Cross-validation Confidence Intervals for Test Error

Pierre Bayle[*]
Princeton University
pbayle@princeton.edu

Alexandre Bayle[*]
Harvard University
alexandre_bayle@g.harvard.edu

Lucas Janson
Harvard University
ljanson@fas.harvard.edu

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

Abstract

This work develops central limit theorems for cross-validation and consistent estimators of its asymptotic variance under weak stability conditions on the learning algorithm. Together, these results provide practical, asymptotically-exact confidence intervals for k-fold test error and valid, powerful hypothesis tests of whether one learning algorithm has smaller k-fold test error than another. These results are also the first of their kind for the popular choice of leave-one-out cross-validation. In our real-data experiments with diverse learning algorithms, the resulting intervals and tests outperform the most popular alternative methods from the literature.

1 Introduction

Cross-validation (CV) [49, 26] is a de facto standard for estimating the test error of a prediction rule. By partitioning a dataset into k equal-sized validation sets, fitting a prediction rule with each validation set held out, evaluating each prediction rule on its corresponding held-out set, and averaging the k error estimates, CV produces an unbiased estimate of the test error with lower variance than a single train-validation split could provide. However, these properties alone are insufficient for high-stakes applications in which the uncertainty of an error estimate impacts decision-making. In predictive cancer prognosis and mortality prediction for instance, scientists and clinicians rely on *test error confidence intervals (CIs)* based on CV and other repeated sample splitting estimators to avoid spurious findings and improve reproducibility [42, 45]. Unfortunately, the CIs most often used have no correctness guarantees and can be severely misleading [30]. The difficulty comes from the dependence across the k averaged error estimates: if the estimates were independent, one could derive an *asymptotically-exact CI* (i.e., a CI with coverage converging exactly to the target level) for test error using a standard central limit theorem. However, the error estimates are seldom independent, due to the overlap amongst training sets and between different training and validation sets. Thus, new tools are needed to develop valid, informative CIs based on CV.

The same uncertainty considerations are relevant when comparing two machine learning methods: before selecting a prediction rule for deployment, one would like to be confident that its test error is better than a baseline or an available alternative. The standard practice amongst both method developers and consumers is to conduct a formal *hypothesis test* for a difference in test error between two prediction rules [22, 38, 43, 13, 19]. Unfortunately, the most popular tests from the literature like the cross-validated t-test [22], the repeated train-validation t-test [43], and the 5×2 CV test [22] have no correctness guarantees and hence can produce misleading conclusions. The difficulty parallels that of the confidence interval setting: standard tests assume independence and do not

[*] Equal contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

appropriately account for the dependencies across CV error estimates. Therefore, new tools are also needed to develop valid, powerful tests for test error improvement based on CV.

Our contributions To meet these needs, we characterize the asymptotic distribution of CV error and develop consistent estimates of its variance under weak stability conditions on the learning algorithm. Together, these results provide practical, asymptotically-exact confidence intervals for test error as well as valid and powerful hypothesis tests of whether one learning algorithm has smaller test error than another. In more detail, we prove in Sec. 2 that k-fold CV error is asymptotically normal around its test error under an abstract asymptotic linearity condition. We then give in Sec. 3 two different stability conditions that hold for large classes of learning algorithms and losses and that individually imply the asymptotic linearity condition. In Sec. 4, we propose two estimators of the asymptotic variance of CV and prove them to be consistent under similar stability conditions; our second estimator accommodates any choice of k and appears to be the first consistent variance estimator for leave-one-out CV. To validate our theory in Sec. 5, we apply our intervals and tests to a diverse collection of classification and regression methods on particle physics and flight delay data and observe consistent improvements in width and power over the most popular alternative methods from the literature.

Related work Despite the ubiquity of CV, we are only aware of three prior efforts to characterize the precise distribution of cross-validation error. The CV central limit theorem (CLT) of Dudoit and van der Laan [23] requires considerably stronger assumptions than our own and is not paired with the consistent estimate of variance needed to construct a valid confidence interval or test. LeDell et al. [36] derive both a CLT and a consistent estimate of variance for CV, but these apply only to the area under the ROC curve (AUC) performance measure. Finally, in very recent work, Austern and Zhou [5] derive a CLT and a consistent estimate of variance for CV under more stringent assumptions than our own. We compare our results with each of these works in detail in Sec. 3.3. We note also that another work [37] aims to test the difference in test error between two learning algorithms using cross-validation but only proves the validity of their procedure for a single train-validation split rather than for CV. Many other works have studied the problem of bounding or estimating the variance of the cross-validation error [11, 43, 9, 39, 31, 34, 16, 2, 3], but none have established the consistency of their variance estimators. Among these, Kale et al. [31], Kumar et al. [34], Celisse and Guedj [16] introduce relevant notions of algorithmic stability to which we link our results in Sec. 3.1. Moreover, non-asymptotic CIs can be derived from the CV concentration inequalities of [14, 18, 16, 2, 3], but these CIs are more difficult to deploy as they require (1) stronger stability assumptions, (2) a known upper bound on stability, and (3) either a known upper bound on the loss or a known uniform bound on the covariates and a known sub-Gaussianity constant for the response variable. In addition, the reliance on somewhat loose inequalities typically leads to overly large, relatively uninformative CIs. For example, we implemented the ridge regression CI of [16, Thm. 3] for the regression experiment of Sec. 5.1 (see App. K.3). When the features are standardized, the narrowest concentration-based interval is 91 times wider than our widest CLT interval in Fig. 5. Without standardization, the narrowest concentration-based interval is 5×10^{14} times wider.

Notation Let $\overset{d}{\to}$, $\overset{p}{\to}$, and $\overset{L^q}{\to}$ for $q > 0$, denote convergence in distribution, in probability, and in L^q norm (i.e., $X_n \overset{L^q}{\to} X \Leftrightarrow \mathbb{E}[|X_n - X|^q] \to 0$), respectively. For each $m, n \in \mathbb{N}$ with $m \leq n$, we define the set $[n] \triangleq \{1, \dots, n\}$ and the vector $m\!:\!n \triangleq (m, \dots, n)$. When considering independent random elements (X, Y), we use $\mathbb{E}_X$ and Var_X to indicate expectation and variance only over X, respectively; that is, $\mathbb{E}_X[f(X,Y)] \triangleq \mathbb{E}[f(X,Y) \mid Y]$ and $\mathrm{Var}_X(f(X,Y)) \triangleq \mathrm{Var}(f(X,Y) \mid Y)$. We will refer to the Euclidean norm of a vector as the ℓ^2 norm in the context of ℓ^2 regularization.

2 A Central Limit Theorem for Cross-validation

In this section, we present a new central limit theorem for k-fold cross-validation. Throughout, any asymptotic statement will take $n \to \infty$, and while we allow the number of folds k_n to depend on the sample size n (e.g., $k_n = n$ for leave-one-out cross-validation), we will write k in place of k_n to simplify our notation. We will also present our main results assuming that k evenly divides n, but we address the indivisible setting in the appendix.

Hereafter, we will refer to a sequence $(Z_i)_{i\geq 1}$ of random datapoints taking values in a set $\mathcal{Z}$. Notably, $(Z_i)_{i\geq 1}$ need not be independent or identically distributed. We let $Z_{1:n}$ designate the first n points, and, for any vector B of indices in $[n]$, we let Z_B denote the subvector of $Z_{1:n}$ corresponding to ordered indices in B. We will also refer to *train-validation splits* (B, B'). These are vectors of indices in $[n]$ representing the ordered points assigned to the training set and validation set.[2] As is typical in CV, we will assume that B and B' partition $[n]$, so that every datapoint is either in the training or validation set.

Given a scalar loss function $h_n(Z_i, Z_B)$ and a set of k train-validation splits $\{(B_j, B'_j)\}_{j=1}^k$ with validation indices $\{B'_j\}_{j=1}^k$ partitioning $[n]$ into k folds, we will use the *k-fold cross-validation error*

$$\hat{R}_n \triangleq \tfrac{1}{n}\textstyle\sum_{j=1}^k \sum_{i\in B'_j} h_n(Z_i, Z_{B_j})$$

to draw inferences about the *k-fold test error*

$$R_n \triangleq \tfrac{1}{n}\textstyle\sum_{j=1}^k \sum_{i\in B'_j} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]. \tag{2.1}$$

A prototypical example of h_n is squared error or 0-1 loss,

$$h_n(Z_i, Z_B) = (Y_i - \hat{f}(X_i; Z_B))^2 \quad \text{or} \quad h_n(Z_i, Z_B) = \mathbb{1}[Y_i \neq \hat{f}(X_i; Z_B)],$$

composed with an algorithm for fitting a prediction rule $\hat{f}(\cdot; Z_B)$ to training data Z_B and predicting the response value of a test point $Z_i = (X_i, Y_i)$.[3] In this setting, the k-fold test error is a standard inferential target [11, 23, 31, 34, 5] and represents the average test error of the k prediction rules $\hat{f}(\cdot; Z_{B_j})$. When comparing the performance of two algorithms in Secs. 4 and 5, we will choose h_n to be the difference between the losses of two prediction rules.

2.1 Asymptotic linearity of cross-validation

The key to our central limit theorem is establishing that the k-fold CV error asymptotically behaves like the k-fold test error plus an average of functions applied to single datapoints. The following proposition provides a convenient characterization of this *asymptotic linearity* property.

Proposition 1 (Asymptotic linearity of k-fold CV)**.** *For any sequence of datapoints* $(Z_i)_{i\geq 1}$*,*

$$\tfrac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n) - \tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{i=1}^n\big(\bar{h}_n(Z_i) - \mathbb{E}\big[\bar{h}_n(Z_i)\big]\big) \overset{p}{\to} \Big(\textit{resp.} \overset{L^q}{\to}\Big)\ 0$$

for a function $\bar{h}_n$ *with* $\sigma_n^2 \triangleq \frac{1}{n}\mathrm{Var}(\sum_{i=1}^n \bar{h}_n(Z_i))$ *if and only if*

$$\begin{aligned}\tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{j=1}^k \sum_{i\in B'_j}\Big(&h_n\big(Z_i, Z_{B_j}\big) - \mathbb{E}\big[h_n\big(Z_i, Z_{B_j}\big) \mid Z_{B_j}\big] \\ &- \big(\bar{h}_n(Z_i) - \mathbb{E}\big[\bar{h}_n(Z_i)\big]\big)\Big) \overset{p}{\to} \Big(\textit{resp.} \overset{L^q}{\to}\Big)\ 0,\end{aligned} \tag{2.2}$$

where the parenthetical convergence indicates that the same statement holds when both convergences in probability are replaced with convergences in L^q for the same $q > 0$.

Typically, one will choose $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:n(1-1/k)})]$. With this choice, we see that the difference of differences in (2.2) is small whenever $h_n(Z_i, Z_{B_j})$ is close to *either* its expectation given Z_i *or* its expectation given Z_{B_j}, but it need not be close to both. As the asymptotic linearity condition (2.2) is still quite abstract, we devote all of Sec. 3 to establishing sufficient conditions for (2.2) that are interpretable, broadly applicable, and simple to verify. Prop. 1 follows from a more general asymptotic linearity characterization for repeated sample-splitting estimators proved in App. A.

2.2 From asymptotic linearity to asymptotic normality

So far, we have assumed nothing about the dependencies amongst the datapoints Z_i. If we additionally assume that the datapoints are i.i.d., the average $\frac{1}{\sigma_n\sqrt{n}}\sum_{i=1}^n\big(\bar{h}_n(Z_i) - \mathbb{E}\big[\bar{h}_n(Z_i)\big]\big)$ converges to a standard normal under a mild integrability condition, and we obtain the following CLT for CV.

[2] We keep track of index order to support asymmetric learning algorithms like stochastic gradient descent.

[3] For randomized learning algorithms (such as random forests or stochastic gradient descent), all statements in this paper should be treated as holding conditional on the external source of randomness.

Theorem 1 (Asymptotic normality of k-fold CV with i.i.d. data)**.** *In the notation of Prop. 1, suppose that the datapoints $(Z_i)_{i \geq 1}$ are i.i.d. copies of a random element Z_0 and that the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ with $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$ is uniformly integrable (UI). If the asymptotic linearity condition (2.2) holds in probability then*

$$\tfrac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n) \xrightarrow{d} \mathcal{N}(0, 1).$$

Thm. 1 is a special case of a more general result, proved in App. B, that applies when the datapoints are independent but not necessarily identically distributed. A simple sufficient condition for the required uniform integrability is that $\sup_n \mathbb{E}[|(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])/\sigma_n|^\alpha] < \infty$ for some $\alpha > 2$. This holds, for example, whenever $\bar{h}_n(Z_0)$ has uniformly bounded α moments (e.g., the 0-1 loss has all moments uniformly bounded) and does not converge to a degenerate distribution. We now turn our attention to the asymptotic linearity condition.

3 Sufficient Conditions for Asymptotic Linearity

3.1 Asymptotic linearity from loss stability

Our first result relates the asymptotic linearity of CV to a specific notion of algorithmic stability, termed *loss stability*.

Definition 1 (Mean-square stability and loss stability)**.** *For $m > 0$, let Z_0 and $Z_0', Z_1, \ldots, Z_m$ be i.i.d. test and training points with $Z_{1:m}^{\setminus i}$ representing $Z_{1:m}$ with Z_i replaced by Z_0'. For any function $h : \mathcal{Z} \times \mathcal{Z}^m \to \mathbb{R}$, the* mean-square stability *[31] is defined as*

$$\gamma_{ms}(h) \triangleq \tfrac{1}{m} \textstyle\sum_{i=1}^m \mathbb{E}[(h(Z_0, Z_{1:m}) - h(Z_0, Z_{1:m}^{\setminus i}))^2] \tag{3.1}$$

and the loss stability *[34] as $\gamma_{loss}(h) \triangleq \gamma_{ms}(h')$, where*

$$h'(Z_0, Z_{1:m}) \triangleq h(Z_0, Z_{1:m}) - \mathbb{E}[h(Z_0, Z_{1:m}) \mid Z_{1:m}].$$

Kumar et al. [34] introduced loss stability to bound the variance of CV in terms of the variance of a single hold-out set estimate. Here we show that a suitable decay in loss stability is also sufficient for L^2 asymptotic linearity, via a non-asymptotic bound on the departure from linearity.

Theorem 2 (Approximate linearity from loss stability)**.** *In the notation of Prop. 1 and Def. 1, suppose that the datapoints $(Z_i)_{i \geq 1}$ are i.i.d. copies of a random element Z_0. Then*

$$\mathrm{Var}(\tfrac{1}{\sqrt{n}} \textstyle\sum_{j=1}^k \sum_{i \in B_j'} (h_n'(Z_i, Z_{B_j}) - \mathbb{E}[h_n'(Z_i, Z_{B_j}) \mid Z_i])) \leq \tfrac{3}{2} n (1 - \tfrac{1}{k}) \gamma_{loss}(h_n). \tag{3.2}$$

Hence the L^2 asymptotic linearity condition (2.2) holds with $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:n(1-1/k)})]$ if the loss stability satisfies $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$.

The proof of Thm. 2 is given in App. C. Recall that in a typical learning context, we have $h_n(Z_0, Z_{1:m}) = \ell(Y_0, \hat{f}(X_0; Z_{1:m}))$ for a fixed loss ℓ, a learned prediction rule $\hat{f}(\cdot; Z_{1:m})$, a test point $Z_0 = (X_0, Y_0)$, and $m = n(1 - 1/k)$. When $\hat{f}(\cdot; Z_{1:m})$ converges to an imperfect prediction rule, we will commonly have $\sigma_n^2 = \mathrm{Var}(\mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_0]) = \Omega(1)$ so that $\gamma_{loss}(h_n) = o(1/n)$ loss stability is sufficient. However, Thm. 2 also accommodates the cases of non-convergent $\hat{f}(\cdot; Z_{1:m})$ and of $\hat{f}(\cdot; Z_{1:m})$ converging to a perfect prediction rule, so that $\sigma_n^2 = o(1)$.

Many learning algorithms are known to enjoy decaying loss stability [14, 25, 29, 16, 4], in part because loss stability is upper-bounded by a variety of algorithmic stability notions studied in the literature. For example, stochastic gradient descent on convex and non-convex objectives [29] and the empirical risk minimization of a strongly convex and Lipschitz objective both have $O(1/n)$ *uniform stability* [14] which implies a loss stability of $O(1/n^2) = o(1/n)$ by [31, Lem. 1] and [34, Lem. 2]. However, we emphasize that the loss h_n need not be convex and need not coincide with a loss function used to train a learning method. Indeed, our stability assumptions also cover k-nearest neighbor methods [20], decision tree methods [4], and ensemble methods [25] and can even hold when training error is a poor proxy for test error due to overfitting (e.g., 1-nearest neighbor has training error 0 but is still suitably stable [20]). In addition, for any loss function, loss stability is upper-bounded by mean-square stability [31] and all L^q stabilities [16] for $q \geq 2$. For bounded loss functions such as the 0-1 loss, loss stability is also weaker than hypothesis stability (also called L^1 stability) [20, 32], weak-hypothesis stability [21], and weak-L^1 stability [35].

3.2 Asymptotic linearity from conditional variance convergence

We can also guarantee asymptotic linearity under weaker moment conditions than Thm. 2 at the expense of stronger requirements on the number of folds k.

Theorem 3 (Asymptotic linearity from conditional variance convergence)**.** *In the notation of Prop. 1, suppose that the datapoints $(Z_i)_{i\geq 1}$ are i.i.d. copies of a random element Z_0. If*

$$\max(k^{q/2}, k^{1-q/2})\mathbb{E}\Big[\big(\tfrac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{1:n(1-1/k)}) - \bar{h}_n(Z_0)\big)\big)^{q/2}\Big] \to 0 \tag{3.3}$$

for a function $\bar{h}_n$ and $q \in (0, 2]$, then $\bar{h}_n$ satisfies the L^q asymptotic linearity condition (2.2). If

$$\mathbb{E}\Big[\min\Big(k, \tfrac{\sqrt{k}}{\sigma_n}\sqrt{\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{1:n(1-1/k)}) - \bar{h}_n(Z_0)\big)}\Big)\Big] \to 0. \tag{3.4}$$

for a function $\bar{h}_n$, then $\bar{h}_n$ satisfies the in-probability asymptotic linearity condition (2.2).

Remark 1. *When $k = O(1)$, (3.4) holds $\Leftrightarrow \frac{1}{\sigma_n}\sqrt{\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{1:n(1-1/k)}) - \bar{h}_n(Z_0)\big)} \xrightarrow{p} 0$.*

Thm. 3 follows from a more general statement proved in App. D. When k is bounded, as in 10-fold CV, the conditions of Thm. 3 are considerably weaker than those of Thm. 2 (see App. E), granting asymptotic linearity whenever the conditional variance converges in probability rather than in L^2. Indeed in App. G, we detail a simple learning problem in which the loss stability is infinite but Thms. 1 and 3 together provide a valid CLT with convergent variance σ_n^2.

3.3 Comparison with prior work

Our sufficient conditions for asymptotic normality are significantly less restrictive and more broadly applicable than the three prior distributional characterizations of CV error [23, 36, 5]. In particular, the CLT of Dudoit and van der Laan [23, Thm. 3] assumes a bounded loss function, excludes the popular case of leave-one-out cross-validation, and requires the prediction rule to be loss-consistent for a risk-minimizing prediction rule. Similarly, the CLT of LeDell et al. [36, Thm. 4.1] applies only to AUC loss, requires the prediction rule to be loss-consistent for a deterministic prediction rule, and requires a bounded number of folds.

Moreover, in our notation, the recent CLT of Austern and Zhou [5, Thm. 1] restricts focus to learning algorithms that treat all training points symmetrically, assumes that its variance parameter

$$\tilde{\sigma}_n^2 \triangleq \mathbb{E}[\mathrm{Var}(h_n(Z_0, Z_{1:m}) \mid Z_{1:m})] \tag{3.5}$$

converges to a non-zero limit, requires mean-square stability $\gamma_{ms}(h_n) = o(1/n)$, and places a $o(1/n^2)$ constraint on the second-order mean-square stability

$$\mathbb{E}[((h_n(Z_0, Z_{1:m}) - h_n(Z_0, Z_{1:m}^{\setminus 1})) - (h_n(Z_0, Z_{1:m}^{\setminus 2}) - h_n(Z_0, Z_{1:m}^{\setminus 1,2})))^2] = o(1/n^2), \tag{3.6}$$

where $Z_{1:m}^{\setminus 1,2}$ represents $Z_{1:m}$ with Z_1, Z_2 replaced by i.i.d. copies Z_1', Z_2'. Kumar et al. [34] showed that the mean-square stability is always an upper bound for the loss stability required by our Thm. 2, and in Apps. F and G we exhibit two simple learning tasks in which $\gamma_{loss}(h_n) = O(1/n^2)$ but $\gamma_{ms}(h_n) = \infty$. Furthermore, when k is constant, as in 10-fold CV, our conditional variance assumptions in Sec. 3.2 are weaker still and hold even for algorithms with infinite loss stability (see App. G). In addition, our results allow for asymmetric learning algorithms (like stochastic gradient descent), accommodate growing, vanishing, and non-convergent variance parameters σ_n^2, and do not require the second-order mean-square stability condition (3.6).

Finally, we note that the asymptotic variance parameter σ_n^2 appearing in Thm. 1 is never larger and sometimes smaller than the variance parameter $\tilde{\sigma}_n^2$ in [5, Thm. 1].

Proposition 2 (Variance comparison)**.** *Let $\sigma_n^2 = \mathrm{Var}(\mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_0])$ be the variance appearing in Thm. 1, with the choice $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:m})]$, and $\tilde{\sigma}_n^2 = \mathbb{E}[\mathrm{Var}(h_n(Z_0, Z_{1:m}) \mid Z_{1:m})]$ be the variance parameter of [5, Eq. (15)] for $m = n(1 - 1/k)$. Then*

$$\sigma_n^2 \leq \tilde{\sigma}_n^2 \leq \sigma_n^2 + \tfrac{m}{2}\gamma_{loss}(h_n),$$

and the first inequality is strict whenever $h(Z_0, Z_{1:m}) - \mathbb{E}[h(Z_0, Z_{1:m}) \mid Z_{1:m}]$ depends on $Z_{1:m}$.

The proof of Prop. 2 can be found in App. H. In App. G, we present a simple learning task for which our central limit theorem provably holds with σ_n^2 converging to a non-zero constant, but the CLT in [5, Eq. (15)] is inapplicable because the variance parameter $\tilde{\sigma}_n^2$ is infinite.

4 Confidence Intervals and Tests for k-fold Test Error

A primary application of our central limit theorems is the construction of asymptotically-exact confidence intervals for the unknown k-fold test error. For example, under the assumptions and notation of Thm. 1, any sample statistic $\hat{\sigma}_n^2$ satisfying relative error consistency, $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$, gives rise to an asymptotically-exact $(1-\alpha)$-confidence interval,

$$C_\alpha \triangleq \hat{R}_n \pm q_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n} \quad \text{satisfying} \quad \lim_{n\to\infty} \mathbb{P}(R_n \in C_\alpha) = 1-\alpha, \tag{4.1}$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution.

A second, related application of our central limit theorems is testing whether, given a dataset $Z_{1:n}$, a k-fold partition $\{B'_j\}_{j=1}^k$, and two algorithms $\mathcal{A}_1$, $\mathcal{A}_2$ for fitting prediction rules, $\mathcal{A}_2$ has larger k-fold test error than $\mathcal{A}_1$. In this circumstance, we may define

$$h_n(Z_0, Z_B) = \ell(Y_0, \hat{f}_1(X_0; Z_B)) - \ell(Y_0, \hat{f}_2(X_0; Z_B))$$

to be the difference of the loss functions of two prediction rules trained on Z_B and tested on $Z_0 = (X_0, Y_0)$. Our aim is to test whether $\mathcal{A}_1$ improves upon $\mathcal{A}_2$ on the fold partition, that is to test the null $H_0 : R_n \geq 0$ against the alternative hypothesis $H_1 : R_n < 0$. Under the assumptions and notation of Thm. 1, an asymptotically-exact level-α test is given by[4]

$$\text{REJECT } H_0 \Leftrightarrow \hat{R}_n < q_\alpha \hat{\sigma}_n/\sqrt{n} \tag{4.2}$$

where q_α is the α-quantile of a standard normal distribution and $\hat{\sigma}_n^2$ is any variance estimator satisfying relative error consistency, $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$. Fortunately, our next theorem describes how to compute such a consistent estimate of σ_n^2 under weak conditions.

