded in probability. At the end of this proof, we analyze the first term, showing that

$$\sup_{\lambda\in\Lambda}\left(\sum_{i=1}^{n}\left\|\left(\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right)-\left(\hat{\theta}_n^{\lambda}-\tilde{\theta}_n^{\lambda}\right)\right\|^2\right)=o_{\mu_n}(1).$$

Combining the above results proves (12).

Step 2:

Next, we show that uniformly over λ, the term

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, \sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda}, Z_n^i\right)\right\rangle \tag{16}$$

in (12) can be approximated by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, -X_n^i\right\rangle .$$

Indeed, under Assumption 5(iii), we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, \sqrt{n}\nabla_\theta l_n\left(\tilde{\theta}_n^{\lambda}, Z_n^i\right)\right\rangle - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, -X_n^i\right\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, n\nabla_\theta^2 l_n\left(\theta_0, Z_n^i\right)\cdot\left(\tilde{\theta}_n^{\lambda}-\theta_0\right)\right\rangle + \frac{1}{n}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, \Delta_n\left(\tilde{\theta}_n^{\lambda}-\theta_0, Z_n^i\right)\right\rangle ,\end{aligned} \tag{17}$$

where

$$\left\|\Delta_n\left(\tilde{\theta}_n^{\lambda}-\theta_0, Z_n^i\right)\right\| \le C_n(Z_n^i)\cdot\left\|\tilde{\theta}_n^{\lambda}-\theta_0\right\|^2 ,$$

and the function $C_n(\cdot)$ is defined in Assumption 5(iii).

By the Cauchy-Schwarz inequality, Assumption 5 and (5),

$$\begin{aligned}&\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}, n\nabla_\theta^2 l_n\left(\theta_0, Z_n^i\right)\cdot\left(\tilde{\theta}_n^{\lambda}-\theta_0\right)\right\rangle\right| \\ &\le n^{-1/2}\cdot\sup_{\lambda\in\Lambda}\left(\sum_{i=1}^{n}\left\|\tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}\right\|^2\right)^{1/2}\cdot\sup_{\lambda\in\Lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|n\nabla_\theta^2 l_n\left(\theta_0, Z_n^i\right)\right\|^2\right)^{1/2}\cdot\sup_{\lambda\in\Lambda}\left\|\tilde{\theta}_n^{\lambda}-\theta_0\right\| \\ &= n^{-1/2}\cdot O_{\mu_n}(1)\cdot O_{\mu_n}(1)\cdot O_{\mu_n}(1) = O_{\mu_n}(n^{-1/2}).\end{aligned}$$

This proves that the first term in the right hand side of (17) is $o_{\mu_n}(1)$ uniformly over λ. By an analogous argument, the second term in the right hand side of (17) is also $o_{\mu_n}(1)$ uniformly over λ.

Step 3:

It thus remains to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda},\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle$$

is asymptotically equivalent to the degrees of freedom term in SURE.

By the definition of $R^\lambda(\cdot)$, we may write

$$\tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda}=-\frac{1}{\sqrt{n}}X_n^i-\frac{1}{\sqrt{n}}\nabla g^\lambda\left(\tilde{\theta}_n\right)^\intercal X_n^i+R^\lambda(\tilde{\theta}_n^{-i}-\tilde{\theta}_n;\tilde{\theta}_n).$$

We can thus expand

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \tilde{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda},\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\left\langle -X_n^i,\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle+\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla g^\lambda\left(\tilde{\theta}_n\right)^\intercal X_n^i,-\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle\\&\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^\lambda(\tilde{\theta}_n^{-i}-\tilde{\theta}_n;\tilde{\theta}_n),\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle.\end{aligned}\tag{18}$$

The first term in (18) is independent of λ and can therefore be neglected.

The second term in (18) is asymptotically equivalent to the degrees of freedom term in SURE. Indeed, since $\tilde{\theta}_n^{-i}-\tilde{\theta}_n=\nabla_\theta l_n\left(\theta_0,Z_n^i\right)$, we can write

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla g^\lambda\left(\tilde{\theta}_n\right)^\intercal X_n^i,-\sqrt{n}\nabla_\theta l_n\left(\theta_0,Z_n^i\right)\right\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla g^\lambda\left(\tilde{\theta}_n\right)^\intercal X_n^i,X_n^i\right\rangle\\&=\mathrm{Tr}\left[\nabla g^\lambda\left(\tilde{\theta}_n\right)^\intercal\hat{\Sigma}_n\right],\end{aligned}$$

where

$$\hat{\Sigma}_n:=\frac{1}{n}\sum_{i=1}^{n}(X_n^i)(X_n^i)^\intercal.$$

But by the law of large of numbers, which can be applied here due to Assumption 5(ii), $\hat{\Sigma}_n = \Sigma + o_{\mu_n}(1)$. We thus conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla g^{\lambda}\left(\tilde{\theta}_n\right)^{\intercal} X_n^i, -\sqrt{n}\nabla_{\theta} l_n\left(\theta_0, Z_n^i\right)\right\rangle = \mathrm{Tr}\left[\nabla g^{\lambda}\left(\tilde{\theta}_n\right)^{\intercal}\Sigma\right] + o_{\mu_n}(1),$$

uniformly over $\lambda \in \Lambda$.

It remains to show the third term in (18) is negligible, i.e.,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^{\lambda}(\tilde{\theta}_n^{-i} - \tilde{\theta}_n; \tilde{\theta}_n), \sqrt{n}\nabla_{\theta} l_n\left(\theta_0, Z_n^i\right)\right\rangle = o_{\mu_n}(1).$$

Recall that $\tilde{\theta}_n^{-i} - \tilde{\theta}_n = -X_n^i/\sqrt{n}$. Fix some $a \in (0, 1/2)$ and $C < \infty$, and expand

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^{\lambda}(\tilde{\theta}_n^{-i} - \tilde{\theta}_n; \tilde{\theta}_n), \sqrt{n}\nabla_{\theta} l_n\left(\theta_0, Z_n^i\right)\right\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^{\lambda}(-X_n^i/\sqrt{n}; \tilde{\theta}_n), -X_n^i\right\rangle \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^{\lambda}(-X_n^i/\sqrt{n}; \tilde{\theta}_n)\cdot \mathbb{I}\left\{\left\|X_n^i\right\| \geq Cn^a\right\}, -X_n^i\right\rangle \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle R^{\lambda}(-X_n^i/\sqrt{n}; \tilde{\theta}_n)\cdot \mathbb{I}\left\{\left\|X_n^i\right\| < Cn^a\right\}, -X_n^i\right\rangle \\
&:= A_{n1}^{\lambda} + A_{n2}^{\lambda}.
\end{aligned}$$

We analyze the terms A_{n1}^{λ} and A_{n2}^{λ} separately. For the term A_{n1}^{λ}, observe that by (11),

$$\sup_{\lambda\in\Lambda}|A_{n1}^{\lambda}| \leq \frac{2}{n}\sum_{i=1}^{n}\left\|X_n^i\right\|^2 \cdot \mathbb{I}\left\{\left\|X_n^i\right\| \geq Cn^a\right\}.$$

Consequently, under Assumption 5(ii) and the given choice of a,

$$\mathbb{E}_{\mu_n}\left[|A_{n1}^{\lambda}|\right] \leq \frac{2}{C^2 n^{2a}}\mathbb{E}_{\mu_n}\left[\left\|X_n^i\right\|^4\right] \to 0$$

as $n \to \infty$. Thus, $\sup_{\lambda\in\Lambda} A_{n1}^{\lambda} = o_{\mu_n}(1)$.

Next, we show $\sup_{\lambda\in\Lambda} A_{n2}^{\lambda} = o_{\mu_n}(1)$. Due to exchangeability over i, this follows if we show that

$$\lim_{n\to\infty} \sqrt{n}\mathbb{E}_{\mu_n}\left[\left|\left\langle R^{\lambda}(-X_n^i/\sqrt{n};\tilde{\theta}_n)\cdot \mathbb{I}_{\Gamma_i}, -X_n^i\right\rangle\right|\right] = 0, \tag{19}$$

where we use $\mathbb{I}_{\Gamma_i}$ as a short-hand for $\mathbb{I}\{\|X_n^i\| < Cn^a\}$. Now, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_{\mu_n}\left[\left|\left\langle R^{\lambda}(-X_n^i/\sqrt{n};\tilde{\theta}_n)\cdot \mathbb{I}_{\Gamma_i}, -X_n^i\right\rangle\right|\right] \\
&\quad \le \mathbb{E}_{\mu_n}^{1/2}\left[\sup_{\lambda\in\Lambda}\frac{\left\|R^{\lambda}\left(-X_n^i/\sqrt{n};\tilde{\theta}_n\right)\right\|^2}{\left\|-X_n^i/\sqrt{n}\right\|^2}\cdot \mathbb{I}_{\Gamma_i}\right]\cdot \mathbb{E}_{\mu_n}^{1/2}\left[\left\|X_n^i\right\|^4\right].
\end{aligned}$$

But $\mathbb{E}_{\mu_n}^{1/2}\left[\left\|X_n^i\right\|^4\right] \le M < \infty$ under Assumption 5(ii), so (19) would follow if we show that

$$\lim_{n\to\infty} \mathbb{E}_{\mu_n}\left[\sup_{\lambda\in\Lambda}\frac{\left\|R^{\lambda}\left(-X_n^i/\sqrt{n};\tilde{\theta}_n\right)\right\|^2}{\left\|-X_n^i/\sqrt{n}\right\|^2}\cdot \mathbb{I}_{\Gamma_i}\right] = 0. \tag{20}$$

