Mitigating multiple descents: Model-agnostic risk monotonization in high-dimensional learning

1 Motivation and summary of contributions

The phenomenon of non-monotonic generalization in the sample size and/or the model complexity is now widely known as "double/multiple descent" and has been empirically as well as theoretically investigated in many models, including linear and logistic regression, random features models, kernel regression, linear classification, decision trees and boosting, among others; see, e.g., [1–9] and the survey papers [10, 11]. In specific instances, the learning algorithm can be modified to guarantee monotonicity of the out-of-sample risk behavior in the sample size (e.g., [12]). However, the ubiquity of the double and multiple descent phenomenon begs the question: *can a general approach be devised that will modify any generic prediction procedure in order to achieve a monotone risk behaviour?* Examples of estimators that are known to exhibit a monotone risk profile are scarce, and the study of their theoretical properties typically requires strong distributional assumptions (e.g., [12, 13]). In this work, we develop a simple, general-purpose methodology that takes as input a generic predictor and, under mild conditions, returns a modified procedure whose out-of-sample risk is *asymptotically monotone* in the sample size under a proportional asymptotic framework. Our approach is based on a carefully designed *cross-validation* procedure, and is applicable to a large class of data generating distributions and learning algorithms. We first present our main result on general cross-validation in Section 2, and then present one of its applications for risk monotonization in Section 3.

2 General cross-validation and model selection

We begin by deriving some general, non-asymptotic oracle risk inequalities for cross-validation that hold under minimal assumptions. While our bounds apply to a wide range of learning problems and may be of independent interest, they are crucial in demonstrating the risk monotonization properties of the procedure presented in Section 3. Setting the stage, let $\mathcal{D}_n = \{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \leqslant i \leqslant n\}$ be a dataset containing i.i.d. observations from a distribution P. For a predictor $\widehat{f} : \mathbb{R}^p \to \mathbb{R}$ fitted on $\mathcal{D}_n$ and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, denote the conditional prediction risk of $\widehat{f}$ by $R(\widehat{f}) := \mathbb{E}[\ell(Y_0, \widehat{f}(X_0))|\mathcal{D}_n]$ where (X_0, Y_0) is a test observation sampled from P, independently of $\mathcal{D}_n$. Consider a random split of $\mathcal{D}_n$ into train and test sets, $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$, containing n_{tr} and n_{te} observations, respectively. We are interested in selecting the predictor with smallest out-of-sample risk among a collection of predictors $\{\widehat{f}^{\xi} : \mathbb{R}^p \to \mathbb{R}, \xi \in \Xi\}$ indexed by a finite set Ξ and trained on $\mathcal{D}_{\text{tr}}$. For each $\xi \in \Xi$, estimate $R(\widehat{f}^{\xi})$ using the *median-of-means* estimator (e.g., [14, 15]) that computes the median of averages of $\ell(Y_j, \widehat{f}(X_j)), (X_j, Y_j) \in \mathcal{D}_{\text{te}}$ over $\lceil 8 \log(|\Xi| n^3) \rceil$ disjoint batches in $\mathcal{D}_{\text{te}}$. Denote by $\widehat{f}^{\text{cv}}$ any of the predictors $\{\widehat{f}^{\xi}, \xi \in \Xi\}$ with the smallest estimated risk. Below we derive two oracle risk inequalities on $R(\widehat{f}^{\text{cv}})$, one in an additive form and the other in a multiplicative form. In preparation for theorem statement to follow, let $\Delta_n^{\text{add}} := \max_{\xi \in \Xi} |R(\widehat{f}^{\xi}) - \widehat{R}(\widehat{f}^{\xi})|$, $\Delta_n^{\text{mul}} := \max_{\xi \in \Xi} |\widehat{R}(\widehat{f}^{\xi})/R(\widehat{f}^{\xi}) - 1|$ denote certain error terms, and $\widehat{\sigma}_{\Xi} := \max_{\xi \in \Xi} \mathbb{E}[|\ell(Y_0, \widehat{f}^{\xi}(X_0))|^2 | \mathcal{D}_n]^{1/2}$, $\widehat{\kappa}_{\Xi} := \max_{\xi \in \Xi} \widehat{\sigma}^{\xi}/R(\widehat{f}^{\xi})$ denote conditional second moment and skewness-like proxies on the loss.

