Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning

Berkeley UNIVERSITY OF CALIFORNIA

Stanford University

Pratik Patil† and Daniel LeJeune§

†Department of Statistics, University of California, Berkeley

§Department of Statistics, Stanford University

Abstract

We study **sketched ridge regression ensembles** built from the general class of sketches **asymptotically free** from the data

- We precisely characterize **asymptotic risk**
- We prove that **generalized cross-validation (GCV)** provides consistent risk estimation for feature sketching ensembles
- We show that GCV also provides consistent **distribution** estimation enabling **prediction intervals**
- We employ an **ensemble trick** for efficiently estimating unsketched ridge regression risk

Freely sketched ridge ensembles

Given: data $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n \times p} \times \mathbb{R}$, feature sketches $\mathbf{S}_1, \ldots, \mathbf{S}_K \in \mathbb{R}^{p \times q}$, and the ensemble predictor at regularization level λ

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda = \frac{1}{K} \sum_{k=1}^{K} \tfrac{1}{n} \mathbf{S}_k \big(\tfrac{1}{n} \mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{S}_k + \lambda \mathbf{I}_q\big)^{-1} \mathbf{S}_k^\top \mathbf{X}^\top \mathbf{y}$$

where $\mathbf{S}_k \mathbf{S}_k^\top$ is **asymptotically free** from the data $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$

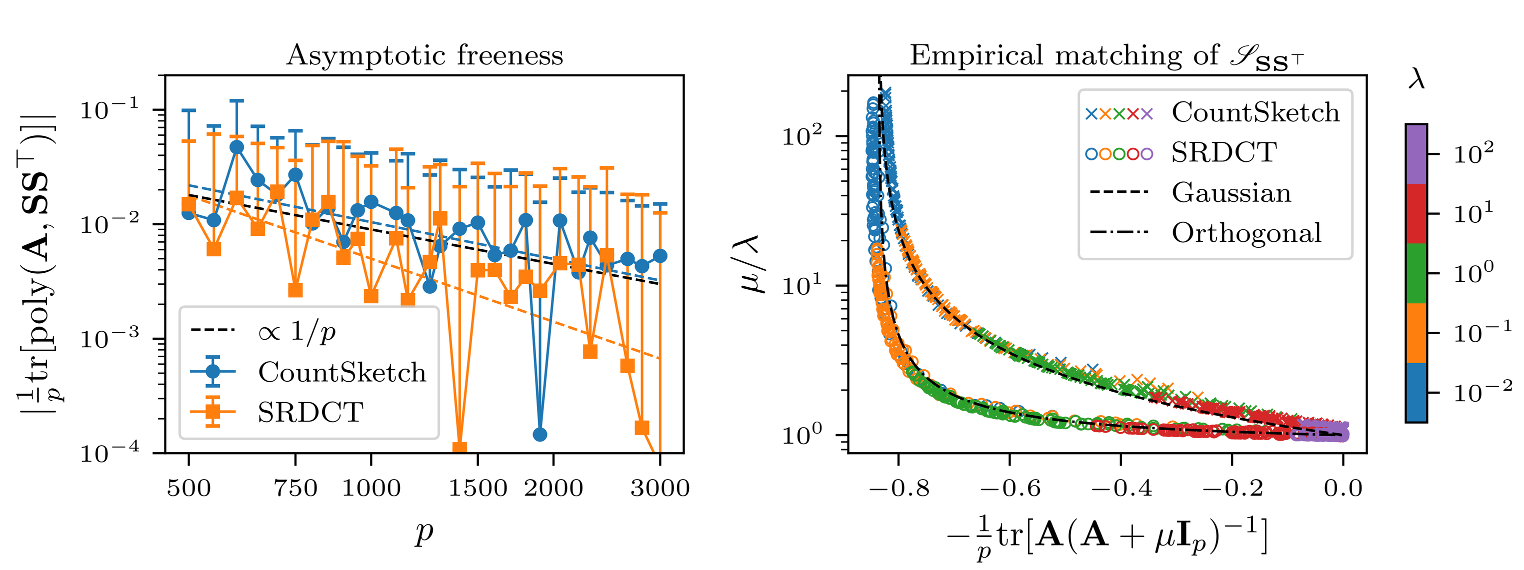
Two matrices $\mathbf{A}$ and $\mathbf{B}$ are almost surely **asymptotically free** if all mixed alternating products of centered polynomials are also centered:

$$\overline{\mathrm{tr}}[p_1(\mathbf{A}) p_2(\mathbf{B}) \ldots p_{L-1}(\mathbf{A}) p_L(\mathbf{B})] \xrightarrow{\text{a.s.}} 0.$$