To prove (20), observe that it is without loss of generality to suppose $R^{\lambda}(\delta;\theta) = R^{\lambda}(\|\delta\|;\theta)$, i.e., that it depends only on $\|\delta\|$, and that $R^{\lambda}(\|\delta\|;\theta)/\|\delta\|$ is increasing in $\|\delta\|$. Otherwise, we can simply define

$$\bar{R}^{\lambda}(\|\delta\|;\theta) = \sup_{\|\delta'\|\le\|\delta\|}\frac{R^{\lambda}(\delta';\theta)}{\|\delta'\|},$$

and this would satisfy these two conditions while still retaining the property (9). Consequently, the left hand side of (20) can be bounded as

$$\begin{aligned}
\mathbb{E}_{\mu_n}\left[\sup_{\lambda\in\Lambda}\frac{\left\|R^{\lambda}\left(\|X_n^i/\sqrt{n}\|;\tilde{\theta}_n\right)\right\|^2}{\left\|X_n^i/\sqrt{n}\right\|^2}\cdot \mathbb{I}_{\Gamma_i}\right] &\le \mathbb{E}_{\mu_n}\left[\sup_{\lambda\in\Lambda}\left\|\frac{R^{\lambda}\left(Cn^{a-\frac{1}{2}};\tilde{\theta}_n\right)}{Cn^{a-\frac{1}{2}}}\right\|^2\right] \\
&= \mathbb{E}_{\mu_n}\left[\sup_{\lambda\in\Lambda}\left\|B^{\lambda}(Cn^{a-\frac{1}{2}},\tilde{\theta}_n)\right\|^2\right],
\end{aligned}$$

where

$$B^\lambda(\delta, \tilde{\theta}_n) := \frac{\left\| R^\lambda \left(\delta; \tilde{\theta}_n \right) \right\|}{\|\delta\|}.$$

We now bound

$$\mathbb{E}_{\mu_n} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(Cn^{a-\frac{1}{2}}, \tilde{\theta}_n) \right\|^2 \right].$$

Fix some $\epsilon > 0$. By the requirement that $R^\lambda(\|\delta\|; \theta)/\|\delta\|$ is increasing in $\|\delta\|$, along with the fact $a < 1/2$, there exists $\bar{n}$ large enough so that $B^\lambda(Cn^{a-\frac{1}{2}}, \theta) \leq B^\lambda(\epsilon, \theta)$ for each $n \geq \bar{n}$, $\theta \in \mathbb{R}^d$ and $\lambda \in \Lambda$. Now, it is straightforward to show

$$\tilde{\theta}_n \xrightarrow[\mu_n]{d} Z \sim N(\theta_0, \Sigma).$$

We then have

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}_{\mu_n} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(Cn^{a-\frac{1}{2}}, \tilde{\theta}_n) \right\|^2 \right] \\ &\leq \lim_{n\to\infty} \mathbb{E}_{\mu_n} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(\epsilon, \tilde{\theta}_n) \right\|^2 \right] = \mathbb{E} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(\epsilon, Z) \right\|^2 \right], \end{aligned}$$

where the equality follows from the properties of weak convergence since equation (11) implies $\sup_{\lambda\in\Lambda} \left\| B^\lambda(\epsilon, \theta) \right\| \leq 2$ uniformly over θ. But (10) implies $\lim_{\epsilon'\to 0} \sup_{\lambda\in\Lambda} B^\lambda(\epsilon', \theta) = 0$ for every $\theta \in \mathbb{R}^d$ excluding a set of Lebesgue measure 0. Since the Gaussian distribution is absolutely continuous with respect to the Lebesgue measure, it then follows by the dominated convergence theorem that

$$\lim_{\epsilon'\to 0} \mathbb{E} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(\epsilon', Z) \right\|^2 \right] = 0.$$

Since $\epsilon > 0$ was arbitrary, we conclude

$$\lim_{n\to\infty} \mathbb{E}_{\mu_n} \left[\sup_{\lambda \in \Lambda} \left\| B^\lambda(Cn^{a-\frac{1}{2}}, \tilde{\theta}_n) \right\|^2 \right] = 0.$$

This proves (20).

It remains to prove the claim, made in Step 2, that

$$\sup_{\lambda\in\Lambda}\left(\sum_{i=1}^{n}\left\|\left(\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right)-\left(\hat{\theta}_n^{\lambda}-\tilde{\theta}_n^{\lambda}\right)\right\|^2\right)=o_{\mu_n}(1).$$

For the remainder of this proof, we make a case distinction between Ridge penalties and Lasso penalties.

Step 4 (Ridge):

We first specialize to the case of quadratic penalties. Define $\tilde{L}_n^{-i}(\theta)=\frac{1}{2}\left\|\theta-\tilde{\theta}_n^{-i}\right\|^2$. Observe that $\tilde{\theta}_n^{\lambda,-i}=\text{argmin}_{\;\theta}\left\{\tilde{L}_n^{-i}(\theta)+\lambda\pi(\theta)\right\}$. Consequently,

$$\begin{aligned}
&\nabla_\theta\left\{\tilde{L}_n^{-i}(\hat{\theta}_n^{\lambda,-i})+\lambda\pi(\hat{\theta}_n^{\lambda,-i})\right\}\\
&=\nabla_\theta\left\{\tilde{L}_n^{-i}(\hat{\theta}_n^{\lambda,-i})+\lambda\pi(\hat{\theta}_n^{\lambda,-i})\right\}-\nabla_\theta\left\{\tilde{L}_n^{-i}(\tilde{\theta}_n^{\lambda,-i})+\lambda\pi(\tilde{\theta}_n^{\lambda,-i})\right\}\\
&=\left\{\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right\}+\lambda\nabla_\theta\left\{\pi(\hat{\theta}_n^{\lambda,-i})-\pi(\tilde{\theta}_n^{\lambda,-i})\right\}.
\end{aligned}$$

But $\nabla_\theta\left\{L_n^{-i}(\hat{\theta}_n^{\lambda,-i})+\lambda\pi(\hat{\theta}_n^{\lambda,-i})\right\}=0$, so we obtain

$$\left\{\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right\}+\lambda\nabla_\theta\left\{\pi(\hat{\theta}_n^{\lambda,-i})-\pi(\tilde{\theta}_n^{\lambda,-i})\right\}=\nabla_\theta\tilde{L}_n^{-i}(\hat{\theta}_n^{\lambda,-i})-\nabla_\theta L_n^{-i}(\hat{\theta}_n^{\lambda,-i}).$$

In a similar vein,

$$\left\{\hat{\theta}_n^{\lambda}-\tilde{\theta}_n^{\lambda}\right\}+\lambda\nabla_\theta\left\{\pi(\hat{\theta}_n^{\lambda})-\pi(\tilde{\theta}_n^{\lambda})\right\}=\nabla_\theta\tilde{L}_n(\hat{\theta}_n^{\lambda})-\nabla_\theta L_n(\hat{\theta}_n^{\lambda}).$$

When $\pi(\theta)$ is the ridge penalty $\frac{1}{2}\theta^\intercal A^{-1}\theta$, we have

$$\left(\hat{\theta}_n^{\lambda,-i}-\tilde{\theta}_n^{\lambda,-i}\right)-\left(\hat{\theta}_n^{\lambda}-\tilde{\theta}_n^{\lambda}\right)=(I+\lambda A^{-1})^{-1}\left\{\nabla_\theta\left(\tilde{L}_n^{-i}-L_n^{-i}\right)(\hat{\theta}_n^{\lambda,-i})-\nabla_\theta\left(\tilde{L}_n-L_n\right)(\hat{\theta}_n^{\lambda})\right\}.$$

By a third order Taylor expansion,

$$\begin{aligned}
&L_n^{-i}(\theta) - L_n^{-i}(\theta_0) \\
&= \frac{1}{\sqrt{n}} \left\{ \sqrt{n} \nabla_\theta L_n^{-i}(\theta_0) \right\}^\intercal (\theta - \theta_0) + \frac{1}{2n} (\theta - \theta_0)^\intercal \left\{ n \nabla_\theta^2 L_n^{-i}(\theta_0) \right\} (\theta - \theta_0) \\
&\quad + \frac{1}{6n^{3/2}} D_\theta^3 L_n^{-i}(\bar{\theta})[\theta - \theta_0, \theta - \theta_0, \theta - \theta_0],
\end{aligned}$$

for some $\bar{\theta}$ between θ and θ_0. At the same time,

$$\tilde{L}_n^{-i}(\theta) - \tilde{L}_n^{-i}(\theta_0) = \frac{1}{\sqrt{n}} \left\{ \sqrt{n} \nabla_\theta L_n^{-i}(\theta_0) \right\}^\intercal (\theta - \theta_0) + \frac{1}{2} (\theta - \theta_0)^\intercal (\theta - \theta_0).$$

Therefore, by Assumption 5(iv), which implies

$$\mathbb{E}_{\mu_n} \left[\sup_{\theta \in \Theta} \left\| n^2 D_\theta^4 l_n \left(\theta, Z_n^i \right) \right\|^4 \right] \leq M,$$

we obtain

$$\nabla_\theta \left(\tilde{L}_n^{-i} - L_n^{-i} \right) (\theta) = \left\{ \nabla_\theta^2 L_n^{-i}(\theta_0) - I \right\} (\theta - \theta_0) + O_{\mu_n}(n^{-3/2}).$$

