Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation

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

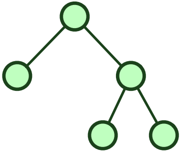
[1] Department of Statistics and Data Science, Carnegie Mellon University
[2] Department of Statistics, University of California, Berkeley
[*] equal contribution

July 2023

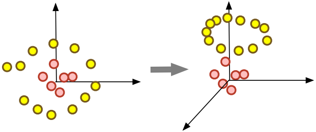
Over-parameterization and regularization

- In the big data era, the success of machine learning and deep learning methods typically have much more parameters than the training samples.

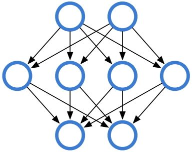
Random forest



Kernel method

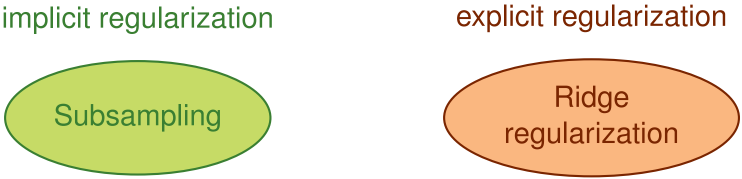


Neural network

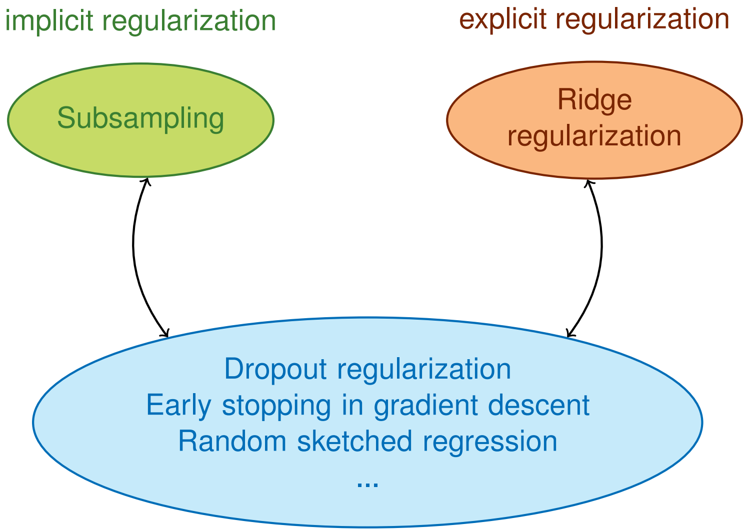


- Optimizing such over-parameterized models requires different types of regularization.

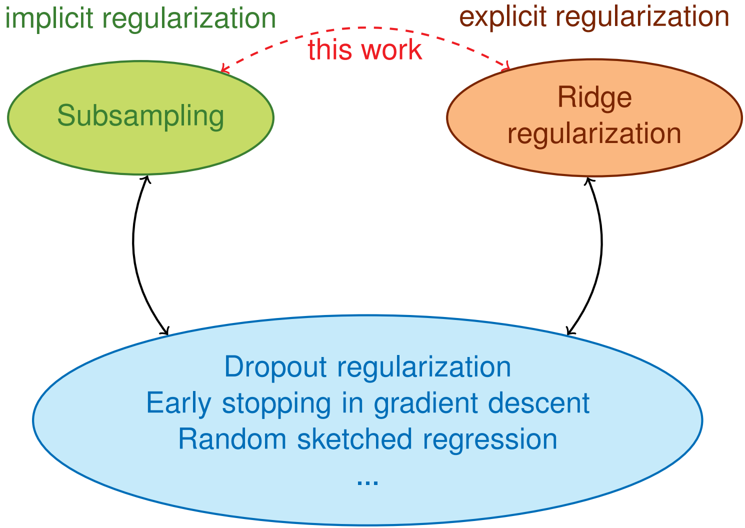
Explicit and implicit regularization



Explicit and implicit regularization



Explicit and implicit regularization



Ridge ensembles

- **Ridge estimator:** Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is defined as:

$$\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_I) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Ridge ensembles

- **Ridge estimator:** Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is defined as:

$$\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_I) = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

- **Ensemble ridge estimator:**

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

with $I_1, \ldots, I_M \sim \mathcal{I}_k := \{\{i_1, \ldots, i_k\} : 1 \leq i_1 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator is defined by letting $M \to \infty$.

Prediction risk

Conditional prediction risk: The goal is to quantify and estimate the prediction risk:

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

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.

Prediction risk

Conditional prediction risk: The goal is to quantify and estimate the prediction risk:

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

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.

Focusing on subsample ridge ensemble, we aim to answer:

(1) What is the role and relationship between implicit subsampling and explicit ridge regularization with regard to prediction risk?

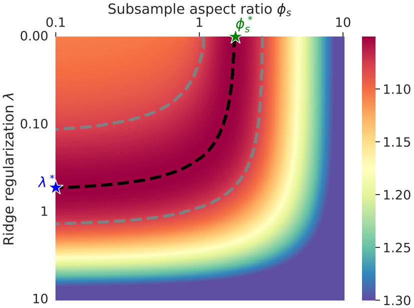
(2) How to tune the subsample aspect ratio ϕ_s and the ridge penalty λ to minimize the prediction risk?

Risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.

