Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation

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

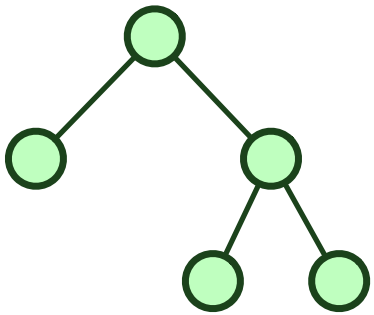
[1] Department of Statistics and Data Science, Carnegie Mellon University
[2] Department of Statistics, University of California, Berkeley
[*] equal contribution

July 2023

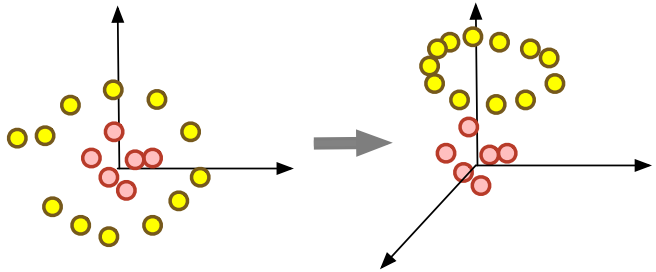
Regularization

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

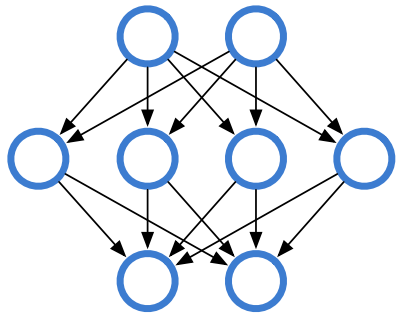
Random forest



Kernel method



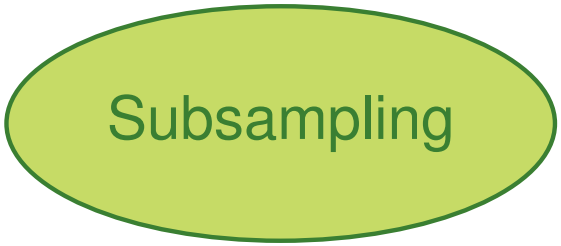
Neural network



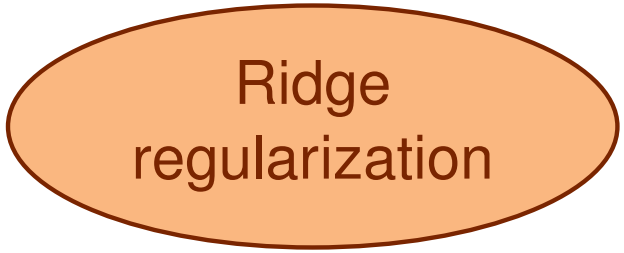
- Optimizing such overparameterized models requires different types of regularization.