Theorem 4 (Consistent within-fold estimate of asymptotic variance)**.** *In the notation of Thm. 1 with $m = n(1-1/k)$, $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:m})]$, and $k < n$, define the within-fold variance estimator*

$$\hat{\sigma}_{n,in}^2 \triangleq \tfrac{1}{k}\sum_{j=1}^k \tfrac{1}{(n/k)-1}\sum_{i\in B'_j}\Big(h_n(Z_i, Z_{B_j}) - \tfrac{k}{n}\sum_{i'\in B'_j} h_n(Z_{i'}, Z_{B_j})\Big)^2.$$

Suppose $(Z_i)_{i\geq 1}$ are i.i.d. copies of a random element Z_0. Then $\hat{\sigma}_{n,in}^2/\sigma_n^2 \xrightarrow{L^1} 1$ whenever $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ and the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable (UI). Moreover, $\hat{\sigma}_{n,in}^2/\sigma_n^2 \xrightarrow{L^2} 1$ whenever $\mathbb{E}[((\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])/\sigma_n)^4] = o(n)$ and the fourth-moment loss stability $\gamma_4(h'_n) \triangleq \frac{1}{m}\sum_{i=1}^m \mathbb{E}[(h'_n(Z_0, Z_{1:m}) - h'_n(Z_0, Z_{1:m}^{\setminus i}))^4] = o(\sigma_n^4/n^2)$*. Here, $Z_{1:m}^{\setminus i}$ denotes $Z_{1:m}$ with Z_i replaced by an identically distributed copy independent of $Z_{0:m}$.*

Thm. 4 follows from explicit error bounds proved in App. I. A notable take-away is that the same two conditions—loss stability $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ and a UI sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$—grant both a central limit theorem for CV (by Thms. 1 and 2) and an L^1-consistent estimate of σ_n^2 (by Thm. 4). Moreover, the L^2-consistency bound of App. I can be viewed as a strengthening of the consistency result of [5, Prop. 1] which analyzes the same variance estimator under more stringent assumptions. In our notation, to establish L^2 consistency, [5, Prop. 1] additionally requires h_n symmetric in its training points, convergence of the variance parameter $\tilde{\sigma}_n^2$ (3.5) to a non-zero constant, control over a fourth-moment analogue of mean-square stability $\gamma_4(h_n) = o(\sigma_n^4/n^2)$ instead of the smaller fourth-moment loss stability $\gamma_4(h'_n)$, and the more restrictive fourth-moment condition $\mathbb{E}[(h_n(Z_0, Z_{1:m})/\sigma_n)^4] = O(1)$.[5] By Prop. 2, their assumptions further imply that σ_n^2 converges to a non-zero constant. In contrast, Thm. 4 accommodates growing, vanishing, and non-convergent variance parameters σ_n^2 and a wider variety of learning procedures and losses.

Since Thm. 4 necessarily excludes the case of leave-one-out CV ($k = n$), we propose a second estimator with consistency guarantees for any k and only slightly stronger stability conditions than Thm. 4 when $k = \Omega(n)$. Notably, Austern and Zhou [5] do not provide a consistent variance estimator for $k = n$, and Dudoit and van der Laan [23] do not establish the consistency of any variance estimator.

[4] The test (4.2) is equivalent to rejecting when the one-sided interval $(-\infty, \hat{R}_n - q_\alpha\hat{\sigma}_n/\sqrt{n}\,]$ excludes 0.

[5] The result [5, Prop. 1] also assumes a fourth moment second-order stability condition similar to (3.6), but this appears to not be used in the proof.

Theorem 5 (Consistent all-pairs estimate of asymptotic variance)**.** *Under the notation of Thm. 1 with $m = n(1 - 1/k)$, and $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:m})]$, define the all-pairs variance estimator*

$$\hat{\sigma}^2_{n,out} \triangleq \frac{1}{k}\sum_{j=1}^{k} \frac{k}{n} \sum_{i \in B'_j} (h_n(Z_i, Z_{B_j}) - \hat{R}_n)^2.$$

If $(Z_i)_{i \geq 1}$ are i.i.d. copies of a random element Z_0, then $\hat{\sigma}^2_{n,out}/\sigma^2_n \xrightarrow{L^1} 1$ whenever $\gamma_{loss}(h_n) = o(\sigma^2_n/n)$, $\gamma_{ms}(h_n) = o(k\sigma^2_n/n)$, and the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma^2_n$ is UI.

Thm. 5 follows from an explicit error bound proved in App. J and differs from the L^1-consistency result of Thm. 4 only in the added requirement $\gamma_{ms}(h_n) = o(k\sigma^2_n/n)$. This mean-square stability condition is especially mild when $k = \Omega(n)$ (as in the case of leave-one-out CV) and ensures that two training sets differing in only n/k points produce prediction rules with comparable test losses.

Importantly, both $\hat{\sigma}^2_{n,in}$ and $\hat{\sigma}^2_{n,out}$ can be computed in $O(n)$ time using just the individual datapoint losses $h_n(Z_i, Z_{B_j})$ outputted by a run of k-fold cross-validation. Moreover, when h_n is binary, as in the case of 0-1 loss, one can compute $\hat{\sigma}^2_{n,out} = \hat{R}_n(1 - \hat{R}_n)$ in $O(1)$ time given access to the overall cross-validation error $\hat{R}_n$ and $\hat{\sigma}^2_{n,in} = \frac{1}{k}\sum_{j=1}^{k} \frac{(n/k)}{(n/k)-1}\hat{R}_{n,j}(1 - \hat{R}_{n,j})$ in $O(k)$ time given access to the k average fold errors $\hat{R}_{n,j} \triangleq \frac{k}{n}\sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$.

5 Numerical Experiments

In this section, we compare our test error confidence intervals (4.1) and tests for algorithm improvement (4.2) with the most popular alternatives from the literature: the hold-out test described in [5, Eq. (17)] based on a single train-validation split, the cross-validated t-test [22], the repeated train-validation t-test [43] (with and without correction), and the 5 × 2-fold CV test [22].[6] These procedures are commonly used and admit both two-sided CIs and one-sided tests, but, unlike our proposals, none except the hold-out method are known to be valid. Our aim is to verify whether our proposed procedures outperform these popular heuristics across a diversity of settings encountered in real learning problems. We fix $k = 10$, use 90-10 train-validation splits for all tests save 5×2-fold CV, and report our results using $\hat{\sigma}^2_{n,out}$ (as $\hat{\sigma}^2_{n,in}$ results are nearly identical).

Evaluating the quality of CIs and tests requires knowledge of the target test error.[7] In each experiment, we use points subsampled from a large real dataset to form a surrogate ground-truth estimate of the test error. Then, we evaluate the CIs and tests constructed from 500 training sets of sample sizes n ranging from 700 to $11,000$ subsampled from the same dataset. Each mean width estimate is displayed with a ± 2 standard error confidence band. The surrounding confidence bands for the coverage, size, and power estimates are 95% Wilson intervals [51], which are known to provide more accurate coverage for binomial proportions than a ± 2 standard error interval [15]. We use the `Higgs` dataset of [6, 7] to study the classification error of random forest, neural network, and ℓ^2-penalized logistic regression classifiers and the Kaggle `FlightDelays` dataset of [1] to study the mean-squared regression error of random forest, neural network, and ridge regression. In each case, we focus on stable settings of these learning algorithms with sufficiently strong ℓ^2 regularization for the neural network, logistic, and ridge learners and small depths for the random forest trees. Complete experimental details are available in App. K.1, and code replicating all experiments can be found at `https://github.com/alexandre-bayle/cvci`.

5.1 Confidence intervals for test error

In App. L.1, we compare the coverage and width of each procedure's 95% CI for each of the described algorithms, datasets, and training set sizes. Two representative examples—logistic regression classification and random forest regression—are displayed in Fig. 1. While the repeated train-validation CI significantly undercovers in all cases, all remaining CIs have coverage near the 95%

[6] We exclude McNemar's test [40] and the difference-of-proportions test which Dietterich [22] found to be less powerful than 5 × 2-fold CV and the conservative Z-test which Nadeau and Bengio [43] found less powerful and more expensive than corrected repeated train-validation splitting.

[7] Generalizing the notion of k-fold test error (2.1), we define the target test error for each testing procedure to be the average test error of the learned prediction rules; see App. K.2 for more details.

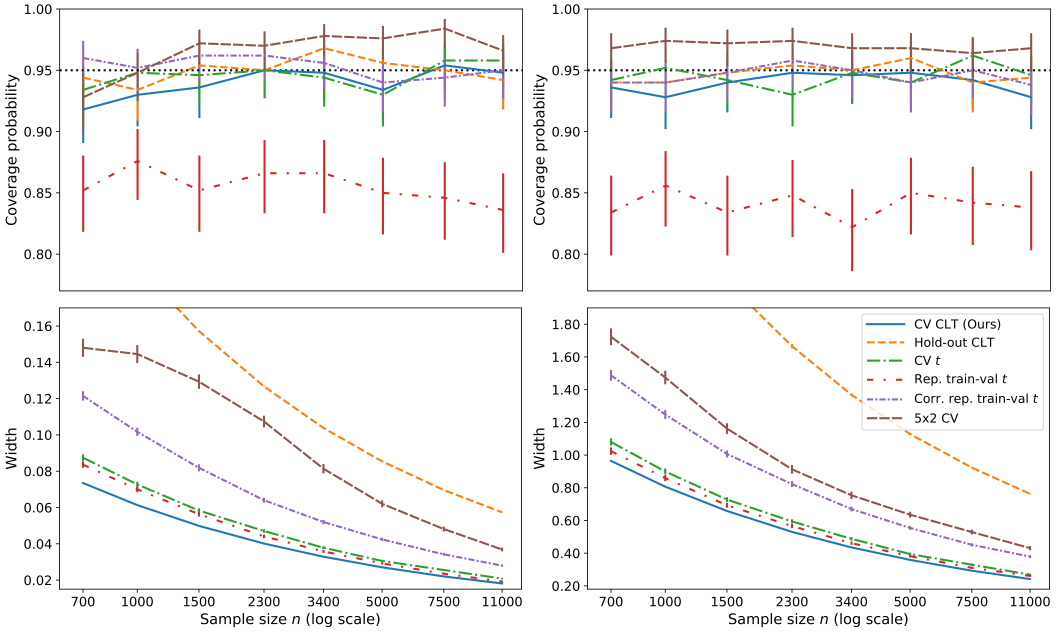


Figure 1: Test error coverage (top) and width (bottom) of 95% confidence intervals (see Sec. 5.1). **Left:** ℓ^2-regularized logistic regression classifier. **Right:** Random forest regression.

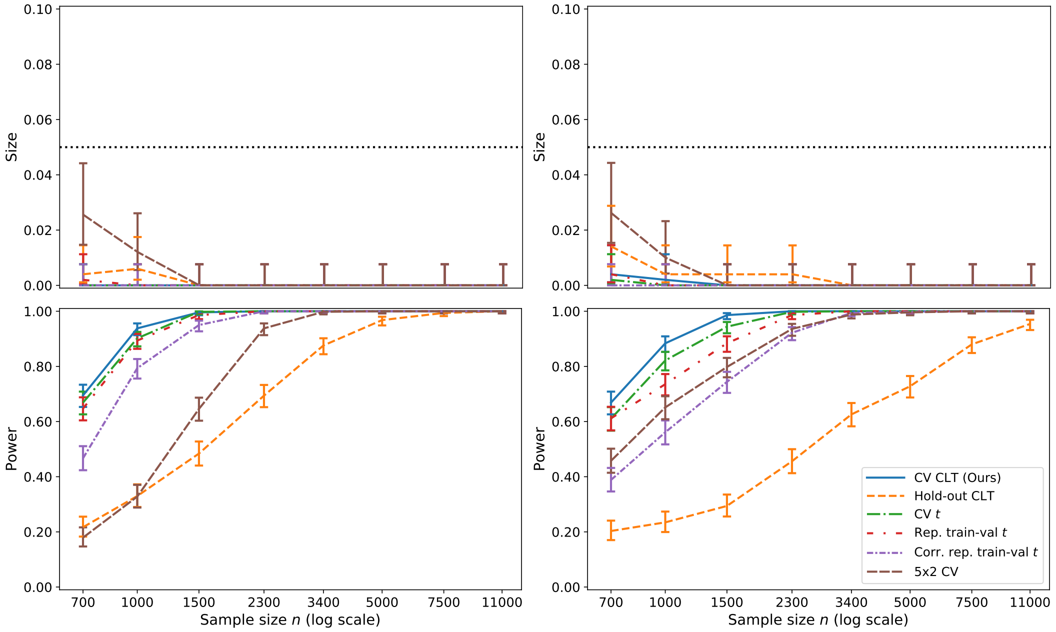


Figure 2: Size when testing $H_1 : \mathrm{Err}(\mathcal{A}_1) < \mathrm{Err}(\mathcal{A}_2)$ (top) and power when testing $H_1 : \mathrm{Err}(\mathcal{A}_2) < \mathrm{Err}(\mathcal{A}_1)$ (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: $\mathcal{A}_1 = \ell^2$-regularized logistic regression, $\mathcal{A}_2 =$ neural network classification. **Right**: $\mathcal{A}_1 =$ random forest, $\mathcal{A}_2 =$ ridge regression.

target, even for the smallest training set size of $n = 700$. The hold-out CI, while valid, is substantially wider and less informative than the other intervals as it is based on only a single train-validation split. Meanwhile, our CLT-based CI delivers the smallest width[8] (and hence greatest precision) for both learning tasks and every dataset size.

[8] All widths in Fig. 1 are displayed with ± 2 standard error bars, but some bars are too small to be visible.

5.2 Testing for improved algorithm performance

Let us write $\mathrm{Err}(\mathcal{A}_1) < \mathrm{Err}(\mathcal{A}_2)$ to signify that the test error of $\mathcal{A}_1$ is smaller than that of $\mathcal{A}_2$. In App. L.2, for each testing procedure, dataset, and pair of algorithms $(\mathcal{A}_1, \mathcal{A}_2)$, we display the size and power of level $\alpha = 0.05$ one-sided tests (4.2) of $H_1 : \mathrm{Err}(\mathcal{A}_1) < \mathrm{Err}(\mathcal{A}_2)$. In each case, we report size estimates for experiments with at least 25 replications under the null and power estimates for experiments with at least 25 replications under the alternative. Here, for representative algorithm pairs, we identify the algorithm $\mathcal{A}_1$ that more often has smaller test error across our simulations and display both the power of the level $\alpha = 0.05$ test of $H_1 : \mathrm{Err}(\mathcal{A}_1) < \mathrm{Err}(\mathcal{A}_2)$ and the size of the level $\alpha = 0.05$ test of $H_1 : \mathrm{Err}(\mathcal{A}_2) < \mathrm{Err}(\mathcal{A}_1)$. Fig. 2 displays these results for $(\mathcal{A}_1, \mathcal{A}_2)$ = (ℓ^2-regularized logistic regression, neural network) classification on the left and $(\mathcal{A}_1, \mathcal{A}_2)$ = (random forest, ridge) regression on the right. The sizes of all testing procedures are below the nominal level of 0.05, and our test is consistently the most powerful for both classification and regression. The hold-out test, while also valid, is significantly less powerful due to its reliance on a single train-validation split. In App. L.3, we observe analogous results when labels are synthetically generated.

5.3 The importance of stability

To illustrate the impact of algorithmic instability on testing procedures, we additionally compare a less stable neural network (with substantially reduced ℓ^2 regularization strength) and a less stable random forest regressor (with larger-depth trees). In Fig. 13 in App. L.4, we observe that the size of every test save the hold-out test rises above the nominal level. In the case of our test, the cause of this size violation is clear. Fig. 15a in App. L.4 demonstrates that the variance of $\frac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n)$ in Thm. 1 is much larger than 1 for this experiment, and Thm. 2 implies this can only occur when the loss stability $\gamma_{loss}(h_n)$ is large. Meanwhile, the variance of the same quantity is close to 1 for the original stable settings of the neural network and random forest regressors. We suspect that instability is also the cause of the other tests' size violations; however, it is difficult to be certain, as these alternative tests have no correctness guarantees. Interestingly, the same destabilized algorithms produce high-quality confidence intervals and relatively stable h_n in the context of single algorithm assessment (see Figs. 14 and 15b in App. L.4), as the variance parameter $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$ is significantly larger for single algorithms. This finding highlights an important feature of our results: it suffices for the loss stability to be negligible relative to the noise level σ_n^2/n.

5.4 Leave-one-out cross-validation

Leave-one-out cross-validation (LOOCV) is often viewed as prohibitive for large datasets, due to the expense of refitting a prediction rule n times. However, for ridge regression, a well-known shortcut based on the Sherman–Morrison–Woodbury formula allows one to carry out LOOCV exactly in the time required to fit a small number of base ridge regressions (see App. K.4 for a derivation of this result). Moreover, recent work shows that, for many learning procedures, LOOCV estimates can be efficiently approximated with only $O(1/n^2)$ error [8, 28, 33, 50] (see also [46, 48, 27] for related guarantees). The $O(1/n^2)$ precision of these inexpensive approximations coupled with the LOOCV consistency of $\hat{\sigma}_{n,out}^2$ (see Thm. 5) allows us to efficiently construct asymptotically-valid CIs and tests for LOOCV, even when n is large. As a simple demonstration, we construct 95% CIs for ridge regression test error based on our LOOCV CLT and compare their coverage and width with those of the procedures described in Sec. 5.1. In Fig. 3 in App. K.4, we see that, like the 10-fold CV CLT intervals, the LOOCV intervals provide coverage near the nominal level and widths smaller than the popular alternatives from the literature; in fact, the 10-fold CV CLT curves are obscured by the nearly identical LOOCV CLT curves. Complete experimental details can be found in App. K.4.

6 Conclusion and Future Work

Our central limit theorems and consistent variance estimators provide new, valid tools for testing algorithm improvement and generating test error intervals under algorithmic stability. An important open question is whether practical valid tests and intervals are also available when our stability conditions are violated. Another promising direction for future work is developing analogous tools for the *expected* test error $\mathbb{E}[R_n]$ instead of the k-fold test error R_n; Austern and Zhou [5] provide significant progress in this direction, but more work, particularly on variance estimation, is needed.

Broader Impact

This work will benefit both users and developers of machine learning methods who want to rigorously assess or compare learning algorithms. Failure of the methods we discuss (which can only happen when the assumptions we state are not satisfied) may lead to the over- or under-estimation of the performance of a learning algorithm on a particular dataset.

Acknowledgments and Disclosure of Funding

We would like to thank Jianqing Fan, Mykhaylo Shkolnikov, Miklos Racz, and Morgane Austern for helpful discussions.

References

[1] (2015). FlightDelays dataset. https://www.kaggle.com/usdot/flight-delays.

[2] Abou-Moustafa, K. and Szepesvári, C. (2019a). An exponential tail bound for Lq stable learning rules. In Garivier, A. and Kale, S., editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 31–63, Chicago, Illinois. PMLR.

[3] Abou-Moustafa, K. and Szepesvári, C. (2019b). An exponential tail bound for the deleted estimate. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 42–50.

[4] Arsov, N., Pavlovski, M., and Kocarev, L. (2019). Stability of decision trees and logistic regression. *arXiv preprint arXiv:1903.00816v1*.

[5] Austern, M. and Zhou, W. (2020). Asymptotics of Cross-Validation. *arXiv preprint arXiv:2001.11111v2*.

[6] Baldi, P., Sadowski, P., and Whiteson, D. (2014a). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5.

[7] Baldi, P., Sadowski, P., and Whiteson, D. (2014b). Higgs dataset. https://archive.ics.uci.edu/ml/datasets/HIGGS.

[8] Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. (2017). On optimal generalizability in parametric learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS17, pages 3455–3465, Red Hook, NY, USA. Curran Associates Inc.

[9] Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross validation. *Journal of Machine Learning Research*, 5:1089–1105.

[10] Billingsley, P. (1995). *Probability and Measure, Third Edition*.

[11] Blum, A., Kalai, A., and Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proc. COLT*, pages 203–208.

[12] Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. (2005). Moment inequalities for functions of independent random variables. *Annals of Probability*, 33(2):514–560.

[13] Bouckaert, R. R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *PAKDD*, pages 3–12. Springer.

[14] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

[15] Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

[16] Celisse, A. and Guedj, B. (2016). Stability revisited: new generalisation bounds for the Leave-one-Out. *arXiv preprint arXiv:1608.06412v1.*

[17] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

[18] Cornec, M. (2010). Concentration inequalities of the cross-validation estimate for stable predictors. *arXiv preprint arXiv:1011.5133v1.*

[19] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[20] Devroye, L. and Wagner, T. (1979a). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207.

[21] Devroye, L. and Wagner, T. (1979b). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604.

[22] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

[23] Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

[24] Durrett, R. (2019). *Probability: Theory and Examples, Version 5.*

[25] Elisseeff, A., Evgeniou, T., and Pontil, M. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79.

[26] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

[27] Ghosh, S., Stephenson, W. T., Nguyen, T. D., Deshpande, S. K., and Broderick, T. (2020). Approximate Cross-Validation for Structured Models. *arXiv preprint arXiv:2006.12669v1.*

[28] Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019). A swiss army infinitesimal jackknife. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1139–1147. PMLR.

[29] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, ICML16, pages 1225–1234. JMLR.org.

[30] Jiang, W., Varma, S., and Simon, R. (2008). Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

[31] Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-validation and mean-square stability. In *Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011).* Citeseer.

[32] Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453.

[33] Koh, P. W., Ang, K.-S., Teo, H. H. K., and Liang, P. (2019). On the Accuracy of Influence Functions for Measuring Group Effects. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS19, pages 5254–5264.

[34] Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35.

[35] Kutin, S. and Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI02, pages 275–282, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[36] LeDell, E., Petersen, M., and van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic Journal of Statistics*, 9(1):1583–1607.

[37] Lei, J. (2019). Cross-validation with confidence. *Journal of the American Statistical Association*, pages 1–20.

[38] Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, 40(3):203228.

[39] Markatou, M., Tian, H., Biswas, S., and Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, 6:1127–1168.

[40] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

[41] Meyer, P.-A. (1966). *Probability and Potentials*. Blaisdell Publishing Co, N.Y.

[42] Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492.

[43] Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3):239–281.

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[45] Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., and van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52.

[46] Rad, K. R. and Maleki, A. (2020). A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[47] Steele, J. M. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14(2):753–758.

[48] Stephenson, W. and Broderick, T. (2020). Approximate cross-validation in high dimensions with guarantees. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2424–2434, Online. PMLR.

[49] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

[50] Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4530–4540, Online. PMLR.

[51] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

A Proof of Prop. 1: Asymptotic linearity of k-fold CV

We first prove a general asymptotic linearity result for repeated sample-splitting estimators. Given a collection $A_n = \{(B_j, B'_j)\}_{j\in[J]}$ of index vector pairs such that for any pair (B_j, B'_j) in A_n, B_j and B'_j are disjoint, and a scalar loss function $\rho_{n,j}(Z_{B'_j}, Z_{B_j})$, define the *cross-validation error* as

$$\hat{R}_n = \tfrac{1}{J}\textstyle\sum_{j=1}^J \rho_{n,j}(Z_{B'_j}, Z_{B_j})$$

and the *multi-fold test error*

$$R_n = \tfrac{1}{J}\textstyle\sum_{j=1}^J \mathbb{E}[\rho_{n,j}(Z_{B'_j}, Z_{B_j}) \mid Z_{B_j}].$$

Note that similarly to the number of folds k in cross-validation, J can depend on the sample size n, but we write J in place of J_n to simplify our notation.