Examples of known asymptotically free sketches:

- Independent: $[\mathbf{S}_k]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, zero mean, bounded moments
- Rotationally invariant: $\mathbf{S}_k = \mathbf{U}_k \mathbf{Q}_k$ with $\mathbf{U}_k$ Haar-distributed
- Randomized Fourier transform: $\mathbf{S}_k = \mathbf{D}_k \boldsymbol{\Phi}_{DFT} \widehat{\mathbf{S}}_k$

We provide **empirical support** for freeness for practical sketches.



Generalized cross-validation (GCV)

Goal: estimate the joint distribution of true labels and predictions $(y_0, x_0^\top \widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda)$ in order to estimate risk $T(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) = \mathbb{E}\left[t(y_0, x_0^\top \widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda)\right]$.

Ensemble predictions are **linear smoothers** $\mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda = \mathbf{L}^{\mathrm{ens}}_\lambda \mathbf{y}$ for

$$\mathbf{L}^{\mathrm{ens}}_\lambda = \frac{1}{K}\sum_{k=1}^{K} \mathbf{L}^k_\lambda = \tfrac{1}{n} \mathbf{X}\mathbf{S}_k (\tfrac{1}{n}\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X}\mathbf{S}_k + \lambda \mathbf{I}_q)^{-1} \mathbf{S}_k^\top \mathbf{X}^\top,$$

giving us the **GCV-corrected empirical distribution**

$$\widehat{P}^{\mathrm{ens}}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \delta\left\{\left(y_i, \frac{x_i^\top \widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda - \frac{1}{n}\mathrm{tr}[\mathbf{L}^{\mathrm{ens}}_\lambda] y_i}{1 - \frac{1}{n}\mathrm{tr}[\mathbf{L}^{\mathrm{ens}}_\lambda]}\right)\right\}.$$

We plug in $\widehat{P}^{\mathrm{ens}}_\lambda$ to obtain **risk estimators**

$$\widehat{T}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) = \int t(y, z) d\widehat{P}^{\mathrm{ens}}_\lambda(y, z) \quad \text{and} \quad \widehat{R}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) = \int (y - z)^2 d\widehat{P}^{\mathrm{ens}}_\lambda(y, z).$$

Sketched ensemble risk

Free sketches satisfy an **asymptotic equivalence**:

$$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q)^{-1} \mathbf{S}^\top \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1}$$

where $\mu \simeq \lambda \mathscr{S}_{\mathbf{SS}^\top}\big(-\frac{1}{p}\mathrm{tr}\,[\mathbf{S}^\top\mathbf{A}\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}]\big)$.

Theoretical results

Theorem 1. For any free sketches $\mathbf{S}_k$,

$$R(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) \simeq R(\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_\mu) + \frac{\mu'\Delta}{K} \quad \text{and} \quad \widehat{R}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) \simeq \widehat{R}(\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_\mu) + \frac{\mu''\Delta}{K},$$

where $\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_\mu = (\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I}_p)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{y}$.

