Cross-Validation Error Dynamics in Smaller Datasets

Bethany Austhof
Department of Mathematics, Statistics, & Computer Science
University of Illinois Chicago
Chicago, USA
bausth2@uic.edu

Lev Reyzin
Department of Mathematics, Statistics, & Computer Science
University of Illinois Chicago
Chicago, USA
lreyzin@uic.edu

***Abstract*—Cross-validation (CV) is the de facto standard for estimating a model's generalization performance, but in smaller datasets it exhibits an underappreciated quirk: across folds, training and test errors are strongly negatively correlated, while training and holdout errors show a moderate anti-correlation and test versus holdout errors are essentially uncorrelated. Herein, we document these phenomena empirically on both real and synthetic datasets under AdaBoost and introduce a simple generative model that explains them. By viewing each CV split as hypergeometric sampling from a finite population and incorporating an overfitting parameter δ that shifts expected errors on train, test, and holdout sets, we derive closed-form expressions for the covariances among observed error rates. Our analysis shows that sampling-induced anti-correlation dominates in small datasets, while overfitting contributes an additional negative term, thus accounting for the observed error dynamics. We discuss the limitations of our approach and suggest directions for more refined models and extensions to regression settings.**

***Index Terms*—cross-validation, small data, error correlations, finite-population sampling, overfitting**

I. Introduction

Cross-validation (CV) is a long-standing widely-used [1], [2] technique to estimate and optimize a model's generalization performance by repeatedly splitting data into training and test sets (see [3] for a survey). While cross-validation has some known pitfalls [4], we herein discuss an underappreciated phenomenon that occurs especially in smaller datasets (containing, say, thousands, not millions of examples): that, across CV splits, training error and test error tend to be negatively correlated—splits where a model fits the training data especially well often yield higher test error, and vice versa. We have also observed a more moderate negative correlation between training and holdout error, as well as almost no correlation between test error and holdout error. These phenomena are of course related to others that have previously been studied, especially in work accounting for the variance in the test and training splits [5] and on correlations across splits [6], [7].

In this short paper, we introduce a simple model that accounts for these patterns. We show how hypergeometric sampling (sampling from a finite population without replacement) of these types induces the observed anti-correlation between training and test splits, and we show how a model's tendency to overfit to its training data can also account for the negative correlation observed between test and holdout error.

II. An Empirical Finding

In experimenting with cross-validation on smaller datasets, we repeatedly observed a strong anti-correlation between training and test errors across runs. This phenomenon seemed surprising at first, though it is easily explainable (as we discuss in the next section). Moreover, we saw a moderate anti-correlation between training and holdout errors and a negligible anti-correlation between test and holdout.

Our experiments first chose a holdout set and then repeatedly split the remaining data into training and test sets, with the holdout set fixed. Fig. 1 shows a graphical display of these phenomena on the 'Adult' UCI dataset, run on AdaBoost. The aim of this short paper is to discuss this experimental phenomenon and come up with a convincing model to explain it. We also investigate the ramifications of such an anti-correlation between test and train performance. Running multiple experiments (Table I) on 11 different datasets testing model selection strategy's effect on mean accuracy, models were selected based on lowest train error and the lowest test error performance measured against a randomly selected model. We saw a general tendency for models selected based on lowest train error to slightly out-perform both randomly selected models and models selected based on lowest test error.

A. Experimental Methods

All experiments were conducted within a single, fully reproducible notebook that implemented data loading, preprocessing, model training, cross-validation, and evaluation. For each dataset, we first sampled a fixed holdout set, which remained unchanged throughout the experiment. The remaining data were repeatedly split into training and test subsets using uniformly random partitions without replacement. Unless otherwise specified, each experiment consisted of 100 independent train–test splits. This design isolates variance arising from cross-validation itself, rather than from fluctuations in the holdout set.