In a similar vein,

$$\nabla_\theta \left(\tilde{L}_n - L_n \right) (\theta) = \left\{ \nabla_\theta^2 L_n(\theta_0) - I \right\} (\theta - \theta_0) + O_{\mu_n}(n^{-3/2}).$$

Taken together, we conclude

$$\begin{aligned}
&\left\{ \nabla_\theta \left(\tilde{L}_n^{-i} - L_n^{-i} \right) (\hat{\theta}_n^{\lambda,-i}) - \nabla_\theta \left(\tilde{L}_n - L_n \right) (\hat{\theta}_n^\lambda) \right\} \\
&= \left\{ \nabla_\theta^2 L_n^{-i}(\theta_0) - I \right\} \left(\hat{\theta}_n^{\lambda,-i} - \theta_0 \right) - \left\{ \nabla_\theta^2 L_n(\theta_0) - I \right\} \left(\hat{\theta}_n^\lambda - \theta_0 \right) + O_{\mu_n}(n^{-3/2}) \\
&= -\frac{1}{n} \left\{ n \nabla_\theta^2 l_n(\theta_0, Z_n^i) \right\} \left(\hat{\theta}_n^{\lambda,-i} - \theta_0 \right) + \left\{ \nabla_\theta^2 L_n(\theta_0) - I \right\} \left(\hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^\lambda \right) + O_{\mu_n}(n^{-3/2}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^{n} \left\| \left(\hat{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^{\lambda,-i} \right) - \left(\hat{\theta}_n^{\lambda} - \tilde{\theta}_n^{\lambda} \right) \right\|^2 \\
&\leq \frac{2}{n^2} \left(\sum_i \left\| n \nabla_\theta^2 l_n(\theta_0, Z_n^i) \right\|^2 \cdot \left\| \hat{\theta}_n^{\lambda,-i} - \theta_0 \right\|^2 \right) \\
&\quad + 2 \left\| \nabla_\theta^2 L_n(\theta_0) - I \right\|^2 \cdot \left(\sum_i \left\| \hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^{\lambda} \right\|^2 \right) + O_{\mu_n}(n^{-3/2}) \\
&= o_{\mu_n}(1).
\end{aligned}$$

Step 4 (Lasso):

We finally prove that for the Lasso penalty, we again have that

$$DD_n := \left(\sum_{i=1}^{n} \left\| \left(\hat{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^{\lambda,-i} \right) - \left(\hat{\theta}_n^{\lambda} - \tilde{\theta}_n^{\lambda} \right) \right\|^2 \right) = o_{\mu_n}(1).$$

We consider the case of fixed λ; taking the supremum over λ is trivial when Λ is finite. We will also assume for notational simplicity that $A = I$, so that $h = \theta$; the general case (where $h = A^{-1} \cdot \theta$) follows immediately.

Let now

$$D_n^i := \left(\hat{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^{\lambda,-i} \right) - \left(\hat{\theta}_n^{\lambda} - \tilde{\theta}_n^{\lambda} \right).$$

Denote $\eta = sign(\tilde{\theta}_n^{\lambda})$, and the set of active coordinates as $J = \{j : \eta_j \neq 0\}$. Define the following three event indicators, where the *sign* function is applied component-wise, and $\rho_n > 0$ for some deterministic sequence ρ_n such that $\rho_n \cdot n^{1/4} \to \infty$ and $\rho_n \to 0$:

$$\begin{aligned}
A_n &= \mathbf{1}\left(sign(\widehat{\theta}_n^{\lambda,-i}) = sign(\tilde{\theta}_n^{\lambda,-i}) = sign(\widehat{\theta}_n^{\lambda}) = \eta \text{ for all } i \right), \\
B_n &= (1 - A_n) \cdot \mathbf{1}\Big(sign(\widehat{\theta}_n^{\lambda}) \neq \eta \text{ or } sign(\tilde{\theta}_n + t + g^{\lambda}(\tilde{\theta}_n + t)) \neq \eta \text{ for some } \|t\| < \rho_n \Big) \\
C_n &= 1 - A_n - B_n.
\end{aligned}$$

Then, by construction, $A_n + B_n + C_n = 1$ for all n, i, so that

$$DD_n = \underbrace{A_n \cdot \sum_{i=1}^{n} \|D_n^i\|^2}_{\text{First sum}} + \underbrace{B_n \cdot \sum_{i=1}^{n} \|D_n^i\|^2}_{\text{Second sum}} + \underbrace{C_n \cdot \sum_{i=1}^{n} \|D_n^i\|^2}_{\text{Third sum}}.$$

To show that $DD_n \to 0$, we will consider each of these three sums separately.

First sum Conditional on $A_n = 1$, the signs for all estimators of θ under consideration coincide with η . The first order conditions for the active coordinates for each of the estimators can therefore be written as

$$\begin{aligned} \nabla_J L_n(\widehat{\theta}_n^\lambda) + \lambda \cdot \eta_J &= 0 \\ \nabla_J L_n(\widehat{\theta}_n^{\lambda,-i}) - \nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n) + \lambda \cdot \eta_J &= 0 \\ (\tilde{\theta}_n^\lambda - \tilde{\theta}_n)_J + \lambda \cdot \eta_J &= 0 \\ (\tilde{\theta}_n^{\lambda,-i} - \tilde{\theta}_n)_J - \nabla_J l_n(\theta_0, Z_i^n) + \lambda \cdot \eta_J &= 0. \end{aligned}$$

Taking differences of the first two and of the second two equations, we get

$$\begin{aligned} \nabla_J L_n(\widehat{\theta}_n^\lambda) - \nabla_J L_n(\widehat{\theta}_n^{\lambda,-i}) &= \nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n), \\ (\tilde{\theta}_n^\lambda - \tilde{\theta}_n^{\lambda,-i})_J &= \nabla_J l_n(\theta_0, Z_i^n). \end{aligned}$$

The first of these equations can be rewritten as

$$(\widehat{\theta}_n^\lambda - \widehat{\theta}_n^{\lambda,-i})_J = (\nabla_J^2 L_n(\bar{\theta}_n^i))^{-1} \cdot \nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n)$$

for some intermediate point $\bar{\theta}^i$. Combining, we get $D_{n,J^c}^i = 0$ (for the inactive coordinates, the double difference vanishes) and

$$\begin{aligned} D_{n,J}^i &= \left[(\nabla_J^2 L_n(\bar{\theta}_n^i))^{-1} \cdot \nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n)\right] - \nabla_J l_n(\theta_0, Z_i^n) \\ &= \left[(\nabla_J^2 L_n(\bar{\theta}_n^i))^{-1} - I_J\right] \cdot \nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n) + \left[\nabla_J l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n) - \nabla_J l_n(\theta_0, Z_i^n)\right]. \end{aligned}$$

and thus, conditional on $A_n = 1$,

$$\begin{aligned}\sum_{i=1}^{n}\|D_n^i\|^2 = \sum_{i=1}^{n}&\|D_{n,J}^i\|^2 \\ &\leq 2\max_i \left\|(\nabla_J^2 L_n(\bar{\theta}_n^i))^{-1} - I_J\right\|^2 \cdot \sum_i \left\|\nabla l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n)\right\|^2 \\ &+ 2\sum_i \left\|\nabla l_n(\widehat{\theta}_n^{\lambda,-i}, Z_i^n) - \nabla l_n(\theta_0, Z_i^n)\right\|^2.\end{aligned}$$

Note that we dropped the J subscript on gradients when taking the upper bound. The max term goes to 0 in probability by Lemma 2. The first sum is $O_p(1)$ given the bound on 4th moments of the score in Assumption 5.2. The second sum is $o_p(1)$ by the Lipschitz condition in Assumption 5.2 and by $\widehat{\theta}_n^{\lambda,-i} - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$. It follows that $A_n \cdot \sum_{i=1}^n \|D_n^i\|^2 = o_p(1)$.

Second sum To control the second sum, we next show that $B_n = o_{\mu_n}(1)$. Recall the following results that were shown previously:

- By Lemma 2, $\widehat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda \to^p 0$.
- Also by Lemma 2, $\sup_{\theta:\|\theta\|\leq C} \nabla\epsilon_n(\theta) = o_{\mu_n}(1)$.
- By Corollary 2, $\widehat{\theta}_n^\lambda \to^d \widehat{\theta} + g^\lambda(\widehat{\theta})$, where $\widehat{\theta} \sim N(\theta_0, \Sigma)$.
- By Lemma 8 below, for all $\delta > 0$ there exists a $\gamma > 0$ such that $g^\lambda(\theta)$ is linear on $S_\gamma(\hat{\theta}) = \{\theta : \|\theta - \hat{\theta}\| < \gamma\}$ with probability greater than $1 - \delta$, where $\hat{\theta} \sim N(\theta_0, \Sigma)$.

We claim that the combination of these results implies that $B_n \to^p 0$ as long as $\rho_n \to 0$. To see this, define $\Theta_\gamma = \{\theta : g^\lambda \text{ is linear (affine) on } S_\gamma(\theta)\}$, for $\gamma > 0$. By Lemma 8, $P(\widehat{\theta} \in \Theta_\gamma) > 1 - \delta$ for γ small enough. By Corollary 2, and the definition of convergence in distribution, we therefore get $P(\widehat{\theta}_n \in \Theta_\gamma) > 1 - \delta$ for n large enough.