Explicit and implicit regularization

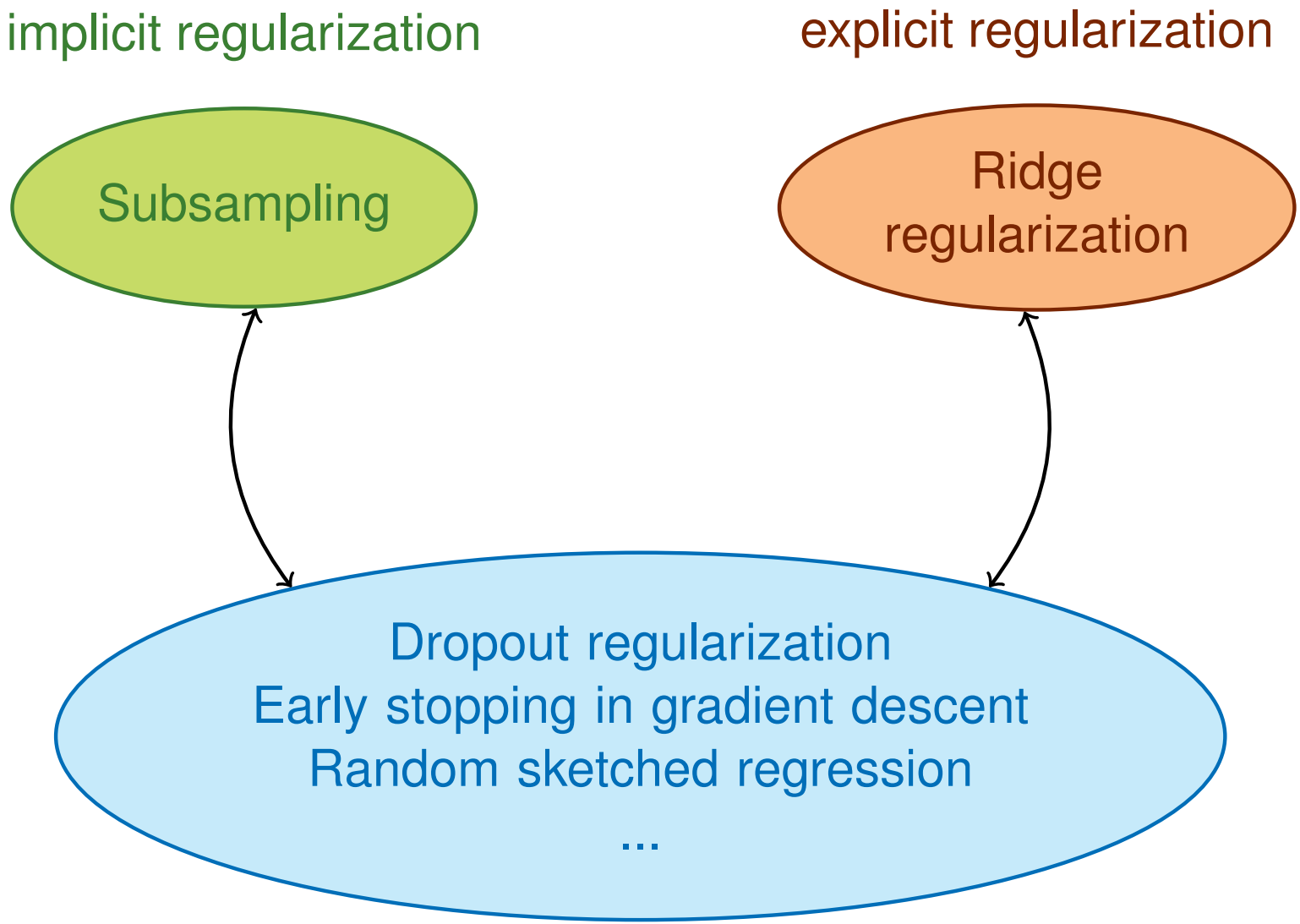
implicit regularization



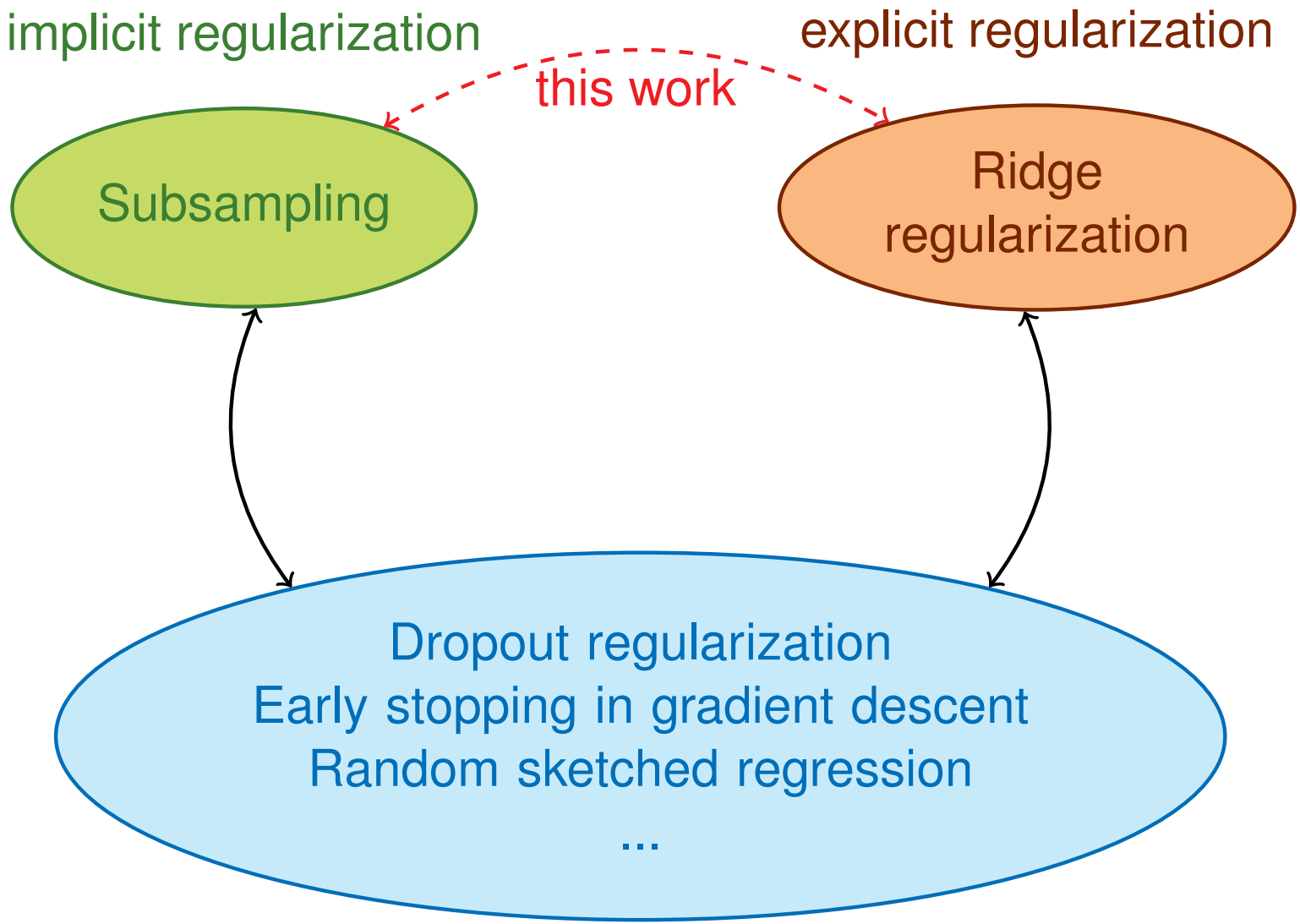
explicit regularization



Explicit and implicit regularization



Explicit and implicit regularization



Ridge ensembles

- **Ridge estimator**: Let $D_n = f(\boldsymbol{x}_j; y_j) \; 2 \; \mathrm{R}^p \quad \mathrm{R} : j \; 2 \; [n]g$ denote a dataset. The ridge estimator fitted on subsampled dataset D_I with $I \quad [n]; jIj = k$ is defined as:

$$\mathrm{b}_k(D_I) = \operatorname*{argmin}_{2\mathrm{R}^p} \frac{1}{k} \sum_{j2I} (y_j \quad \boldsymbol{x}_j^{>} \quad)^2 + \quad k \quad k_2^2:$$