Theorem 2. Under random data assumptions on $\mathbf{X}$ and $\mathbf{y}$,

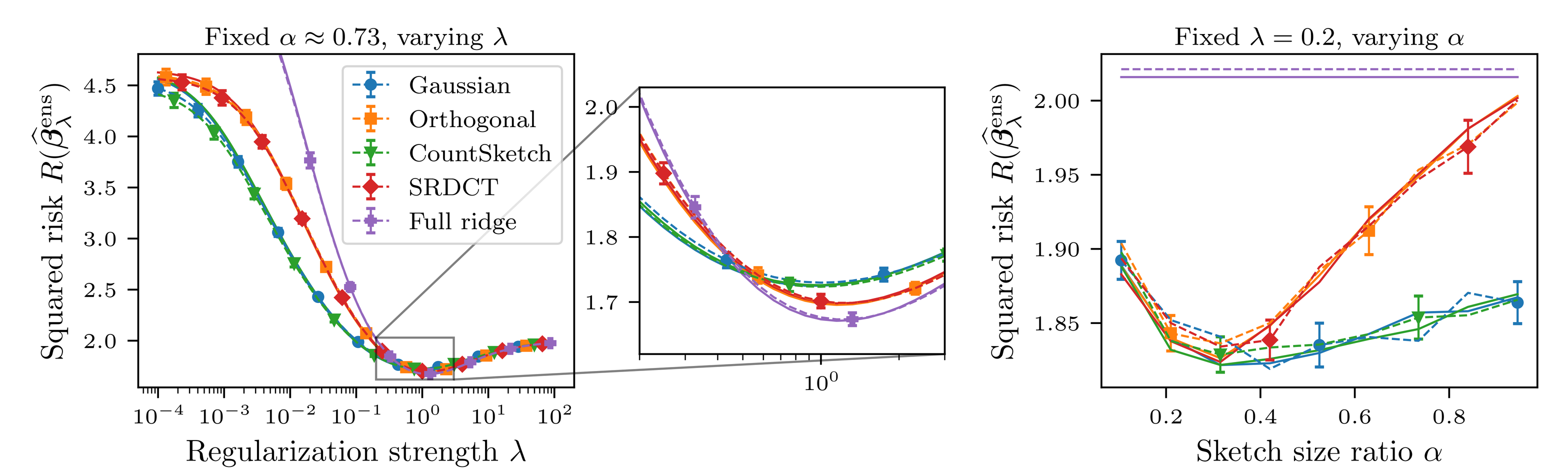
$$\mu' \simeq \mu'', \quad \text{and therefore} \quad \widehat{R}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) \simeq R(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda).$$

Theorem 3. For any t pseudo-Lipshitz of order 2,

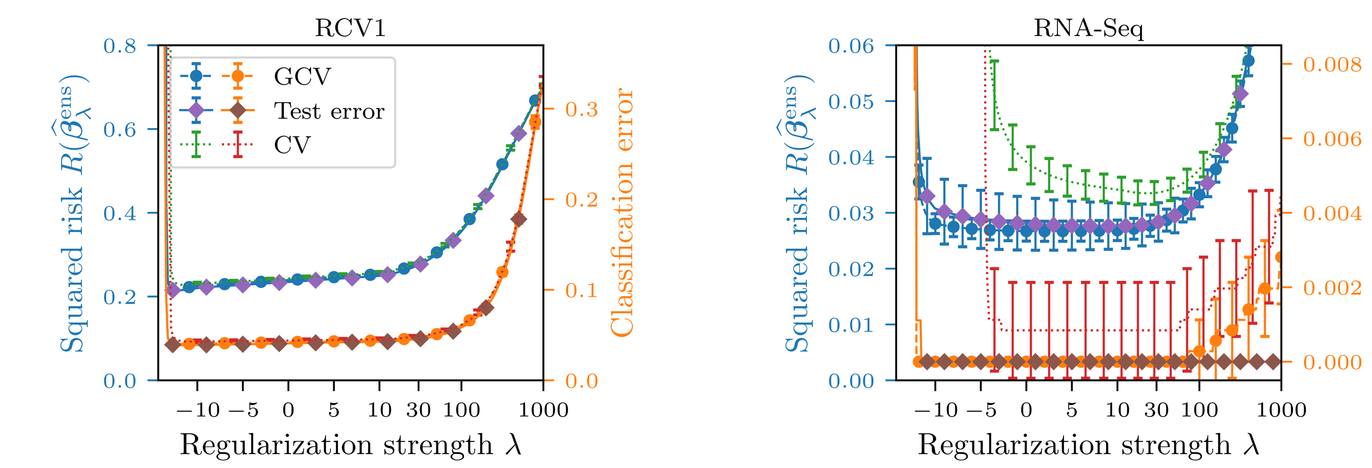
$$\widehat{T}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda) \simeq T(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_\lambda), \quad \text{and therefore} \quad \widehat{P}^{\mathrm{ens}}_\lambda \overset{2}{\Rightarrow} P^{\mathrm{ens}}_\lambda.$$