By Lemma 2, $\|\widehat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\| < \gamma$ with probability greater than $1 - \delta$ for n large enough. This implies that $sign(\widehat{\theta}_n^\lambda)_J = \eta_J$, because $|\tilde{\theta}_{n,j}^\lambda|$ is bounded away from 0 for $j \in J$, by definition of Θ_γ. This takes care of the active coordinates.

Let now $j \in J^c$ be one of the inactive coordinates. If $\tilde{\theta}_n \in \Theta_\gamma$, then necessarily $|\tilde{\theta}_{j,n}| + \gamma < \lambda$, by definition of Θ_γ, since the mapping from $\tilde{\theta}_{j,n}$ to $\tilde{\theta}^\lambda_{j,n}$ has a kink at $\pm\lambda$.

Let θ, θ' be equal in all coordinates except j, where $\theta_j = t$ and $\theta'_j = 0$. Then, by Lemma 2, by $|\tilde{\theta}_{j,n}| + \gamma < \lambda$, and by the Lipschitz continuity of $\epsilon(\theta)$ with constant $\gamma_n < \gamma$, for n large enough, which follows again from Lemma 2,

$$\begin{aligned}
&(L_n(\theta) + \lambda\|\theta\|_1) - (L_n(\theta') + \lambda\|\theta'\|_1) \\
=&\tfrac{1}{2}(t - \tilde{\theta}_{j,n})^2 + \epsilon(\theta) + \lambda|t| - \tfrac{1}{2}\tilde{\theta}^2_{j,n} - \epsilon(\theta') \\
\geq&\tfrac{1}{2}t^2 + |t| \cdot (-(\lambda - \gamma) + \lambda - \gamma_n) \\
=&\tfrac{1}{2}t^2 + |t| \cdot (\gamma - \gamma_n) \\
\geq&0,
\end{aligned}$$

with equality only for $t = 0$. It follows that the minimizer of $L_n(\theta) + \lambda\|\theta\|_1$ necessarily has jth component equal to zero. This takes care of the inactive coordinates.

Third sum To control the third sum, we show that $C_n \to^p 0$. Since $\sum_{i=1}^n \|D_n^i\|^2 = O_p(1)$ (which follows from $\|D_n^i\| \leq \|\hat{\theta}_n^{\lambda,-i} - \hat{\theta}_n^\lambda\| + \|\tilde{\theta}_n^{\lambda,-i} - \tilde{\theta}_n^\lambda\|$ and the bounds (4)–(5), by the same argument as for (13)), this implies $C_n \cdot \sum_i \|D_n^i\|^2 = o_p(1)$.

Recall that $C_n = 1$ requires: (a) $sign(\widehat{\theta}_n^\lambda) = \eta$ and g^λ is linear on $B_{\rho_n}(\tilde{\theta}_n)$ (the negation of B_n's condition), but (b) there exists some i such that $sign(\widehat{\theta}_n^{\lambda,-i}) \neq \eta$ or $sign(\tilde{\theta}_n^{\lambda,-i}) \neq \eta$. We show that on the event described in (a), neither type of sign disagreement occurs, with probability approaching 1.

Preliminary: rate for the influence-function approximation error. We claim

$$\sup_\lambda \|\widehat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\| = O_{\mu_n}(n^{-1/2}) = o_{\mu_n}(\rho_n), \tag{21}$$

where the second equality uses $\rho_n n^{1/2} = (\rho_n n^{1/4}) \cdot n^{1/4} \to \infty$. By Lemma 2, $\widehat{\theta}_n^\lambda$ and $\tilde{\theta}_n^\lambda$ are respectively the minimizers of $\frac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \epsilon_n(\theta) + \lambda\pi(\theta)$ and $\frac{1}{2}\|\theta - \tilde{\theta}_n\|^2 + \lambda\pi(\theta)$. By Lemma 1 in Wilson et al. (2020) applied to the perturbation ϵ_n,

$$\sup_\lambda \|\widehat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\| \;\leq\; \frac{1}{\mu} \sup_{\|\theta\| \leq C} \|\nabla \epsilon_n(\theta)\|. \tag{22}$$

Since $\nabla L_n(\theta_0) = \theta_0 - \tilde{\theta}_n$ exactly (by the definition $\tilde{\theta}_n = \theta_0 - \sum_i \nabla_\theta l_n(\theta_0, Z_n^i)$ in Lemma 2) and $\nabla \epsilon_n(\theta) = \nabla L_n(\theta) - (\theta - \tilde{\theta}_n)$, we have $\nabla \epsilon_n(\theta_0) = 0$ exactly. Therefore, for $\|\theta\| \leq C$,

$$\|\nabla \epsilon_n(\theta)\| \;=\; \|\nabla \epsilon_n(\theta) - \nabla \epsilon_n(\theta_0)\| \;\leq\; C \cdot \sup_{\|\theta'\| \leq C} \|\nabla^2 \epsilon_n(\theta')\|.$$

Now $\nabla^2 \epsilon_n(\theta) = \nabla^2 L_n(\theta) - I$. The CLT applied to $n\nabla_\theta^2 l_n(\theta_0, Z_n^i)$ (finite second moments from Assumption 5(iii)) gives $\|\nabla^2 L_n(\theta_0) - I\| = O_{\mu_n}(n^{-1/2})$, using that $\|E_{\mu_n}[n\nabla_\theta^2 l_n(\theta_0, Z_n^i)] - I\| = O(n^{-1/2})$ by Assumption 2 and the Lipschitz condition in Assumption 5(iii). The same Lipschitz condition and the functional CLT extend this rate uniformly over bounded neighborhoods:

$$\sup_{\|\theta\| \leq C} \|\nabla \epsilon_n(\theta)\| \;=\; O_{\mu_n}(n^{-1/2}) \;=\; o_{\mu_n}(\rho_n). \tag{23}$$

Substituting into (22) establishes (21).

Sign agreement for $\tilde{\theta}_n^{\lambda,-i}$. Since $\tilde{\theta}_n^{-i} - \tilde{\theta}_n = -\frac{1}{\sqrt{n}} X_n^i$ where $X_n^i = -\nabla_\beta l(\theta_0/\sqrt{n}, Z_n^i)$, the event $\|\tilde{\theta}_n^{-i} - \tilde{\theta}_n\| > \rho_n$ for some i has probability bounded by

$$\sum_{i=1}^n P\left(\frac{1}{\sqrt{n}} \|X_n^i\| > \rho_n\right) \leq n \cdot \frac{E_{\mu_n}[\|X_n^i\|^4]}{(\rho_n \sqrt{n})^4} = \frac{M}{(\rho_n n^{1/4})^4} \to 0,$$

by the 4th-moment bound in Assumption 5(ii) and $\rho_n \cdot n^{1/4} \to \infty$. On the complement of this event, $\tilde{\theta}_n^{-i} \in B_{\rho_n}(\tilde{\theta}_n)$ for every i, so g^λ is linear on a neighborhood of each $\tilde{\theta}_n^{-i}$, and therefore $sign(\tilde{\theta}_n^{\lambda,-i}) = \eta$ for all i.

Sign agreement for $\widehat{\theta}_n^{\lambda,-i}$*: active coordinates.* On the event $C_n = 1$, g^λ is linear on $B_{\rho_n}(\tilde{\theta}_n)$. For $j \in J$, this forces $|\tilde{\theta}_{n,j}| > \lambda + \rho_n$, hence $|\tilde{\theta}_{n,j}^\lambda| = |\tilde{\theta}_{n,j}| - \lambda > \rho_n$. By (21), $\sup_\lambda \|\widehat{\theta}_n^\lambda - \tilde{\theta}_n^\lambda\| = o_{\mu_n}(\rho_n)$, so with probability approaching 1,

$$|\widehat{\theta}_{n,j}^\lambda| \;\geq\; \rho_n/2 \qquad \text{for all } j \in J. \tag{24}$$

It remains to show $|\widehat{\theta}^{\lambda,-i}_{n,j} - \widehat{\theta}^{\lambda}_{n,j}| < \rho_n/4$ for all $j \in J$ and all i simultaneously. By (4), $\sup_\lambda \|\widehat{\theta}^{\lambda,-i}_n - \widehat{\theta}^{\lambda}_n\| \leq \frac{1}{\mu}\|\nabla_\theta l_n(\widehat{\theta}_n, Z^i_n)\|$. The Lipschitz condition in Assumption 5(ii) gives

$$\|\nabla_\theta l_n(\widehat{\theta}_n, Z^i_n)\| \;\leq\; \frac{1}{\sqrt{n}}\|X^i_n\| + \frac{B_n(Z^i_n)}{\sqrt{n}}\|\widehat{\theta}_n - \theta_0\|.$$

A Chebyshev union bound on the first term (4th-moment bound in Assumption 5(ii)) and a Chebyshev union bound on the second term (2nd-moment bound $E_{\mu_n}[B^2_n] < \infty$ in Assumption 5(ii), together with $\widehat{\theta}_n - \theta_0 = O_{\mu_n}(1)$) give

$$P\left(\exists\, i :\ \sup_\lambda \|\widehat{\theta}^{\lambda,-i}_n - \widehat{\theta}^{\lambda}_n\| > \frac{\rho_n}{4}\right) \;\leq\; \frac{C_1 M}{(\rho_n n^{1/4})^4} + \frac{C_2 E_{\mu_n}[B^2_n]}{n\rho^2_n} \to 0, \tag{25}$$

for absolute constants C_1, C_2, using $(\rho_n n^{1/4})^4 \to \infty$ and $n\rho^2_n \geq (\rho_n n^{1/4})^2 \cdot n^{1/2} \to \infty$. On the complement of (25) and given (24), we have $|\widehat{\theta}^{\lambda,-i}_{n,j}| \geq \rho_n/4 > 0$ with sign η_j, for all $j \in J$ and all i.

