Cross-validation Confidence Intervals for Test Error

Pierre Bayle[1], **Alexandre Bayle**[2], Lucas Janson[2] and Lester Mackey[3]

[1]Princeton University [2]Harvard University [3]Microsoft Research New England

Overview

Cross-validation (CV) is a de facto standard for estimating the test error of a prediction rule. 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. The tools most often used have no correctness guarantees and can be severely misleading, and thus accurately quantifying the uncertainty of the CV estimate is essential yet challenging because of its complex dependence structure.

1. We characterize the **asymptotic distribution** of the CV error and develop two **consistent** estimators of the asymptotic variance under weak stability conditions on the learning algorithm.
2. We provide **practical**, **asymptotically-exact** confidence intervals (CIs) for test error as well as **powerful**, **asymptotically-valid** hypothesis tests of whether one learning algorithm has smaller test error than another.
3. We observe consistent **improvements** in width and power over the most popular alternative methods from the literature.

Setting and Notation

Datapoints $(Z_i)_{i\geq 1}$ i.i.d. copies of a random element Z_0. $Z_{1:n}$ designates the first n points, and, for any vector B of indices in $[n]$, Z_B denotes the subvector of $Z_{1:n}$ corresponding to ordered indices in B. The number of folds k can be either fixed or dependent on n.

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**

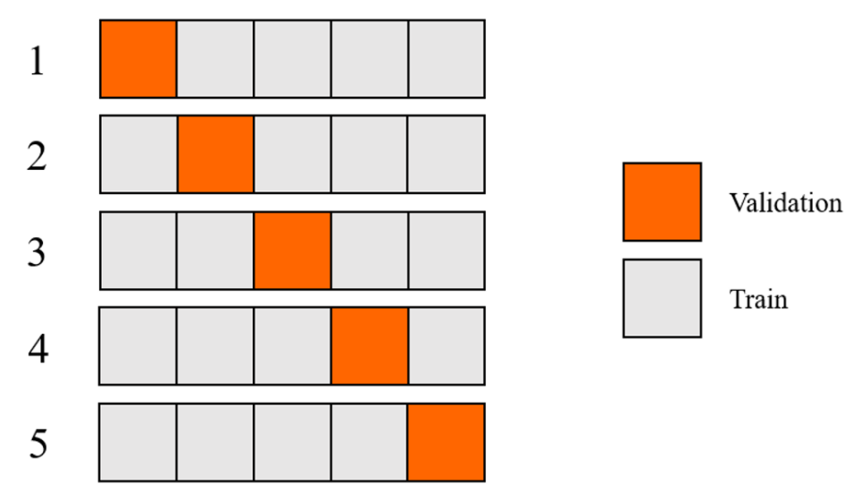

Figure 1: k-fold cross-validation

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

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

$$R_n \triangleq \tfrac{1}{n}\textstyle\sum_{j=1}^k \sum_{i\in B'_j} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]. \tag{2}$$

Standard inferential target representing the average test error of the k prediction rules $\hat{f}(\cdot; Z_{B_j})$ for $j = 1, \dots, k$. 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 $Z_i = (X_i, Y_i)$.

Algorithmic Stability

Definition 1: Mean-square stability and loss stability

For $m > 0$, let Z_0 and $Z'_0, Z_1, \dots, Z_m$ 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** [2] is defined as

$$\gamma_{ms}(h) \triangleq \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}[(h(Z_0, Z_{1:m}) - h(Z_0, Z_{1:m}^{\setminus i}))^2] \tag{3}$$

and the **loss stability** [3] as $\gamma_{loss}(h) \triangleq \gamma_{ms}(h')$, where

$$h'(Z_0, Z_{1:m}) \triangleq h(Z_0, Z_{1:m}) - \mathbb{E}[h(Z_0, Z_{1:m}) \mid Z_{1:m}]. \tag{4}$$

Many learning algorithms are known to enjoy decaying loss stability (e.g., SGD, ERM, k-NN, decision trees, ensemble methods), in part because loss stability is upper-bounded by a variety of algorithmic stability notions studied in the literature. In particular, $\gamma_{loss}(h) \leq \gamma_{ms}(h)$.