Proposition 3 (Asymptotic linearity of CV)**.** *For any sequence of datapoints $(Z_i)_{i\geq 1}$,*

$$\tfrac{\sqrt{n}}{\sigma_n}\Big(\hat{R}_n - R_n\Big) - \tfrac{\sqrt{n}}{\sigma_n J}\textstyle\sum_{j=1}^J(\bar\rho_{n,j}(Z_{B'_j}) - \mathbb{E}[\bar\rho_{n,j}(Z_{B'_j})]) \overset{p}{\to} \Big(\textit{resp. } \overset{L^q}{\to}\Big)\ 0$$

for functions $\bar\rho_{n,1}, \dots, \bar\rho_{n,J}$ with $\sigma_n^2 \triangleq \frac{1}{J}\mathrm{Var}(\sum_{j=1}^J \bar\rho_{n,j}(Z_{B'_j}))$ if and only if

$$\begin{aligned}\tfrac{\sqrt{n}}{\sigma_n J}\textstyle\sum_{j=1}^J\Big(&\rho_{n,j}(Z_{B'_j}, Z_{B_j}) - \mathbb{E}\Big[\rho_{n,j}(Z_{B'_j}, Z_{B_j}) \mid Z_{B_j}\Big] \\ &- \Big(\bar\rho_{n,j}(Z_{B'_j}) - \mathbb{E}\Big[\bar\rho_{n,j}(Z_{B'_j})\Big]\Big)\Big) \overset{p}{\to} \Big(\textit{resp. } \overset{L^q}{\to}\Big)\ 0\end{aligned}$$

where the parenthetical convergence indicates that the same statement holds when both convergences in probability are replaced with convergences in L^q for the same $q > 0$.

Proof For each $(B_j, B'_j) \in A_n$, let

$$L_j = \rho_{n,j}(Z_{B'_j}, Z_{B_j}) - \mathbb{E}[\rho_{n,j}(Z_{B'_j}, Z_{B_j}) \mid Z_{B_j}] - \Big(\bar\rho_{n,j}(Z_{B'_j}) - \mathbb{E}[\bar\rho_{n,j}(Z_{B'_j})]\Big).$$

Then

$$\begin{aligned}\tfrac{\sqrt{n}}{\sigma_n}\Big(\hat{R}_n - R_n\Big) &= \tfrac{\sqrt{n}}{\sigma_n J}\textstyle\sum_{j=1}^J(\rho_{n,j}(Z_{B'_j}, Z_{B_j}) - \mathbb{E}[\rho_{n,j}(Z_{B'_j}, Z_{B_j}) \mid Z_{B_j}]) \\ &= \tfrac{\sqrt{n}}{\sigma_n J}\textstyle\sum_{j=1}^J L_j + \tfrac{\sqrt{n}}{\sigma_n J}\sum_{j=1}^J(\bar\rho_{n,j}(Z_{B'_j}) - \mathbb{E}[\bar\rho_{n,j}(Z_{B'_j})]).\end{aligned}$$

The result now follows from the assumption that $\frac{\sqrt{n}}{\sigma_n J}\sum_{j=1}^J L_j \overset{p}{\to} \Big(\text{resp. } \overset{L^q}{\to}\Big)\ 0$. □

Prop. 1 now follows directly from Prop. 3 with the choices:

- $A_n = \{(B_\ell, i) : \ell \in [k], i \in B'_\ell\}$,
- for all $j \in [J]$, $\rho_{n,j}(Z_i, Z_{B_\ell}) = h_n(Z_i, Z_{B_\ell})$ and $\bar\rho_{n,j}(Z_i) = \bar h_n(Z_i)$ for the associated $\ell \in [k]$ and $i \in B'_\ell$.

Note that for these choices, we have $J = |A_n| = \sum_{\ell=1}^k |B'_\ell| = n$.

B Proof of Thm. 1: Asymptotic normality of k-fold CV with i.i.d. data

Thm. 1 follows from the next more general result, which establishes the asymptotic normality of k-fold CV with independent (not necessarily identically distributed) data.

Theorem 6 (Asymptotic normality of k-fold CV with independent data)**.** *Under the notation of Prop. 1, suppose that the datapoints $(Z_i)_{i\geq 1}$ are independent. If the triangular array $\big(\bar h_n(Z_i) - \mathbb{E}[\bar h_n(Z_i)]\big)_{n,i}$ satisfies* Lindeberg's condition*,*

$$\forall\varepsilon > 0,\ \tfrac{1}{n\sigma_n^2}\textstyle\sum_{i=1}^n \mathbb{E}\Big[\big(\bar h_n(Z_i) - \mathbb{E}[\bar h_n(Z_i)]\big)^2 \mathbb{1}\big[|\bar h_n(Z_i) - \mathbb{E}[\bar h_n(Z_i)]| > \varepsilon\,\sigma_n\sqrt{n}\big]\Big] \to 0, \quad \text{(B.1)}$$

then

$$\tfrac{1}{\sigma_n\sqrt{n}}\sum_{i=1}^n\big(\bar h_n(Z_i)-\mathbb{E}\big[\bar h_n(Z_i)\big]\big)\xrightarrow{d}\mathcal{N}(0,1).$$

Additionally, if (2.2) *holds in probability, then*

$$\tfrac{\sqrt{n}}{\sigma_n}\Big(\hat R_n-R_n\Big)\xrightarrow{d}\mathcal{N}(0,1).$$

Proof By independence of the datapoints $(Z_i)_{i\geq 1}$, $(\bar h_n(Z_i))_{n,i}$ are independent, and $n\sigma_n^2=\mathrm{Var}(\sum_{i=1}^n\bar h_n(Z_i))$. Under Lindeberg's condition, we get the first convergence result thanks to Lindeberg's Central Limit Theorem (see [10, Thm. 27.2]). Additionally, if assumption (2.2) holds, we apply Prop. 1 and Slutsky's theorem to get the second convergence result. □

If the $(Z_i)_{i\geq 1}$ are i.i.d., then $\sigma_n^2=\frac{1}{n}\mathrm{Var}(\sum_{i=1}^n\bar h_n(Z_i))=\mathrm{Var}(\bar h_n(Z_0))$, and Lindeberg's condition (B.1) reduces to

$$\forall\varepsilon>0,\ \tfrac{1}{\sigma_n^2}\mathbb{E}\Big[\big(\bar h_n(Z_0)-\mathbb{E}\big[\bar h_n(Z_0)\big]\big)^2\mathbb{1}\big[|\bar h_n(Z_0)-\mathbb{E}\big[\bar h_n(Z_0)\big]|>\varepsilon\,\sigma_n\sqrt{n}\big]\Big]\to 0.$$

We will show that this follows from the assumed uniform integrability of the sequence $X_n=(\bar h_n(Z_0)-\mathbb{E}[\bar h_n(Z_0)])^2/\sigma_n^2$. Indeed, for any $\varepsilon>0$ and all n,

$$\mathbb{E}[X_n\mathbb{1}\big[X_n>n\varepsilon^2\big]]\leq\sup_m\mathbb{E}[X_m\mathbb{1}\big[X_m>n\varepsilon^2\big]]\to 0,$$

as $n\to\infty$ by the uniform integrability of the sequence of X_n. Thm. 1 therefore follows from Thm. 6.

C Proof of Thm. 2: Approximate linearity from loss stability

Thm. 2 will follow from the following more general result.

Theorem 7 (Approximate linearity from loss stability)**.** *Under the notation of App. A, with* $\{(B_j,B'_j)\}_{j\in[J]}$ *a collection of disjoint index vector pairs where* $(B'_j)_{j\in[J]}$ *is a pairwise disjoint family, and* $\rho_{n,j}(Z_{B'_j},Z_{B_j})\triangleq\frac{1}{|B'_j|}\sum_{i\in B'_j}h_{n,j}(Z_i,Z_{B_j})$*, suppose that the datapoints* $(Z_i)_{i\geq 1}$ *are i.i.d. copies of a random element* Z_0*. Define* $\rho'_{n,j}(Z_{B'_j},Z_{B_j})\triangleq\rho_{n,j}(Z_i,Z_{B_j})-\mathbb{E}[\rho_{n,j}(Z_i,Z_{B_j})\mid Z_{B_j}]$ *and* $\rho''_{n,j}(Z_{B'_j},Z_{B_j})\triangleq\rho'_{n,j}(Z_i,Z_{B_j})-\mathbb{E}[\rho'_{n,j}(Z_i,Z_{B_j})\mid Z_i]$*. Then*

$$\mathbb{E}[(\tfrac{1}{J}\sum_{j=1}^J\rho''_{n,j}(Z_{B'_j},Z_{B_j}))^2]\leq\tfrac{1}{J^2}\Big(\sum_{j\neq j'}\sqrt{\gamma_{loss}(h_{n,j})\gamma_{loss}(h_{n,j'})}$$
$$+\sum_{j=1}^J\tfrac{1}{|B'_j|}\tfrac{1}{2}|B_j|\gamma_{loss}(h_{n,j})\Big).\qquad\text{(C.1)}$$

Proof

Define $h'_{n,j}$ and $h''_{n,j}$ as:

$h'_{n,j}(Z_i,Z_{B_j})\triangleq h_{n,j}(Z_i,Z_{B_j})-\mathbb{E}[h_{n,j}(Z_i,Z_{B_j})\mid Z_{B_j}]$,

$h''_{n,j}(Z_i,Z_{B_j})\triangleq h'_{n,j}(Z_i,Z_{B_j})-\mathbb{E}[h'_{n,j}(Z_i,Z_{B_j})\mid Z_i]$.

Therefore, we have $\rho'_{n,j}(Z_{B'_j},Z_{B_j})=\frac{1}{|B'_j|}\sum_{i\in B'_j}h'_{n,j}(Z_i,Z_{B_j})$ and $\rho''_{n,j}(Z_{B'_j},Z_{B_j})=\frac{1}{|B'_j|}\sum_{i\in B'_j}h''_{n,j}(Z_i,Z_{B_j})$.

Thus

$$(\tfrac{1}{J}\sum_{j=1}^J\rho''_{n,j}(Z_{B'_j},Z_{B_j}))^2=\tfrac{1}{J^2}\sum_{j,j'=1}^J\rho''_{n,j}(Z_{B'_j},Z_{B_j})\rho''_{n,j'}(Z_{B'_{j'}},Z_{B_{j'}})$$
$$=\tfrac{1}{J^2}\sum_{j,j'=1}^J\tfrac{1}{|B'_j|}\tfrac{1}{|B'_{j'}|}\sum_{i\in B'_j}\sum_{i'\in B'_{j'}}h''_{n,j}(Z_i,Z_{B_j})h''_{n,j'}(Z_{i'},Z_{B_{j'}}).$$

In what follows, $Z^{\setminus i}_{B_{j'}}$ is $Z_{B_{j'}}$ with Z_i replaced by Z'_0, an i.i.d. copy of Z_0, independent of $(Z_i)_{i\geq 1}$. Note that if $i\notin B_{j'}$, $Z^{\setminus i}_{B_{j'}}$ is just $Z_{B_{j'}}$. We similarly define $Z^{\setminus i'}_{B_j}$.

If $j \neq j'$, we have $\mathbb{E}_{Z_i}[\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} h''_{n,j}(Z_i, Z_{B_j}) h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}})] = 0$, because (i) $h''_{n,j}(Z_i, Z_{B_j})$ and $h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}})$ are conditionally independent given everything but Z_i, and (ii) $\mathbb{E}_{Z_i}[h''_{n,j}(Z_i, Z_{B_j})] = 0$.

Similarly, if $j \neq j'$,

$$\mathbb{E}_{Z_{i'}}[\textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j}) h''_{n,j'}(Z_{i'}, Z_{B_{j'}})] = 0,$$

$$\mathbb{E}_{Z_{i'}}[\textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j}) h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}})] = 0.$$

Therefore, if $j \neq j'$,

$$\begin{aligned}
&\mathbb{E}[\tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} h''_{n,j}(Z_i, Z_{B_j}) h''_{n,j'}(Z_{i'}, Z_{B_{j'}})] \\
&= \mathbb{E}[\tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} \big((h''_{n,j}(Z_i, Z_{B_j}) - h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j})) \\
&\qquad\qquad\qquad\qquad \times (h''_{n,j'}(Z_{i'}, Z_{B_{j'}}) - h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}}))\big)] \\
&= \tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} \mathbb{E}\big[\big((h''_{n,j}(Z_i, Z_{B_j}) - h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j})) \\
&\qquad\qquad\qquad\qquad \times (h''_{n,j'}(Z_{i'}, Z_{B_{j'}}) - h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}}))\big)\big] \\
&\leq \tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} \sqrt{\mathbb{E}\Big[(h''_{n,j}(Z_i, Z_{B_j}) - h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j}))^2\Big]} \\
&\qquad\qquad\qquad\qquad \times \sqrt{\mathbb{E}\Big[(h''_{n,j'}(Z_{i'}, Z_{B_{j'}}) - h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}}))^2\Big]} \\
&\leq \Big(\tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} \mathbb{E}\Big[(h''_{n,j}(Z_i, Z_{B_j}) - h''_{n,j}(Z_i, Z^{\setminus i'}_{B_j}))^2\Big] \\
&\qquad\qquad\qquad\qquad \times \mathbb{E}\Big[(h''_{n,j'}(Z_{i'}, Z_{B_{j'}}) - h''_{n,j'}(Z_{i'}, Z^{\setminus i}_{B_{j'}}))^2\Big]\Big)^{1/2} \\
&= \Big(\tfrac{1}{|B'_j|} \tfrac{1}{|B'_{j'}|} \textstyle\sum_{i \in B'_j} \sum_{i' \in B'_{j'}} \mathbb{E}\Big[(h''_{n,j}(Z_0, Z_{B_j}) - h''_{n,j}(Z_0, Z^{\setminus i'}_{B_j}))^2\Big] \\
&\qquad\qquad\qquad\qquad \times \mathbb{E}\Big[(h''_{n,j'}(Z_0, Z_{B_{j'}}) - h''_{n,j'}(Z_0, Z^{\setminus i}_{B_{j'}}))^2\Big]\Big)^{1/2} \\
&= \Big(\tfrac{1}{|B'_{j'}|} \textstyle\sum_{i' \in B'_{j'}} \mathbb{E}\Big[(h''_{n,j}(Z_0, Z_{B_j}) - h''_{n,j}(Z_0, Z^{\setminus i'}_{B_j}))^2\Big]\Big)^{1/2} \\
&\qquad\qquad\qquad\qquad \times \Big(\tfrac{1}{|B'_j|} \textstyle\sum_{i \in B'_j} \mathbb{E}\Big[(h''_{n,j'}(Z_0, Z_{B_{j'}}) - h''_{n,j'}(Z_0, Z^{\setminus i}_{B_{j'}}))^2\Big]\Big)^{1/2} \\
&= \sqrt{\gamma_{ms}(h''_{n,j}) \gamma_{ms}(h''_{n,j'})} = \sqrt{\gamma_{ms}(h'_{n,j}) \gamma_{ms}(h'_{n,j'})} = \sqrt{\gamma_{loss}(h_{n,j}) \gamma_{loss}(h_{n,j'})},
\end{aligned}$$

where we have applied Cauchy–Schwarz inequality and Jensen's inequality, used that the datapoints are i.i.d. copies of Z_0 and applied the definitions of mean-square stability and loss stability.

If $j = j'$ and $i \neq i'$, then $\mathbb{E}_{Z_i}[h''_{n,j}(Z_i, Z_{B_j}) h''_{n,j}(Z_{i'}, Z_{B_j})] = 0$.

If $j = j'$ and $i = i'$, then $\mathbb{E}[h''_{n,j}(Z_i, Z_{B_j})^2] = \mathbb{E}[\mathrm{Var}(h'_{n,j}(Z_i, Z_{B_j}) \mid Z_i)]$.

We now state a conditional application of a version of the Efron–Stein inequality due to Steele [47].

Lemma 1 (Conditional Efron–Stein inequality). *Suppose that, given W, the random vectors $X_{1:m}$ and $X'_{1:m}$ are conditionally independent and identically distributed and that the components of $X_{1:m}$ are conditionally independent given W. Then, for any suitably measurable function f*

$$\begin{aligned}
\tfrac{1}{2}\mathbb{E}[(f(X_{1:m}, W) - f(X'_{1:m}, W))^2 \mid W] &= \mathrm{Var}(f(X_{1:m}, W) \mid W) \\
&\leq \tfrac{1}{2} \textstyle\sum_{i=1}^m \mathbb{E}[(f(X_{1:m}, W) - f(X^{\setminus i}_{1:m}, W))^2 \mid W]
\end{aligned}$$

where, for each $i \in [m]$, $X^{\setminus i}_{1:m}$ represents $X_{1:m}$ with X_i replaced with X'_i.

Using Lemma 1, we get $\mathbb{E}[\mathrm{Var}(h'_{n,j}(Z_i, Z_{B_j}) \mid Z_i)] \le \frac{1}{2}|B_j|\gamma_{ms}(h'_{n,j}) = \frac{1}{2}|B_j|\gamma_{loss}(h_{n,j})$. Combining everything, we get

$$\mathbb{E}[(\tfrac{1}{J}\textstyle\sum_{j=1}^{J} \rho''_{n,j}(Z_{B'_j}, Z_{B_j}))^2] \le \tfrac{1}{J^2}\Big(\textstyle\sum_{j\neq j'} \sqrt{\gamma_{loss}(h_{n,j})\gamma_{loss}(h_{n,j'})} \\ + \textstyle\sum_{j=1}^{J} \tfrac{1}{|B'_j|}\tfrac{1}{2}|B_j|\gamma_{loss}(h_{n,j})\Big).$$

□

In the case of k-fold cross-validation with equal-sized folds and i.i.d. data, the left-hand side of (C.1) becomes

$$\mathrm{Var}\Big(\tfrac{1}{n}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i, Z_{B_j}) - \mathbb{E}\big[h'_n(Z_i, Z_{B_j}) \mid Z_i\big])\Big),$$

and its right-hand side simplifies to

$$\tfrac{1}{k^2}\Big(k(k-1)\sqrt{\gamma_{loss}(h_n)^2} + k\tfrac{k}{n}\tfrac{1}{2}n(1-\tfrac{1}{k})\gamma_{loss}(h_n)\Big) = \tfrac{3}{2}(1-\tfrac{1}{k})\gamma_{loss}(h_n).$$

Hence,

$$\tfrac{1}{n}\mathrm{Var}\Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i, Z_{B_j}) - \mathbb{E}\big[h'_n(Z_i, Z_{B_j}) \mid Z_i\big])\Big) \le \tfrac{3}{2}(1-\tfrac{1}{k})\gamma_{loss}(h_n).$$

We then note that the asymptotic linearity condition (2.2) in L^2-norm with the choice $\bar{h}_n(z) = \mathbb{E}\big[h_n(z, Z_{1:n(1-1/k)})\big]$ can be written as

$$\tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i, Z_{B_j}) - \mathbb{E}\big[h'_n(Z_i, Z_{B_j}) \mid Z_i\big]) \xrightarrow{L^2} 0,$$

which is implied by (3.2) when $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$. Therefore, Thm. 2 follows from Thm. 7.

D Proof of Thm. 3: Asymptotic linearity from conditional variance convergence

Thm. 3 will follow from the following more general statement.

Theorem 8 (Asymptotic linearity from conditional variance convergence)**.** *Under the notation of Prop. 1, suppose that the datapoints $(Z_i)_{i\ge 1}$ are i.i.d. copies of a random element Z_0. If a function $\bar{h}_n$ satisfies*

$$\max(k^{q-1}, 1)\textstyle\sum_{j=1}^{k}\mathbb{E}\Big[\Big(\tfrac{|B'_j|}{n\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{B_j}) - \bar{h}_n(Z_0)\big)\Big)^{q/2}\Big] \to 0$$

for some $q \in (0, 2]$, then $\bar{h}_n$ satisfies the L^q asymptotic linearity condition (2.2). If a function $\bar{h}_n$ satisfies

$$\textstyle\sum_{j=1}^{k}\mathbb{E}\Bigg[\min\Bigg(1, \tfrac{\sqrt{|B'_j|}}{\sigma_n\sqrt{n}}\sqrt{\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{B_j}) - \bar{h}_n(Z_0)\big)}\Bigg)\Bigg] \to 0,$$

then $\bar{h}_n$ satisfies the in-probability asymptotic linearity condition (2.2).

Proof In the notation of Prop. 1, for each $j \in [k]$, let

$$L_j = \tfrac{1}{|B'_j|}\textstyle\sum_{i\in B'_j}(h_n(Z_i, Z_{B_j}) - \bar{h}_n(Z_i)) - \mathbb{E}_{Z_0}[h_n(Z_0, Z_{B_j}) - \bar{h}_n(Z_0)].$$

We first note that for any non-decreasing concave ψ satisfying the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\psi\left(\left|\tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{j=1}^k |B'_j| L_j\right|\right)\right] &\leq \mathbb{E}\left[\psi\left(\tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{j=1}^k |B'_j||L_j|\right)\right] \\
&\leq \textstyle\sum_{j=1}^k \mathbb{E}\left[\psi\left(\tfrac{1}{\sigma_n\sqrt{n}}|B'_j||L_j|\right)\right] \\
&= \textstyle\sum_{j=1}^k \mathbb{E}\left[\mathbb{E}_{Z_{B'_j}}\left[\psi\left(\tfrac{1}{\sigma_n\sqrt{n}}|B'_j||L_j|\right)\right]\right] \\
&\leq \textstyle\sum_{j=1}^k \mathbb{E}\left[\psi\left(\tfrac{1}{\sigma_n\sqrt{n}}|B'_j|\mathbb{E}_{Z_{B'_j}}[|L_j|]\right)\right] \\
&\leq \textstyle\sum_{j=1}^k \mathbb{E}\left[\psi\left(\tfrac{1}{\sigma_n\sqrt{n}}|B'_j|\sqrt{\mathrm{Var}_{Z_{B'_j}}(L_j)}\right)\right] \\
&= \textstyle\sum_{j=1}^k \mathbb{E}\left[\psi\left(\tfrac{\sqrt{|B'_j|}}{\sigma_n\sqrt{n}}\sqrt{\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{B_j}) - \bar{h}_n(Z_0)\big)}\right)\right],
\end{aligned}$$

where we have applied the triangle inequality twice, the tower property once, and Jensen's inequality twice. The advertised L^q result for $q \in (0, 1]$ now follows by taking $\psi(x) = x^q$, and the in-probability result follows by taking $\psi(x) = \min(1, x)$ and invoking the following lemma.

Lemma 2. *For any sequence of random variables $(X_n)_{n\geq 1}$, $X_n \overset{p}{\to} 0$ if and only if $\mathbb{E}[\psi(|X_n|)] \to 0$, where $\psi(x) = \min(1, x)$.*

Proof If $X_n \overset{p}{\to} 0$, then as $X_n \overset{d}{\to} 0$ and ψ is bounded and continuous for nonnegative x, $\mathbb{E}[\psi(|X_n|)] \to 0$. Now suppose $\mathbb{E}[\psi(|X_n|)] \to 0$. Since ψ is nonnegative and non-decreasing for nonnegative x, we have $\mathbb{P}(|X_n| > \epsilon) \leq \mathbb{E}[\psi(|X_n|)]/\psi(\epsilon) \to 0$ for every $\epsilon > 0$ by Markov's inequality. Hence, $X_n \overset{p}{\to} 0$. □

Now fix any $q \in (1, 2]$, and note that as $x \mapsto x^q$ is non-decreasing and convex on the nonnegative reals, we have

$$\begin{aligned}
\mathbb{E}\left[\left|\tfrac{1}{\sigma_n\sqrt{n}}\textstyle\sum_{j=1}^k |B'_j| L_j\right|^q\right] &\leq \mathbb{E}\left[\left(\tfrac{k}{k}\textstyle\sum_{j=1}^k \tfrac{1}{\sigma_n\sqrt{n}}|B'_j||L_j|\right)^q\right] \\
&\leq \tfrac{k^q}{k}\textstyle\sum_{j=1}^k \mathbb{E}\left[\left(\tfrac{1}{\sigma_n\sqrt{n}}|B'_j||L_j|\right)^q\right] \\
&= k^{q-1}\mathbb{E}\left[\textstyle\sum_{j=1}^k \mathbb{E}_{Z_{B'_j}}\left[\left(\tfrac{1}{n\sigma_n^2}|B'_j|^2|L_j|^2\right)^{q/2}\right]\right] \\
&\leq k^{q-1}\mathbb{E}\left[\textstyle\sum_{j=1}^k \left(\tfrac{|B'_j|^2}{n\sigma_n^2}\mathrm{Var}_{Z_{B'_j}}(L_j)\right)^{q/2}\right] \\
&= k^{q-1}\mathbb{E}\left[\textstyle\sum_{j=1}^k \left(\tfrac{|B'_j|}{n\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{B_j}) - \bar{h}_n(Z_0)\big)\right)^{q/2}\right],
\end{aligned}$$

where we have applied the triangle inequality, Jensen's inequality using the convexity of $x \mapsto x^q$, the tower property, and Jensen's inequality using the concavity of $x \mapsto x^{q/2}$. Hence, the L^q result for $q \in (1, 2]$ follows from our convergence assumption. □

Thm. 3 then follows from Thm. 8 by replacing $|B'_j|$ with $\frac{n}{k}$ and Z_{B_j} with $Z_{1:n(1-1/k)}$ since folds are equal-sized and the Z_i's are i.i.d.