To ensure comparability across models and datasets, we used a standardized preprocessing pipeline based on

`scikit-learn`'s `ColumnTransformer`: continuous features were scaled using standardization, and categorical features were one-hot encoded. We evaluated five commonly used classification algorithms:

- AdaBoost with 50 decision-stump base learners,
- Random Forests with 100 trees and maximum depth 10,
- RBF-kernel SVM with $C = 1.0$,
- Logistic Regression with a maximum of 1000 iterations,
- A feedforward neural network (MLP) with hidden layers of sizes (100, 50), early stopping, and maximum 300 training iterations.

For each train–test split, we recorded the training error, test error, and holdout error, producing a triple of error quantities for every model instantiation. These values were used to compute empirical correlations, generate scatterplots, and evaluate model-selection strategies.

To study the effect of the observed train–test anti-correlation on model choice, we repeated the procedure across 11 datasets. For each dataset and algorithm, we compared three selection rules: choosing the model with the lowest training error, choosing the model with the lowest test error, and selecting a model uniformly at random. The holdout accuracy of the selected model was then used as the evaluation metric. Aggregated results across datasets are reported in Table I. Our methodological goal here was not hyperparameter optimization but rather to examine the structural behavior of cross-validation under repeated random splits in small data regimes.

B. Analysis of Results

Repeated experimentation revealed mild patterns in test–train anti-correlation and best model performance that we examine here. In Fig. 2 we see that outside of cross-validation trained with a neural network, there is a consistent pattern of anti-correlation between test and train error rates. This deviation in anti-correlation cannot be explained solely by differences in training error, i.e., anti-correlation wasn't lost due to no variability in training errors, as several of the other models also consistently achieve near-zero training error. Instead, we hypothesize that anti-correlation relies on a form of bias that links the difficulty of the training fold to the effective complexity of the learned classifier. For models such as AdaBoost or SVMs, a "hard" fold forces the learner toward a simpler or more biased decision boundary, which tends to generalize better; conversely, an "easy" fold allows the model to fit more idiosyncratic structure, often harming test performance. This coupling creates the systematic inversion of training and testing errors.

Neural networks, by contrast, do not exhibit such behavior in the small-data regime. One plausible hypothesis is that when faced with harder folds, they do not systematically revert to simpler effective hypotheses, but instead rely on high-dimensional representations whose effective complexity does not contract in a predictable way. As a result, variations in training-set difficulty do not induce correspondingly structured changes in generalization performance, and the train–test anti-correlation mechanism breaks down. This interpretation is consistent with a broader body of work documenting counterintuitive and stability-related phenomena in neural networks [8].

We now address where this observed anti-correlation points to in terms of optimal model selection. Unfortunately, we failed to consistently find a method of model selection that yielded a statistically significant hold-out performance. However, we note that there is an observed mild improvement on hold-out performance in models selected based on yielding lowest train error. While often not statistically significantly different at $n = 75$ trials of cross-validation, under a chi-square test of significance lowest train error models wins at holdout performance most often among tested models at a statistically significant $\alpha = .05$ level. Thus, when there is no obvious model choice, our recommendation is to select a model based on lowest train error performance.

III. A Theoretical Explanation

This section presents a minimal probabilistic model consistent with our experimental findings: (i) training and CV-test errors are strongly negatively correlated across splits, (ii) training and holdout errors exhibit a nontrivial (typically moderate) correlation, and (iii) CV-test and holdout errors are only weakly correlated. The key point is that these three phenomena arise from *distinct* sources of variability that affect the observed errors in different ways.

We model three conceptually separate contributors to across-split variability:

1) **Fold composition (difficulty) variability** induced by sampling without replacement from a finite population, which affects *train* and *test* in opposite directions and is absent from the fixed holdout set.
2) **Overfitting variability**, summarized by a scalar δ, which lowers training error but increases out-of-sample (test and holdout) error.
3) **Overall generalization-quality variability**, summarized by a scalar γ, which captures split-to-split fluctuations in how well the learned hypothesis generalizes across *all* evaluation sets.

