Generalized equivalences between subsampling and ridge regularization

Pratik Patil[1] Jin-Hong Du[2]

[1]Department of Statistics, UCB [2]Department of Statistics, CMU

Ensemble predictors and generalized risks

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

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

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

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

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 $\widehat{\boldsymbol{\beta}}_{k,\infty}^{\lambda}(\mathcal{D}_n)$ is obtained with $M \to \infty$.

Generalized risk. For a linear functional $L_{\boldsymbol{A},\boldsymbol{b}}(\boldsymbol{\beta}) = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{b}$, we study

$$R(\widehat{\boldsymbol{\beta}}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0) = \frac{1}{\operatorname{nrow}(\boldsymbol{A})}\|L_{\boldsymbol{A},\boldsymbol{b}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2, \tag{3}$$

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

Data assumptions. Each feature vector $\boldsymbol{x}_i$ for $i \in [n]$ can be decomposed as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ contains i.i.d. entries z_{ij} for $j \in [p]$ with mean 0, variance 1, and bounded $4+\mu$ moments for some $\mu > 0$. Response distribution: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4+\mu$ moments.

Summary of results

Table 1: Comparison with related work. "✓°" indicates a partial equivalence result connecting the *optimal* prediction risk of the ridge predictor and the full ridgeless ensemble.

	Type of equivalence results			Type of data assumptions		
	pred. risk	gen. risk	estimator	response	feature	lim. spectrum
Lejeune 2020	✓°			linear	isotropic Gaussian	exists
Patil 2022	✓°			linear	isotropic RMT	exists
Du 2023	✓			linear	anisotropic RMT	exists
This work	✓	✓	✓	arbitrary	anisotropic RMT	need not exist

- **Risk equivalences.** We establish asymptotic equivalences of the full-ensemble ridge estimators at different ridge penalties λ and subsample ratios ψ along specific paths in the (λ, ψ)-plane for a variety of generalized risk functionals.
- **Structural equivalences.** We establish structural equivalences for linear functionals of the ensemble ridge estimators that hold for all ensemble sizes.
- **Equivalence implications.** The prediction risk of an optimally tuned ridge estimator is monotonically increasing in p/n under mild regularity conditions.
- **Generality of equivalences.** The results apply to arbitrary responses with bounded $4+\mu$ moments, as well as features with general covariance structures.

Generalized risk equivalences

Equivalence paths. Given $\phi \in (0,\infty)$ and $\bar{\psi} \in [\phi, \infty]$, our statement of equivalences between different ensemble estimators is defined through certain paths characterized by two endpoints $(0, \bar{\psi})$ and $(\bar{\lambda}, \phi)$. Let H_p be the empirical spectral distribution of $\boldsymbol{\Sigma}$: $H_p(r) = p^{-1}\sum_{i=1}^p \mathbb{1}_{\{r_i \leq r\}}$, where r_i's are the eigenvalues of $\boldsymbol{\Sigma}$. Consider the following system of equations in $\bar{\lambda}$ and v:

$$\frac{1}{v} = \bar{\lambda} + \phi\int \frac{r}{1+vr}\,\mathrm{d}H_p(r), \quad \text{and} \quad \frac{1}{v} = \bar{\psi}\int\frac{r}{1+vr}\,\mathrm{d}H_p(r). \tag{4}$$

Now, define a path $\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi})$ that passes through the endpoints $(0, \bar{\psi})$ and $(\bar{\lambda}, \phi)$:

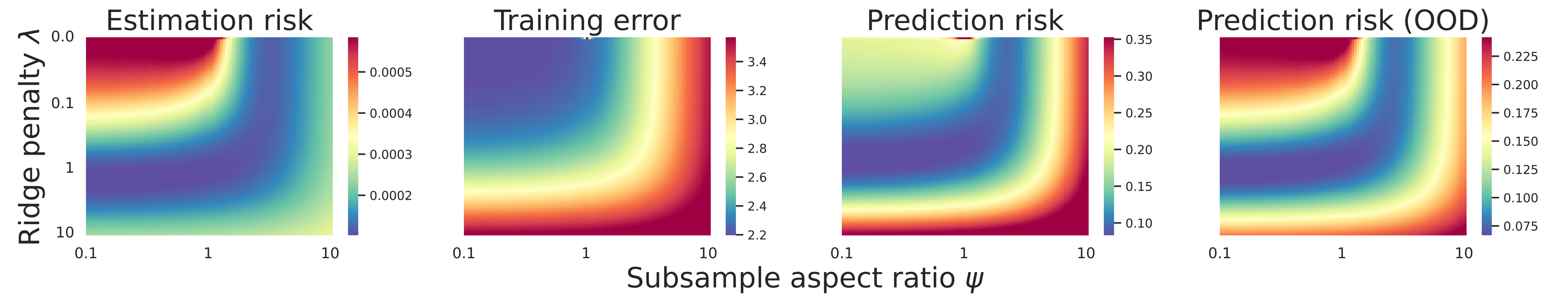
$$\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi}) = \{(1-\theta)\cdot(\bar{\lambda}, \phi) + \theta\cdot(0, \bar{\psi}) \mid \theta \in [0,1]\}. \tag{5}$$