Sign agreement for $\widehat{\theta}^{\lambda,-i}_n$: inactive coordinates. It remains to show that $\widehat{\theta}^{\lambda,-i}_{n,j} = 0$ for $j \in J^c$ and all i, with probability approaching 1. The argument parallels the one used for the second sum (showing $sign(\widehat{\theta}^{\lambda}_n) = \eta$), now applied to L^{-i}_n and $\widehat{\theta}^{\lambda,-i}_n$.

By Lemma 2,

$$L^{-i}_n(\theta) = \tfrac{1}{2}\|\theta - \tilde{\theta}^{-i}_n\|^2 + \epsilon^{(i)}_n(\theta),$$

where

$$\begin{aligned}\epsilon^{(i)}_n(\theta) = \epsilon_n(\theta) - \big[&l_n(\theta, Z^i_n) - l_n(\theta_0, Z^i_n) - \langle \nabla_\theta l_n(\theta_0, Z^i_n),\, \theta - \theta_0\rangle \\ &+ \tfrac{1}{2n}(\theta - \theta_0)^\intercal \left\{n\nabla^2_\theta l_n(\theta_0, Z^i_n)\right\}(\theta - \theta_0)\big].\end{aligned}$$

Consider a candidate minimizer θ of $L^{-i}_n(\theta) + \lambda\|\theta\|_1$ with $\theta_j = t \neq 0$ for some $j \in J^c$, and let θ' agree with θ except $\theta'_j = 0$. On the event that g^λ is linear on $B_{\rho_n}(\tilde{\theta}_n)$, the KKT conditions imply $|\tilde{\theta}_{n,j}| + \rho_n < \lambda$ for $j \in J^c$. Therefore, by the same calculation as in the second sum,

$$\big(L_n^{-i}(\theta) + \lambda|t|\big) - L_n^{-i}(\theta') \geq \tfrac{1}{2}t^2 + |t|\left(\lambda - |\tilde{\theta}_{n,j}^{-i}| - \sup_{\|\theta\|\leq C} \|\nabla \epsilon_n^{(i)}(\theta)\|\right).$$

Since $|\tilde{\theta}_{n,j}^{-i}| \leq |\tilde{\theta}_{n,j}| + n^{-1/2}\|X_n^i\|$ and the slack is $\lambda - |\tilde{\theta}_{n,j}| > \rho_n$ on the event under consideration, it suffices to show simultaneously for all i:

$$n^{-1/2}\|X_n^i\| \; + \; \sup_{\|\theta\|\leq C} \|\nabla \epsilon_n^{(i)}(\theta)\| \; < \; \rho_n. \tag{26}$$

Write $\nabla \epsilon_n^{(i)}(\theta) = \nabla \epsilon_n(\theta) - R_3(\theta, Z_n^i)$, where

$$R_3(\theta, Z_n^i) := \nabla_\theta l_n(\theta, Z_n^i) - \nabla_\theta l_n(\theta_0, Z_n^i) - \tfrac{1}{n}\big\{n\nabla_\theta^2 l_n(\theta_0, Z_n^i)\big\}(\theta - \theta_0)$$

is the second-order Taylor remainder of $\nabla_\theta l_n(\cdot, Z_n^i)$ around θ_0. We bound the three contributions to (26) in turn.

Control of $\nabla \epsilon_n$. By (23), $\sup_{\|\theta\|\leq C}\|\nabla \epsilon_n(\theta)\| = O_{\mu_n}(n^{-1/2}) = o_{\mu_n}(\rho_n)$; this term is the same for all i.

Simultaneous control of $R_3(\cdot, Z_n^i)$. By the Lipschitz condition in Assumption 5(iii), $\|R_3(\theta, Z_n^i)\| \leq \frac{C_n(Z_n^i)}{2n}\|\theta - \theta_0\|^2$, so $\sup_{\|\theta\|\leq C}\|R_3(\theta, Z_n^i)\| \leq \frac{C_n(Z_n^i)C^2}{2n}$. A Chebyshev union bound gives

$$P\left(\exists\, i : \sup_{\|\theta\|\leq C}\|R_3(\theta, Z_n^i)\| > \frac{\rho_n}{3}\right) \; \leq \; n{\cdot}P\left(C_n(Z_n^i) > \frac{2n\rho_n}{3C^2}\right) \; \leq \; \frac{9C^4 \sup_n E_{\mu_n}[C_n(Z_n^i)^2]}{4\,n\rho_n^2} \to 0, \tag{27}$$

since $n\rho_n^2 \geq (\rho_n n^{1/4})^2 \cdot n^{1/2} \to \infty$ and $\sup_n E_{\mu_n}[C_n(Z_n^i)^2] < \infty$ by Assumption 5(iii).

Simultaneous control of $n^{-1/2}\|X_n^i\|$. By the same Chebyshev argument as for the first sign-agreement step,

$$P\Big(\exists\, i : \; n^{-1/2}\|X_n^i\| > \frac{\rho_n}{3}\Big) \; \leq \; \frac{81\, M}{(\rho_n n^{1/4})^4} \to 0. \tag{28}$$

On the complement of the three events above, (26) holds for all i: $n^{-1/2}\|X_n^i\| +$

$\sup_{\|\theta\|\leq C} \|\nabla \epsilon_n^{(i)}(\theta)\| \leq \rho_n/3 + o_{\mu_n}(\rho_n) + \rho_n/3 < \rho_n$ for n large enough. The slack is therefore strictly positive, so $\widehat{\theta}_{n,j}^{\lambda,-i} = 0$ for all $j \in J^c$ and all i.

Conclusion. Combining the three cases above, we conclude that on the event described in condition (a) (which has probability approaching 1 given $B_n \to^p 0$), all signs agree: $sign(\widehat{\theta}_n^{\lambda,-i}) = sign(\tilde{\theta}_n^{\lambda,-i}) = \eta$ for every i. This contradicts condition (b), so $C_n \to^p 0$. Since $\sum_i \|D_n^i\|^2 = O_p(1)$, we obtain $C_n \cdot \sum_i \|D_n^i\|^2 = o_p(1)$.

Combining the bounds for all three sums, we conclude $DD_n = o_{\mu_n}(1)$, which completes the proof of Lemma 4. □

C Characterizing SURE for Ridge

To prove Lemma 5 for the Ridge penalty $\frac{1}{2}\theta \cdot A^{-1} \cdot \theta$, we start by deriving a series of properties of the function $SURE(\lambda, \hat{\theta}, \Sigma) = \text{trace}(\Sigma) + \|g^\lambda\|^2 + 2\,\text{trace}\left(\nabla g^\lambda \cdot \Sigma\right)$. The properties derived in this section are purely analytic, not probabilistic, and concern the behavior of $SURE$ as a function, not any asymptotic limits.

Recall that $g^\lambda(\theta) = \text{argmin}_{\,g}\, \frac{1}{2}\|g\|^2 + \lambda \cdot \pi(\theta + g)$, and that g^λ satisfies the first-order condition,

$$g^\lambda(\theta) = -\lambda \cdot \nabla\pi(\theta + g^\lambda(\theta)),$$

for a suitable sub-gradient $\nabla\pi$ of π. Ridge corresponds to penalties of the form $\pi(\theta) = \frac{1}{2}\theta \cdot A^{-1} \cdot \theta$, where A is positive definite. Denote $C_\lambda = -(\frac{1}{\lambda}A + I)^{-1}$. The first order condition for $g^\lambda(\theta)$ then implies

$$g^\lambda(\theta) = C_\lambda \cdot \theta, \qquad\qquad \nabla g^\lambda(\theta) = C_\lambda.$$

and thus

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

The following change of coordinates will be convenient for some of our arguments. Denote $R_n = \|\hat{\theta}_n\|$ and $\nu_n = \hat{\theta}_n / R_n$, and similarly for R and ν. In a slight abuse of notation, we shall write

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

The following lemma characterizes the behavior of $SURE$ for Ridge. Property 1 and supermodularity are derived directly from the expression for $SURE$. Properties 2a, 2b, and 3 are then consequences of supermodularity.

Lemma 6 (Properties of $SURE$ for Ridge)**.** *Suppose that $\pi(\theta) = \frac{1}{2}\theta \cdot A^{-1} \cdot \theta$, where A is positive definite. Then the following holds:*

1. *For every point θ, the function $SURE(\lambda, \theta)$ satisfies*

$$\sup_{\lambda \in \Lambda} |SURE(\lambda, \theta') - SURE(\lambda, \theta)| \to 0$$

   as $\theta' \to \theta$.

2. *The function $SURE(\lambda, R, \nu)$ is strictly supermodular in λ and R. This implies:*

   (a) $\lambda(R, \nu) = \text{argmin}_{\lambda \in \mathbb{R}^+} SURE(\lambda, R, \nu)$ is monotonically decreasing in R, given ν.