This separation allows training–test error to be strongly anti-correlated while training–holdout and test–holdout correlations remain weaker.

A. Setup and Latent "Difficulty" Field

Let N be the total dataset size. We first draw a fixed holdout set of size H, leaving $M = N - H$ points for repeated random train–test splitting. On each replicate, we draw a training set of size n uniformly at random without replacement from these M points, and the remaining $m = M - n$ points form the CV test fold.

To model heterogeneity in example difficulty, assign each of the M CV points a fixed (but unobserved) difficulty score u_i, $i = 1, \ldots, M$, with finite-population mean zero:

$$\frac{1}{M} \sum_{i=1}^{M} u_i \;=\; 0.$$

Fig. 1. AdaBoost results on UCI 'Adult' dataset [9]. We used decision stumps with 50 base learners. We performed 100-fold cross-validation. There are slightly over 48,000 instances and 15 features. The data was partitioned into train, test, and holdout as follows: initially 10% was reserved as holdout, then 67.5% of the data was devoted to training and the remaining 22.5% to test.

Fig. 2. Mean and standard deviation of correlation between test and train error rates by dataset and algorithm, averaged over 75 separate runs with different random seeds. Aside from neural networks, the anti-correlation pattern persists consistently.

Let the finite-population variance be

$$s_u^2 = \frac{1}{M-1}\sum_{i=1}^{M} u_i^2.$$

For a given split, define the average difficulty in the training and test folds as

$$\bar{u}_{\text{train}} = \frac{1}{n}\sum_{i\in\text{Train}} u_i, \qquad \bar{u}_{\text{test}} = \frac{1}{m}\sum_{i\in\text{Test}} u_i.$$

Because Train and Test are complementary subsets of the same centered finite population,

$$n\,\bar{u}_{\text{train}} + m\,\bar{u}_{\text{test}} = 0 \quad\Longrightarrow\quad \bar{u}_{\text{test}} = -\frac{n}{m}\,\bar{u}_{\text{train}}. \tag{1}$$

Thus, fold-composition effects are perfectly negatively aligned between training and test splits.

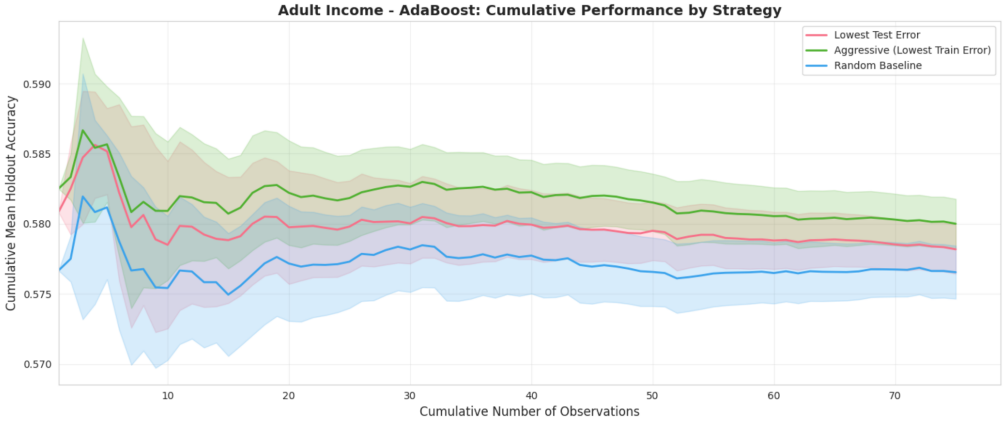

Fig. 3. Seventy-five runs of AdaBoost 100-fold cross-validation on UCI 'Adult' dataset [9], using the same setup as in Fig. 1 with different random seeds. This illustrates how the aggregated numbers in Table I were obtained.

