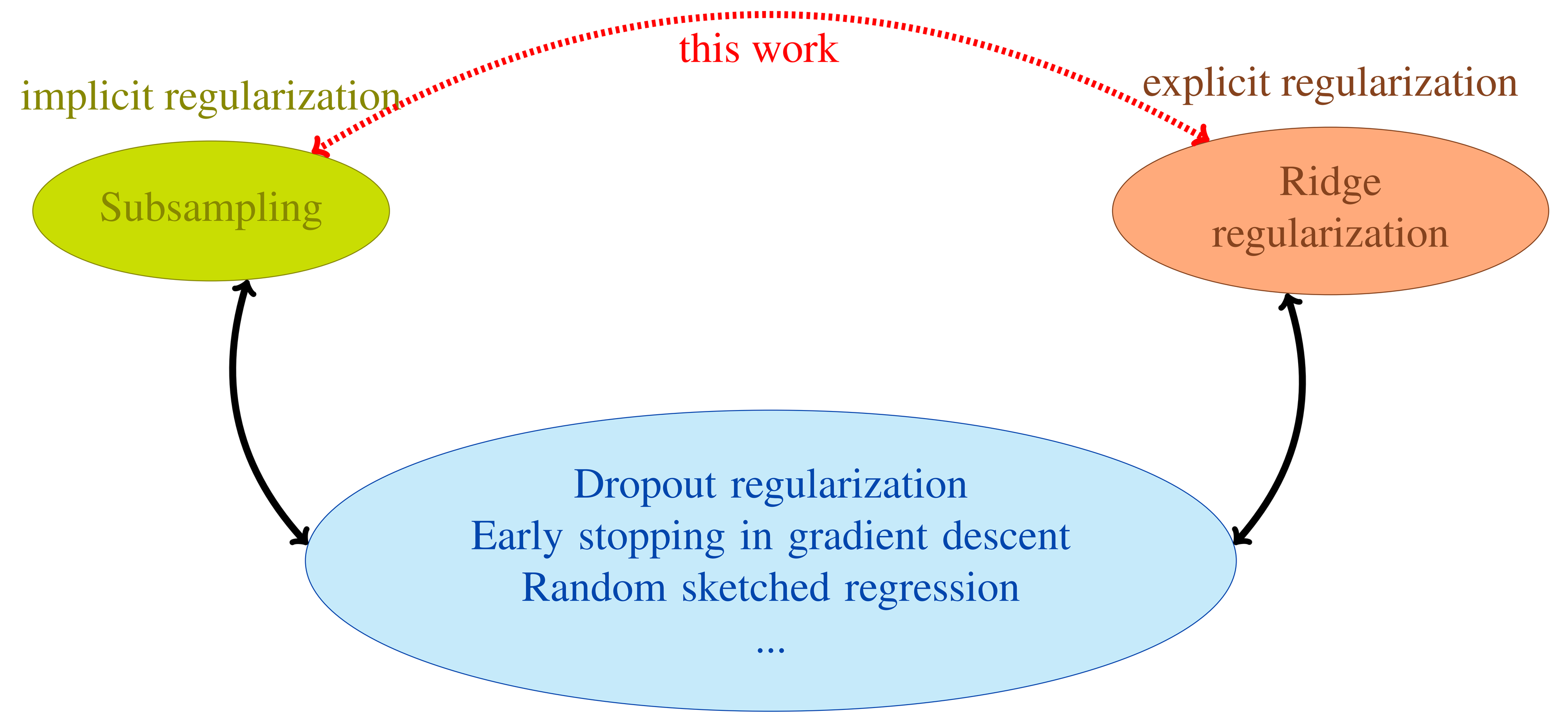
Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation

Jin-Hong Du[1*] Pratik Patil[2*] Arun Kumar Kuchibhotla[1]

[1]Department of Statistics and Data Science, Carnegie Mellon University

[2]Department of Statistics, University of California, Berkeley [*]equal contribution

Background



Ridge estimator: Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) : j \in [n]\}$ denote a dataset containing i.i.d. random vectors in $\mathbb{R}^p \times \mathbb{R}$. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ is defined as:

$$\widehat{\beta}_k^{\lambda}(\mathcal{D}_I) = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 / k + \lambda \|\boldsymbol{\beta}\|_2^2, \qquad I \subseteq [n], |I| = k \tag{1}$$

Ensemble ridge estimator: For $\lambda \geq 0$, the ensemble estimator is then defined as:

$$\widetilde{\beta}_{k,M}^{\lambda}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^{M}) := \frac{1}{M} \sum_{\ell \in [M]} \widehat{\beta}_k^{\lambda}(\mathcal{D}_{I_\ell}), \tag{2}$$

where $I_1, \ldots, I_M$ are samples from $\mathcal{I}_k := \{\{i_1, i_2, \ldots, i_k\} : 1 \leq i_1 < i_2 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator $\widetilde{\beta}_{k,\infty}^{\lambda}(\mathcal{D}_n)$ is obtained with $M \to \infty$.