Theorem 1. *The conditional prediction risk of $\widehat{f}^{\text{cv}}$ satisfies following deterministic oracle risk inequalities:*

$$R(\widehat{f}^{\text{cv}}) \leqslant \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) + 2\Delta_n^{\text{add}} \quad \textit{and} \quad R(\widehat{f}^{\text{cv}}) \leqslant (1 + \Delta_n^{\text{mul}})/(1 - \Delta_n^{\text{mul}})_+ \cdot \min_{\xi \in \Xi} R(\widehat{f}^{\xi}).$$

Furthermore, there exist absolute constants $c^{\text{add}} > 0$ and $c^{\text{mul}} > 0$ such that, with probability at least $1 - n^{-3}$,

$$\Delta_n^{\text{add}} \leqslant c^{\text{add}} \sqrt{\log(|\Xi| n)/n_{\text{te}}} \cdot \widehat{\sigma}_{\Xi} \quad \textit{and} \quad \Delta_n^{\text{mul}} \leqslant c^{\text{mul}} \sqrt{\log(|\Xi| n)/n_{\text{te}}} \cdot \widehat{\kappa}_{\Xi}.$$

Remarks: The above bounds hold for arbitrary, even highly unbalanced, splits of the data into train and test sets, an essential feature that will be useful for our application to follow. Theorem 1 extends on existing results on cross-validation and model selection (e.g., [16–18]) in two important ways: (1) We derive two forms of inequalities: the additive form that proves to be useful when analyzing bounded loss functions (especially, classification losses); while the multiplicative form is useful for unbounded loss functions (especially, regression losses). (2) Instead of using sample mean to estimate the prediction risk as it common in most cross-validation procedures, we employ the median-of-means estimator that proves to be useful when the number of predictors under comparison grows with the sample size.

3 Risk monotonization in high dimensions

Theorem 1 allows to monotonize the risk function of an arbitrary procedure. Let $R(\widetilde{f}(\cdot;\mathcal{D}_n))$ be the prediction risk of the generic procedure $\widetilde{f}$ trained on a dataset $\mathcal{D}_n$ with sample size n and feature size p. We will rely on the proportional asymptotic framework in which $n, p \to \infty$ with the aspect ratio p/n converging to a constant $\gamma \in (0, \infty]$. As noted above, in such regime the asymptotic risk profile of $\widetilde{f}$ has been recently shown to be non-monotonic for a wide variety of problems and procedures. Leveraging the risk bound for cross-validation from Theorem 1, it is possible to modify the original procedure $\widetilde{f}$ and obtain a new procedure $\widetilde{f}^{\text{mon}}$ whose asymptotic risk profile is provably monotonic in γ (or $1/\gamma$; see Figure 1[1]) The basic idea is simple:

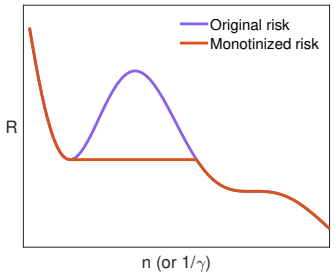

Figure 1: Risk monotonization

through sub-sampling we estimate the value of the asymptotic risk profile of $\widetilde{f}$ over a grid of aspect ratios larger than p/n, and then identify the aspect ratio for which the estimated prediction risk is the smallest. This is accomplished by fitting the original procedure $\widetilde{f}$ over sub-samples of varying size, with the size of the sub-sample treated as a tuning parameter to optimize over. The final procedure $\widetilde{f}^{\text{mon}}$ is then obtained though cross-validation, as described in Section 2. In detail, define the index set $\Xi := \{1, 2, \ldots, \lceil n_{\text{tr}}/\lfloor n^\nu \rfloor - 2 \rceil\}$ for some $\nu \in [0, 1)$ and, for each $\xi \in \Xi$, set $\widehat{f}^{\xi}(\cdot) = \widetilde{f}(\cdot; \mathcal{D}^{\xi}_{\text{tr}})$, where $\mathcal{D}^{\xi}_{\text{tr}}$ is a random subset of $\mathcal{D}_{\text{tr}}$ of size $n_\xi := n_{\text{tr}} - \xi\lfloor n^\nu \rfloor$. Finally, we let $\widetilde{f}^{\text{mon}}(\cdot; \mathcal{D}_n) = \widehat{f}^{\text{cv}}$. Our main result below shows that the resulting risk function for $\widetilde{f}^{\text{mon}}(\cdot; \mathcal{D}_n)$ is asymptotically non-decreasing in γ, under some mild assumptions on $\widetilde{f}$ and ℓ.