TABLE I

MEAN ACCURACY ± STANDARD DEVIATION ACROSS 75 RUNS FOR EACH DATASET, ALGORITHM, AND SELECTION STRATEGY FOR MODEL (HERE TRAIN INDICATES MODEL WITH LOWEST TRAINING ERROR AMONGST ALL CV FOLDS, TEST IS SIMILAR AND RAND IS A RANDOMLY SELECTED MODEL). BOLD INDICATES BEST; * INDICATES SIGNIFICANCE AT $\alpha = .05$.

Dataset	Alg.	Train	Test	Rand
Adult	AB	**0.580±0.008***	0.578±0.008	0.577±0.008
	LR	0.575±0.009	0.576±0.009	**0.576±0.008**
	NN	0.553±0.011	**0.566±0.009***	0.565±0.011
	RF	0.581±0.007	0.580±0.007	**0.581±0.008**
	SVM	0.572±0.009	**0.573±0.007**	0.572±0.008
Bank	AB	**0.895±0.005**	0.894±0.005	0.894±0.006
	LR	**0.898±0.006**	0.897±0.005	0.897±0.005
	NN	0.897±0.007	**0.900±0.006**	0.898±0.007
	RF	**0.897±0.004**	0.896±0.004	0.896±0.004
	SVM	**0.898±0.006**	0.897±0.005	0.897±0.005
Heart	AB	0.803±0.059	**0.810±0.055**	0.807±0.060
	LR	0.821±0.057	**0.823±0.055**	0.820±0.051
	NN	**0.816±0.047**	0.808±0.057	0.802±0.055
	RF	0.810±0.057	0.807±0.057	**0.815±0.058**
	SVM	0.816±0.056	0.808±0.055	**0.821±0.057**
Iono	AB	0.883±0.039	0.890±0.038	**0.892±0.041**
	LR	0.878±0.041	**0.887±0.045**	0.881±0.043
	NN	**0.894±0.048***	0.886±0.049	0.860±0.051
	RF	0.934±0.033	0.936±0.031	**0.938±0.034**
	SVM	0.938±0.035	0.938±0.033	**0.939±0.033**
Landsat	AB	**0.933±0.012**	0.932±0.011	0.931±0.011
	LR	**0.930±0.011**	**0.930±0.011**	0.929±0.010
	NN	**0.950±0.010***	0.946±0.010	0.945±0.012
	RF	**0.954±0.009**	**0.954±0.009**	**0.954±0.009**
	SVM	**0.952±0.009**	0.951±0.010	0.951±0.010
Letter	AB	**0.851±0.008***	0.847±0.009	0.844±0.008
	LR	**0.808±0.007***	0.805±0.007	0.805±0.007
	NN	**0.948±0.009***	0.945±0.009	0.940±0.010
	RF	**0.919±0.008***	0.916±0.009	0.914±0.010
	SVM	**0.922±0.009**	0.919±0.009	0.919±0.009
Shop	AB	**0.896±0.007**	0.895±0.007	0.894±0.007
	LR	0.887±0.007	0.886±0.006	**0.887±0.006**
	NN	0.898±0.007	**0.900±0.007**	0.899±0.007
	RF	**0.905±0.007**	**0.905±0.007**	**0.905±0.006**
	SVM	**0.894±0.006**	**0.894±0.006**	**0.894±0.006**
Spam	AB	0.900±0.041	**0.908±0.036**	0.898±0.044
	LR	**0.926±0.010**	0.924±0.010	0.924±0.010
	NN	**0.941±0.008**	0.939±0.009	0.938±0.008
	RF	0.940±0.008	0.939±0.009	**0.940±0.008**
	SVM	**0.930±0.009**	**0.930±0.010**	**0.930±0.009**
Shuttle	AB	**0.997±0.002**	**0.997±0.002**	**0.997±0.002**
	LR	**0.962±0.007**	0.961±0.007	0.961±0.007
	NN	**0.997±0.002***	0.996±0.002	0.996±0.003
	RF	**0.998±0.001**	**0.998±0.001**	**0.998±0.001**
	SVM	**0.996±0.002**	**0.996±0.002**	**0.996±0.002**
Wine	AB	**0.818±0.008**	**0.818±0.008**	0.816±0.008
	LR	**0.818±0.008**	**0.818±0.008**	**0.818±0.008**
	NN	**0.839±0.010***	0.835±0.010	0.831±0.010
	RF	**0.866±0.008**	0.864±0.008	0.864±0.008
	SVM	**0.834±0.007**	0.832±0.008	0.833±0.007
Yeast	AB	**0.772±0.025**	0.765±0.028	0.762±0.027
	LR	**0.761±0.022**	0.756±0.021	0.755±0.020
	NN	**0.772±0.027**	0.766±0.023	0.766±0.025
	RF	**0.777±0.024**	0.776±0.025	**0.777±0.025**
	SVM	**0.771±0.025**	**0.771±0.023**	**0.771±0.023**