   (b) $\lambda(R, \nu)$ 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 $\epsilon > 0$,*

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

Proof of Lemma 6 (Properties of $SURE$ for Ridge):

1. We have

$$\begin{aligned}
&SURE(\lambda,\theta') - SURE(\lambda,\theta)\\
&= \|C_\lambda \cdot \theta'\|^2 - \|C_\lambda \cdot \theta\|^2\\
&\leq \|C_\lambda\|^2 \cdot \|\theta' - \theta\| \cdot \|\theta' + \theta\|\\
&\leq \|\theta' - \theta\| \cdot \|\theta' + \theta\|,
\end{aligned}$$

   where in the third line we have used Cauchy-Schwartz ($\|\theta'\|^2 - \|\theta\|^2 = \langle \theta' - \theta, \theta' + \theta \rangle \leq \|\theta' - \theta\| \cdot \|\theta' + \theta\|$), and in the last line we have used that A is positive definite, which implies $\|C_\lambda\| \leq 1$. The claim follows.

2. The first and last terms in the expression for $SURE$ do not depend on θ. The middle term can be written as $R^2 \cdot \|C_\lambda \cdot \nu\|^2$, and thus

$$\frac{\partial^2}{\partial\lambda\partial R} SURE(\lambda, R, \nu) = 2R \cdot \frac{\partial}{\partial\lambda} \|C_\lambda \cdot \nu\|^2 > 0,$$

   using again the positive definiteness of A. This implies that $SURE(\lambda, R, \nu)$ is strictly supermodular in λ and R, i.e., whenever $\epsilon > 0$ and $\delta > 0$

$$\begin{aligned}
&[SURE(\lambda+\epsilon, R, \nu) - SURE(\lambda, R, \nu)]\\
&\quad - [SURE(\lambda+\epsilon, R-\delta, \nu) - SURE(\lambda, R-\delta, \nu)] > 0.
\end{aligned}$$

   (a) Monotonicity of $\lambda(R,\nu)$ in R follows from supermodularity of $SURE$, by Topkis's theorem.

   (b) That $\lambda(R,\nu)$ is continuous in R almost everywhere holds because monotonic functions are continuous almost everywhere: The set of discontinuity points is at most countable, by Theorem 4.30 of Rudin (1991).

3. For the given R, ν, let δ be such that $|\lambda(R',\nu) - \lambda(R,\nu)| < \epsilon/2$ whenever $|R' - R| \leq \delta$; such a δ exists by continuity.

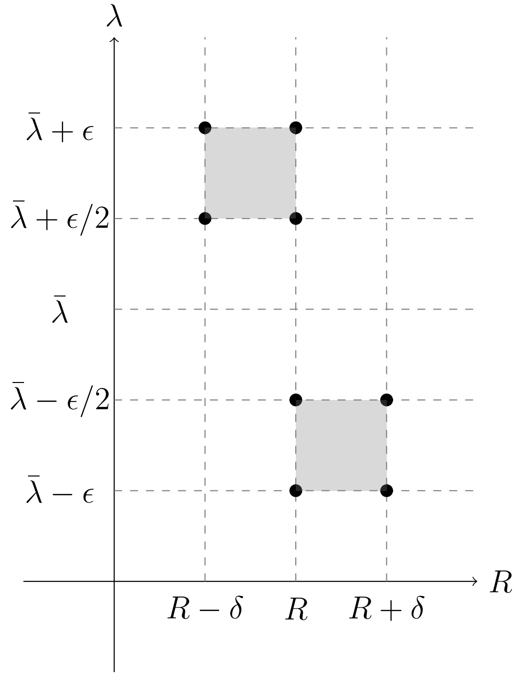
Define

$$A(\lambda_1, \lambda_2, R) = SURE(\lambda_1, R, \nu) - SURE(\lambda_2, R, \nu),$$

let $\bar{\lambda} = \lambda(R, \nu)$ and

$$\begin{aligned}\overline{\Delta} = \min \Big(& A(\bar{\lambda} + \epsilon, \bar{\lambda} + \epsilon/2, R) - A(\bar{\lambda} + \epsilon, \bar{\lambda} + \epsilon/2, R - \delta), \\ & A(\bar{\lambda} - \epsilon/2, \bar{\lambda} - \epsilon, R + \delta) - A(\bar{\lambda} - \epsilon/2, \bar{\lambda} - \epsilon, R)\Big).\end{aligned}$$

By strict supermodularity, both of the "double differences" in this definition are positive, and thus $\overline{\Delta} > 0$. The following figure illustrates the definition of $\overline{\Delta}$. The differences defining A are taken over vertical segments for different λ and fixed R. The double differences defining Δ are taken over of the grey rectangles in the figure:


λ
$\bar{\lambda} + \epsilon$
$\bar{\lambda} + \epsilon/2$
$\bar{\lambda}$
$\bar{\lambda} - \epsilon/2$
$\bar{\lambda} - \epsilon$
R
$R - \delta$
R
$R + \delta$

We claim that

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

The following argument will correspond to the "top left" rectangle in the figure; the "bottom right" case is symmetric. Fix $\tilde{\lambda} > \bar{\lambda} + \epsilon$.

- Given our assumptions and definitions, $\tilde{\lambda} > \bar{\lambda} + \epsilon > \bar{\lambda} + \epsilon/2 > \lambda(R-\delta) \geq \bar{\lambda}$. Therefore, by supermodularity

  $$\begin{aligned} &A(\tilde{\lambda}, \lambda(R-\delta), R) - A(\tilde{\lambda}, \lambda(R-\delta), R-\delta) \\ &\geq A(\bar{\lambda}+\epsilon, \bar{\lambda}+\epsilon/2, R) - A(\bar{\lambda}+\epsilon, \bar{\lambda}+\epsilon/2, R-\delta). \end{aligned}$$

- By definition of $\overline{\Delta}$,

  $$A(\bar{\lambda}+\epsilon, \bar{\lambda}+\epsilon/2, R) - A(\bar{\lambda}+\epsilon, \bar{\lambda}+\epsilon/2, R-\delta) \geq \overline{\Delta}.$$

- By optimality of $\lambda(R-\delta)$ for $R-\delta$,

  $$A(\tilde{\lambda}, \lambda(R-\delta), R-\delta) \geq 0.$$

- Combining the preceding three items, we get

  $$A(\tilde{\lambda}, \lambda(R-\delta), R) \geq \overline{\Delta}.$$

  and thus

  $$SURE(\tilde{\lambda}, R, \nu) \geq SURE(\lambda(R-\delta), R, \nu) + \overline{\Delta} \geq SURE(\bar{\lambda}, R, \nu) + \overline{\Delta}.$$

  The argument for $\tilde{\lambda} < \bar{\lambda} - \epsilon$ is analogous, and the claim follows.

□

D Characterizing SURE for Lasso

Lasso corresponds to penalties of the form $\pi(\theta) = \|A^{-1} \cdot \theta\|_1$, where A is an invertible matrix, and $\|\cdot\|_1$ is the L_1 norm.[8] Given λ, we can characterize $g^\lambda(\theta)$ as follows. Denote $h^\lambda(\theta) = A^{-1}(\theta + g^\lambda(\theta))$. The optimal $h^\lambda(\theta)$ solves

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

The solution to this convex optimization problem is of the form

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

where $\eta_j = sign(h_j^\lambda)$, $J = \{j : \eta_j \neq 0\}$, and A_J is the subset of columns corresponding to the index set J. These conditions follow immediately from the first order conditions for $h^\lambda(\theta)$.

Lemma 7 (Properties of $SURE$ for Lasso)**.** *Suppose that $\pi(\theta) = \|A^{-1} \cdot \theta\|_1$, where A is an invertible matrix, and $\|\cdot\|_1$ is the L_1 norm. Let $k = \dim(\theta)$. Then the following holds:*

1. *As a function of λ, for every R, ν, the graph of $SURE(\lambda, R, \nu)$ consists of at most 3^k continuous segments indexed by $\eta \in \{-1, 0, 1\}^k$. On each of these segments η and $J = \{j : \eta_j \neq 0\}$ are constant,*

$$\nabla g^\lambda = A_J \cdot (A_J' A_J)^{-1} \cdot A_J' - I,$$
$$\|g^\lambda\|^2 = \|\nabla g^\lambda \cdot \theta\|^2 + \lambda^2 \cdot \eta_J'(A_J' A_J)^{-1}\eta_J,$$

and $SURE(\lambda, R, \nu)$ is a monotonically increasing quadratic polynomial in λ of the form

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

[8] For Lasso, it is *not* without loss of generality to assume that A is diagonal.

2. *SURE for Lasso scales with R as follows:*

$$SURE(R \cdot \lambda, R, \nu) = \text{trace}(\Sigma) + R^2 \cdot \|g^\lambda(\nu)\|^2 + 2\,\text{trace}\left(\nabla g^\lambda(\nu) \cdot \Sigma\right).$$

   Given ν, let $\lambda_1, \lambda_2, \ldots, \lambda_m$ ($m \leq 3^k$) be the local minimizers of $SURE(\lambda, 1, \nu)$, corresponding to values of λ where η changes. The local minimizers of $SURE(\lambda, R, \nu)$ are then given by $R \cdot \lambda_1, R \cdot \lambda_2, \ldots, R \cdot \lambda_m$.