Theorem 1. For any $\bar{\psi} \in [\phi, +\infty]$, let $\bar{\lambda}$ be as defined in (4). Then, for any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path $\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi})$ as defined in (5), the generalized risk functionals (3) of the full-ensemble estimator are asymptotically equivalent:

$$R\big(\widehat{\boldsymbol{\beta}}^{\lambda_1}_{\lfloor p/\psi_1\rfloor,\infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big) \simeq R\big(\widehat{\boldsymbol{\beta}}^{\lambda_2}_{\lfloor p/\psi_2\rfloor,\infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big). \tag{6}$$

Table 2: Summary of asymptotic equivalences between subsampling and ridge regularization for generalized risks and their corresponding statistical learning problems.

Statistical learning problem	$L_{\boldsymbol{A},\boldsymbol{b}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$	$\boldsymbol{A}$	$\boldsymbol{b}$	$\operatorname{nrow}(\boldsymbol{A})$
vector coefficient estimation	$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$	$\boldsymbol{I}_p$	0	p
projected coefficient estimation	$\boldsymbol{a}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$	$\boldsymbol{a}^\top$	0	1
training error estimation	$\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}$	$\boldsymbol{X}$	$-\boldsymbol{f}_{\mathrm{NL}}$	n
in-sample prediction	$\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$	$\boldsymbol{X}$	0	n
out-of-sample prediction	$\boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}} - y_0$	$\boldsymbol{x}_0^\top$	$-\varepsilon_0$	1



Structural equivalences

Theorem 3. For any $\bar{\psi} \in [\phi, +\infty]$, let $\bar{\lambda}$ be as in (4). Then, for any $M \in \mathbb{N}\cup\{\infty\}$ and any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path (5), the M-ensemble estimators are asymptotically equivalent:

$$\widehat{\boldsymbol{\beta}}^{\lambda_1}_{\lfloor p/\psi_1\rfloor, M} \simeq \widehat{\boldsymbol{\beta}}^{\lambda_2}_{\lfloor p/\psi_2\rfloor, M}, \qquad \forall (\lambda_1,\psi_1), (\lambda_2,\psi_2) \in \mathcal{P}(\bar{\lambda}; \phi, \bar{\psi}). \tag{7}$$

Data-dependent paths. For any $M \in \mathbb{N}\cup\{\infty\}$, let $\bar{\lambda}_n$ be the value that satisfies the following equation in ensemble ridgeless and ridge gram matrices:

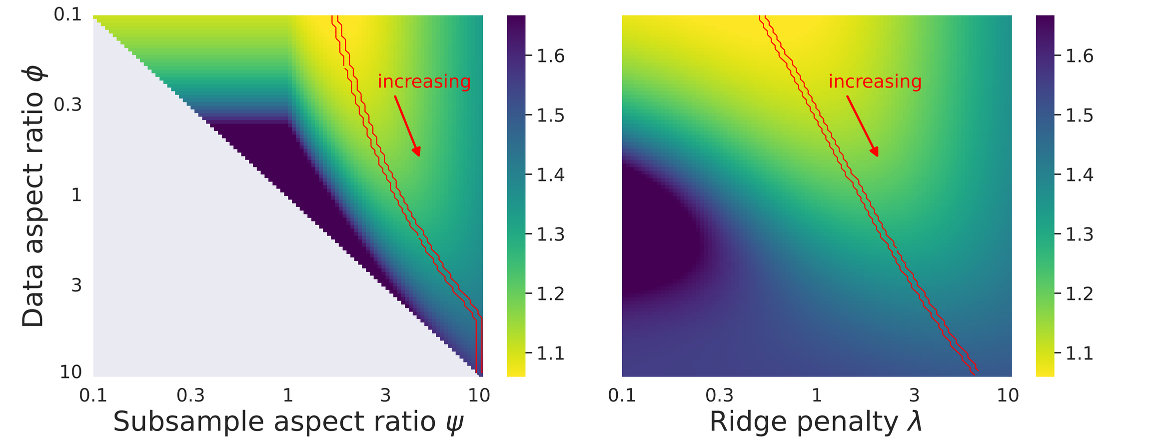
$$\frac{1}{M}\sum_{\ell=1}^M \frac{1}{k}\operatorname{tr}\left[\left(\frac{1}{k}\boldsymbol{L}_{I_\ell}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{L}_{I_\ell}\right)^+\right] = \frac{1}{n}\operatorname{tr}\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top + \bar{\lambda}_n\boldsymbol{I}_n\right)^{-1}\right]. \tag{8}$$

Define the data-dependent path $\mathcal{P}_n = \mathcal{P}(\bar{\lambda}_n; \phi_n, \bar{\psi}_n)$. Theorems 1 & 3 hold with $\mathcal{P}_n$.