Alg. abbreviations: AB=AdaBoost, RF=Random Forest, LR=Logistic Regression, NN=Neural Network.

Standard sampling-without-replacement identities yield

$$\mathrm{Var}(\bar{u}_{\mathrm{train}}) = \left(1 - \frac{n}{M}\right)\frac{s_u^2}{n} = \frac{m}{M}\cdot\frac{s_u^2}{n}, \tag{2}$$

$$\mathrm{Var}(\bar{u}_{\mathrm{test}}) = \left(1 - \frac{m}{M}\right)\frac{s_u^2}{m} = \frac{n}{M}\cdot\frac{s_u^2}{m}, \tag{3}$$

$$\mathrm{Cov}(\bar{u}_{\mathrm{train}}, \bar{u}_{\mathrm{test}}) = -\frac{s_u^2}{M}. \tag{4}$$

B. Model-Level Variability

a) Overfitting variability.: Let δ capture split-to-split variation in effective model complexity or overfitting. Larger δ lowers training error but increases out-of-sample error. Assume

$$\mathbb{E}[\delta] = 0, \qquad \mathrm{Var}(\delta) = \sigma_\delta^2.$$

b) Generalization-quality variability.: Let γ capture variation in the overall quality of the learned hypothesis across splits, reflecting factors such as how informative or representative the training fold is. Larger γ corresponds to uniformly better generalization. Assume

$$\mathbb{E}[\gamma] = 0, \qquad \mathrm{Var}(\gamma) = \sigma_\gamma^2.$$

For simplicity, we treat δ, γ, and the fold-composition averages $(\bar{u}_{\mathrm{train}}, \bar{u}_{\mathrm{test}})$ as mutually uncorrelated. The assumption of independence is made for analytical clarity; allowing moderate dependence does not change the qualitative conclusions.

C. Observed Errors

We model the observed errors as additive combinations of baseline error, fold difficulty, model-level effects, and evaluation noise:

$$\hat{p}_{\mathrm{train}} = \mu \;+\; a\,\bar{u}_{\mathrm{train}} \;-\; \alpha\,\delta \;-\; \gamma \;+\; \varepsilon_{\mathrm{train}}, \tag{5}$$

$$\hat{p}_{\mathrm{test}} = \mu \;+\; a\,\bar{u}_{\mathrm{test}} \;+\; \beta\,\delta \;-\; \gamma \;+\; \varepsilon_{\mathrm{test}}, \tag{6}$$

$$\hat{p}_{\mathrm{hold}} = \mu_{\mathrm{hold}} \;+\; \beta\,\delta \;-\; \gamma \;+\; \varepsilon_{\mathrm{holdout}}. \tag{7}$$

Here $a > 0$ scales the effect of fold difficulty; $\alpha, \beta > 0$ control the strength of overfitting effects (often $\alpha > \beta$); and $\varepsilon_{\mathrm{train}}, \varepsilon_{\mathrm{test}}, \varepsilon_{\mathrm{holdout}}$ are mean-zero noise terms, mutually uncorrelated and uncorrelated with all latent variables.