Empirical results

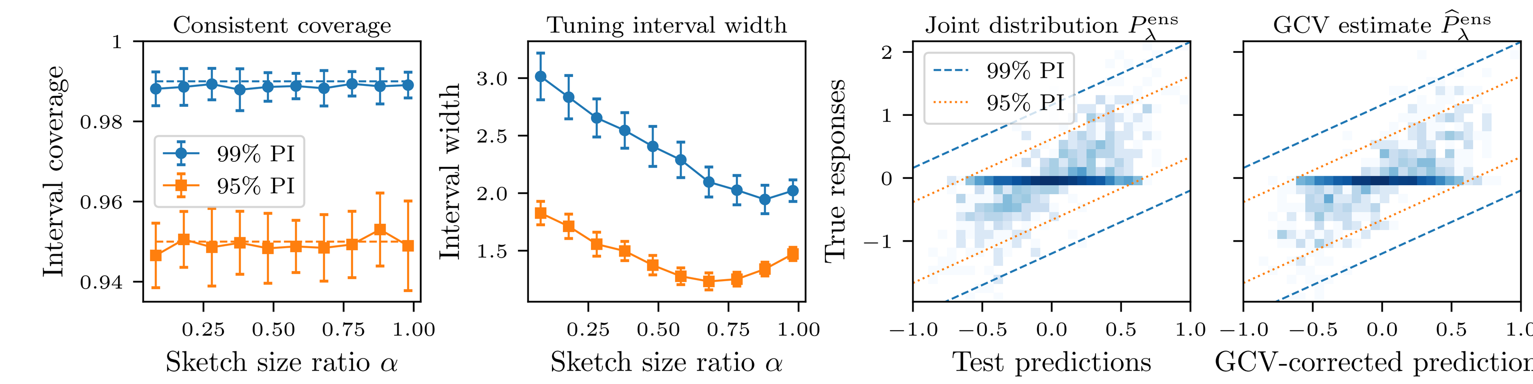
Consistency across sketching families



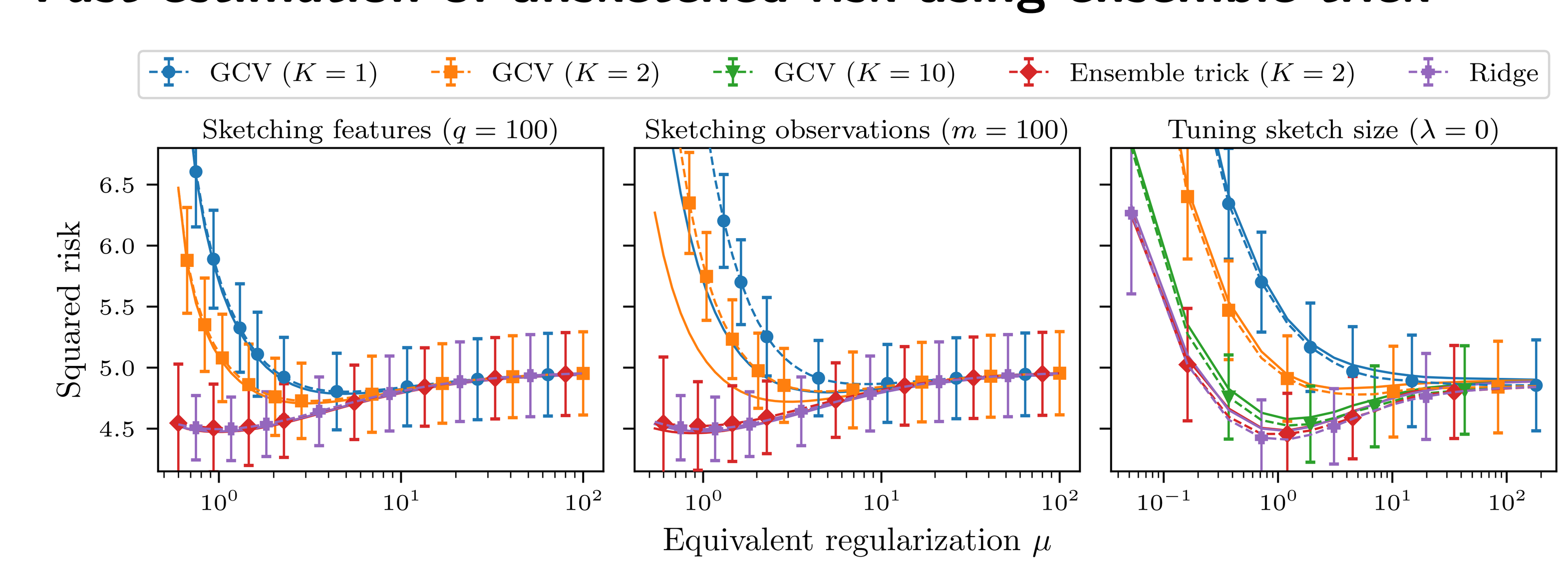
Risk estimation for real data



Consistent prediction intervals



Fast estimation of unsketched risk using ensemble trick



Acknowledgements

This collaboration was partially supported by Office of Naval Research MURI grant N00014-20-1-2787. DL was supported by Army Research Office grant 2003514594.