Theorem 2. *Suppose there exists a deterministic function $R^{\text{det}}(\cdot; \widetilde{f}) : (0, \infty] \to [0, \infty]$ such that for every $\gamma \in (0, \infty]$, $R(\widetilde{f}(\cdot; \mathcal{D}_n)) \xrightarrow{p} R^{\text{det}}(\gamma; \widetilde{f})$, whenever $n, p \to \infty$ and $p/n \to \gamma$, and the loss function ℓ is such that $\widehat{\sigma}_\Xi = O_p(1)$ or $\widehat{\kappa}_\Xi = O_p(1)$. Then, $|R(\widetilde{f}^{\text{mon}}) - \min_{\zeta \geqslant \gamma} R^{\text{det}}(\zeta; \widetilde{f})| \xrightarrow{p} 0$ as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

Remarks: (1) Since $\min_{\zeta \geqslant \gamma} R^{\text{det}}(\zeta; \widetilde{f}) \leqslant R^{\text{det}}(\gamma; \widetilde{f})$, the asymptotic risk of $\widetilde{f}^{\text{mon}}$ is no worse than that of $\widetilde{f}$. (2) The assumptions imposed for Theorem 2 are quite mild and our results are broadly applicable. Indeed, the risk profile $R^{\text{det}}(\cdot; \widetilde{f})$ of several estimators have been recently identified under under proportional asymptotics, [20–24], including the min ℓ_2-norm least squares estimator (MNLS) for linear and kernel regression [5, 13, 25, 26], min ℓ_1-norm interpolator [8], min ℓ_1-norm classifier [7], max-margin linear classifiers [27], among others. Our results are directly applicable to all these cases with minimal modifications. The requirements on the loss functions can be verified for common loss functions: e.g., for bounded losses, we can show that $\widehat{\sigma}_\Xi = O_p(1)$, while for unbounded squared regression loss, assuming $L_4 - L_2$ equivalence [28–30], we can show that $\widehat{\kappa}_\Xi = O_p(1)$ for linear predictors, even those with diverging risks such as MNLS when $n \approx p$.

Figure 2 illustrates our procedure with MNLS [13]. To reduce external randomness in choosing $\mathcal{D}^{\xi}_{\text{tr}}$ of size n_ξ, we also consider a variant where we set $\widehat{f}^{\xi} = \sum_{j=1}^{M} \widetilde{f}(x; \mathcal{D}^{\xi,j}_{\text{tr}})/M$ where $\mathcal{D}^{\xi,j}_{\text{tr}}, 1 \leqslant j \leqslant M$ are M randomly drawn sets of size n_ξ. We do not know the exact risk behavior of the resulting predictor for $M > 1$. From the theory of U-statistics [31], one can show that the risk for $M > 1$ is at most the risk for the predictor with $M = 1$. We do observe that the risk for $M > 1$ is monotone in the limiting aspect ratio in our experiments.

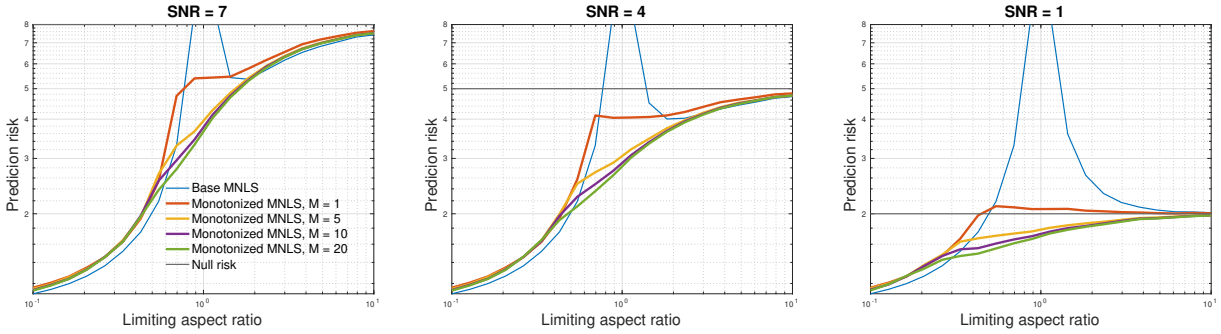

Figure 2: Our cross-validated procedure applied to MNLS as the base procedure with varying M, and high, moderate, and low SNR regimes. Here, $n = 1000$, $n_{\text{tr}} = 900$, $n_{\text{te}} = 100$, $n^\nu = 50$. Note that MNLS has unbounded risk near $\gamma = 1$, while risk of monotonized MNLS remains bounded for all $M \geqslant 1$ and all $\gamma > 0$.