Asymptotic Normality

Theorem 1: Asymptotic normality of cross-validation

Let $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_0]$ and $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$. If the following conditions hold:

Stability $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$,

Uniform integrability (UI) $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is a UI sequence,

then

$$\boldsymbol{\tfrac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n) \xrightarrow{d} \mathcal{N}(0,1)}. \tag{5}$$

These assumptions are very mild and satisfied by many algorithms.

CV Confidence Intervals & Tests for k-fold Test Error

- Goal 1: construct an asymptotically-exact $(1-\alpha)$-confidence interval for the unknown k-fold test error R_n.
- Proposal: a 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,

$$C_\alpha \triangleq \hat{R}_n \pm q_{1-\alpha/2}\,\hat{\sigma}_n/\sqrt{n} \quad \text{satisfying} \quad \lim_{n\to\infty} \mathbb{P}(R_n \in C_\alpha) = 1 - \alpha, \tag{6}$$

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

- Goal 2: test whether algorithm $\mathcal{A}_1$ improves upon algorithm $\mathcal{A}_2$ on the fold partition.
- Proposal: 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 $Z_0 = (X_0, Y_0)$. An asymptotically-exact level-α test is given by

$$\text{REJECT } H_0 \Leftrightarrow \hat{R}_n < q_\alpha \hat{\sigma}_n/\sqrt{n}, \tag{7}$$

  for $H_0 : R_n \geq 0$ against $H_1 : R_n < 0$.

Consistent Variance Estimators

For $k < n$, define the within-fold variance estimator

$$\hat{\sigma}_{n,in}^2 \triangleq \tfrac{1}{k}\textstyle\sum_{j=1}^k \frac{1}{(n/k)-1} \sum_{i\in B'_j}\Big(h_n(Z_i, Z_{B_j}) - \frac{k}{n}\sum_{i'\in B'_j} h_n(Z_{i'}, Z_{B_j})\Big)^2. \tag{8}$$

As this estimator necessarily excludes the case of leave-one-out CV ($k = n$), we propose a second estimator with consistency guarantees for any k: the all-pairs variance estimator

$$\hat{\sigma}_{n,out}^2 \triangleq \tfrac{1}{k}\textstyle\sum_{j=1}^k \frac{k}{n} \sum_{i\in B'_j}(h_n(Z_i, Z_{B_j}) - \hat{R}_n)^2. \tag{9}$$

Theorem 2: Consistent estimators of asymptotic variance

If $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$ and $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is UI, then $\boldsymbol{\hat{\sigma}_{n,in}^2/\sigma_n^2 \xrightarrow{L^1} 1}$. If additionally $\gamma_{ms}(h_n) = o(k\sigma_n^2/n)$, then $\boldsymbol{\hat{\sigma}_{n,out}^2/\sigma_n^2 \xrightarrow{L^1} 1}$.

- The same two conditions—loss stability and UI sequence—grant both a central limit theorem for CV and an L^1-consistent estimate of σ_n^2.
- MSS condition $\gamma_{ms}(h_n) = o(k\sigma_n^2/n)$ especially mild when $k = \Omega(n)$ (as in LOOCV).
- Both $\hat{\sigma}_{n,in}^2$ and $\hat{\sigma}_{n,out}^2$ 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 CV. When h_n is binary, as in the case of 0-1 loss, one can compute $\hat{\sigma}_{n,out}^2$ in $O(1)$ time given access to the overall cross-validation error $\hat{R}_n$ and $\hat{\sigma}_{n,in}^2$ in $O(k)$ time given access to the k average fold errors.

Numerical Experiments

We compare on Higgs boson and flight delay data our test error CIs (6) and tests for algorithm improvement (7) with the most popular alternatives from the literature. 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 procedure (solid blue line) achieves the target coverage probability for the CIs and target level for the constructed tests, and consistently delivers the **smallest width** for the CIs and the **best power** for the tests, across all sample sizes.