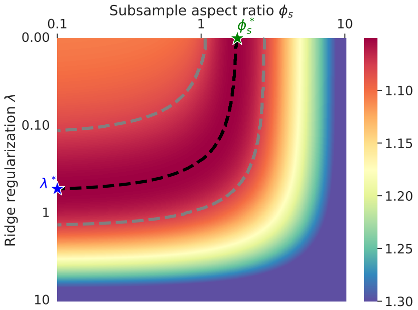


Risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- Risk equivalence (Theorem 2.3):

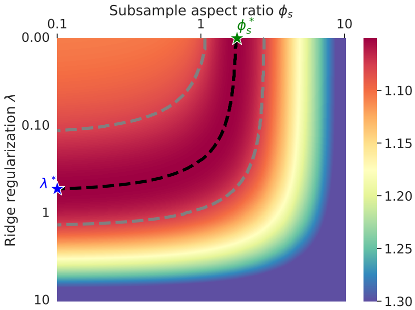
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{0}_{\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda \geq 0}} \mathscr{R}^{\lambda}_{\infty}(\phi,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

Risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- 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).

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (λ, ϕ_s) to minimize the prediction risk?

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (λ, ϕ_s) to minimize the prediction risk?
- For ordinary ridge ($M = 1$ or $k = n$), the **generalized cross-validation (GCV)** estimator is known to be consistent.

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (λ, ϕ_s) to minimize the prediction risk?
- For ordinary ridge ($M = 1$ or $k = n$), the **generalized cross-validation (GCV)** estimator is known to be consistent.
- For general M, the GCV estimator is defined as

$$\mathrm{gcv}_{k,M}^{\lambda} = \frac{T_{k,M}^{\lambda}}{D_{k,M}^{\lambda}} \quad \longleftarrow$$

training error

degree of freedom correction

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (λ, ϕ_s) to minimize the prediction risk?
- For ordinary ridge ($M = 1$ or $k = n$), the **generalized cross-validation (GCV)** estimator is known to be consistent.
- For general M, the GCV estimator is defined as

$$\mathrm{gcv}_{k,M}^{\lambda} = \frac{T_{k,M}^{\lambda}}{D_{k,M}^{\lambda}} = \frac{\frac{1}{|I_{1:M}|}\sum_{i \in I_{1:M}} (y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{k,M}^{\lambda})^2}{(1 - |I_{1:M}|^{-1}\,\mathrm{tr}(\boldsymbol{S}_{k,M}^{\lambda}))^2},$$

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 degree of freedom.

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (λ, ϕ_s) to minimize the prediction risk?
- For ordinary ridge ($M = 1$ or $k = n$), the **generalized cross-validation (GCV)** estimator is known to be consistent.
- For general M, the GCV estimator is defined as

$$\mathrm{gcv}_{k,M}^{\lambda} = \frac{T_{k,M}^{\lambda}}{D_{k,M}^{\lambda}} = \frac{\frac{1}{|I_{1:M}|} \sum_{i \in I_{1:M}} (y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{k,M}^{\lambda})^2}{(1 - |I_{1:M}|^{-1} \operatorname{tr}(\boldsymbol{S}_{k,M}^{\lambda}))^2},$$

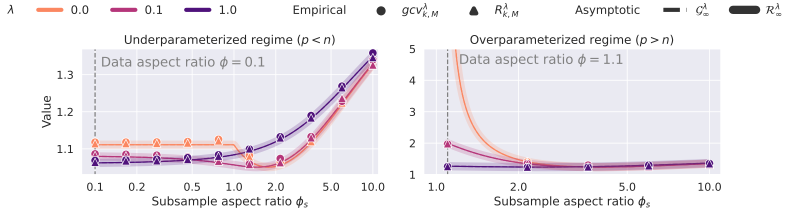
  where $\boldsymbol{S}_{k,M}^{\lambda} = \frac{1}{M} \sum_{\ell=1}^{M} \boldsymbol{X}_{I_\ell} (\boldsymbol{X}_{I_\ell}^\top \boldsymbol{X}_{I_\ell}/k + \lambda \boldsymbol{I}_p)^{+} \boldsymbol{X}_{I_\ell}^\top / k$ is the smoothing matrix that represents the degree of freedom.
- The GCV for full ensemble is defined by letting M tend to infinity.

Uniform consistency of GCV for full-ensemble ridge

- (Theorem 3.1, informal) For all $\lambda \geq 0$, we have

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

- This allows selecting the optimal ensemble and subsample sizes in a data-dependent manner:



Coupled with the risk equivalence result, it suffices to fix λ and only tune the subsample size k or subsample aspect ratio ϕ_s.

Inconsistency on finite ensembles

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

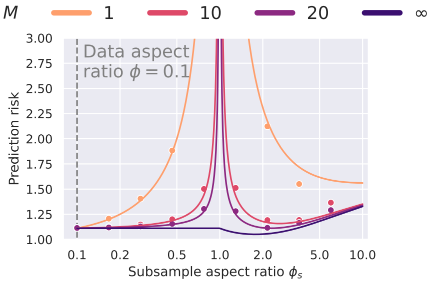
$$|\mathrm{gcv}_{k,2}^{0} - R_{k,2}^{0}| \overset{\mathrm{p}}{\not\to} 0.$$

Inconsistency on finite ensembles

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

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

- The bias scales as $1/M$, which is negligible for large M:



point - empirical GCV line - theoretical risk

Summary

- This work [1] reveals the connections between the *implicit regularization* induced by *subsampling* and *explicit ridge regularization* for subsample ridge ensembles.
- We establish the *uniform consistency* of GCV for full ridge ensembles.
- We show that GCV can be *inconsistent* even for ridge ensembles when $M = 2$.
- 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)