[1] We thank [19] for figure, enabling our simultaneous attempt at illustration and comedic relief in the form of this footnote.

References

[1] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In search of the real inductive bias: On the role of implicit regularization in deep learning". In: *arXiv preprint arXiv:1412.6614* (2014).

[2] Preetum Nakkiran et al. "Deep double descent: Where bigger models and more data hurt". In: *arXiv preprint arXiv:1912.02292* (2019).

[3] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. "To understand deep learning we need to understand kernel learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 541–549.

[4] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[5] Song Mei and Andrea Montanari. "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". In: *Communications on Pure and Applied Mathematics* (2019).

[6] Peter L. Bartlett et al. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[7] Tengyuan Liang and Pragya Sur. "A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum ℓ_1-Norm Interpolated Classifiers". In: *arXiv preprint arXiv:2002.01586* (2020).

[8] Yue Li and Yuting Wei. "Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent". In: *arXiv preprint arXiv:2110.09502* (2021).

[9] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification". In: *arXiv preprint arXiv:1911.05822* (2019).

[10] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *arXiv preprint arXiv:2103.09177* (2021).

[11] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. "A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning". In: *arXiv preprint arXiv:2109.02355* (2021).

[12] Preetum Nakkiran et al. "Optimal regularization can mitigate double descent". In: *arXiv preprint arXiv:2003.01897* (2020).

[13] Trevor Hastie et al. "Surprises in high-dimensional ridgeless least squares interpolation". In: *arXiv preprint arXiv:1903.08560* (2019).

[14] Luc Devroye et al. "Sub-Gaussian mean estimators". In: *The Annals of Statistics* 44.6 (2016), pp. 2695–2725.

[15] Gábor Lugosi and Shahar Mendelson. "Mean estimation and regression under heavy-tailed distributions: A survey". In: *Foundations of Computational Mathematics* 19.5 (2019), pp. 1145–1190.

[16] Aad W. Van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. "Oracle inequalities for multi-fold cross validation". In: *Statistics & Decisions* 24.3 (2006), pp. 351–371.

[17] Mark J. Van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[18] Yuhong Yang. "Consistency of cross validation for comparing regression procedures". In: *The Annals of Statistics* 35.6 (2007), pp. 2450–2473.

[19] TOPML Organizing Committee. *Workshop of the Theory of Overparameterized Machine Learning*. 2022. URL: https://cpb-us-e1.wpmucdn.com/blogs.rice.edu/dist/c/12072/files/2021/12/cropped-TOPML2022_logo.png (visited on 02/17/2022).

[20] Noureddine El Karoui. "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results". In: *arXiv preprint arXiv:1311.2445* (2013).

[21] Noureddine El Karoui. "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators". In: *Probability Theory and Related Fields* 170.1 (2018), pp. 95–175.

[22] David Donoho and Andrea Montanari. "High dimensional robust m-estimation: Asymptotic variance via approximate message passing". In: *Probability Theory and Related Fields* 166.3 (2016), pp. 935–969.

[23] Léo Miolane and Andrea Montanari. "The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning". In: *The Annals of Statistics* 49.4 (2021), pp. 2313–2335.

[24] Michael Celentano, Andrea Montanari, and Yuting Wei. "The Lasso with general Gaussian designs with applications to hypothesis testing". In: *arXiv preprint arXiv:2007.13716* (2020).

[25] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features". In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1167–1180.

[26] Tengyuan Liang and Alexander Rakhlin. "Just interpolate: Kernel "ridgeless" regression can generalize". In: *The Annals of Statistics* 48.3 (2020), pp. 1329–1347.

[27] Andrea Montanari et al. "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime". In: *arXiv preprint arXiv:1911.01544* (2019).

[28] Stanislav Minsker and Xiaohan Wei. "Robust modifications of U-statistics and applications to covariance estimation problems". In: *Bernoulli* 26.1 (2020), pp. 694–727.

[29] Stanislav Minsker. "Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries". In: *The Annals of Statistics* 46.6A (2018), pp. 2871–2903.

[30] Shahar Mendelson and Nikita Zhivotovskiy. "Robust covariance estimation under L_4-L_2 norm equivalence". In: *The Annals of Statistics* 48.3 (2020), pp. 1648–1664.

[31] Robert J. Serfling. *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons, 2009.