Ridge ensembles

- **Ridge estimator**: Let $D_n = f(\boldsymbol{x}_j; y_j)\ 2\ \mathrm{R}^p \quad \mathrm{R} : j\ 2\ [n]g$ denote a dataset. The ridge estimator fitted on subsampled dataset D_I with $I \quad [n]; jIj = k$ is defined as:

$$\mathrm{b}_k(D_I) = \operatorname*{argmin}_{2\mathrm{R}^p} \frac{1}{k} \sum_{j2I} (y_j \quad \boldsymbol{x}_j^{>} \quad)^2 + \quad k \quad k_2^2:$$

- **Ensemble ridge estimator**:

$$\mathrm{e}_{k;M}(D_n; fI_{`}g_{`=1}^{M}) := \frac{1}{M} \sum_{`2[M]} \mathrm{b}_k(D_{I_{`}});$$

with $I_1; : : : ; I_M \quad I_k := f fi_1; : : : ; i_k g : 1 \quad i_1 < : : : < i_k \quad ng$. The *full-ensemble* ridge estimator is defined by letting $M \ ! \ 1$.

Risk equivalence

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

$$R_{k;M} := \mathrm{E}_{(\boldsymbol{x};y)}[(y \quad \boldsymbol{x}^{>} \, \mathrm{e}_{k;M})^2 \; j \; D_n; \; fI \cdot g_{\cdot=1}^{M}]; \tag{1}$$

under proportional asymptotics where $n; p; k \; ! \; 1$, $p{=}n \; ! \;$ and $p{=}k \; ! \; {}_s$. Here, and ${}_s$ are the data and subsample aspect ratios, respectively.

Risk equivalence

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

$$R_{k;\infty} \xrightarrow{\text{a.s.}} R_{k;\infty}(\gamma; \gamma_s):$$

- For $\sigma = 1$, the risk profile as a function of $(\gamma; \gamma_s)$ is shown in the figure in the log-log scale.

Risk equivalence

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

$$R_{k;\infty} \xrightarrow{\text{a.s.}} R_{k;\infty}(\gamma; \gamma_s).$$

- For $\gamma = 1$, the risk profile as a function of $(\gamma; \gamma_s)$ is shown in the figure in the log-log scale.

- Risk equivalence (Theorem 2.3):

$$\underbrace{\min_s R^0_\infty(\gamma; \gamma_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} R_\infty(\gamma; \lambda)}_{\text{opt. ridge predictor}} = \underbrace{\min_{s;\, \lambda \geq 0} R_\infty(\gamma; \gamma_s)}_{\text{opt. ridge ensemble}}.$$

Risk equivalence

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

$$R_{k;\infty} \xrightarrow{\text{a.s.}} R_{k;\infty}(\gamma; \gamma_s):$$

- For $\sigma = 1$, the risk profile as a function of $(\gamma; \gamma_s)$ is shown in the figure in the log-log scale.

- Implication: the implicit regularization provided by the subsample ensemble (a larger γ_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 (; $_s$) to minimize the prediction risk?

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (; $_s$) 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 (; $_s$) 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

$$\text{gcv}_{k;M} = \frac{T_{k;M}}{D_{k;M}}$$

training error

degree of freedom correction

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick $(\quad ; \quad _s)$ 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

$$\text{gcv}_{k;M} = \frac{T_{k;M}}{D_{k;M}} = \frac{\frac{1}{jI_{1:M}j} \sum_{i2I_{1:M}} (y_i \quad \boldsymbol{x}_i^{>} \, \mathrm{e}_{k;M})^2}{(1 \quad jI_{1:M}j \quad ^1 \operatorname{tr}(\boldsymbol{S}_{k;M}))^2};$$

where $\boldsymbol{S}_{k;M} = \frac{1}{M} \sum_{`=1}^{M} \boldsymbol{X}_{I`} (\boldsymbol{X}_{I`}^{>} \boldsymbol{X}_{I`} =k + \quad \boldsymbol{I}_p)^{+} \boldsymbol{X}_{I`}^{>} =k$ is the smoothing matrix that represents the degree of freedom.