Conditional prediction risk: The goal is to study the prediction risk:

$$R_{k,M}^{\lambda} := \mathbb{E}_{(\boldsymbol{x},y)}[(y - \boldsymbol{x}^\top \widetilde{\beta}_{k,M}^{\lambda})^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}], \tag{3}$$

under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. Here, ϕ and ϕ_s are the *data* and *subsample aspect ratios*, respectively.

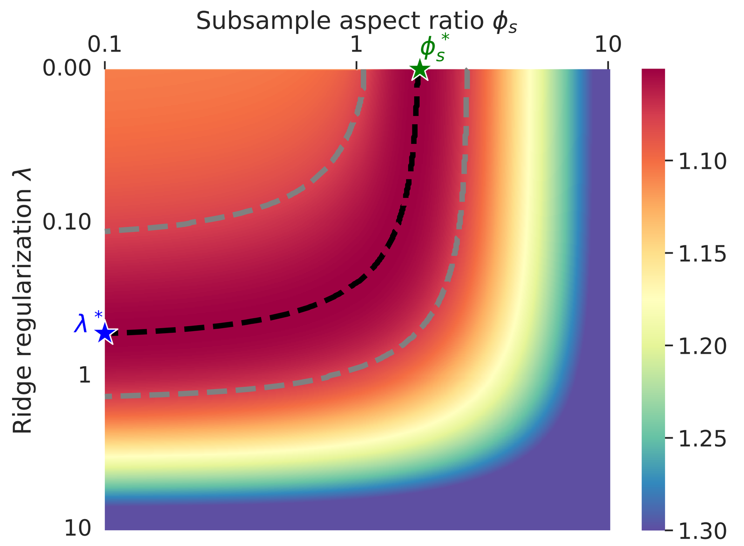
Summary

- **General risk equivalences.** We establish prediction risk equivalences between *implicit* regularization of subsampling and *explicit* ridge regularization for the subsample ridge ensemble. For any $\tau \geq 0$, we provide the set $\mathcal{C}_\tau$ of pairs (λ, ϕ_s) such that the risk of the full ridge ensemble with ridge penalty λ and subsample aspect ratio ϕ_s is equal to the risk of the ridge predictor with ridge penalty τ.
- **Uniform consistency of GCV.** For full ridge ensembles, we establish the uniform consistency of GCV across all possible subsample sizes k. Notably, this result is also applicable to the ridgeless regression ($\lambda = 0$). This enables tuning the subsample size in a data-dependent manner.
- **Finite-ensemble surprises.** Even though GCV is consistent for $M = 1$ and $M = \infty$, interestingly, this is the first paper that proves GCV *can* be inconsistent even for ridge ensembles when the ensemble size $M = 2$. Nevertheless, GCV is applicable for tuning subsample sizes, even with moderate ensemble sizes in practice.

Assumptions

- **Feature model:** $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ contains i.i.d. entries with bounded $4 + \delta$ moments, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ has bounded eigenvalues and limiting spectral distribution.
- **Response model:** $\boldsymbol{y} = \boldsymbol{X}\beta_0 + \boldsymbol{\epsilon}$, where $\beta_0 \in \mathbb{R}^p$ satisfies $\|\beta_0\|_2^2 \xrightarrow{\text{a.s.}} \rho^2$, and $\boldsymbol{\epsilon}$ contains i.i.d. entries with variance σ^2 and bounded $4 + \delta$ moments. The limiting spectral distribution of β_0's (squared) projection onto $\boldsymbol{\Sigma}$ exists.

Risk equivalence in the full ensemble



Previous work has shown that the limiting risk exists: $R_{k,\infty}^{\lambda} \xrightarrow{\text{a.s.}} \mathscr{R}_{\infty}^{\lambda}(\phi, \phi_s)$. When $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure.

Theorem 2.3

$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}_{\infty}^{0}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}_{\infty}^{\lambda}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi,\\ \lambda \geq 0}} \mathscr{R}_{\infty}^{\lambda}(\phi, \phi_s)}_{\text{opt. ridge ensemble}}.$$

- Implication: the implicit regularization provided by the subsample ensemble (a larger ϕ_s, or a smaller k) amounts to adding more explicit ridge regularization (a larger λ).
- Usage: tuning ridge penalty λ for optimal ridge predictors ($\phi_s = \phi$) by tuning subsample aspect ratio ϕ_s for ridgeless ensembles ($\lambda = 0$).