3. *Let $\lambda(R, \nu) = \text{argmin}_{\lambda \in \mathbb{R}^+} SURE(\lambda, R, \nu)$. Then, given ν, $\lambda(R, \nu) = R \cdot \lambda_{j(R)}$, where $j(R)$ is a monotonically decreasing mapping from R to $1, 2, \ldots, m$, and $\lambda_1, \lambda_2, \ldots, \lambda_m$ are as before. The graph $\lambda^*(R, \nu)$ thus follows a piecewise linear "sawtooth" pattern with at most 3^k jumps.*

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 $\epsilon > 0$,*

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

Proof of Lemma 7 (Properties of SURE for Lasso):

1. Consider two values λ_1, λ_2 of λ such that η is the same for these two values. It follows from the optimality conditions for h^λ that for any intermediate value of λ between λ_1, λ_2, the optimal h^λ is a linear interpolation between h^{λ_1} and h^{λ_2}, and in particular η remains the same in between λ_1, λ_2. For details, see Mairal and Yu (2012), Lemma 2.

   The vector $\eta \in \{-1, 0, 1\}^k$ can take 3^k possible values. The gradient of g^λ with respect to θ is a function of J, which is a function of η, but it does not depend on λ otherwise:

$$\nabla g^\lambda = A \cdot \nabla h^\lambda - I = A_J \cdot (A_J' A_J)^{-1} \cdot A_J' - I.$$

   It follows that the penalty term $2\,\text{trace}\left(\nabla g^\lambda \cdot \Sigma\right)$ in the expression for $SURE$ has at most $3^k - 1$ jumps, as a function of λ for fixed θ, and is constant in between these jumps.[9]

Consider now the term $\|g^\lambda\|^2$, which is the other term in the expression for $SURE$. Fixing λ, and the corresponding set J of active coordinates, we get that $\|g^\lambda\|^2 = \|A_J h_J^\lambda - \theta\|^2$, which is minimized at $\lambda = 0$, holding J fixed. For $\lambda = 0$ we get

$$\|A_J h_J^0 - \theta\|^2 = \|A_J((A_J' A_J)^{-1} A_J' - I) \cdot \theta\|^2 = \|\nabla g^\lambda \cdot \theta\|^2.$$

This is the sum of squared errors for an OLS regression of the elements of θ on the rows of A_J.

Given J, $\|g^\lambda\|^2$ is a quadratic function of λ with second derivative $\partial_\lambda^2 \|A_J h_J^\lambda - \theta\|^2 = 2 \cdot \eta_J'(A_J' A_J)^{-1} \eta_J$. It follows that

$$\|g^\lambda\|^2 = \|\nabla g^\lambda \cdot \theta\|^2 + \lambda^2 \cdot \eta_J'(A_J' A_J)^{-1} \eta_J.$$

This last result implies that $\|g^\lambda\|^2$ is monotonically increasing in λ on each segment defined by η. Since g^λ is continuous in λ (cf. Lemma 2 in Mairal and Yu 2012), this also implies that $\|g^\lambda\|^2$ is monotonically increasing across $\lambda \in \mathbb{R}^+$; we will use this fact below.

2. Multiplying the objective of the optimization problem $h^\lambda(\theta) = \text{argmin}_h \frac{1}{2}\|A \cdot h - \theta\|^2 + \lambda \cdot \|h\|_1$ by a factor $1/R^2$ yields

$$h^\lambda(R \cdot \nu) = \underset{h}{\text{argmin}} \ \tfrac{1}{2}\|A \cdot (h/R) - \nu\|^2 + \lambda/R \cdot \|h/R\|_1 = R \cdot h^{\lambda/R}(\nu),$$

and thus also

$$\begin{aligned} g^\lambda(R \cdot \nu) &= R \cdot g^{\lambda/R}(\nu), \text{ and} \\ \nabla g^\lambda(R \cdot \nu) &= \tfrac{1}{R} \cdot \partial_\nu g^\lambda(R \cdot \nu) = \nabla g^{\lambda/R}(\nu). \end{aligned}$$

[9] This bound can be refined, cf. Mairal and Yu (2012), but it is enough for our purposes.

which immediately implies

$$SURE(R\cdot\lambda, R, \nu) = \text{trace}(\Sigma) + R^2 \cdot \|g^\lambda(\nu)\|^2 + 2\,\text{trace}\left(\nabla g^\lambda(\nu)\cdot\Sigma\right).$$

Turning to the characterization of local minima, since $\|g^\lambda(\nu)\|^2$ is monotonically increasing in λ (cf. item 1), the local minima of $SURE(R\cdot\lambda, R, \nu)$ are exactly the values of λ where $2\,\text{trace}\left(\nabla g^\lambda(\nu)\cdot\Sigma\right)$ jumps down. These values are independent of R, and the claim follows.

3. That $\lambda(R,\nu) = R\cdot\lambda_{j(R)}$ follows immediately from the preceding item. It remains to show that $j(R)$ is monotonically decreasing. To see this, consider any pair of values $j > j'$. Then $\|g^{\lambda_j}(\nu)\|^2 > \|g^{\lambda_{j'}}(\nu)\|^2$ by monotonicity of $\|g^\lambda(\nu)\|^2$ in λ (cf. item 1), and we get that

$$SURE(R\cdot\lambda_j, R, \nu) - SURE(R\cdot\lambda_{j'}, R, \nu) = R^2\cdot\left(\|g^{\lambda_j}(\nu)\|^2 - \|g^{\lambda_{j'}}(\nu)\|^2\right)$$

is increasing in R. The claim follows.

4. By the preceding argument, at a point of continuity in R

$$\inf_{\lambda\in\{R\cdot\lambda_1, R\cdot\lambda_2,\ldots,R\cdot\lambda_m\}\backslash\bar\lambda} SURE(\lambda, R, \nu) - SURE(\bar\lambda, R, \nu) > 0.$$

The same holds for λ to the right of any of the local minimizers $\{R\cdot\lambda_1, R\cdot\lambda_2, \ldots, R\cdot\lambda_m\}\backslash\bar\lambda$, since $SURE$ is monotonically increasing in λ away from the local minimizers.

It only remains to verify the condition for λ immediately to the right of the global minimizer $\bar\lambda$. This holds because $\|g^\lambda\|^2 = const. + \lambda^2\cdot\eta_J'(A_J'A_J)^{-1}\eta_J$ is strictly monotonically increasing in λ for $\lambda > 0$ and $\eta \neq 0$.

□

Lemma 8 (Local linearity of g^λ)**.**

1. *For all $\gamma > 0$ there exists an $\epsilon > 0$ such that $g^\lambda(\theta)$ is linear on $B_\epsilon(\hat\theta) = \{\theta : \|\theta - \hat\theta\| < \epsilon\}$ with probability greater than $1-\gamma$, where $\hat\theta \sim N(\theta_0, \Sigma)$.*

2. For almost every point θ, the function $SURE(\lambda, \theta)$ satisfies

$$\sup_{\lambda \in \Lambda} |SURE(\lambda, \theta') - SURE(\lambda, \theta)| \to 0$$

as $\theta' \to \theta$.

Proof of Lemma 8: To show the first claim, denote

$$\Theta_\eta = \{\theta : \ sign(h^\lambda(\theta)) = \eta\}.$$

By the KKT conditions for h^λ, the set Θ_η is convex for every $\eta \in \{-1, 0, 1\}^k$, and its boundary $\partial\Theta_\eta$ is a finite union of subsets of hyperplanes. Furthermore $\mathbb{R}^k = \bigcup_\eta \Theta_\eta$, and $h^\lambda(\theta)$ is linear in θ on each of the sets Θ_η.

The claim of the lemma therefore follows if we can show that the probability of an ϵ-band around the boundary $\partial\Theta_\eta$,

$$\partial\Theta_\eta^\epsilon = \{\theta : \ d(\theta, \partial\Theta_\eta) < \epsilon\},$$

where d is the Euclidean distance, has vanishing measure for small ϵ. Because $\partial\Theta_\eta$ is a subset of a finite union of hyperplanes, $P(\hat{\theta} \in \partial\Theta_\eta) = 0$. By the properties of probability measures, since

$$\partial\Theta_\eta = \bigcap_{\epsilon > 0} \partial\Theta_\eta^\epsilon$$

we get

$$0 = P(\hat{\theta} \in \partial\Theta_\eta) = \lim_{\epsilon \to 0} P(\hat{\theta} \in \partial\Theta_\eta^\epsilon).$$

Therefore, for ϵ small enough, $\hat{\theta}$ is more than ϵ away from the boundary of any of the sets Θ_η with probability bigger than $1 - \gamma$, and the claim follows.

We now turn to the second claim. Fix $\lambda \in \Lambda$. By the preceding argument, for almost all points θ, θ is in the interior of Θ_η, for some η. Thus $\theta' \in \Theta_\eta$, as well, for $\|\theta' - \theta\|$ small enough. By the characterization of $SURE$ for Lasso in item 1 of Lemma 7, given λ and η, ∇g^λ is constant and $SURE = const. +$

$\|g^\lambda\|^2 = const. + \|\nabla g^\lambda \cdot \theta\|^2$. This is continuous in θ, and thus $|SURE(\lambda, \theta') - SURE(\lambda, \theta)| \to 0$ as $\theta' \to \theta$. Since almost everywhere continuity of $SURE$ in θ thus holds for fixed λ, it also holds simultaneously for any finite set of λ, and the claim follows. □

E Convergence of risk

The remainder of our proof will draw on some standard results, which we recall here, including the following results from van der Vaart (2000):

1. **Joint convergence (item (v) of Theorem 2.6)**
   If $W_n^1 \to^d W^1$, and $W_n^2 \to^p 0$, then $W_n \to^d W$,
   where $W_n = (W_n^1, W_n^2)$, $W = (W^1, 0)$.