E Conditional Variance Convergence from Loss Stability

We show that the quantity appearing in (3.3) is controlled by the loss stability, for any $q \in (0, 2]$. Note however that (3.3) can be satisfied even in a case where the loss stability is infinite (see App. G).

Proposition 4 (Conditional variance convergence from loss stability)**.** *Suppose that k divides n evenly. Under the notation of Thm. 3 with $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:n(1-1/k)})]$,*

$$\mathbb{E}\Big[\Big(\tfrac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{1:n(1-1/k)}) - \bar{h}_n(Z_0)\big)\Big)^{q/2}\Big] \le \Big(\tfrac{1}{\sigma_n^2}\tfrac{1}{2}n(1-1/k)\gamma_{loss}(h_n)\Big)^{q/2},$$

for any $q \in (0, 2]$. Consequently, the condition (3.3) is verified whenever $\gamma_{loss}(h_n) = o\Big(\frac{\sigma_n^2}{n(1-1/k)\max(k, k^{(2/q)-1})}\Big)$.

Remark 2. *If $k = O(1)$, this loss stability assumption simplifies to $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ for any $q \in (0, 2]$.*

Proof Write $m = n(1 - 1/k)$. Then

$$\mathrm{Var}_{Z_0}(h_n(Z_0, Z_{1:m}) - \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_0]) = \mathrm{Var}_{Z_0}(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0]),$$

since the difference $h_n(Z_0, Z_{1:m}) - h'_n(Z_0, Z_{1:m}) = \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_{1:m}]$ is a $Z_{1:m}$-measurable function. For $0 < q \le 2$, using Jensen's inequality,

$$\begin{aligned} &\mathbb{E}\Big[(\mathrm{Var}_{Z_0}(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0]))^{q/2}\Big] \\ &\le \mathbb{E}[\mathrm{Var}_{Z_0}(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0])]^{q/2}. \end{aligned}$$

We can bound it using loss stability.

$$\begin{aligned} &\mathrm{Var}_{Z_0}(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0]) \\ &= \mathbb{E}_{Z_0}\big[\big((h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0]) \\ &\qquad - (\mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_{1:m}] - \mathbb{E}[h'_n(Z_0, Z_{1:m})])\big)^2\big] \\ &= \mathbb{E}[(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0])^2 \mid Z_{1:m}], \end{aligned}$$

so that

$$\begin{aligned} \mathbb{E}[\mathrm{Var}_{Z_0}(h'_n(Z_0, Z_{1:m}) - \mathbb{E}[h'_n(Z_0, Z_{1:m}) \mid Z_0])] &= \mathbb{E}[(h'_n(Z_0, Z_{1:m}) - \mathbb{E}\big[h'_n(Z_0, Z_{1:m}) \mid Z_0])^2\big] \\ &= \mathbb{E}[\mathrm{Var}(h'_n(Z_0, Z_{1:m}) \mid Z_0)] \\ &\le \tfrac{1}{2}m\gamma_{loss}(h_n), \end{aligned}$$

where the last inequality comes from Lemma 1. Consequently,

$$\mathbb{E}\Big[\Big(\tfrac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}(h_n(Z_0, Z_{1:m}) - \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_0])\Big)^{q/2}\Big] \le \Big(\tfrac{1}{\sigma_n^2}\tfrac{1}{2}m\gamma_{loss}(h_n)\Big)^{q/2}.$$

□

F Excess Loss of Sample Mean: $o(\frac{\sigma_n^2}{n})$ loss stability, constant $\sigma_n^2 \in (0, \infty)$, infinite mean-square stability

Here we present a very simple learning task in which (i) the CLT conditions of Thms. 1 and 2 hold and (ii) mean-square stability (3.1) is infinite.

Example 1 (Excess loss of sample mean: $o(\frac{\sigma_n^2}{n})$ loss stability, constant $\sigma_n^2 \in (0, \infty)$, infinite mean-square stability)**.** *Suppose $(Z_i)_{i\ge1}$ are independent and identically distributed copies of a random element Z_0 with $\mathbb{E}[Z_0] = 0$ and $\mathbb{E}[Z_0^2] < \infty$. Consider k-fold cross-validation of the excess loss of the sample mean relative to a constant prediction rule:*

$$h_n(z, \mathcal{D}) = (z - \hat{f}(\mathcal{D}))^2 - (z - a)^2 \quad \textit{where} \quad \hat{f}(\mathcal{D}) \triangleq \tfrac{1}{|\mathcal{D}|}\textstyle\sum_{Z_0 \in \mathcal{D}} Z_0 \quad \textit{and} \quad a \neq 0.$$

The variance parameter of Thm. 1 $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0)) = 4a^2\mathrm{Var}(Z_0)$ when $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:n(1-1/k)})]$, and the loss stability $\gamma_{loss}(h_n) = \frac{8\mathrm{Var}(Z_0)^2}{n^2(1-1/k)^2} = o(\sigma_n^2/n)$. Consequently Thm. 2 implies asymptotic linearity. The uniform integrability condition of Thm. 1 also holds. Together, these results imply that the CLT of Thm. 1 is applicable. However, whenever Z_0 does not have a fourth moment, the mean-square stability (3.1) is infinite.

Proof Introduce the shorthand $m = n(1-1/k)$, fix any $\mathcal{D}$ with $|\mathcal{D}| = m$, and suppose $\mathcal{D}^{Z_0'}$ is formed by swapping Z_0' for an independent point Z_0'' in $\mathcal{D}$. For any z we have

$$\begin{aligned}(z-\hat{f}(\mathcal{D}))^2-(z-\hat{f}(\mathcal{D}^{Z_0'}))^2 &= (\hat{f}(\mathcal{D})-\hat{f}(\mathcal{D}^{Z_0'}))(2z-\hat{f}(\mathcal{D})-\hat{f}(\mathcal{D}^{Z_0'}))\\ &= \tfrac{1}{m}(Z_0''-Z_0')(2z-\hat{f}(\mathcal{D})-\hat{f}(\mathcal{D}^{Z_0'}))\\ &= \tfrac{1}{m}(Z_0''-Z_0')(2z-\tfrac{1}{m}(Z_0''+Z_0')-2(\hat{f}(\mathcal{D})-\tfrac{1}{m}Z_0''))\\ &= \tfrac{2}{m}(Z_0''-Z_0')(z-(\hat{f}(\mathcal{D})-\tfrac{1}{m}Z_0''))-\tfrac{1}{m^2}(Z_0''^2-Z_0'^2).\end{aligned}$$

Hence, the mean-square stability equals

$$\begin{aligned}&\mathbb{E}[((Z_0-\hat{f}(\mathcal{D}))^2-(Z_0-\hat{f}(\mathcal{D}^{Z_0'}))^2)^2]\\ &= \tfrac{1}{m^4}\mathbb{E}[(Z_0''^2-Z_0'^2)^2]+\tfrac{4}{m^2}\mathbb{E}[(Z_0''-Z_0')^2]\mathbb{E}[(Z_0-(\hat{f}(\mathcal{D})-\tfrac{1}{m}Z_0''))^2]\\ &- \tfrac{4}{m^3}\mathbb{E}[(Z_0''-Z_0')(Z_0''^2-Z_0'^2)]\mathbb{E}[Z_0-(\hat{f}(\mathcal{D})-\tfrac{1}{m}Z_0'')]\\ &= \tfrac{1}{m^4}\mathbb{E}[(Z_0''^2-Z_0'^2)^2]+\tfrac{4}{m^2}\mathbb{E}[(Z_0''-Z_0')^2]\mathbb{E}[(Z_0-(\hat{f}(\mathcal{D})-\tfrac{1}{m}Z_0''))^2]\\ &\geq \tfrac{2}{m^4}\mathrm{Var}(Z_0^2)\end{aligned}$$

since $\mathbb{E}[Z_0] = 0$, and $Z_0, Z_0', Z_0'', \hat{f}(\mathcal{D}) - \frac{1}{m}Z_0''$ are mutually independent.

Moreover, the loss stability equals

$$\begin{aligned}&\mathbb{E}[((Z_0-\hat{f}(\mathcal{D}))^2-(Z_0-\hat{f}(\mathcal{D}^{Z_0'}))^2-\mathbb{E}_{Z_0}[(Z_0-\hat{f}(\mathcal{D}))^2-(Z_0-\hat{f}(\mathcal{D}^{Z_0'}))^2])^2]\\ &= \tfrac{4}{m^2}\mathbb{E}[(Z_0''-Z_0')^2]\mathbb{E}[(Z_0-\mathbb{E}[Z_0])^2] = \tfrac{8}{m^2}\mathrm{Var}(Z_0)^2.\end{aligned}$$

Finally, for any z, $\mathbb{E}[(z-\hat{f}(\mathcal{D}))^2-(z-a)^2] = \frac{1}{m}\mathrm{Var}(Z_0)+2za-a^2$. Consequently, for $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0,\mathcal{D}) \mid Z_0]$, we get the following equalities:

$$\begin{aligned}&\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0)) = 4a^2\mathrm{Var}(Z_0), \quad \text{and}\\ &(\bar{h}_n(Z_0)-\mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2 = Z_0^2/\mathrm{Var}(Z_0).\end{aligned}$$

The distribution of $Z_0^2/\mathrm{Var}(Z_0)$ does not depend on n and is integrable, so the sequence of $(\bar{h}_n(Z_0)-\mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable. □

G Loss of Surrogate Mean: constant $\sigma_n^2 \in (0,\infty)$, infinite $\tilde{\sigma}_n^2$, vanishing conditional variance

The following example details a simple task in which (i) the CLT conditions of Thm. 1 and Thm. 3 hold and (ii) mean-square stability, $\tilde{\sigma}_n^2$, and loss stability are infinite.

Example 2 (Loss of surrogate mean: constant $\sigma_n^2 \in (0,\infty)$, infinite $\tilde{\sigma}_n^2$, vanishing conditional variance)**.** *Suppose $(Z_i)_{i\geq 1}$ are independent and identically distributed copies of a random element $Z_0 = (X_0, Y_0)$ with $Z_i = (X_i, Y_i)$ and $\mathbb{E}[X_0] = \mathbb{E}[Y_0]$. Consider k-fold cross-validation of the following prediction rule under squared error loss:*

$$h_n((x,y),\mathcal{D}) = (y-\hat{f}(\mathcal{D}))^2 \quad \textit{where} \quad \hat{f}(\mathcal{D}) \triangleq \tfrac{1}{|\mathcal{D}|}\textstyle\sum_{(X_0,Y_0)\in\mathcal{D}} X_0.$$

The loss stability $\gamma_{loss}(h_n) = \frac{8\mathrm{Var}(X_0)\mathrm{Var}(Y_0)}{n^2(1-1/k)^2}$, and the variance parameter of Thm. 1

$$\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0)) = \mathrm{Var}((Y_0-\mathbb{E}[Y_0])^2), \tag{G.1}$$

when $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:n(1-1/k)})]$. Hence, if $\mathbb{E}[X_0^2], \mathbb{E}[Y_0^4] < \infty$, then $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ and Thm. 2 implies asymptotic linearity. The uniform integrability condition of Thm. 1 also holds. Together, these results imply that the CLT of Thm. 1 is applicable.

If X_0 has no fourth moment, then the mean-square stability (3.1) is infinite.

If X_0 has no second moment, then the loss stability and the [5, Theorem 1] variance parameter

$$\tilde{\sigma}_n^2 = \mathbb{E}[\mathrm{Var}(h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_{1:n(1-1/k)})] = \mathrm{Var}((Y_0 - \mathbb{E}[Y_0])^2) + \tfrac{8\mathrm{Var}(X_0)\mathrm{Var}(Y_0)}{n(1-1/k)}.$$

are infinite. However,

$$\begin{aligned}&\sqrt{k}\mathbb{E}\Big[\sqrt{\tfrac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, Z_{1:n(1-1/k)}) - \bar{h}_n(Z_0)\big)}\Big]\\ &= 2\sqrt{k}\sqrt{\tfrac{\mathrm{Var}(Y_0)}{\mathrm{Var}((Y_0-\mathbb{E}[Y_0])^2)}}\mathbb{E}[|\hat{f}(Z_{1:n(1-1/k)}) - \mathbb{E}[X_0]|].\end{aligned}$$

Hence, if $\mathbb{E}[Y_0^4] < \infty$ and $k = O(1)$, L^1 asymptotic linearity follows from Thm. 3, the uniform integrability condition of Thm. 1 still holds, and the CLT of Thm. 1 holds with the finite variance parameter (G.1).

Proof Without loss of generality, we will assume $\mathbb{E}[X_0] = \mathbb{E}[Y_0] = 0$; the formulas in the general case are obtained by replacing X_0 with $X_0 - \mathbb{E}[X_0]$ and similarly for Y_0. Introduce the shorthand $m = n(1-1/k)$, fix any $\mathcal{D}$ with $|\mathcal{D}| = m$, and suppose $\mathcal{D}^{Z_0'}$ is formed by swapping Z_0' for an independent point Z_0'' in $\mathcal{D}$. For any $z = (x, y)$ we have

$$\begin{aligned}(y - \hat{f}(\mathcal{D}))^2 - (y - \hat{f}(\mathcal{D}^{Z_0'}))^2 &= (\hat{f}(\mathcal{D}) - \hat{f}(\mathcal{D}^{Z_0'}))(2y - \hat{f}(\mathcal{D}) - \hat{f}(\mathcal{D}^{Z_0'}))\\ &= \tfrac{1}{m}(X_0'' - X_0')(2y - \hat{f}(\mathcal{D}) - \hat{f}(\mathcal{D}^{Z_0'}))\\ &= \tfrac{1}{m}(X_0'' - X_0')(2y - \tfrac{1}{m}(X_0'' + X_0') - 2(\hat{f}(\mathcal{D}) - \tfrac{1}{m}X_0''))\\ &= \tfrac{2}{m}(X_0'' - X_0')(y - (\hat{f}(\mathcal{D}) - \tfrac{1}{m}X_0'')) - \tfrac{1}{m^2}(X_0''^2 - X_0'^2).\end{aligned}$$

Hence, the mean-square stability equals

$$\begin{aligned}&\mathbb{E}[((Y_0 - \hat{f}(\mathcal{D}))^2 - (Y_0 - \hat{f}(\mathcal{D}^{Z_0'}))^2)^2]\\ &= \tfrac{1}{m^4}\mathbb{E}[(X_0''^2 - X_0'^2)^2] + \tfrac{4}{m^2}\mathbb{E}[(X_0'' - X_0')^2]\mathbb{E}[(Y_0 - (\hat{f}(\mathcal{D}) - \tfrac{1}{m}X_0''))^2]\\ &- \tfrac{4}{m^3}\mathbb{E}[(X_0'' - X_0')(X_0''^2 - X_0'^2)]\mathbb{E}[Y_0 - (\hat{f}(\mathcal{D}) - \tfrac{1}{m}X_0'')]\\ &= \tfrac{1}{m^4}\mathbb{E}[(X_0''^2 - X_0'^2)^2] + \tfrac{4}{m^2}\mathbb{E}[(X_0'' - X_0')^2]\mathbb{E}[(Y_0 - (\hat{f}(\mathcal{D}) - \tfrac{1}{m}X_0''))^2]\\ &\geq \tfrac{2}{m^4}\mathrm{Var}((X_0 - \mathbb{E}[X_0])^2)\end{aligned}$$

since $\mathbb{E}[Y_0] = 0$, and $Z_0, Z_0', Z_0'', \hat{f}(\mathcal{D}) - \frac{1}{m}X_0''$ are mutually independent.

Moreover, the loss stability equals

$$\begin{aligned}&\mathbb{E}[((Y_0 - \hat{f}(\mathcal{D}))^2 - (Y_0 - \hat{f}(\mathcal{D}^{Z_0'}))^2 - \mathbb{E}_Y[(Y_0 - \hat{f}(\mathcal{D}))^2 - (Y_0 - \hat{f}(\mathcal{D}^{Z_0'}))^2])^2]\\ &= \tfrac{4}{m^2}\mathbb{E}[(X_0'' - X_0')^2]\mathbb{E}[(Y_0 - \mathbb{E}[Y_0])^2] = \tfrac{8}{m^2}\mathrm{Var}(X_0)\mathrm{Var}(Y_0).\end{aligned}$$

Next note that, for any y, y',

$$\begin{aligned}\mathbb{E}[(y - \hat{f}(\mathcal{D}))^2 - (y' - \hat{f}(\mathcal{D}))^2] &= (y - y')(y + y' - 2\mathbb{E}[\hat{f}(\mathcal{D})])\\ &= (y^2 - y'^2) - 2(y - y')\mathbb{E}[\hat{f}(\mathcal{D})] = y^2 - y'^2\end{aligned}$$

since $\mathbb{E}[X_0] = 0$. Therefore, $\mathrm{Var}(\mathbb{E}[h_n(Z_0, \mathcal{D}) \mid Z_0]) = \frac{1}{2}\mathbb{E}[(Y_0^2 - Y_0'^2)^2] = \mathrm{Var}(Y_0^2)$.

For any y, $\mathbb{E}[(y - \hat{f}(\mathcal{D}))^2] = y^2 + \mathbb{E}[\hat{f}(\mathcal{D})^2]$ since $\mathbb{E}[X_0] = 0$. Consequently, for $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, \mathcal{D}) \mid Z_0]$, we get the following equalities:

$$\begin{aligned}&\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0)) = \mathrm{Var}(Y_0^2), \quad \text{and}\\ &(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2 = (Y_0^2 - \mathbb{E}[Y_0^2])^2/\mathrm{Var}(Y_0^2).\end{aligned}$$

The distribution of $(Y_0^2 - \mathbb{E}[Y_0^2])^2/\mathrm{Var}(Y_0^2)$ does not depend on n and is integrable, so the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable.

Since

$$(y - \hat{f}(\mathcal{D}))^2 - (y' - \hat{f}(\mathcal{D}))^2 = (y^2 - y'^2) - 2(y - y')\hat{f}(\mathcal{D})$$

we can compute the variance parameter of [5],

$$\begin{aligned}\tilde{\sigma}_n^2 &= \mathbb{E}[\mathrm{Var}(h_n(Z_0, \mathcal{D}) \mid \mathcal{D})] = \mathbb{E}[(h_n(Z_0, \mathcal{D}) - \mathbb{E}[h_n(Z_0, \mathcal{D})|\mathcal{D}])^2] \\ &= \tfrac{1}{2}\mathbb{E}[(h_n(Z_0, \mathcal{D}) - h_n(Z_0', \mathcal{D}))^2] \\ &= \tfrac{1}{2}\mathbb{E}((Y_0^2 - Y_0'^2)^2] + 4\mathbb{E}[(Y_0 - Y_0')^2]\mathbb{E}[\hat{f}(\mathcal{D})^2] - \mathbb{E}[(y^2 - y'^2)(y - y')]\mathbb{E}[\hat{f}(\mathcal{D})] \\ &= \mathrm{Var}(Y_0^2) + 8\mathrm{Var}(Y_0)\tfrac{1}{m}\mathrm{Var}(X_0),\end{aligned}$$

since $\mathbb{E}[\hat{f}(\mathcal{D})] = 0$ and $\hat{f}(\mathcal{D}), Y_0, Y_0'$ are mutually independent.

Finally, let's compute $\sqrt{k}\,\mathbb{E}\Big[\sqrt{\frac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, \mathcal{D}) - \bar{h}_n(Z_0)\big)}\Big]$.

For any y, y',

$$\begin{aligned}&((y - \hat{f}(\mathcal{D}))^2 - \mathbb{E}[(y - \hat{f}(\mathcal{D}))^2]) - ((y' - \hat{f}(\mathcal{D}))^2 - \mathbb{E}[(y' - \hat{f}(\mathcal{D}))^2]) \\ &= (y^2 - y'^2) - 2(y - y')\hat{f}(\mathcal{D}) - (y^2 - y'^2) \\ &= -2(y - y')\hat{f}(\mathcal{D}),\end{aligned}$$

so that $\mathrm{Var}_{Z_0}\big(h_n(Z_0, \mathcal{D}) - \bar{h}_n(Z_0)\big) = \frac{1}{2}\mathbb{E}[(-2(Y_0 - Y_0')\hat{f}(\mathcal{D}))^2 \mid \mathcal{D}] = 4\mathrm{Var}(Y_0)(\hat{f}(\mathcal{D}))^2$.

Then

$$\sqrt{k}\mathbb{E}\Big[\sqrt{\tfrac{1}{\sigma_n^2}\mathrm{Var}_{Z_0}\big(h_n(Z_0, \mathcal{D}) - \bar{h}_n(Z_0)\big)}\Big] = 2\sqrt{k}\sqrt{\tfrac{\mathrm{Var}(Y_0)}{\mathrm{Var}(Y_0^2)}}\mathbb{E}[|\hat{f}(\mathcal{D})|]. \tag{G.2}$$

If $\mathbb{E}[|X_0|] < \infty$, the family of empirical averages $\{\frac{1}{m}\sum_{i=1}^m X_i : m \geq 1\}$ is uniformly integrable and the weak law of large numbers implies that $\hat{f}(\mathcal{D})$ converges to 0 in probability. Hence, $\hat{f}(\mathcal{D}) \xrightarrow{L^1} 0$. The quantity (G.2) then goes to zero when $k = O(1)$. □

H Proof of Prop. 2: Variance comparison

Prop. 2 will follow from the following more general result.

Proposition 5. *Fix any $j \in [k]$, and define $\sigma_{n,j}^2 \triangleq \mathrm{Var}(\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_0])$ and $\tilde{\sigma}_{n,j}^2 \triangleq \mathbb{E}[\mathrm{Var}(h_n(Z_0, Z_{B_j}) \mid Z_{B_j})]$. Then*

$$\sigma_{n,j}^2 \leq \tilde{\sigma}_{n,j}^2 \leq \sigma_{n,j}^2 + \tfrac{|B_j|}{2}\gamma_{loss}(h_n),$$

where the first inequality is strict whenever $h_n'(Z_0, Z_{B_j}) = h_n(Z_0, Z_{B_j}) - \mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}]$ depends on Z_{B_j}.

Proof For all $j \in [k]$, we can rewrite both variance parameters.

$$\begin{aligned}\tilde{\sigma}_{n,j}^2 &= \mathbb{E}[\mathrm{Var}(h_n(Z_0, Z_{B_j}) \mid Z_{B_j})] \\ &= \mathbb{E}[(h_n(Z_0, Z_{B_j}) - \mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}])^2] \\ &= \mathbb{E}[h_n'(Z_0, Z_{B_j})^2] = \mathrm{Var}(h_n'(Z_0, Z_{B_j})).\end{aligned}$$

$$\begin{aligned}\sigma_{n,j}^2 &= \mathrm{Var}(\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_0]) \\ &= \mathbb{E}[(\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_0] - \mathbb{E}[h_n(Z_0, Z_{B_j})])^2] \\ &= \mathbb{E}[\mathbb{E}[h_n'(Z_0, Z_{B_j}) \mid Z_0]^2] = \mathrm{Var}(\mathbb{E}[h_n'(Z_0, Z_{B_j}) \mid Z_0]) \\ &= \mathrm{Var}(h_n'(Z_0, Z_{B_j})) - \mathbb{E}[\mathrm{Var}(h_n'(Z_0, Z_{B_j}) \mid Z_0)] \\ &= \tilde{\sigma}_{n,j}^2 - \mathbb{E}[\mathrm{Var}(h_n'(Z_0, Z_{B_j}) \mid Z_0)] \leq \tilde{\sigma}_{n,j}^2,\end{aligned}$$

where the final inequality is strict whenever $\mathbb{E}[\mathrm{Var}(h'_n(Z_0, Z_{B_j}) \mid Z_0)]$ is non-zero.