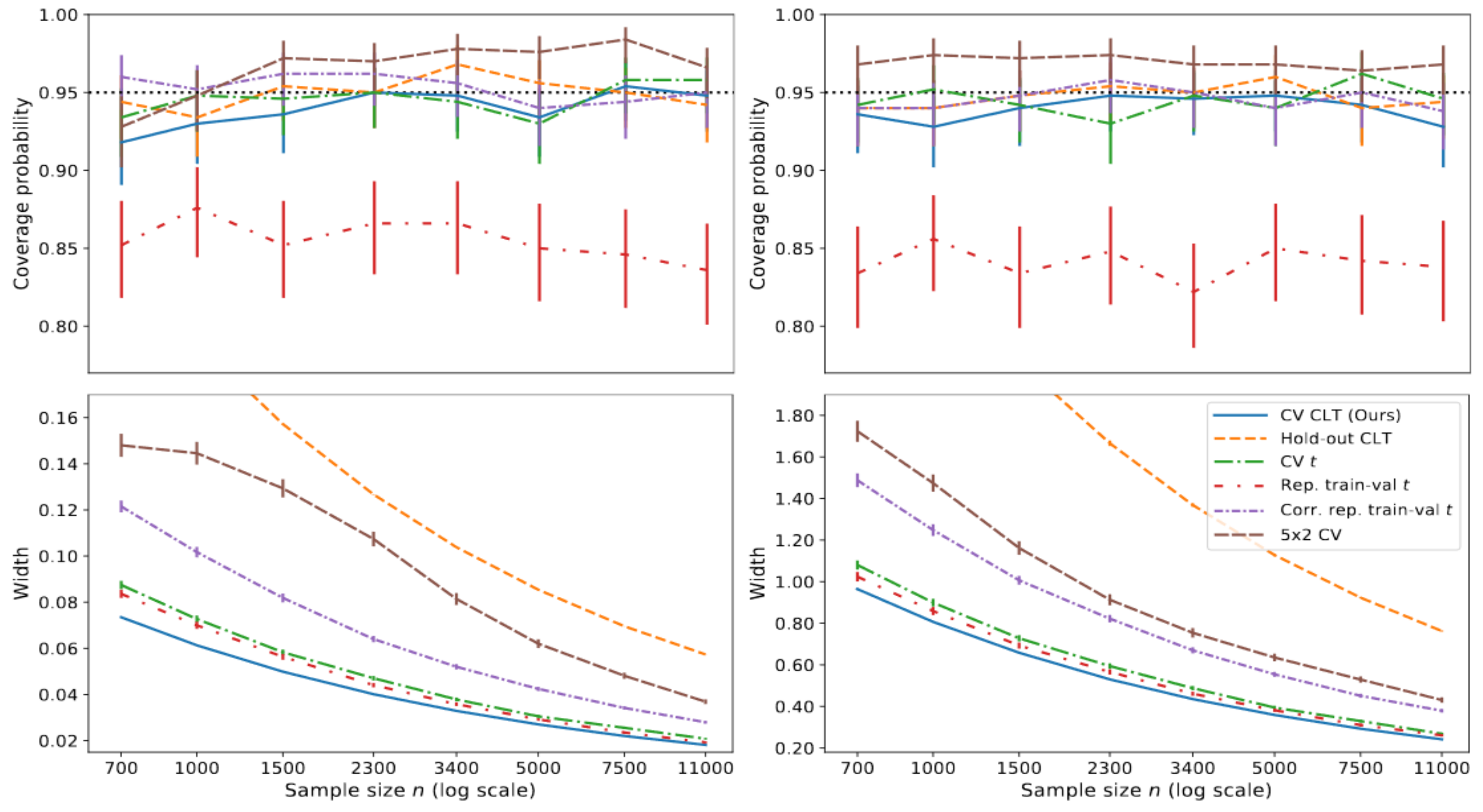

Figure 2: Test error coverage (top) and width (bottom) of 95% confidence intervals. **Left:** ℓ^2-regularized logistic regression classifier. **Right:** Random forest regression.