Generalized cross-validation for ridge ensembles

- Beyond quantitative analysis, how can one pick (; $_s$) 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

$$\text{gcv}_{k;M} = \frac{T_{k;M}}{D_{k;M}} = \frac{\frac{1}{jI_{1:M}j} \sum_{i2I_{1:M}} (y_i \quad \boldsymbol{x}_i^{>} \, \mathrm{e}_{k;M})^2}{(1 \quad jI_{1:M}j \quad ^1 \operatorname{tr}(\boldsymbol{S}_{k;M}))^2};$$

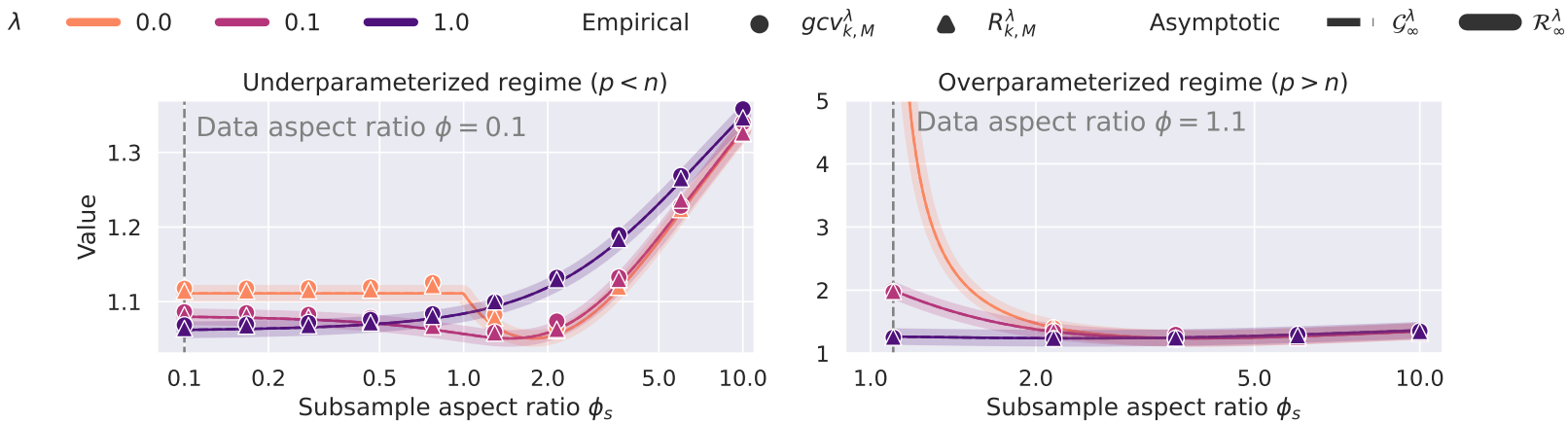
  where $\boldsymbol{S}_{k;M} = \frac{1}{M} \sum_{`=1}^{M} \boldsymbol{X}_{I`} (\boldsymbol{X}_{I`}^{>} \boldsymbol{X}_{I`} =k + \quad \boldsymbol{I}_p)^{+} \boldsymbol{X}_{I`}^{>} =k$ is the smoothing matrix that represents the degree of freedom.
- The GCV for full ensemble is defined by letting M tend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

Uniform consistency of GCV for full-ensemble ridge

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

$$\max_{k2K_n} j\text{gcv}_{k;1} \quad R_{k;1}\, j \xrightarrow{\text{a.s.}} 0:$$

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



Coupled with the risk equivalence result, it suffices to fix $\quad$ and only tune the subsample size k or subsample aspect ratio $\ _s$.

Inconsistency on finite ensembles

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

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

Inconsistency on finite ensembles

- (Proposition 3.3, informal) For ensemble size $M = 2$, ridge penalty $= 0$, and any $\in (0, 1)$,

$$|\mathrm{gcv}^{0}_{k,2} \quad R^{0}_{k,2}| \not\xrightarrow{p} 0.$$

- 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)