Since every non-constant variable has either infinite or strictly positive variance, $\mathbb{E}[\mathrm{Var}(h'_n(Z_0, Z_{B_j}) \mid Z_0)] = 0 \Leftrightarrow h'_n(Z_0, Z_{B_j}) = \mathbb{E}[h'_n(Z_0, Z_{B_j}) \mid Z_0]$, that is, if and only if $h'_n(Z_0, Z_{B_j}) = h_n(Z_0, Z_{B_j}) - \mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}]$ is independent of Z_{B_j}.

Finally, we know from Lemma 1 that the difference $\tilde{\sigma}^2_{n,j} - \sigma^2_{n,j} = \mathbb{E}[\mathrm{Var}(h'_n(Z_0, Z_{B_j}) \mid Z_0)] \leq \frac{1}{2}|B_j|\gamma_{loss}(h_n)$. □

Prop. 2 then follows from Prop. 5 since the Z_i's are i.i.d. and, when k divides n, the only possible size for B_j is $n(1 - 1/k)$.

I Proof of Thm. 4: Consistent within-fold estimate of asymptotic variance

We will prove the following more detailed statement from which Thm. 4 will follow.

Theorem 9 (Consistent within-fold estimate of asymptotic variance)**.** *Suppose that $k \leq n/2$ and that k divides n evenly. Under the notation of Thm. 1 with $m = n(1 - 1/k)$, $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:m})]$, $h'_n(Z_0, Z_{1:m}) = h_n(Z_0, Z_{1:m}) - \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_{1:m}]$ and $\bar{h}'_n(z) = \mathbb{E}[h'_n(z, Z_{1:m})]$, define the within-fold variance estimate*

$$\hat{\sigma}^2_{n,in} \triangleq \tfrac{1}{k} \textstyle\sum_{j=1}^{k} \tfrac{1}{(n/k)-1} \sum_{i \in B'_j} \Big(h_n(Z_i, Z_{B_j}) - \tfrac{k}{n} \sum_{i' \in B'_j} h_n(Z_{i'}, Z_{B_j})\Big)^2.$$

If $(Z_i)_{i \geq 1}$ are i.i.d. copies of a random element Z_0, then

$$\mathbb{E}[|\hat{\sigma}^2_{n,in} - \sigma^2_n|] \leq \tfrac{2n^2}{n-k}\gamma_{loss}(h_n) + 2\sqrt{\tfrac{2n^2}{n-k}\gamma_{loss}(h_n)\sigma^2_n} + \sqrt{\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n}$$

and there exists an absolute constant C specified in the proof such that

$$\begin{aligned}\mathbb{E}[(\hat{\sigma}^2_{n,in} - \sigma^2_n)^2] \leq 4\tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n) + 8\sqrt{\tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n)(\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n + \sigma^4_n)} \\ + 2(\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n) \qquad \text{(I.1)}\end{aligned}$$

where $\gamma_4(h'_n) \triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[(h'_n(Z_0, Z_{1:m}) - h'_n(Z_0, Z^{\setminus i}_{1:m}))^4]$. Here, $Z^{\setminus i}_{1:m}$ denotes $Z_{1:m}$ with Z_i replaced by an i.i.d. copy independent of $Z_{0:m}$.

Moreover,

$$\mathbb{E}[|\hat{\sigma}^2_{n,in} - \sigma^2_n|] \leq \tfrac{2n^2}{n-k}\gamma_{loss}(h_n) + 2\sqrt{\tfrac{2n^2}{n-k}\gamma_{loss}(h_n)\sigma^2_n} + \sqrt{\tfrac{2}{n(n/k-1)}\sigma^4_n} + o(\sigma^2_n) \qquad \text{(I.2)}$$

whenever the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma^2_n$ is uniformly integrable.

Proof

Eliminating training set randomness We begin by approximating our variance estimate

$$\begin{aligned}\hat{\sigma}^2_{n,in} &= \tfrac{1}{k} \textstyle\sum_{j=1}^{k} \tfrac{1}{(n/k)-1} \sum_{i \in B'_j} \Big(h_n(Z_i, Z_{B_j}) - \tfrac{k}{n} \sum_{i' \in B'_j} h_n(Z_{i'}, Z_{B_j})\Big)^2 \\ &= \tfrac{1}{k} \textstyle\sum_{j=1}^{k} \tfrac{1}{(n/k)-1} \sum_{i \in B'_j} \Big(h'_n(Z_i, Z_{B_j}) - \tfrac{k}{n} \sum_{i' \in B'_j} h'_n(Z_{i'}, Z_{B_j})\Big)^2\end{aligned}$$

by a quantity eliminating training set randomness in each summand,

$$\hat{\sigma}^2_{n,in,approx} \triangleq \tfrac{1}{k} \textstyle\sum_{j=1}^{k} \tfrac{1}{(n/k)-1} \sum_{i \in B'_j} \Big(\bar{h}'_n(Z_i) - \tfrac{k}{n} \sum_{i' \in B'_j} \bar{h}'_n(Z_{i'})\Big)^2,$$

where $\bar{h}'_n(z) = \mathbb{E}[h'_n(z, Z_{1:m})]$. Note that $\bar{h}'_n(Z_0)$ has expectation 0.

By Cauchy–Schwarz, we have

$$|\hat{\sigma}^2_{n,in} - \hat{\sigma}^2_{n,in,approx}| \leq \Delta + 2\sqrt{\Delta}\hat{\sigma}_{n,in,approx}$$

for the error term

$$\begin{aligned}
\Delta &\triangleq \tfrac{1}{k}\textstyle\sum_{j=1}^{k}\tfrac{1}{(n/k)-1}\sum_{i\in B'_j}\Big(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i)+\tfrac{k}{n}\sum_{i'\in B'_j}\bar{h}'_n(Z_{i'})-\tfrac{k}{n}\sum_{i'\in B'_j}h'_n(Z_{i'},Z_{B_j})\Big)^2\\
&\le 2\tfrac{1}{k}\textstyle\sum_{j=1}^{k}\tfrac{1}{(n/k)-1}\sum_{i\in B'_j}(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^2\\
&\quad+2\tfrac{1}{k}\textstyle\sum_{j=1}^{k}\tfrac{1}{(n/k)-1}\sum_{i\in B'_j}\Big(\tfrac{k}{n}\sum_{i'\in B'_j}(\bar{h}'_n(Z_{i'})-h'_n(Z_{i'},Z_{B_j}))\Big)^2\\
&\le \tfrac{2}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^2\\
&\quad+2\tfrac{1}{k}\textstyle\sum_{j=1}^{k}\tfrac{1}{(n/k)-1}\sum_{i\in B'_j}\tfrac{k}{n}\sum_{i'\in B'_j}(\bar{h}'_n(Z_{i'})-h'_n(Z_{i'},Z_{B_j}))^2\\
&= \tfrac{4}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^2 \qquad\qquad \text{(I.3)}
\end{aligned}$$

where we have used Jensen's inequality twice.

Thus,

$$\begin{aligned}
\Delta^2 &\le \tfrac{16n^2}{(n-k)^2}\tfrac{1}{n}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^4\\
&= \tfrac{16n}{(n-k)^2}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^4 \qquad\qquad \text{(I.4)}
\end{aligned}$$

by Jensen's inequality.

Controlling the error Δ We will first control the error term Δ. By the bound (I.3) and the conditional Efron–Stein inequality (Lemma 1), we have

$$\begin{aligned}
\mathbb{E}[\Delta] &\le \tfrac{4}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^2]\\
&= \tfrac{4}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\mathbb{E}[h'_n(Z_i,Z_{B_j})\mid Z_i])^2]\\
&= \tfrac{4}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\mathbb{E}[h'_n(Z_i,Z_{B_j})\mid Z_i])^2\mid Z_i]]\\
&= \tfrac{4}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[\mathrm{Var}(h'_n(Z_i,Z_{B_j})\mid Z_i)]\\
&\le \tfrac{2}{n-k}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}m\gamma_{ms}(h'_n)\\
&\le \tfrac{2n^2}{n-k}\gamma_{ms}(h'_n)=\tfrac{2n^2}{n-k}\gamma_{loss}(h_n). \qquad\qquad \text{(I.5)}
\end{aligned}$$

Controlling Δ^2 In the following, for any $j\in[k]$ and $i'\in B_j$, $Z_{B_j}^{\setminus i'}$ is Z_{B_j} with $Z_{i'}$ replaced by Z_0. By the bound (I.4) and Boucheron et al. [12, Thm. 2], and by noting that $x^4=x_+^4+x_-^4$ where $x_+=\max(x,0)$ and $x_-=\max(-x,0)$, we have

$$\begin{aligned}
\mathbb{E}[\Delta^2] &\le \tfrac{16n}{(n-k)^2}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\bar{h}'_n(Z_i))^4]\\
&= \tfrac{16n}{(n-k)^2}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\mathbb{E}[\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\mathbb{E}[h'_n(Z_i,Z_{B_j})\mid Z_i])^4\mid Z_i]]\\
&= \tfrac{16n}{(n-k)^2}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\Big(\mathbb{E}[\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\mathbb{E}[h'_n(Z_i,Z_{B_j})\mid Z_i])_+^4\mid Z_i]]\\
&\qquad\qquad+\mathbb{E}[\mathbb{E}[(h'_n(Z_i,Z_{B_j})-\mathbb{E}[h'_n(Z_i,Z_{B_j})\mid Z_i])_-^4\mid Z_i]]\Big)\\
&\le \tfrac{16n}{(n-k)^2}(1-\tfrac{1}{4})^2 4(\tfrac{8}{7})^2 16\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\Big(\mathbb{E}[(\mathbb{E}[\sum_{i'\in B_j}(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_+^2\mid Z_{B_j}])^2]\\
&\qquad\qquad+\mathbb{E}[(\mathbb{E}[\textstyle\sum_{i'\in B_j}(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_-^2\mid Z_{B_j}])^2]\Big)\\
&\le \tfrac{16n}{(n-k)^2}\tfrac{2304}{49}\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\Big(\mathbb{E}[(\sum_{i'\in B_j}(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_+^2)^2]\\
&\qquad\qquad+\mathbb{E}[(\textstyle\sum_{i'\in B_j}(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_-^2)^2]\Big)\\
&\le \tfrac{36864n}{49(n-k)^2}m\textstyle\sum_{j=1}^{k}\sum_{i\in B'_j}\Big(\sum_{i'\in B_j}\mathbb{E}[(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_+^4]\\
&\qquad\qquad+\textstyle\sum_{i'\in B_j}\mathbb{E}[(h'_n(Z_i,Z_{B_j})-h'_n(Z_i,Z_{B_j}^{\setminus i'}))_-^4]\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \tfrac{36864n}{49(n-k)^2} m \textstyle\sum_{j=1}^k \sum_{i\in B'_j} \sum_{i'\in B_j} \mathbb{E}[(h'_n(Z_i, Z_{B_j}) - h'_n(Z_i, Z_{B_j}^{\setminus i'}))^4] \\
&\le \tfrac{36864n^4}{49(n-k)^2}\gamma_4(h'_n) = \tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n) \qquad \text{(I.6)}
\end{aligned}
$$

where $\gamma_4(h'_n) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[(h'_n(Z_0, Z_{1:m}) - h'_n(Z_0, Z_{1:m}^{\setminus i}))^4]$ and $C = 36864/49$.

Controlling the error $\hat\sigma^2_{n,in,approx} - \sigma^2_n$ To control the error $\hat\sigma^2_{n,in,approx} - \sigma^2_n$, we first rewrite $\hat\sigma^2_{n,in,approx}$ as

$$
\begin{aligned}
\hat\sigma^2_{n,in,approx} &= \tfrac{1}{k}\textstyle\sum_{j=1}^k \tfrac{1}{(n/k)-1}\sum_{i\in B'_j}\Big(\bar h'_n(Z_i) - \tfrac{k}{n}\sum_{i'\in B'_j}\bar h'_n(Z_{i'})\Big)^2 \\
&= \tfrac{1}{k}\textstyle\sum_{j=1}^k \tfrac{n}{n-k}\tfrac{k}{n}\sum_{i\in B'_j}\Big(\bar h'_n(Z_i) - \tfrac{k}{n}\sum_{i'\in B'_j}\bar h'_n(Z_{i'})\Big)^2 \\
&= \tfrac{n}{n-k}\tfrac{1}{k}\textstyle\sum_{j=1}^k\Big(\tfrac{k}{n}\sum_{i\in B'_j}\bar h'_n(Z_i)^2 - \Big(\tfrac{k}{n}\sum_{i\in B'_j}\bar h'_n(Z_i)\Big)^2\Big) \\
&= \tfrac{n}{n-k}\tfrac{1}{n}\textstyle\sum_{j=1}^k\sum_{i\in B'_j}\bar h'_n(Z_i)^2 - \tfrac{n}{n-k}\tfrac{1}{k}\sum_{j=1}^k\Big(\tfrac{k}{n}\sum_{i\in B'_j}\bar h'_n(Z_i)\Big)^2.
\end{aligned}
$$

We rewrite it once again to find

$$
\begin{aligned}
\hat\sigma^2_{n,in,approx} &= \tfrac{1}{n}\textstyle\sum_{j=1}^k\sum_{i\in B'_j}\bar h'_n(Z_i)^2 - \tfrac{1}{k}\sum_{j=1}^k W_j \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^n \bar h'_n(Z_i)^2 - \tfrac{1}{k}\sum_{j=1}^k W_j \qquad \text{(I.7)}
\end{aligned}
$$

where

$$
W_j \triangleq \tfrac{1}{\binom{n/k}{2}}\textstyle\sum_{\substack{i,i'\in B'_j\\ i<i'}}\bar h'_n(Z_i)\bar h'_n(Z_{i'}).
$$

Since $(W_j)_{j\in[k]}$ are i.i.d. with mean 0 and for $i_1 < i'_1$ and $i_2 < i'_2$

$$
\mathbb{E}[\bar h'_n(Z_{i_1})\bar h'_n(Z_{i'_1})\bar h'_n(Z_{i_2})\bar h'_n(Z_{i'_2})] = 0
$$

whenever $i_1 \neq i_2$ or $i'_1 \neq i'_2$, we have

$$
\begin{aligned}
\mathbb{E}[(\tfrac{1}{k}\textstyle\sum_{j=1}^k W_j)^2] = \tfrac{1}{k}\mathrm{Var}(W_1) &= \tfrac{1}{k}\tfrac{1}{\binom{n/k}{2}^2}\textstyle\sum_{\substack{i,i'\in B'_j\\ i<i'}}\mathbb{E}[\bar h'_n(Z_i)^2\bar h'_n(Z_{i'})^2] \\
&= \tfrac{1}{k}\tfrac{1}{\binom{n/k}{2}^2}\textstyle\sum_{\substack{i,i'\in B'_j\\ i<i'}}\mathbb{E}[\bar h'_n(Z_i)^2]\mathbb{E}[\bar h'_n(Z_{i'})^2] \\
&= \tfrac{1}{k}\tfrac{1}{\binom{n/k}{2}^2}\textstyle\sum_{\substack{i,i'\in B'_j\\ i<i'}}\sigma_n^4 = \tfrac{1}{k}\tfrac{1}{\binom{n/k}{2}}\sigma_n^4 = \tfrac{2}{n(n/k-1)}\sigma_n^4 \qquad \text{(I.8)}
\end{aligned}
$$

by noticing that $\mathbb{E}[\bar h'_n(Z_0)^2] = \mathrm{Var}(\bar h'_n(Z_0)) = \mathrm{Var}(\bar h_n(Z_0)) = \sigma_n^2$.

Moreover, by the independence of our datapoints, we have

$$
\mathbb{E}[\bar h'_n(Z_{i_1})^2\bar h'_n(Z_{i_2})\bar h'_n(Z_{i'_2})] = 0
$$

for all $i_1, i_2, i'_2 \in [n]$ such that $i_2 < i'_2$, and thus

$$
\begin{aligned}
\mathbb{E}[(\hat\sigma^2_{n,in,approx} - \sigma_n^2)^2] &= \mathrm{Var}(\hat\sigma^2_{n,in,approx}) \\
&= \mathrm{Var}(\tfrac{1}{n}\textstyle\sum_{i=1}^n \bar h'_n(Z_i)^2) + \mathbb{E}[(\tfrac{1}{k}\sum_{j=1}^k W_j)^2] \\
&= \tfrac{1}{n}\mathrm{Var}(\bar h'_n(Z_0)^2) + \tfrac{2}{n(n/k-1)}\sigma_n^4 \\
&= \tfrac{1}{n}\mathbb{E}[\bar h'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma_n^4. \qquad \text{(I.9)}
\end{aligned}
$$

Putting the pieces together We have

$$
\mathbb{E}[\sqrt{\Delta}\hat\sigma_{n,in,approx}] \le \sqrt{\mathbb{E}[\Delta]\mathbb{E}[\hat\sigma^2_{n,in,approx}]} \le \sqrt{\tfrac{2n^2}{n-k}\gamma_{loss}(h_n)\sigma_n^2}
$$

by Cauchy–Schwarz and the bound (I.5).

We also have

$$\begin{aligned}\mathbb{E}[\Delta\hat{\sigma}^2_{n,in,approx}] &\le \sqrt{\mathbb{E}[\Delta^2]\mathbb{E}[\hat{\sigma}^4_{n,in,approx}]} \\ &= \sqrt{\mathbb{E}[\Delta^2](\mathrm{Var}(\hat{\sigma}^2_{n,in,approx}) + \mathbb{E}[\hat{\sigma}^2_{n,in,approx}]^2)} \\ &\le \sqrt{\tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n)(\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n + \sigma^4_n)}\end{aligned}$$

by Cauchy-Schwarz, (I.6) and (I.9).

Assembling our results with the triangle inequality and Cauchy–Schwarz for the L^1 bound and with Jensen's inequality for the L^2 bound, we find that

$$\begin{aligned}\mathbb{E}[|\hat{\sigma}^2_{n,in} - \sigma^2_n|] &\le \mathbb{E}[|\hat{\sigma}^2_{n,in} - \hat{\sigma}^2_{n,in,approx}|] + \mathbb{E}[|\hat{\sigma}^2_{n,in,approx} - \sigma^2_n|] \\ &\le \mathbb{E}[\Delta] + 2\mathbb{E}[\sqrt{\Delta}\hat{\sigma}_{n,in,approx}] + \sqrt{\mathbb{E}[(\hat{\sigma}^2_{n,in,approx} - \sigma^2_n)^2]} \\ &\le \tfrac{2n^2}{n-k}\gamma_{loss}(h_n) + 2\sqrt{\tfrac{2n^2}{n-k}\gamma_{loss}(h_n)\sigma^2_n} + \sqrt{\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[(\hat{\sigma}^2_{n,in} - \sigma^2_n)^2] &\le 2\mathbb{E}[(\hat{\sigma}^2_{n,in} - \hat{\sigma}^2_{n,in,approx})^2] + 2\mathbb{E}[(\hat{\sigma}^2_{n,in,approx} - \sigma^2_n)^2] \\ &\le 4\mathbb{E}[\Delta^2] + 8\mathbb{E}[\Delta\hat{\sigma}^2_{n,in,approx}] + 2\mathbb{E}[(\hat{\sigma}^2_{n,in,approx} - \sigma^2_n)^2] \\ &\le 4\tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n) + 8\sqrt{\tfrac{Cn^4}{(n-k)^2}\gamma_4(h'_n)(\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n + \sigma^4_n)} \\ &\quad + 2(\tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^4] + \tfrac{3k-n}{n(n-k)}\sigma^4_n)\end{aligned}$$

as advertised.

In order to get the bound

$$\mathbb{E}[|\hat{\sigma}^2_{n,in} - \sigma^2_n|] \le \tfrac{2n^2}{n-k}\gamma_{loss}(h_n) + 2\sqrt{\tfrac{2n^2}{n-k}\gamma_{loss}(h_n)\sigma^2_n} + \sqrt{\tfrac{2}{n(n/k-1)}\sigma^4_n} + o(\sigma^2_n)$$

whenever the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma^2_n$ is uniformly integrable, i.e., the sequence of $\bar{h}'_n(Z_0)^2/\sigma^2_n$ is uniformly integrable, we need to argue that $\frac{1}{n}\sum_{i=1}^n \bar{h}'_n(Z_i)^2/\sigma^2_n \xrightarrow{L^1} 1$. Indeed, thanks to (I.7) and (I.8), this will lead to $\mathbb{E}[|\hat{\sigma}^2_{n,in,approx} - \sigma^2_n|] \le \sqrt{\frac{2}{n(n/k-1)}\sigma^4_n} + o(\sigma^2_n)$.

To this end, we show that for any triangular i.i.d. array $(X_{n,i})_{n,i}$ such that $(X_{n,1})_{n\ge1}$ is uniformly integrable, then the two conditions in the weak law of large numbers for triangular arrays of [24, Thm. 2.2.11] (stated below) are satisfied. We will also show that for such $(X_{n,i})_{n,i}$, $(S_n \triangleq \frac{1}{n}\sum_{i=1}^n X_{n,i})_{n\ge1}$ is uniformly integrable. Together, these results will imply L^1 convergence. We will then choose $X_{n,i} = \bar{h}'_n(Z_i)^2/\sigma^2_n$ to get the desired result in our specific case.

Theorem 10 (Weak law for triangular arrays [24, Thm. 2.2.11])**.** *For each n, let $X_{n,i}$, $1 \le i \le n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{n,i} = X_{n,i}\mathbb{1}[|X_{n,i}| \le b_n]$. Suppose that as $n \to \infty$*

$$\textstyle\sum_{i=1}^n \mathbb{P}(|X_{n,i}| > b_n) \to 0, \quad \textit{and} \tag{I.10}$$

$$b_n^{-2}\textstyle\sum_{i=1}^n \mathbb{E}[\bar{X}^2_{n,i}] \to 0. \tag{I.11}$$

If we let $S_n = \sum_{i=1}^n X_{n,i}$ and $a_n = \sum_{i=1}^n \mathbb{E}[\bar{X}_{n,i}]$, then $(S_n - a_n)/b_n \xrightarrow{p} 0$.

To prove our result, we specify the case of interest $b_n = n$. First, $n\mathbb{P}(|X_{n,1}| > n) \le \mathbb{E}[|X_{n,1}|\mathbb{1}[|X_{n,1}| > n]] \le \sup_{m\ge1}\mathbb{E}[|X_{m,1}|\mathbb{1}[|X_{m,1}| > n]] \to 0$ as $n \to \infty$, because $(X_{n,1})_{n\ge1}$ is uniformly integrable. Thus the first condition (I.10) holds.

Note that we then get $\mathbb{E}[X_{n,1}\mathbb{1}[|X_{n,1}| \le n]] \to 1$ as $n \to \infty$, for our choice $X_{n,i} = \bar{h}'_n(Z_i)^2/\sigma^2_n$ which satisfies $\mathbb{E}[X_{n,i}] = 1$.

To verify the second condition (I.11), we will show that $n^{-1}\mathbb{E}[X^2_{n,1}\mathbb{1}[X_{n,1} \le n]] \to 0$. To this end, we need the following lemma, which gives a useful formulation of uniform integrability.

Lemma 3 (De la Vallée Poussin Theorem [41, Thm. 22]). *If $(X_n)_{n\geq 1}$ is uniformly integrable, then there exists a nonnegative increasing function G such that $G(t)/t \to \infty$ as $t \to \infty$ and* $\sup_n \mathbb{E}[G(X_n)] < \infty$.