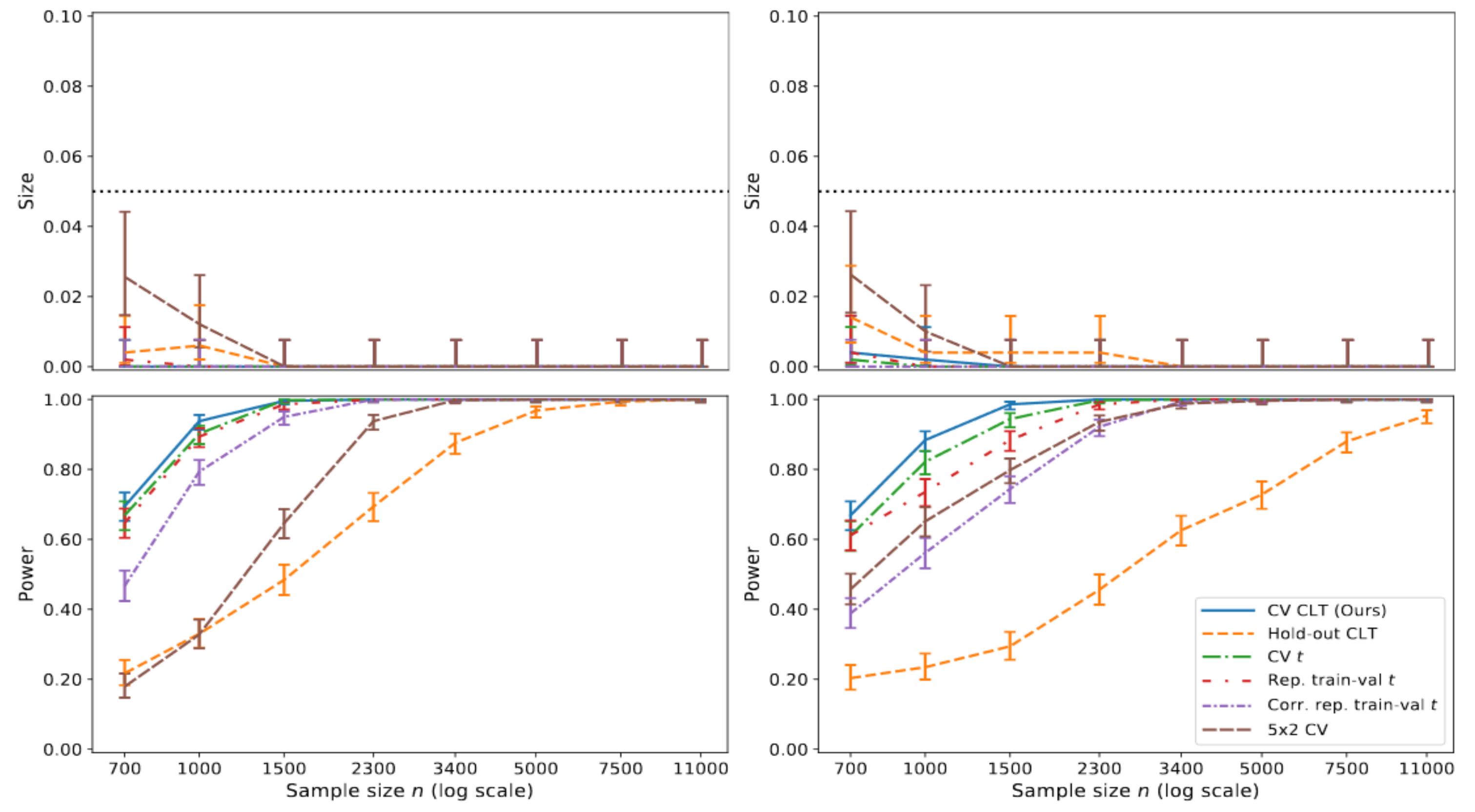

Figure 3: Size when testing $H_1 : \mathrm{Err}(\mathcal{A}_1) < \mathrm{Err}(\mathcal{A}_2)$ (top) and power when testing $H_1 : \mathrm{Err}(\mathcal{A}_2) < \mathrm{Err}(\mathcal{A}_1)$ (bottom) of level-0.05 tests for improved test error. **Left**: $\mathcal{A}_1 = \ell^2$-regularized logistic regression, $\mathcal{A}_2 =$ neural network classification. **Right**: $\mathcal{A}_1 =$ random forest, $\mathcal{A}_2 =$ ridge regression.

Many more experiments in the paper! [1] https://arxiv.org/abs/2007.12671 https://github.com/alexandre-bayle/cvci

References

[1] Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[2] 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 (ICS)*. Citeseer.

[3] 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 (ICML)*, pages 27–35.