Implications: Monotonicity of optimal ridge

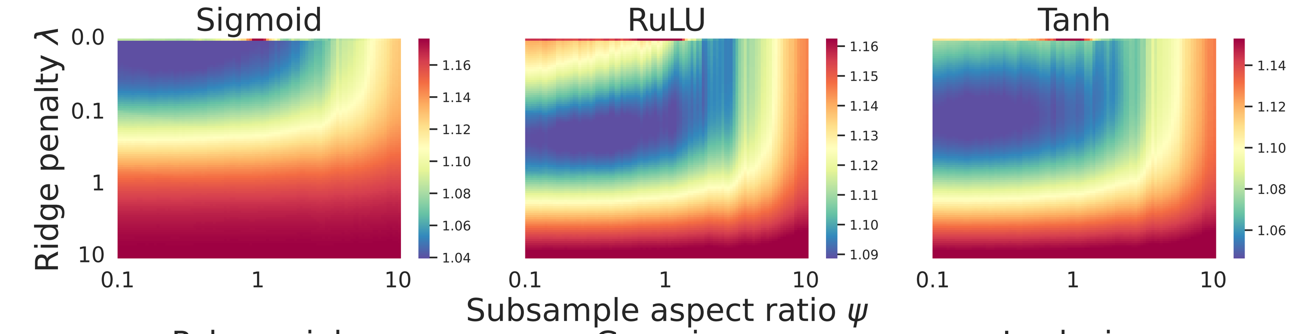
Risk monotonicity. Many common methods, such as ridgeless or lassoless predictors, exhibit non-monotonic behavior in the sample size or the limiting aspect ratio. An open problem raised by Nakkiran et al. (2021) asks whether the prediction risk of ridge regression with optimal ridge penalty λ^* is monotonically increasing in the data aspect ratio $\phi = p/n$. Our equivalences imply that the prediction risk of an optimally-tuned ridge estimator is monotonically increasing in the data aspect ratio under mild regularity conditions. Under proportional asymptotics, our result settles a recent open question raised by Conjecture 1 of Nakkiran et al. (2021) concerning the monotonicity of optimal ridge regression under anisotropic features and general data models while maintaining a regularity condition that preserves the linearized signal-to-noise ratios across regression problems.

Theorem 6. Let $k, n, p \to \infty$ such that $p/n \to \phi \in (0,\infty)$ and $p/k \to \psi \in [\phi, \infty]$. Then, for $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{b} = \boldsymbol{0}$, the optimal risk of the ridgeless ensemble, $\min_{\psi\geq\phi}\mathscr{R}(0; \phi, \psi)$, is monotonically increasing in ϕ. Consequently, the optimal risk of the ridge predictor, $\min_{\geq 0}\mathscr{R}(; \phi, \phi)$, is also monotonically increasing in ϕ.

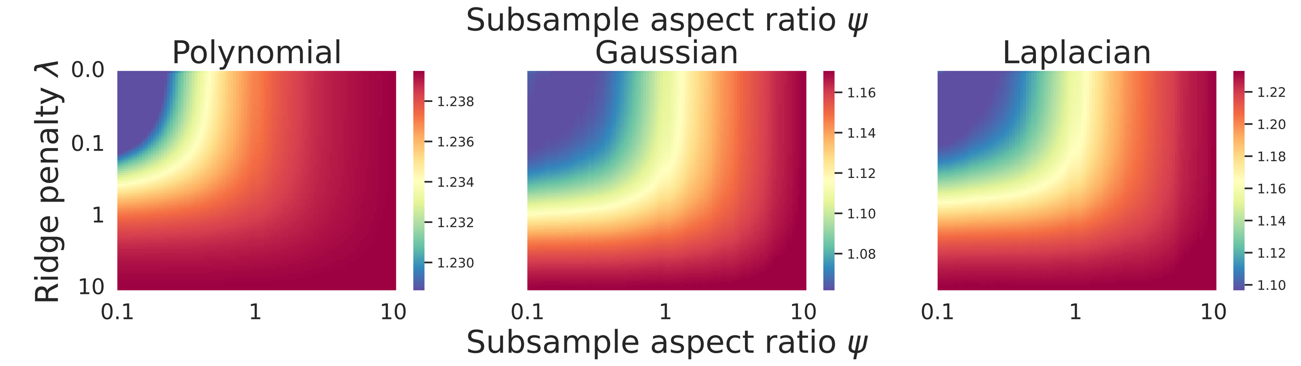


Equivalences for random and kernel features

- Equivalences for random features (Conjecture 7)



- Equivalences for kernel features (Conjecture 8)



References:

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. "The implicit regularization of ordinary least squares ensembles". In: *International Conference on Artificial Intelligence and Statistics*. 2020

Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. "Bagging in overparameterized learning: Risk characterization and risk monotonization". In: *arXiv preprint arXiv:2210.11445* (2022)

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

Preetum Nakkiran et al. "Optimal Regularization can Mitigate Double Descent". In: *International Conference on Learning Representations*. 2021