With such a function G, fix any T such that $G(t)/t \geq 1$ for all $t \geq T$, so that $t/G(t) \leq 1$ for all $t \geq T$. Using [24, Lem. 2.2.13] for the first equality, we can write

$$\begin{aligned}\tfrac{1}{n}\mathbb{E}[X_{n,1}^2 \mathbb{1}[X_{n,1} \leq n]] &= \tfrac{2}{n}\textstyle\int_0^\infty y\mathbb{P}(X_{n,1}\mathbb{1}[X_{n,1} \leq n] > y)dy \\ &\leq \tfrac{2}{n}\textstyle\int_0^n y\mathbb{P}(X_{n,1} > y)dy \\ &= \tfrac{2}{n}(\textstyle\int_0^T y\mathbb{P}(X_{n,1} > y)dy + \int_T^n y\mathbb{P}(X_{n,1} > y)dy) \\ &\leq \tfrac{T^2}{n} + \tfrac{2}{n}\textstyle\int_T^n y\mathbb{P}(X_{n,1} > y)dy) \\ &\leq \tfrac{T^2}{n} + \mathbb{E}[G(X_{n,1})]\tfrac{2}{n}\textstyle\int_T^n y/G(y)dy \\ &= \tfrac{T^2}{n} + o(1),\end{aligned}$$

where the penultimate line follows from Markov's inequality and the last line comes from the following lemma since $\sup_{y\geq T} y/G(y) \leq 1$ and $y/G(y) \to 0$.

Lemma 4. *If $f(y) \to 0$ as $y \to \infty$ and $\sup_{y\geq T}|f(y)| \leq M$, then $\frac{1}{n}\int_T^n f(y)dy \to 0$.*

Proof Let $f_n(z) = f(nz)\mathbb{1}[z > T/n]$, and note that, for any $z \geq 0$, $f_n(z) \to 0$ as $n \to \infty$. Then

$$\begin{aligned}\tfrac{1}{n}\textstyle\int_T^n f(y)dy &= \textstyle\int_0^1 \mathbb{1}[z > T/n]f(nz)dz \\ &= \textstyle\int_0^1 f_n(z)dz \\ &\to 0\end{aligned}$$

by the bounded convergence theorem. □

Consequently, the second condition (I.11) holds.

Moreover, $(S_n \triangleq \frac{1}{n}\sum_{i=1}^n X_{n,i})_{n\geq 1}$ is uniformly integrable whenever $(X_{n,i})_{n,i}$ is a triangular i.i.d. array such that $(X_{n,1})_{n\geq 1}$ is uniformly integrable for the following reasons:

1. $\sup_n \mathbb{E}[|S_n|] \leq \sup_n \mathbb{E}[|X_{n,1}|] < \infty$ by triangle inequality and because $(X_{n,1})_{n\geq 1}$ is uniformly integrable.
2. For any $\varepsilon > 0$, let $\delta > 0$ such that for any event A satisfying $\mathbb{P}(A) \leq \delta$, $\sup_n \mathbb{E}[|X_{n,1}|\mathbb{1}[A]] \leq \varepsilon$. Such δ exists because $(X_{n,1})_{n\geq 1}$ is uniformly integrable. Then $\sup_n \mathbb{E}[|S_n|\mathbb{1}[A]] \leq \varepsilon$ by triangle inequality.

The combination of convergence in probability and uniform integrability implies convergence in L^1. As a result, $\frac{1}{n}\sum_{i=1}^n \bar{h}'_n(Z_i)^2/\sigma_n^2 \xrightarrow{L^1} 1$ as long as the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2 = \bar{h}'_n(Z_0)^2/\sigma_n^2$ is uniformly integrable.

Therefore, $\mathbb{E}[|\hat{\sigma}^2_{n,in,approx}/\sigma_n^2 - 1|] \leq \sqrt{\frac{2}{n(n/k-1)}} + o(1)$, and we get the result advertised. □

If $k \leq n/2$, which is the case here since $k < n$ and k divides n, then $\frac{2}{n(n/k-1)} \to 0$ and $\frac{3k-n}{n(n-k)} \to 0$. Therefore, by (I.2), we have $(\hat{\sigma}^2_{n,in} - \sigma_n^2)/\sigma_n^2 \xrightarrow{L^1} 0$, i.e. $\hat{\sigma}^2_{n,in}/\sigma_n^2 \xrightarrow{L^1} 1$, whenever the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable and $\gamma_{loss}(h_n) = o(\frac{n-k}{n^2}\sigma_n^2)$, or equivalently $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ since $k \leq n/2$, and, by (I.1), we have $(\hat{\sigma}^2_{n,in} - \sigma_n^2)/\sigma_n^2 \xrightarrow{L^2} 0$, i.e. $\hat{\sigma}^2_{n,in}/\sigma_n^2 \xrightarrow{L^2} 1$, whenever $\mathbb{E}[\bar{h}'_n(Z_0)^4] = \mathbb{E}[(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^4] = o(n\sigma_n^4)$ and $\gamma_4(h'_n) = o(\frac{(n-k)^2}{n^4}\sigma_n^4)$, or equivalently $\gamma_4(h'_n) = o(\sigma_n^4/n^2)$ since $k \leq n/2$.

Thm. 4 thus follows from Thm. 9.

Strengthening of the consistency result of [5, Prop. 1] We provide more details about the comparison of our L^2-consistency result with [5, Prop. 1]. We have $\gamma_4(h'_n) \leq 16\gamma_4(h_n)$ and

$\mathbb{E}[(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^4] \leq 16\mathbb{E}[h_n(Z_0, Z_{1:m})^4]$ by Jensen's inequality. Moreover, if $\tilde{\sigma}_n^2$ converges to a non-zero constant, since $\gamma_{loss}(h_n) \leq \gamma_{ms}(h_n) \leq \sqrt{\gamma_4(h_n)}$, then $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ whenever $\gamma_4(h_n) = o(\sigma_n^4/n^2)$ and thus σ_n^2 converges to the same non-zero constant as $\tilde{\sigma}_n^2$ does by Prop. 2.

J Proof of Thm. 5: Consistent all-pairs estimate of asymptotic variance

We will prove the following more detailed statement from which Thm. 5 will follow.

Theorem 11 (Consistent all-pairs estimate of asymptotic variance)**.** *Suppose that k divides n evenly. Under the notation of Thm. 1 with $m = n(1 - 1/k)$, $\bar{h}_n(z) = \mathbb{E}[h_n(z, Z_{1:m})]$, $h_n'(Z_0, Z_{1:m}) = h_n(Z_0, Z_{1:m}) - \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_{1:m}]$ and $\bar{h}_n'(z) = \mathbb{E}[h_n'(z, Z_{1:m})]$, define the all-pairs variance estimate*

$$\hat{\sigma}_{n,out}^2 \triangleq \tfrac{1}{k} \textstyle\sum_{j=1}^k \tfrac{k}{n} \sum_{i \in B_j'} \Big(h_n(Z_i, Z_{B_j}) - \hat{R}_n\Big)^2.$$

If $(Z_i)_{i \geq 1}$ are i.i.d. copies of a random element Z_0 and $\tilde{\sigma}_n^2 = \mathbb{E}[h_n'(Z_0, Z_{1:m})^2]$, then

$$\begin{aligned}\mathbb{E}[|\hat{\sigma}_{n,out}^2 - \sigma_n^2|] \leq (1 + \tfrac{n}{k})\gamma_{ms}(h_n) + 2\sqrt{2(1 + \tfrac{n}{k})\gamma_{ms}(h_n)\tilde{\sigma}_n^2} + m\gamma_{loss}(h_n) \\ + 2\sqrt{m\gamma_{loss}(h_n)(1 - \tfrac{1}{n})\sigma_n^2} + \sqrt{\tfrac{1}{n}(\mathbb{E}[\bar{h}_n'(Z_0)^4] - \sigma_n^4)} + \tfrac{1}{n}\sigma_n^2.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}[|\hat{\sigma}_{n,out}^2 - \sigma_n^2|] \leq (1 + \tfrac{n}{k})\gamma_{ms}(h_n) + 2\sqrt{2(1 + \tfrac{n}{k})\gamma_{ms}(h_n)\tilde{\sigma}_n^2} + m\gamma_{loss}(h_n) \\ + 2\sqrt{m\gamma_{loss}(h_n)(1 - \tfrac{1}{n})\sigma_n^2} + \tfrac{1}{n}\sigma_n^2 + o(\sigma_n^2).\end{aligned} \tag{J.1}$$

whenever the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable.

Proof

A common training set for each validation point pair We begin by approximating our variance estimate

$$\begin{aligned}\hat{\sigma}_{n,out}^2 &= \tfrac{1}{n} \textstyle\sum_{j=1}^k \sum_{i \in B_j'} \Big(h_n(Z_i, Z_{B_j}) - \hat{R}_n\Big)^2 \\ &= \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} \tfrac{1}{2}(h_n(Z_i, Z_{B_j}) - h_n(Z_{i'}, Z_{B_{j'}}))^2\end{aligned}$$

by a quantity that employs the same training set for each pair of validation points $Z_{(i,i')}$,

$$\begin{aligned}\hat{\sigma}_{n,out,approx,1}^2 &\triangleq \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} \tfrac{1}{2}(h_n(Z_i, Z_{B_j}^{\backslash i'}) - h_n(Z_{i'}, Z_{B_j}^{\backslash i'}))^2 \\ &= \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} \tfrac{1}{2}(h_n'(Z_i, Z_{B_j}^{\backslash i'}) - h_n'(Z_{i'}, Z_{B_j}^{\backslash i'}))^2.\end{aligned}$$

Here, for any $j \in [k]$ and $i' \in [n]$, $Z_{B_j}^{\backslash i'}$ is Z_{B_j} with $Z_{i'}$ replaced by Z_0. By Cauchy–Schwarz, we have

$$|\hat{\sigma}_{n,out}^2 - \hat{\sigma}_{n,out,approx,1}^2| \leq \Delta_1 + 2\sqrt{\Delta_1}\hat{\sigma}_{n,out,approx,1}$$

for the error term

$$\begin{aligned}\Delta_1 &\triangleq \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} \tfrac{1}{2}(h_n(Z_i, Z_{B_j}) - h_n(Z_i, Z_{B_j}^{\backslash i'}) + h_n(Z_{i'}, Z_{B_j}^{\backslash i'}) - h_n(Z_{i'}, Z_{B_{j'}}))^2 \\ &\leq \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} (h_n(Z_i, Z_{B_j}) - h_n(Z_i, Z_{B_j}^{\backslash i'}))^2 \\ &\quad + \tfrac{1}{n^2} \textstyle\sum_{j,j'=1}^k \sum_{i \in B_j', i' \in B_{j'}'} (h_n(Z_{i'}, Z_{B_j}^{\backslash i'}) - h_n(Z_{i'}, Z_{B_{j'}}))^2,\end{aligned} \tag{J.2}$$

where we have used Jensen's inequality in the final display.

Controlling the error Δ_1 We will first control the error term Δ_1. Note that, for $B_{j'} \neq B_j$, $|B_{j'} \backslash (B_{j'} \cap B_j)| = \frac{n}{k}$. Hence, by the bound (J.2) and the conditional Efron-Stein inequality (Lemma 1), we have

$$\begin{aligned}\mathbb{E}[\Delta_1] &\leq \gamma_{ms}(h_n) + \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} \mathbb{E}[(h_n(Z_{i'}, Z_{B_j}^{\backslash i'}) - h_n(Z_{i'}, Z_{B_{j'}}))^2] \\ &\leq \gamma_{ms}(h_n) + \tfrac{n}{k}\gamma_{ms}(h_n) = (1+\tfrac{n}{k})\gamma_{ms}(h_n). \end{aligned} \tag{J.3}$$

Eliminating training set randomness We then approximate $\hat{\sigma}^2_{n,out,approx,1}$ by a quantity eliminating training set randomness in each summand,

$$\hat{\sigma}^2_{n,out,approx,2} \triangleq \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} \tfrac{1}{2}(\bar{h}'_n(Z_i) - \bar{h}'_n(Z_{i'}))^2$$

where $\bar{h}'_n(z) = \mathbb{E}[h'_n(z, Z_{1:m})]$. Note that $\bar{h}'_n(Z_0)$ has expectation 0.

By Cauchy–Schwarz, we have

$$|\hat{\sigma}^2_{n,out,approx,1} - \hat{\sigma}^2_{n,out,approx,2}| \leq \Delta_2 + 2\sqrt{\Delta_2}\hat{\sigma}_{n,out,approx,2}$$

for the error term

$$\begin{aligned}\Delta_2 &\triangleq \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} \tfrac{1}{2}(h'_n(Z_i, Z_{B_j}^{\backslash i'}) - \bar{h}'_n(Z_i) + \bar{h}'_n(Z_{i'}) - h'_n(Z_{i'}, Z_{B_j}^{\backslash i'}))^2 \\ &\leq \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} (h'_n(Z_i, Z_{B_j}^{\backslash i'}) - \bar{h}'_n(Z_i))^2 \\ &\quad + \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} (\bar{h}'_n(Z_{i'}) - h'_n(Z_{i'}, Z_{B_j}^{\backslash i'}))^2, \end{aligned} \tag{J.4}$$

where we have used Jensen's inequality in the final display.

Controlling the error Δ_2 We will control the error term Δ_2. By the bound (J.4) and the conditional Efron-Stein inequality (Lemma 1), we have

$$\mathbb{E}[\Delta_2] \leq 2\tfrac{m}{2}\gamma_{ms}(h'_n) = m\gamma_{loss}(h_n). \tag{J.5}$$

Controlling the error $\hat{\sigma}^2_{n,out,approx,2} - \sigma^2_n$ To control the error $\hat{\sigma}^2_{n,out,approx,2} - \sigma^2_n$, we first rewrite $\hat{\sigma}^2_{n,out,approx,2}$ as

$$\begin{aligned}\hat{\sigma}^2_{n,out,approx,2} &= \tfrac{1}{n^2}\textstyle\sum_{j,j'=1}^{k}\sum_{i\in B'_j, i'\in B'_{j'}} \tfrac{1}{2}(\bar{h}'_n(Z_i) - \bar{h}'_n(Z_{i'}))^2 \\ &= \tfrac{1}{n^2}\textstyle\sum_{i,i'=1}^{n} \tfrac{1}{2}(\bar{h}'_n(Z_i) - \bar{h}'_n(Z_{i'}))^2 \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\big(\bar{h}'_n(Z_i) - \tfrac{1}{n}\sum_{i'=1}^{n}\bar{h}'_n(Z_{i'})\big)^2 \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\bar{h}'_n(Z_i)^2 - \big(\tfrac{1}{n}\sum_{i=1}^{n}\bar{h}'_n(Z_i)\big)^2. \end{aligned} \tag{J.6}$$

Since $\mathbb{E}[\bar{h}'_n(Z_i)\bar{h}'_n(Z_{i'})] = 0$ for all $i, i' \in [n]$ with $i \neq i'$ due to independence, we have

$$\mathbb{E}[\big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\bar{h}'_n(Z_i)\big)^2] = \tfrac{1}{n}\mathbb{E}[\bar{h}'_n(Z_0)^2] = \tfrac{1}{n}\sigma^2_n. \tag{J.7}$$

Furthermore,

$$\mathbb{E}[(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\bar{h}'_n(Z_i)^2 - \sigma^2_n)^2] = \mathrm{Var}(\tfrac{1}{n}\sum_{i=1}^{n}\bar{h}'_n(Z_i)^2) = \tfrac{1}{n}\mathrm{Var}(\bar{h}'_n(Z_0)^2) = \tfrac{1}{n}(\mathbb{E}[\bar{h}'_n(Z_0)^4] - \sigma^4_n)$$

by independence. Hence,

$$\mathbb{E}[|\hat{\sigma}^2_{n,out,approx,2} - \sigma^2_n|] \leq \sqrt{\tfrac{1}{n}(\mathbb{E}[\bar{h}'_n(Z_0)^4] - \sigma^4_n)} + \tfrac{1}{n}\sigma^2_n.$$

Putting the pieces together Since each

$$\tfrac{1}{2}(h'_n(Z_i, Z_{B_j}^{\backslash i'}) - h'_n(Z_{i'}, Z_{B_j}^{\backslash i'}))^2 \leq h'_n(Z_i, Z_{B_j}^{\backslash i'})^2 + h'_n(Z_{i'}, Z_{B_j}^{\backslash i'})^2,$$

we have

$$\mathbb{E}[\hat{\sigma}^2_{n,out,approx,1}] \leq 2\mathbb{E}[h'_n(Z_0, Z_{1:m})^2] = 2\tilde{\sigma}^2_n$$

and hence

$$\mathbb{E}[\sqrt{\Delta_1}\hat{\sigma}_{n,out,approx,1}] \leq \sqrt{\mathbb{E}[\Delta_1]\mathbb{E}[\hat{\sigma}^2_{n,out,approx,1}]} \leq \sqrt{2(1+\tfrac{n}{k})\gamma_{ms}(h_n)\tilde{\sigma}^2_n}$$

by Cauchy–Schwarz and the bound (J.3).

Moreover, $\mathbb{E}[\hat{\sigma}^2_{n,out,approx,2}] = (1-\frac{1}{n})\sigma^2_n$, hence

$$\mathbb{E}[\sqrt{\Delta_2}\hat{\sigma}_{n,out,approx,2}] \leq \sqrt{\mathbb{E}[\Delta_2]\mathbb{E}[\hat{\sigma}^2_{n,out,approx,2}]} \leq \sqrt{m\gamma_{loss}(h_n)(1-\tfrac{1}{n})\sigma^2_n}$$

by Cauchy–Schwarz and the bound (J.5).

Assembling our results with the triangle inequality, we find that

$$\begin{aligned}
\mathbb{E}[|\hat{\sigma}^2_{n,out} - \sigma^2_n|] &\leq \mathbb{E}[|\hat{\sigma}^2_{n,out} - \hat{\sigma}^2_{n,out,approx,1}|] + \mathbb{E}[|\hat{\sigma}^2_{n,out,approx,1} - \hat{\sigma}^2_{n,out,approx,2}|] \\
&\quad + \mathbb{E}[|\hat{\sigma}^2_{n,out,approx,2} - \sigma^2_n|] \\
&\leq \mathbb{E}[\Delta_1] + 2\mathbb{E}[\sqrt{\Delta_1}\hat{\sigma}_{n,out,approx,1}] \\
&\quad + \mathbb{E}[\Delta_2] + 2\mathbb{E}[\sqrt{\Delta_2}\hat{\sigma}_{n,out,approx,2}] \\
&\quad + \sqrt{\tfrac{1}{n}(\mathbb{E}[\bar{h}'_n(Z_0)^4] - \sigma^4_n)} + \tfrac{1}{n}\sigma^2_n \\
&\leq (1+\tfrac{n}{k})\gamma_{ms}(h_n) + 2\sqrt{2(1+\tfrac{n}{k})\gamma_{ms}(h_n)\tilde{\sigma}^2_n} \\
&\quad + m\gamma_{loss}(h_n) + 2\sqrt{m\gamma_{loss}(h_n)(1-\tfrac{1}{n})\sigma^2_n} \\
&\quad + \sqrt{\tfrac{1}{n}(\mathbb{E}[\bar{h}'_n(Z_0)^4] - \sigma^4_n)} + \tfrac{1}{n}\sigma^2_n
\end{aligned}$$

as advertised.

We showed in the proof of Thm. 9 that $\frac{1}{n}\sum_{i=1}^n \bar{h}'_n(Z_i)^2/\sigma^2_n \overset{L^1}{\to} 1$ whenever the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma^2_n = \bar{h}'_n(Z_0)^2/\sigma^2_n$ is uniformly integrable. Thus, with (J.6) and (J.7), we get $\mathbb{E}[|\hat{\sigma}^2_{n,out,approx,2}/\sigma^2_n - 1|] \leq 1/n + o(1)$, and the final bound advertised

$$\begin{aligned}
\mathbb{E}[|\hat{\sigma}^2_{n,out} - \sigma^2_n|] &\leq (1+\tfrac{n}{k})\gamma_{ms}(h_n) + 2\sqrt{2(1+\tfrac{n}{k})\gamma_{ms}(h_n)\tilde{\sigma}^2_n} \\
&\quad + m\gamma_{loss}(h_n) + 2\sqrt{m\gamma_{loss}(h_n)(1-\tfrac{1}{n})\sigma^2_n} \\
&\quad + \tfrac{1}{n}\sigma^2_n + o(\sigma^2_n).
\end{aligned}$$

□

By the bound (J.1), $(\hat{\sigma}^2_{n,out} - \sigma^2_n)/\sigma^2_n \overset{L^1}{\to} 0$, i.e. $\hat{\sigma}^2_{n,out}/\sigma^2_n \overset{L^1}{\to} 1$, if the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma^2_n$ is uniformly integrable, $\gamma_{loss}(h_n) = o(\sigma^2_n/n)$ and $\gamma_{ms}(h_n) = o(\min(\frac{k\sigma^2_n}{n}, \frac{k\,\sigma^4_n}{n\,\tilde{\sigma}^2_n}))$. By noticing that $\tilde{\sigma}^2_n/\sigma^2_n \to 1$ when $\gamma_{loss}(h_n) = o(\sigma^2_n/n)$ thanks to Prop. 2, the last condition becomes $\gamma_{ms}(h_n) = o(k\sigma^2_n/n)$. Therefore, Thm. 5 follows from Thm. 11.

K Experimental Setup Details

Here, we provide more details about the experimental setup of Sec. 5.

K.1 General experimental setup details

Learning algorithms and hyperparameters To illustrate the performance of our confidence intervals and tests in practice, we carry out our experiments with a diverse collection of popular learning algorithms. For classification, we use the `xgboost XGBRFClassifier` with `n_estimators=100`, `subsample=0.5` and `max_depth=1`, the `scikit-learn MLPClassifier` neural network with `hidden_layer_sizes=(8,4,)` defining the architecture and `alpha=1e2`, and the `scikit-learn` ℓ^2-penalized `LogisticRegression` with `solver='lbfgs'` and `C=1e-3`. For regression, we use the `xgboost XGBRFRegressor` with `n_estimators=100`, `subsample=0.5` and `max_depth=1`, the `scikit-learn MLPRegressor` neural network with `hidden_layer_sizes=(8,4,)` defining the architecture and `alpha=1e2`, and the `scikit-learn Ridge` regressor with `alpha=1e6`. The random forest `max_depth` hyperparameter and neural network, logistic, and ridge ℓ^2 regularization strengths were selected to ensure the stability of each algorithm. All remaining hyperparameters are set to their defaults, and we set random seeds for all algorithms' random states for reproducibility. We use `scikit-learn` [44] version 0.22.1 and `xgboost` [17] version 1.0.2.

Training set sample sizes n For both datasets, we work with the following training set sample sizes n: 700, 1,000, 1,500, 2,300, 3,400, 5,000, 7,500, 11,000. Up to some rounding, this corresponds to a geometric sequence with growth rate 50%.

Details on the `Higgs` dataset The target variable has value either 0 or 1 and there are 28 features. We initially shuffle the rows of the dataset uniformly at random and then, starting at the 5,000,001-th instance, we take 500 consecutive chunks of the largest sample size, that is 11,000. For each n, we take the first n instances of these 500 chunks to play the role of our 500 independent replications of size n. The features are standardized during training in the following way: for each iteration of k-fold CV ($k = 10$ here), we rescale the validation fold and the remaining folds, used as training, with the mean and standard deviation of the training data. The features for the training folds then have mean 0 and variance 1.

Details on the `FlightsDelay` dataset To avoid the temporal dependence issues inherent to time series datasets, we treat the complete `FlightsDelay` dataset as the population and thus process it differently from the `Higgs` dataset. For this dataset, we predict the signed log transform ($y \mapsto \mathrm{sign}(y)\log(1 + |y|)$; this addressed the very heavy tails of y on its original scale) of the delay at arrival using 4 features: the scheduled time of the journey from the origin airport to the destination airport (taxi included), the distance between the two airports, the scheduled time of departure in minutes (converted from a time to a number between 0 and 1,439) and the airline operating the plane (that we one-hot encode). We drop the instances that have missing values for at least one of these variables. Then, we perform 500 times the sampling with replacement of 11,000 points, that is the largest sample size. For each n, we take the first n instances of these 500 chunks to play the role of our 500 independent replications of size n. The features are standardized during training in the same way we do for the `Higgs` dataset.