Generalized Cross-Validation (GCV)

For general M, the GCV estimator is defined as

$$\mathrm{gcv}_{k,M}^{\lambda} = \frac{T_{k,M}^{\lambda}}{D_{k,M}^{\lambda}} = \frac{\frac{1}{|I_{1:M}|} \sum_{i \in I_{1:M}} (y_i - \boldsymbol{x}_i^\top \widetilde{\beta}_{k,M}^{\lambda})^2}{(1 - |I_{1:M}|^{-1} \operatorname{tr}(\boldsymbol{S}_{k,M}^{\lambda}))^2} \quad \begin{array}{l} \text{training error} \\ \text{degrees-of-freedom correction} \end{array}$$

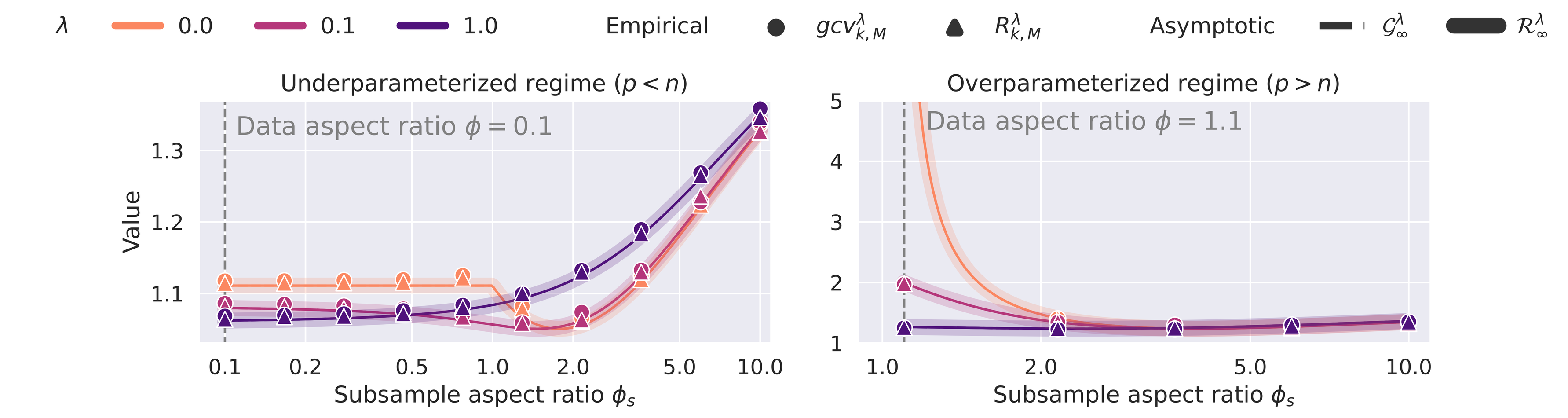
where $\boldsymbol{S}_{k,M}^{\lambda} = \frac{1}{M} \sum_{\ell=1}^{M} \boldsymbol{X}_{I_\ell} (\boldsymbol{X}_{I_\ell}^\top \boldsymbol{X}_{I_\ell} / k + \lambda \boldsymbol{I}_p)^{+} \boldsymbol{X}_{I_\ell}^\top / k$ is the smoothing matrix that represents the degrees-of-freedom.

Theorem 3.1 For all $\lambda \geq 0$, we have

$$\max_{k \in \mathcal{K}_n} |\mathrm{gcv}_{k,\infty}^{\lambda} - R_{k,\infty}^{\lambda}| \xrightarrow{\text{a.s.}} 0. \tag{4}$$

Coupled with the risk equivalence, we further have data-dependent tuning:

Corollary 3.2 $\quad \mathrm{gcv}_{\widehat{k}^0,\infty}^{0} \xrightarrow{\text{a.s.}} \min_{\phi_s \geq \phi, \lambda \geq 0} \mathscr{R}_{\infty}^{\lambda}(\phi, \phi_s).$

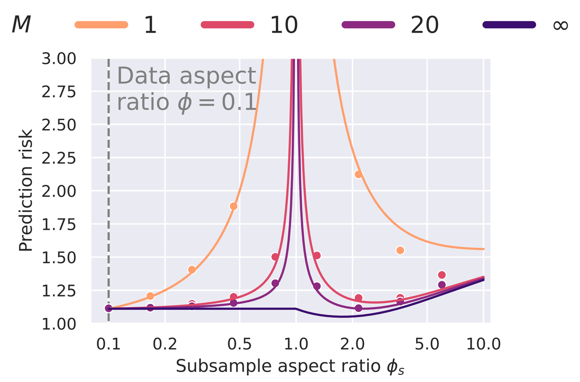


Inconsistency

Proposition 3.3 For ensemble size $M = 2$, ridge penalty $\lambda = 0$, and any $\phi \in (0, \infty)$,

$$|\mathrm{gcv}_{k,2}^{0} - R_{k,2}^{0}| \overset{\text{p}}{\not\to} 0.$$

The bias scales as $1/M$ and is negligible for large M.



Future directions

- Bias correction of GCV for finite M;
- Extension to other metrics [2];
- Extension to other base predictors.

[1] Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. "Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation". In: *International Conference on Machine Learning* (2023)

[2] Pratik Patil and Jin-Hong Du. "Generalized equivalences between subsampling and ridge regularization". In: *arXiv preprint arXiv:2305.18496* (2023)