2. **Almost everywhere CMT (item (ii) of Theorem 2.3):** Suppose that $W_n \to^d W$ (converges in distribution), and that $s(W)$ is almost everywhere continuous. Then $s(W_n) \to^d s(W)$.

3. **Uniform integrability and convergence of expectations (Theorem 2.20):** Suppose that $s(W_n) \to^d s(W)$, and that

   $$\lim_{M \to \infty} \limsup_{n \to \infty} E\left[|s(W_n)|\mathbf{1}(|s(W_n)| > M)\right] = 0. \tag{29}$$

   Then $E[s(W_n)] \to E[s(W)]$.

E.1 Convergence in distribution

Consider some arbitrary function $c(\lambda)$ that is minimized to choose the tuning parameter λ (in due time, we will substitute CV_n for this function). Define

$$\Delta(\lambda) = c(\lambda) - SURE(\lambda, \theta, \Sigma).$$

We think of Δ as an element of the space of bounded functions on $\Lambda \subset \mathbb{R}^+$, endowed with the sup norm. Define furthermore

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

We have proven for Ridge (in Lemma 6), and for Lasso (in Lemma 7), that the minimum of $SURE$ with respect to λ is well separated for almost all θ (with the exception of Lasso when λ is so large that $\hat{\theta}^\lambda = 0$). This implies the following lemma. The "min" in the definition of $\tilde{\lambda}$ serves as a tie-breaking rule in the case of non-uniqueness of the minimizer.

Lemma 9. *For almost every θ, the mapping from $w = (\theta, \Delta)$ to $g^{\tilde{\lambda}(\theta,\Delta)}$, where*

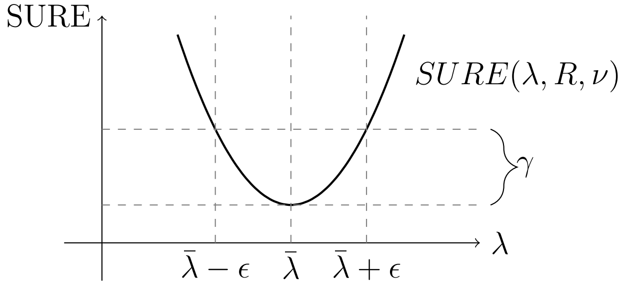
$$\tilde{\lambda}(\theta, \Delta) = \min \left(\operatorname*{argmin}_{\lambda \in \Lambda} \, [SURE(\lambda, \theta, \Sigma) + \Delta(\lambda)] \right),$$

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

Proof. We first prove the claim for Ridge, before discussing the necessary modifications for the argument to apply to Lasso. Fix ν. By item 2a of Lemma 6, for almost every R, $\lambda(R, \nu)$ is continuous in R. Fix such an R, and let $\theta = R\nu$. Fix $\epsilon > 0$ and let

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

where $\bar{\lambda} = \tilde{\lambda}(\theta, 0) \in \operatorname{argmin}_\lambda SURE(\lambda, R, \nu)$. The following figure illustrates the definition of γ:



By item 3 of Lemma 6, $\gamma > 0$. By item 1 of Lemma 6, there exists a δ such that if $\|\theta' - \theta\| < \delta$ then $\sup_{\lambda \in \mathbb{R}^+} |SURE(\lambda, \theta') - SURE(\lambda, \theta)| < \gamma/4$.

Let $w' = (\theta', \Delta)$ and $w = (\theta, 0)$. Suppose that $\|w' - w\| < \min(\delta, \gamma/4)$, so that $\|\theta' - \theta\| < \delta$ and $\sup_\lambda |\Delta(\lambda)| < \gamma/4$. Denote $c(\lambda) = SURE(\lambda, \theta') + \Delta(\lambda)$, so that $\tilde{\lambda}(\theta', \Delta)$ is a minimizer of $c(\lambda)$. Then

$$\begin{aligned} &SURE(\tilde{\lambda}(\theta', \Delta), \theta) \\ &< SURE(\tilde{\lambda}(\theta', \Delta), \theta') + \tfrac{1}{4}\gamma \\ &< c(\tilde{\lambda}(\theta', \Delta)) + \tfrac{1}{2}\gamma \\ &\leq c(\bar{\lambda}) + \tfrac{1}{2}\gamma \\ &< SURE(\bar{\lambda}, \theta') + \tfrac{3}{4}\gamma \\ &< SURE(\bar{\lambda}, \theta) + \gamma. \end{aligned}$$

It follows that $|\tilde{\lambda}(\theta', \Delta) - \bar{\lambda}| < \epsilon$. This proves that $\tilde{\lambda}(\theta, \Delta)$ is continuous at $(\theta, 0)$ The claim for Ridge follows, since continuity of $g^\lambda(\theta) = (\frac{1}{\lambda} A + I)^{-1} \cdot \theta$ in both λ and θ is immediate.

Turning to Lasso, most of this argument holds verbatim, with the following modifications: Consider first values of θ such that $\hat{\theta}^{\tilde{\lambda}} \neq 0$.

1. By item 4 of Lemma 7, $\gamma > 0$ for almost all such θ.

2. By item 2 of Lemma 8, for *almost* all θ there exists a δ such that if $\|\theta' - \theta\| < \delta$ then $\sup_{\lambda \in \Lambda} |SURE(\lambda, \theta') - SURE(\lambda, \theta)| < \gamma/4$.

3. By the characterization of g^λ and h^λ given at the outset of Appendix D, $g^\lambda(\theta)$ is continuous in both λ and θ. The claim thus follows for θ such that $\hat{\theta}^{\tilde{\lambda}} \neq 0$.

Consider now values of θ such that $\hat{\theta}^{\tilde{\lambda}} = 0$. For such values, $SURE$ is flat in λ for values of λ greater than $\tilde{\lambda}$, because $\hat{\theta}^\lambda = 0$ and $\nabla g^\lambda = 0$ for all such λ (see Figure 2 for an example). Because $SURE$ is flat to the right, continuity of $\tilde{\lambda}(\theta, \Delta)$ does not necessarily hold at $(\theta, 0)$; small perturbations of Δ can lead to large changes of $\tilde{\lambda}$.

By the same arguments used to prove item 4 of Lemma 7, we obtain however (for almost all such θ) that

$$\inf_{\lambda < \lambda_m} SURE(\lambda, R, \nu) - SURE(\bar{\lambda}, R, \nu) > 0,$$

where λ_m is defined as in Lemma 7. This, in combination with item 2 of Lemma 8 (for almost all θ, $\sup_{\lambda \in \Lambda} |SURE(\lambda, \theta') - SURE(\lambda, \theta)| \to 0$ as $\theta' \to \theta$), implies that $\tilde{\lambda}(\theta', \Delta)$ is such that $g^{\tilde{\lambda}(\theta,0)}(\theta') = -\theta'$ for all (θ', Δ) in a neighborhood of $(\theta, 0)$, and the claim follows. □

E.2 Proof of convergence in distribution

We can now prove Lemma 5, drawing on our preceding Lemmas.

Proof of Lemma 5:

- By Lemma 2,
$$\hat{\theta}_n \to^d \hat{\theta} \sim N(\theta_0, \Sigma).$$

- Let $\Delta_n(\lambda) = CV_n(\lambda) - SURE(\lambda, \hat{\theta}_n, \Sigma)$. By Lemma 4,
$$\sup_{\lambda \in \Lambda} |\Delta_n(\lambda)| \to^p 0.$$

- By joint convergence (van der Vaart 2000, item (v) of Theorem 2.6), $W_n = (\hat{\theta}_n, \Delta_n) \to^d W = (\hat{\theta}, 0)$.

- By Lemma 9, the mapping from W_n to $\hat{\theta}_n + g^{\tilde{\lambda}(\hat{\theta}_n, \Delta_n)}(\hat{\theta}_n)$ is almost surely continuous on the support of W.

- By definition, $\tilde{\lambda}(\hat{\theta}_n, \Delta_n) = \lambda_n^* = \text{argmin}_{\lambda} CV_n(\lambda)$, and $\tilde{\lambda}(\hat{\theta}, 0) = \lambda^* = \text{argmin}_{\lambda} SURE(\lambda, \hat{\theta}, \Sigma)$.

- The almost surely continuous mapping theorem (van der Vaart 2000, Theorem 2.3) then implies

$$\hat{\theta}_n + g^{\lambda_n^*}(\hat{\theta}_n) \to_d \hat{\theta} + g^{\lambda^*}(\hat{\theta}) = \hat{\theta}^{\lambda^*}.$$

- By Lemma 2 $\hat{\theta}_n^* = \hat{\theta}_n + g^{\lambda_n^*}(\hat{\theta}_n) + o_p(1)$, and the claim follows.

□

E.3 Convergence of loss and risk

Proof of Theorem 1:
By Lemma 3,

$$\bar{L}_n(\theta, \theta_0) \to \tfrac{1}{2}\|\theta - \theta_0\|^2$$

uniformly in any bounded neighborhood of θ_0. By Lemma 5,

$$\hat{\theta}_n^* \to_d \hat{\theta}^*.$$

Combining these two results gives

$$\bar{L}_n(\hat{\theta}_n^*, \theta_0) \to_d \tfrac{1}{2}\|\hat{\theta}^* - \theta_0\|^2.$$

The distributional convergence claim of Theorem 1 follows. The claim of Corollary 1 is then immediate. □