Computing target test errors For the `FlightDelays` experiments, since training datapoints are sampled with replacement, the population distribution is the entirety of the `FlightDelays` dataset, and we use this exact population distribution to compute all test errors. For the `Higgs` experiments, we form a surrogate ground-truth estimate of the target test errors using the first 5,000,000 datapoints of the shuffled `Higgs` dataset. As an illustration, for our method where the target test error is the k-fold test error $R_n = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in B'_j}\mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}] = \frac{1}{k}\sum_{j=1}^{10}\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}]$, we use these instances to compute the k conditional expectations by a Monte Carlo approximation. Practically, for each training set Z_B, we compute the average loss on these instances of the fitted prediction rule learned on Z_B. Then, we evaluate the CIs and tests constructed from the 500 training sets of varying sizes n sampled from the datasets.

Random seeds Seeds are set in the code to ensure reproducibility. They are used for the initial random shuffling of the datasets, the sampling with replacement for the regression dataset, the random partitioning of samples in each replication, and the randomized algorithms.

K.2 List of procedures

In our numerical experiments, we compare our procedures with the most popular alternatives from the literature. For each procedure, we give its target test error R_n, the estimator $\hat{R}_n$ of this target, the variance estimator $\hat{\sigma}_n^2$, the two-sided CI used in Sec. 5.1, and the one-sided test used in Sec. 5.2.

In the following, q_α is the α-quantile of a standard normal distribution and $t_{\nu,\alpha}$ is the α-quantile of a t distribution with ν degrees of freedom.

1. Our 10-fold CV CLT-based test, with $\hat{\sigma}_n$ being either $\hat{\sigma}_{n,in}$ (Thm. 4) or $\hat{\sigma}_{n,out}$ (Thm. 5). The curve with $\hat{\sigma}_{n,in}$ is not displayed in our plots since the results are almost identical to those for $\hat{\sigma}_{n,out}$ and the curves are overlapping.
   - Target test error: $R_n = \frac{1}{10}\sum_{j=1}^{10}\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}]$.
   - Estimator: $\hat{R}_n = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in B'_j} h_n(Z_i, Z_{B_j})$.
   - Variance estimator: $\hat{\sigma}_n^2$, either $\hat{\sigma}_{n,in}^2$ or $\hat{\sigma}_{n,out}^2$.
   - Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm q_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n}$.
   - One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < q_\alpha\hat{\sigma}_n/\sqrt{n}$.
2. Hold-out test described, for instance, in Austern and Zhou [5, Eq. (17)].
   - Target test error: $R_n = \mathbb{E}[h_n(Z_0, Z_S) \mid Z_S]$, where S is a subset of size $\lfloor n(1-1/10)\rfloor$ of $[n]$. Since we already have a partition for our 10-fold CV, we can use the first fold B_1 for S.
   - Estimator: $\hat{R}_n = \frac{1}{|S^c|}\sum_{i\in S^c} h_n(Z_i, Z_S)$.
   - Variance estimator: $\hat{\sigma}_n^2 = \frac{1}{|S^c|}\sum_{i\in S^c}(h_n(Z_i, Z_S) - \hat{R}_n)^2$.
   - Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm q_{1-\alpha/2}\hat{\sigma}_n\sqrt{10}/\sqrt{n}$.
   - One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < q_\alpha\hat{\sigma}_n\sqrt{10}/\sqrt{n}$.
3. Cross-validated t-test of Dietterich [22], 10 folds.
   - Target test error: $R_n = \frac{1}{10}\sum_{j=1}^{10}\mathbb{E}[h_n(Z_0, Z_{B_j}) \mid Z_{B_j}]$.
   - Estimator: $\hat{R}_n = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in B'_j} h_n(Z_i, Z_{B_j})$.
   - Variance estimator: $\hat{\sigma}_n^2 = \frac{1}{10-1}\sum_{j=1}^{10}(p_j - \hat{R}_n)^2$, where $p_j \triangleq \frac{1}{|B'_j|}\sum_{i\in B'_j} h_n(Z_i, Z_{B_j})$.
   - Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm t_{10-1,1-\alpha/2}\hat{\sigma}_n/\sqrt{10}$.
   - One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < t_{10-1,\alpha}\hat{\sigma}_n/\sqrt{10}$.
4. Repeated train-validation t-test of Nadeau and Bengio [43], 10 repetitions of 90-10 train-validation splits.
   - Target test error: $R_n = \frac{1}{10}\sum_{j=1}^{10}\mathbb{E}[h_n(Z_0, Z_{S_j}) \mid Z_{S_j}]$, where for any $j \in [10]$, S_j is a subset of size $\lfloor n(1-1/10)\rfloor$ of $[n]$, and these 10 subsets are chosen independently.
   - Estimator: $\hat{R}_n = \frac{1}{10}\sum_{j=1}^{10} p_j$, where $p_j \triangleq \frac{1}{|S_j^c|}\sum_{i\in S_j^c} h_n(Z_i, Z_{S_j})$.
   - Variance estimator: $\hat{\sigma}_n^2 = \frac{1}{10-1}\sum_{j=1}^{10}(p_j - \hat{R}_n)^2$.
   - Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm t_{10-1,1-\alpha/2}\hat{\sigma}_n/\sqrt{10}$.
   - One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < t_{10-1,\alpha}\hat{\sigma}_n/\sqrt{10}$.
5. Corrected repeated train-validation t-test of Nadeau and Bengio [43], 10 repetitions of 90-10 train-validation splits.
   - Target test error: $R_n = \frac{1}{10}\sum_{j=1}^{10}\mathbb{E}[h_n(Z_0, Z_{S_j}) \mid Z_{S_j}]$, where for any $j \in [10]$, S_j is the same as in the previous procedure.
   - Estimator: $\hat{R}_n = \frac{1}{10}\sum_{j=1}^{10} p_j$, where p_j is the same as in the previous procedure.
   - Variance estimator: $\hat{\sigma}_n^2 = (\frac{1}{10} + \frac{0.1}{1-0.1})\frac{10}{10-1}\sum_{j=1}^{10}(p_j - \hat{R}_n)^2$.
   - Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm t_{10-1,1-\alpha/2}\hat{\sigma}_n/\sqrt{10}$.
   - One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < t_{10-1,\alpha}\hat{\sigma}_n/\sqrt{10}$.
6. 5×2-fold CV test of Dietterich [22].
   - Target test error: $R_n = \frac{1}{5}\sum_{j=1}^{5}\frac{1}{2}(\mathbb{E}[h_n(Z_0, Z_{B_{1,j}}) \mid Z_{B_{1,j}}] + \mathbb{E}[h_n(Z_0, Z_{B_{2,j}}) \mid Z_{B_{2,j}}])$, where for any $j \in [5]$, $\{B_{1,j}^c, B_{2,j}^c\}$ is a partition of $[n]$ into 2 folds of size $n/2$, and these 5 partitions are chosen independently.
   - Estimator: $\hat{R}_n = \frac{1}{|B_{1,1}^c|}\sum_{i\in B_{1,1}^c} h_n(Z_i, Z_{B_{1,1}})$.

- Variance estimator: $\hat{\sigma}_n^2 = \frac{1}{5}\sum_{j=1}^{5} s_j^2$, where $s_j^2 \triangleq (p_{1,j} - \bar{p}_j)^2 + (p_{2,j} - \bar{p}_j)^2$ with $\bar{p}_j \triangleq (p_{1,j} + p_{2,j})/2$ and $p_{k,j} \triangleq \frac{1}{|B_{k,j}^c|}\sum_{i \in B_{k,j}^c} h_n(Z_i, Z_{B_{k,j}})$ for $k \in [2], j \in [5]$.
- Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm t_{5,1-\alpha/2}\hat{\sigma}_n$.
- One-sided test: REJECT $H_0 \Leftrightarrow \hat{R}_n < t_{5,\alpha}\hat{\sigma}_n$.

K.3 Concentration-based confidence intervals

For comparison in Sec. 1, we also implemented the ridge regression CI from [16, Thm. 3] for the `FlightDelays` experiment (an implementable CI is not provided for any other learning algorithm in [16]). This CI takes as input a uniform bound B_Y on the absolute value of the target variable Y and a uniform bound B_X on the ℓ^2 norm of the feature vector X. After mean-centering, we find the maximum absolute value of Y across the `FlightDelays` dataset to be $B_Y = 8.03$. After mean-centering, we find the maximum ℓ^2 norm of a feature vector X across the `FlightDelays` to be $B_X = 13.17$ if each feature is normalized to have standard deviation 1 or $B_X = 4200$ if the features are left unnormalized. When normalizing features as in Fig. 5, the smallest width produced by [16, Thm. 3] for any value of n is 90.2; that is 91 times larger than the largest width of our CLT intervals (equal to 0.99). When not normalizing as in Fig. 3, our maximum width is 0.98, but the minimum [16, Thm. 3] width is 5×10^{14}.

K.4 Leave-one-out cross-validation

To evaluate the LOOCV CLT-based CIs discussed in Sec. 5.4 we follow the ridge regression experimental setup of App. K.1 except that we regress onto the raw feature values instead of the standardized features values described in App. K.1. For our LOOCV CLT-based CIs, the quantities of interest are the following.

- Target test error: $R_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[h_n(Z_0, Z_{\{i\}^c}) \mid Z_{\{i\}^c}]$.
- Estimator: $\hat{R}_n = \frac{1}{n}\sum_{i=1}^{n} h_n(Z_i, Z_{\{i\}^c})$ computed efficiently using the Sherman–Morrison–Woodbury derivation below.
- Variance estimator: $\hat{\sigma}_{n,out}^2$ with $k = n$ folds.
- Two-sided $(1-\alpha)$-CI: $\hat{R}_n \pm q_{1-\alpha/2}\hat{\sigma}_{n,out}/\sqrt{n}$.

Results We construct 95% CIs for ridge regression test error based on our LOOCV CLT and compare their coverage and width with those of the procedures described in Sec. 5.1. We see that, like the 10-fold CV CLT intervals, the LOOCV intervals provide coverage near the nominal level and widths smaller than the popular alternatives from the literature; in fact, the 10-fold CV CLT curves are obscured by the nearly identical LOOCV CLT curves.

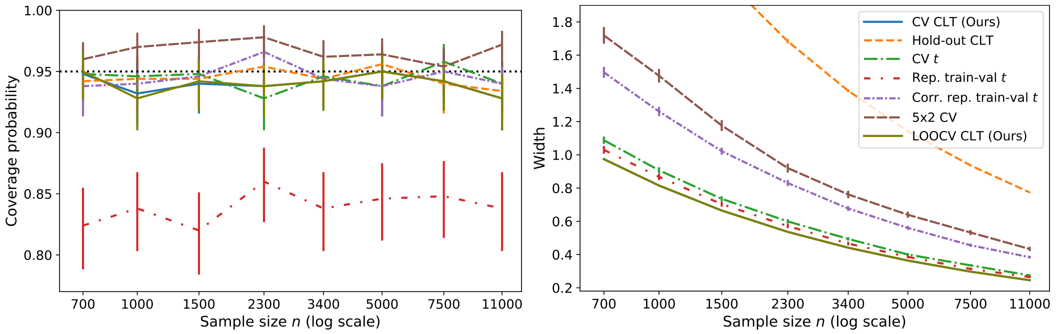


Figure 3: Test error coverage (left) and width (right) of 95% confidence intervals for ridge regression, including leave-one-out CV intervals (see Sec. 5.4). The CV CLT curves are obscured by the nearly identical LOOCV CLT curves.

Efficient computation We explain here how the Sherman–Morrison–Woodbury formula can be used to efficiently compute the individual losses $h_n(Z_i, Z_{\{i\}^c})$, and therefore $\hat{R}_n$ as well as $\hat{\sigma}_{n,out}$, and the loss on the instances used to form a surrogate ground-truth estimate of the target error R_n. Let $X \in \mathbb{R}^{n\times p}$ be the matrix of predictors, whose i-th row is $x_i^\top$, and $Y \in \mathbb{R}^n$ be the target variable. The weight vector estimate $\hat{w}$ minimizes $\min_{w\in\mathbb{R}^p} \|Y - Xw\|_2^2 + \lambda\|w\|_2^2$, and is given by the closed-form formula

$$\hat{w} = (X^\top X + \lambda I_p)^{-1} X^\top Y.$$

We precompute $M \triangleq (X^\top X + \lambda I_p)^{-1}$ and $v \triangleq X^\top Y$, that satisfy $\hat{w} = Mv$. Suppose that we have an additional set with covariate matrix $\tilde{X}$ and target variable $\tilde{Y}$, representing the instances used to form a surrogate ground-truth estimate of R_n. We also precompute $q \triangleq \tilde{X}\hat{w}$ and $A \triangleq \tilde{X}M$.

For the datapoint i, let $X^{(-i)}$ denote X without its i-th row and $Y^{(-i)}$ denote Y without its i-th element. Let $M_i \triangleq ({X^{(-i)}}^\top X^{(-i)} + \lambda I_p)^{-1}$, $v_i \triangleq {X^{(-i)}}^\top Y^{(-i)}$ and $w_i \triangleq M_i v_i$. We can efficiently compute M_i from M based on the Sherman–Morrison–Woodbury formula.

$$\begin{aligned} M_i &= ({X^{(-i)}}^\top X^{(-i)} + \lambda I_p)^{-1} \\ &= (X^\top X - x_i x_i^\top + \lambda I_p)^{-1} \\ &= M - M x_i x_i^\top M / (-1 + h_i), \end{aligned}$$

where $h_i \triangleq x_i^\top M x_i$.

We can compute v_i from v, with $v_i = {X^{(-i)}}^\top Y^{(-i)} = v - x_i y_i$.

Therefore, $w_i = M_i v_i = (M - M x_i x_i^\top M(-1+h_i)^{-1})(v - x_i y_i) = \hat{w} + M x_i(\langle \hat{w}, x_i\rangle - y_i)/(1 - h_i)$ can be computed without fitting any additional prediction rule. Then $h_n(Z_i, Z_{\{i\}^c}) = (y_i - \langle w_i, x_i\rangle)^2$, and we use them to compute $\hat{R}_n$ and $\hat{\sigma}_{n,out}$. To make predictions for the covariate matrix $\tilde{X}$, we efficiently compute $\tilde{X} w_i$ as

$$\begin{aligned} \tilde{X} w_i &= \tilde{X}\hat{w} + \tilde{X} M x_i(\langle \hat{w}, x_i\rangle - y_i)/(1 - h_i) \\ &= q + A x_i(\langle \hat{w}, x_i\rangle - y_i)/(1 - h_i), \end{aligned}$$

and $\frac{1}{N}\|\tilde{Y} - \tilde{X} w_i\|_2^2$ is an estimate of $\mathbb{E}[h_n(Z_0, Z_{\{i\}^c}) \mid Z_{\{i\}^c}]$, where N is the size of the whole dataset. An estimate of R_n is then $\frac{1}{n}\sum_{i=1}^n \frac{1}{N}\|\tilde{Y} - \tilde{X} w_i\|_2^2$.

L Additional Experimental Results

This section reports the additional results of the experiments described in Sec. 5.

L.1 Additional results from Sec. 5.1: Confidence intervals for test error

The remaining results of the experiments described in Sec. 5.1 are provided in Figs. 4 and 5. We remind that each mean width estimate is displayed with a ± 2 standard error confidence band, while the confidence band surrounding each coverage estimate is a 95% Wilson interval. For all 6 learning tasks, all procedures except the repeated train-validation t interval provide near-nominal coverage, and our CV CLT intervals provide the smallest widths.

L.2 Additional results from Sec. 5.2: Testing for improved algorithm performance

In this section, we provide additional experimental details and results for the testing for improved algorithm performance experiments of Sec. 5.2. We highlight that the aim of this assessment is not to establish power convergence or to assess power in an absolute sense but rather to verify whether, for a diversity of settings encountered in real learning problems, our proposed tests provide power comparable to or better than the most popular heuristics from the literature. For all testing experiments, we estimate size as $\frac{\text{\# of rejections in } H_0 \text{ replications}}{\text{\# } H_0 \text{ replications}}$ and power as $\frac{\text{\# of rejections in } H_1 \text{ replications}}{\text{\# } H_1 \text{ replications}}$, where each simulation is classified as H_0 or H_1 depending on which algorithm has smaller test

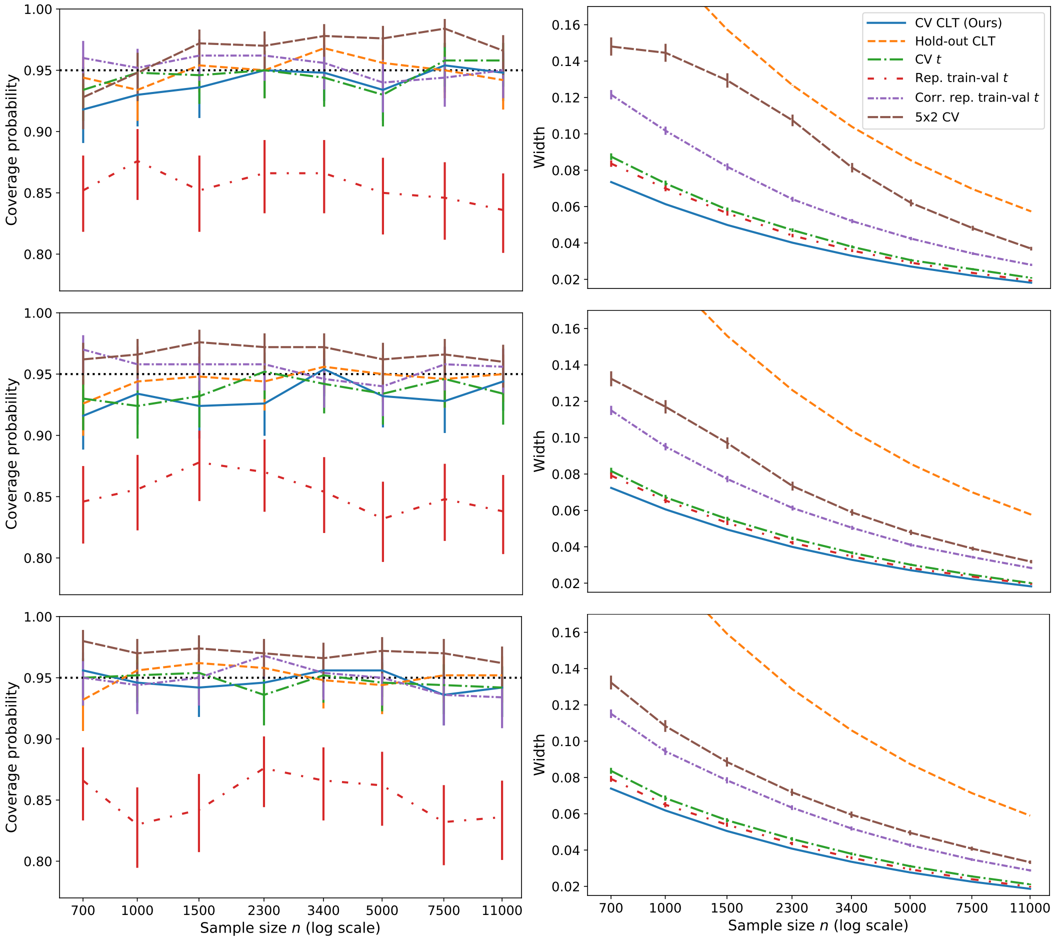


Figure 4: Test error coverage (left) and width (right) of 95% confidence intervals (see Sec. 5.1). **Top:** ℓ^2-regularized logistic regression classifier. **Middle:** Random forest classifier. **Bottom:** Neural network classifier.

error. Moreover, a size point is only displayed if at least 25 replications were classified as H_0, and a power point is only displayed if at least 25 replications were classified as H_1.

The remaining results of the testing experiments described in Sec. 5.2 are provided in Figs. 6 to 11. In contrast to Fig. 2,[9] in Figs. 6 to 11 we plot the size and power of the level $\alpha = 0.05$ test of $H_1 : \text{Err}(\mathcal{A}_1) < \text{Err}(\mathcal{A}_2)$ in the left column of each figure and the size and power of the level α test of $H_1 : \text{Err}(\mathcal{A}_2) < \text{Err}(\mathcal{A}_1)$ in the right column. Notably, we only observe size estimates exceeding the level when the number of H_0 replications is very small (that is, when one algorithm improves upon the other so infrequently that the Monte Carlo error in the size estimate is large).

L.3 Testing with synthetically generated labels

We complement the real-data hypothesis testing experiments of Sec. 5.2 with a controlled experiment in which class labels are synthetically generated from a known logistic regression distribution. Specifically, we replicate the exact classification experimental setup of Sec. 5.2 to compare logistic regression and random forest classification and use the same `Higgs` dataset covariates, but we replace each datapoint label Y_i with an independent draw from the logistic regression distribution $Y_i \sim \text{Ber}(\frac{1}{1+\exp(-\langle X_i, \beta \rangle)})$ for β a 28-dimensional vector with odd entries equal to 1 and even en-

[9] Recall that in Fig. 2 we identified the algorithm $\mathcal{A}_1$ that more often had smaller test error across our simulations and displayed the power of $H_1 : \text{Err}(\mathcal{A}_1) < \text{Err}(\mathcal{A}_2)$ and the size of the level $\alpha = 0.05$ test of $H_1 : \text{Err}(\mathcal{A}_2) < \text{Err}(\mathcal{A}_1)$.

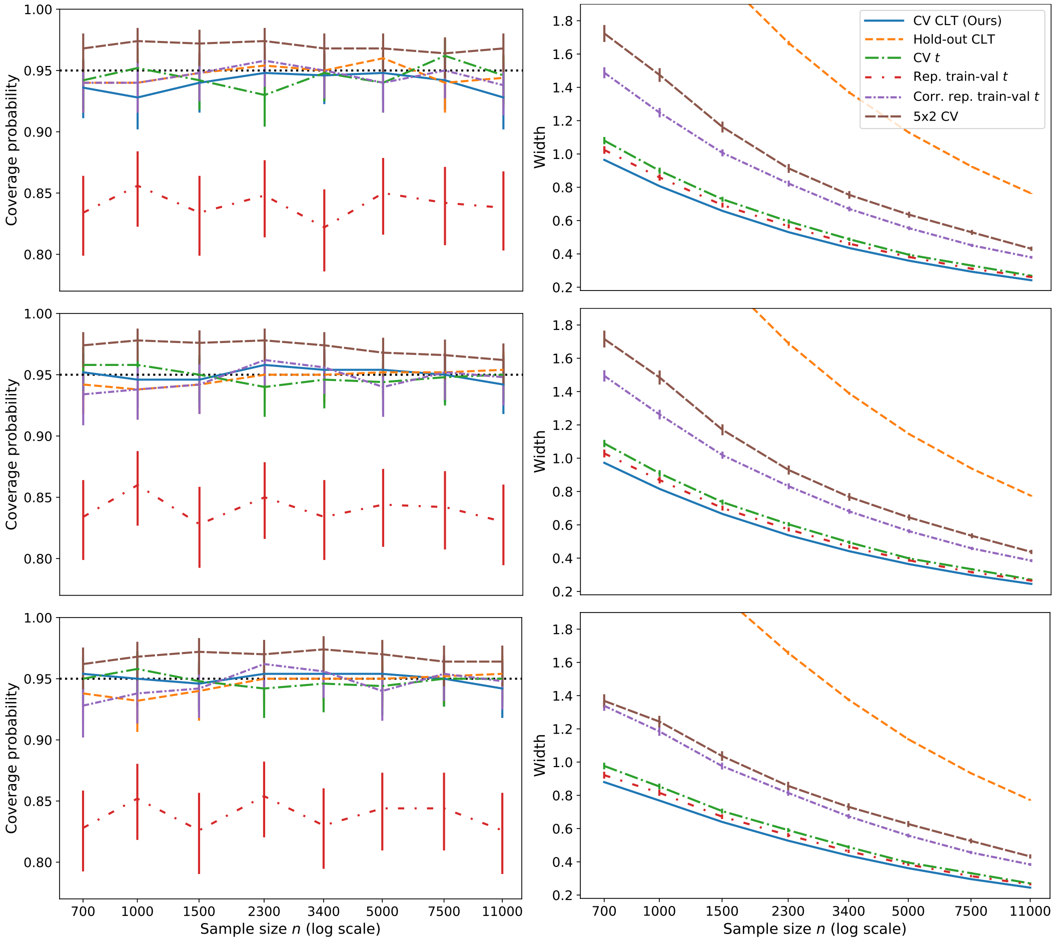


Figure 5: Test error coverage (left) and width (right) of 95% confidence intervals (see Sec. 5.1). **Top:** Random forest regression. **Middle:** Ridge regression. **Bottom:** Neural network regression.

tries equal to -1. This experiment enables us to evaluate our hypothesis tests in a *realizable* setting in which the true label generating distribution belongs to the logistic regression model family. In Fig. 12, we plot in the left column the size and power of the level $\alpha = 0.05$ test of H_1: random forest improves upon ℓ^2-regularized logistic regression classifier, and in the right column the size and power of the level α test of H_1: ℓ^2-regularized logistic regression classifier improves upon random forest. As expected, almost all replications satisfy that ℓ^2-regularized logistic regression improves upon random forest and we observe that in this setting as well, our method consistently outperforms other alternatives.

L.4 Results for Sec. 5.3: Importance of stability

In this section, we provide the figures (Figs. 13 to 15) and experimental details supporting the importance of stability experiment of Sec. 5.3. Compared to the chosen hyperparameters described in App. K, for this example, we used the default value of `max_depth` for `XGBRFRegressor`, that is 6, and the default value of `alpha` for `MLPRegressor`, that is 1e-4. For Figs. 15a and 15b, we obtain an estimate of $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$ by computing a Monte Carlo approximation of $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:m}) \mid Z_0]$ for each of 10,000 Z_0 values and then reporting the empirical variance of these 10,000 approximated values. For each value of Z_0 we employ the Monte Carlo approximation of

$$\bar{h}_n(Z_0) \approx \frac{1}{500} \sum_{\ell=1}^{500} \frac{1}{k} \sum_{j=1}^{k} h_n(Z_0, Z_{B_j}^{(\ell)})$$

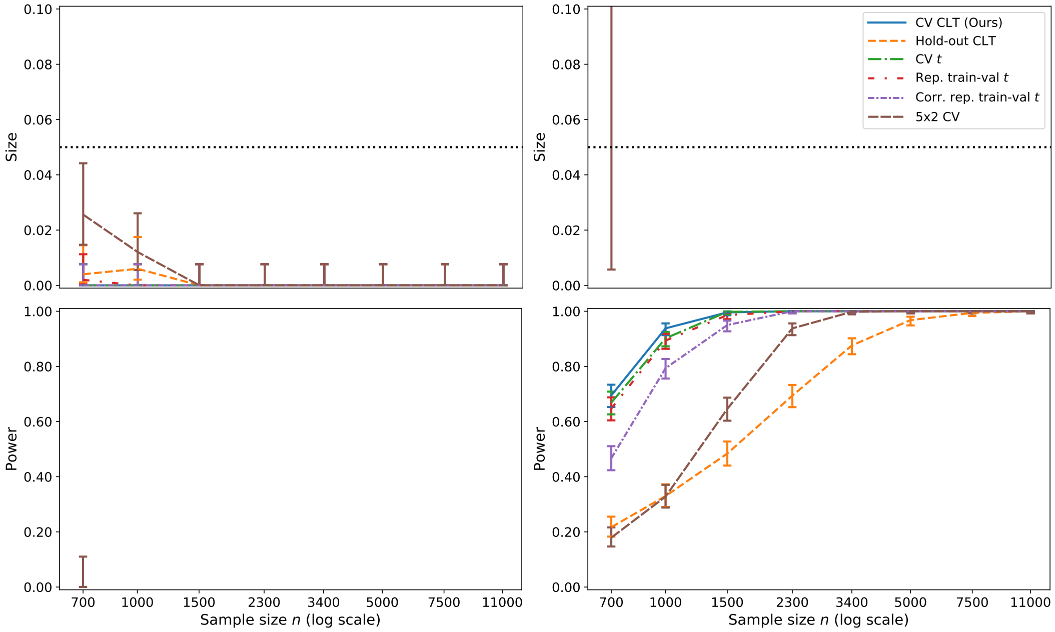


Figure 6: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: neural network improves upon ℓ^2-regularized logistic regression classifier. **Right**: Testing H_1: ℓ^2-regularized logistic regression classifier improves upon neural network.

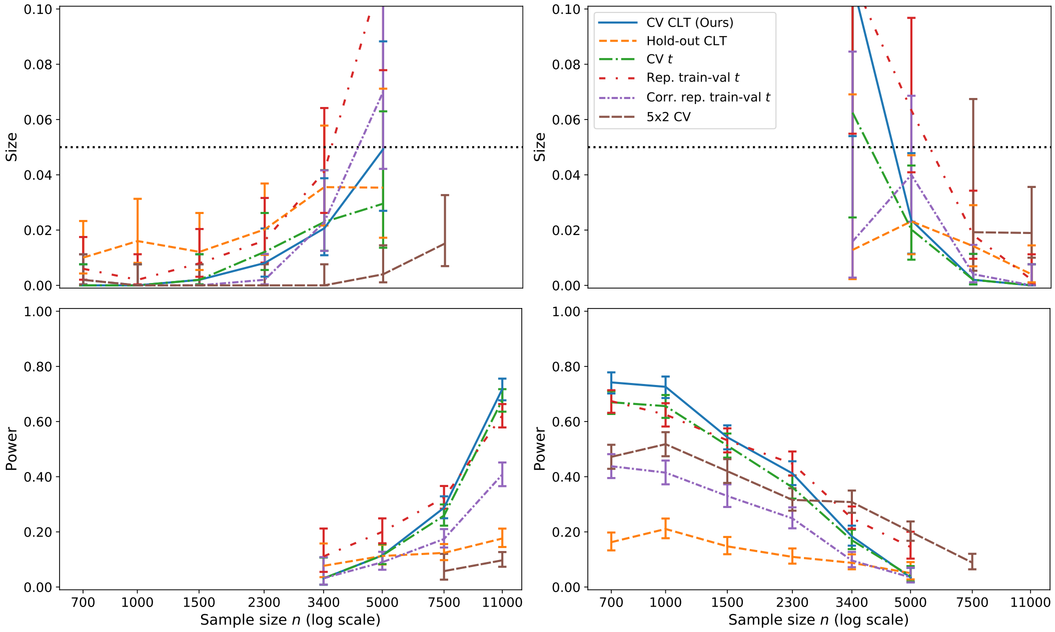


Figure 7: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: ℓ^2-regularized logistic regression classifier improves upon random forest. **Right**: Testing H_1: random forest improves upon ℓ^2-regularized logistic regression classifier.

where $(Z_{1:n}^{(\ell)})_{\ell=1}^{500}$ are the 500 datasets of size n described in App. K.

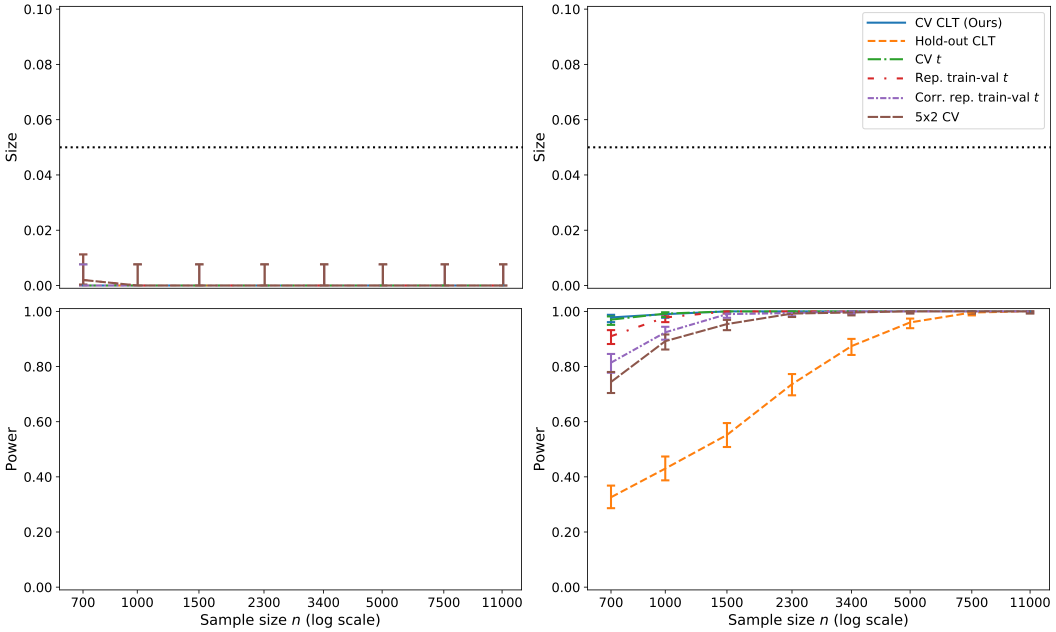


Figure 8: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: neural network classifier improves upon random forest. **Right**: Testing H_1: random forest classifier improves upon neural network.

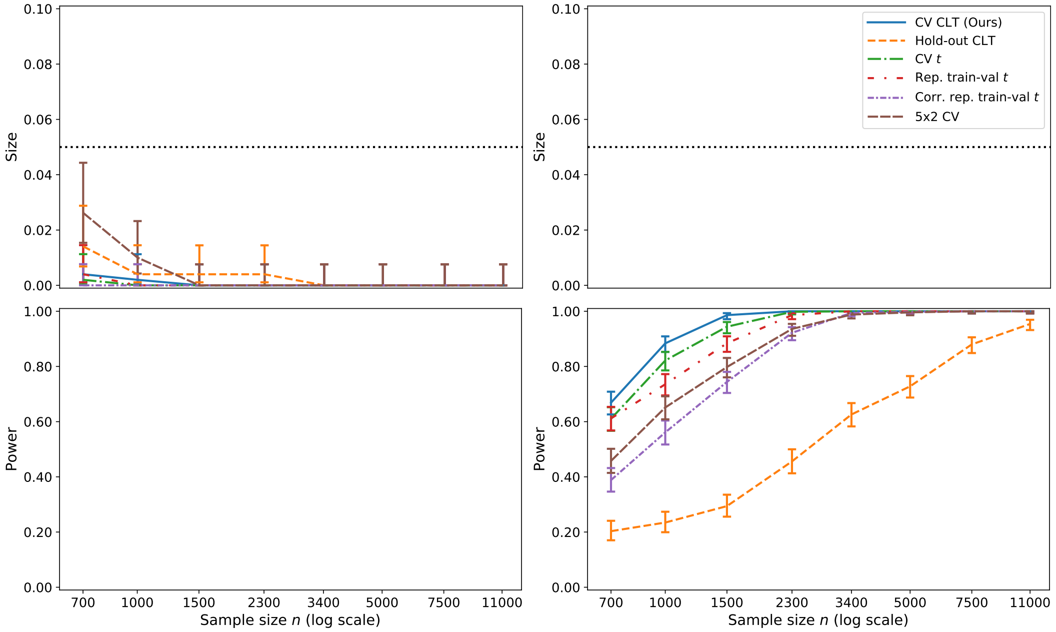


Figure 9: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: ridge regression improves upon random forest. **Right**: Testing H_1: random forest improves upon ridge regression.

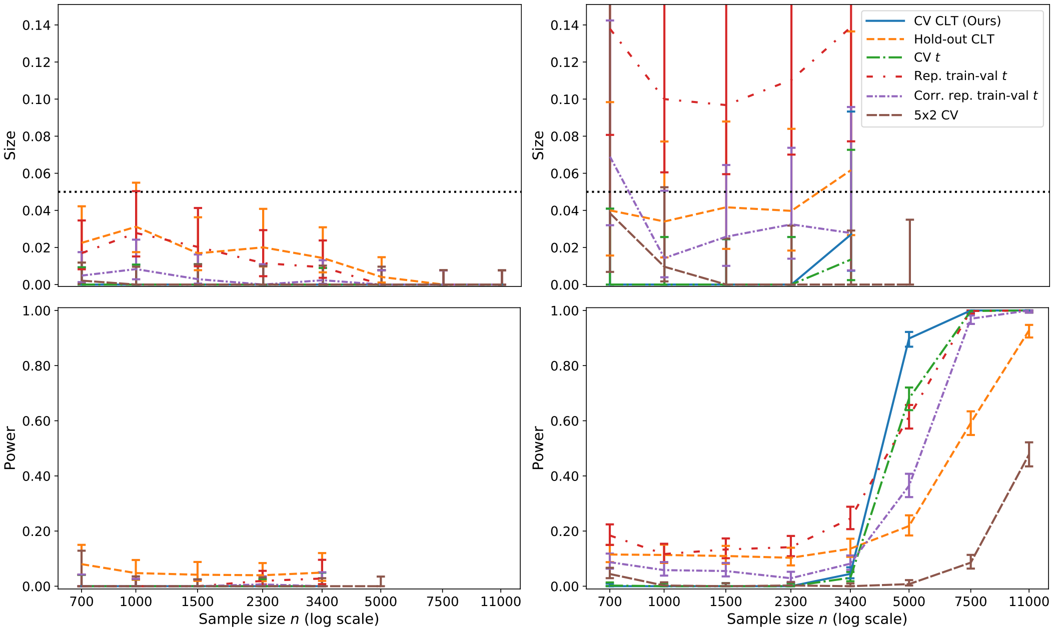


Figure 10: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: neural network improves upon ridge regression. **Right**: Testing H_1: ridge regression improves upon neural network.

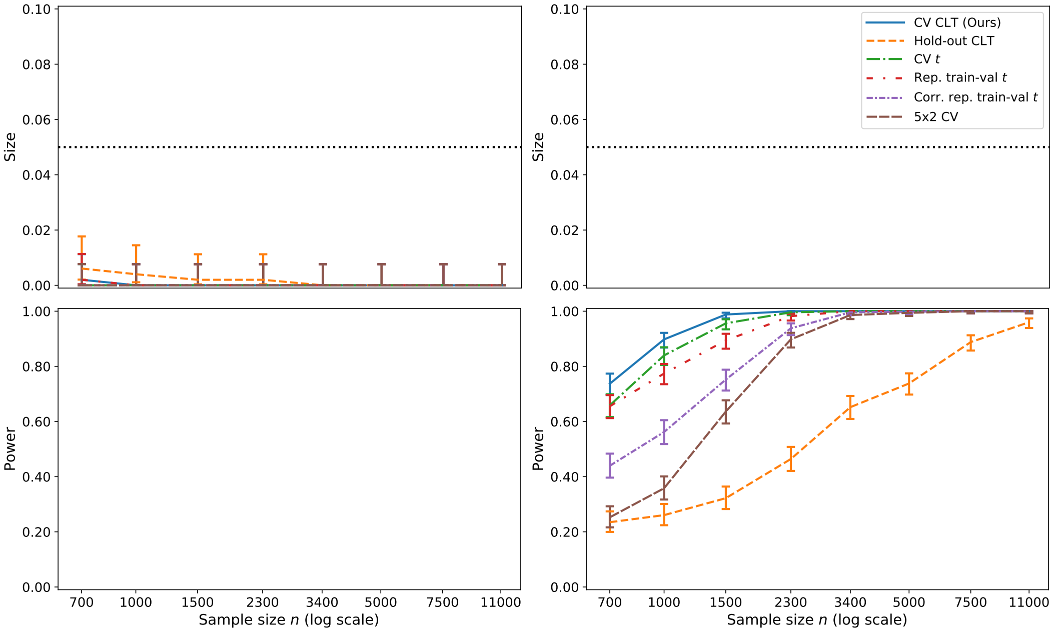


Figure 11: Size (top) and power (bottom) of level-0.05 tests for improved test error (see Sec. 5.2). **Left**: Testing H_1: neural network regression improves upon random forest. **Right**: Testing H_1: random forest regression improves upon neural network.

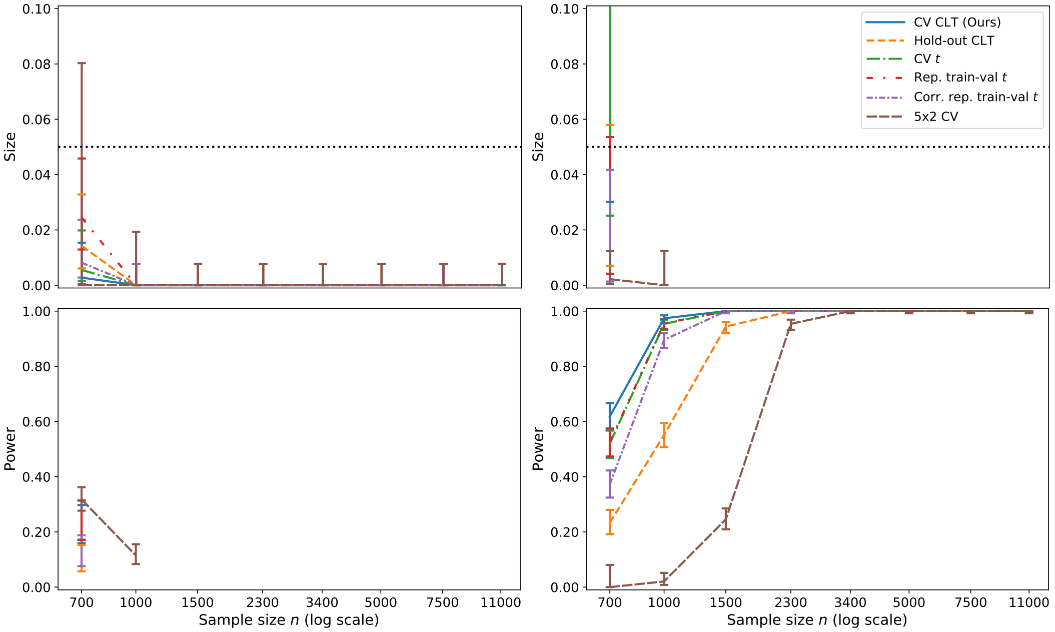


Figure 12: Size (top) and power (bottom) of level-0.05 tests for improved test error **with synthetic logistic regression labels** (see App. L.3). **Left**: Testing H_1: random forest improves upon ℓ^2-regularized logistic regression classifier. **Right**: Testing H_1: ℓ^2-regularized logistic regression classifier improves upon random forest.

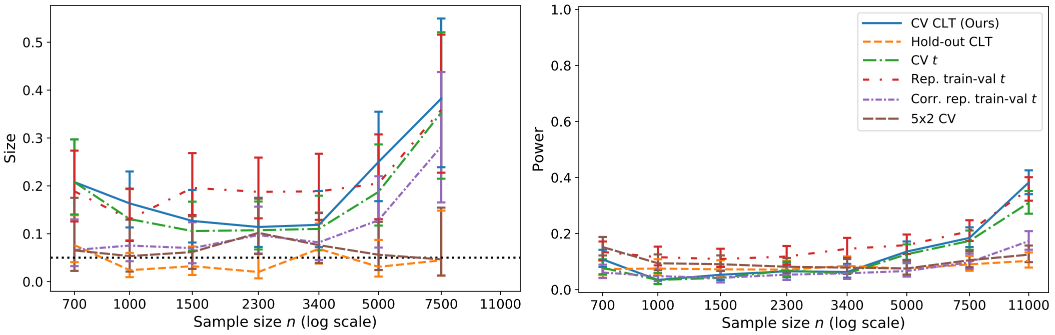


Figure 13: **Impact of instability** on size (left) and power (right) of level-0.05 tests for improved test error (see Sec. 5.3). Testing H_1: less stable neural network regression improves upon less stable random forest.

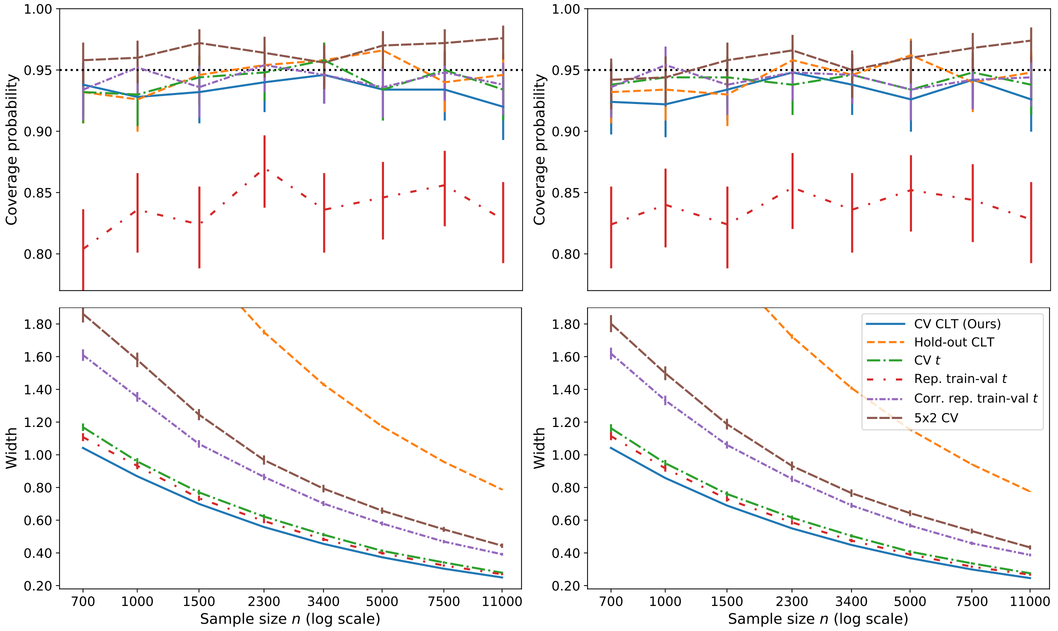
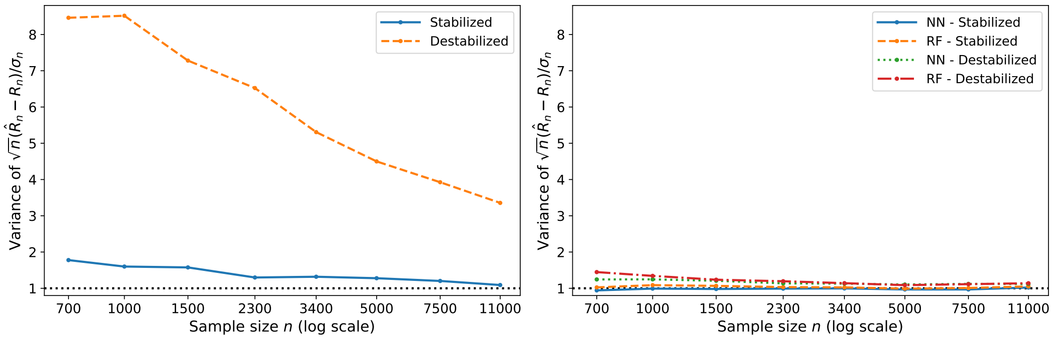


Figure 14: **Impact of instability** on test error coverage (top) and width (bottom) of 95% confidence intervals (see Sec. 5.3). **Left:** Less stable neural network regression. **Right:** Less stable random forest regression.



(a) Algorithm comparison

(b) Single algorithm assessment

Figure 15: **Impact of instability** on variance of $\frac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n)$ (see Sec. 5.3). **Left:** $h_n(Z_0, Z_B) = (Y_0 - \hat{f}_1(X_0; Z_B))^2 - (Y_0 - \hat{f}_2(X_0; Z_B))^2$ for neural network and random forest prediction rules, $\hat{f}_1$ and $\hat{f}_2$. As predicted in Thms. 1 and 2, the variance is close to 1 when h_n is stable, but the variance can be much larger when h_n is unstable. **Right:** $h_n(Z_0, Z_B) = (Y_0 - \hat{f}(X_0; Z_B))^2$ for neural network or random forest prediction rule, $\hat{f}$. The same destabilized algorithms produce relatively stable h_n in the context of single algorithm assessment, as the variance parameter $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$ is larger.