D. Predicted Covariances and Correlations

Using (5)–(7) and (4), we obtain:

a) Train vs. Test.:

$$\mathrm{Cov}(\hat{p}_{\mathrm{train}}, \hat{p}_{\mathrm{test}}) = -\,a^2\,\frac{s_u^2}{M} \;-\; \alpha\beta\,\sigma_\delta^2 \;+\; \sigma_\gamma^2. \tag{8}$$

In small datasets, the finite-population term dominates, yielding a strong negative correlation despite the shared γ component.

b) Train vs. Holdout.:

$$\mathrm{Cov}(\hat{p}_{\mathrm{train}}, \hat{p}_{\mathrm{hold}}) = -\,\alpha\beta\,\sigma_\delta^2 \;+\; \sigma_\gamma^2. \tag{9}$$

Thus, training and holdout errors can exhibit a moderate correlation whose sign and magnitude depend on the balance between overfitting variability and overall generalization quality.

c) Test vs. Holdout.:

$$\mathrm{Cov}(\hat{p}_{\mathrm{test}}, \hat{p}_{\mathrm{hold}}) = \beta^2\,\sigma_\delta^2 \;+\; \sigma_\gamma^2 \;>\; 0. \tag{10}$$

The correlation between test and holdout errors will usually remain weak because $\hat{p}_{\mathrm{test}}$ includes an additional variance term from fold composition:

$$\mathrm{Var}(\hat{p}_{\mathrm{test}}) = a^2\,\mathrm{Var}(\bar{u}_{\mathrm{test}}) + \beta^2\sigma_\delta^2 + \sigma_\gamma^2 + \mathrm{Var}(\varepsilon_{\mathrm{test}}),$$

which is absent from $\hat{p}_{\mathrm{hold}}$. This extra variability dilutes the test–holdout correlation even when the covariance is positive.

E. Interpretation and "Small Data" Behavior

The model isolates three distinct mechanisms: exact complementarity of finite-population sampling (which drives strong train–test anti-correlation), overfitting variability (which weakly couples training to out-of-sample performance), and generalization-quality variability (which induces shared movement across all error estimates). In small datasets, fold-composition effects dominate test-set variance, naturally explaining why test error is strongly anti-correlated with training error yet only weakly informative about holdout performance. As dataset sizes increase and learning algorithms become more stable, these effects diminish and the correlations collapse toward zero.

IV. Conclusions

We documented an underappreciated empirical pattern in repeated cross-validation on smaller datasets: training and test errors across splits tend to be strongly negatively correlated, training and holdout errors exhibit a nontrivial (often moderate) correlation, and test and holdout errors are only weakly correlated. Beyond being a curiosity, this behavior matters because it complicates the interpretation of cross-validation outcomes and can subtly affect model selection when one repeatedly trains models on different random splits.

To explain these observations, we introduced a minimal generative model in Section III that separates three sources of across-split variability: (i) finite-population sampling effects from drawing complementary train/test folds without replacement, captured by the latent fold-difficulty field u_i; (ii) split-to-split variation in effective overfitting, summarized by δ, which reduces training error but worsens out-of-sample error; and (iii) split-to-split variation in overall generalization quality, summarized by γ, which shifts errors on train, test, and holdout in a shared direction. In this model, the strong train–test anti-correlation is primarily a consequence of complementarity in finite-population sampling, while the weaker relationships involving holdout error arise from the shared model-level terms δ and γ, with test–holdout correlation additionally diluted by test-fold composition variance.

The model is intentionally coarse and leaves several natural refinements open. Most notably, it does not explicitly model how particular "hard" examples migrate between folds and influence the learned hypothesis, nor does it attempt to predict when particular learning algorithms (e.g., neural networks in our experiments) will fail to exhibit the train–test anti-correlation pattern. A more detailed treatment might allow δ and γ to depend on fold difficulty, incorporate heterogeneity in example-level noise, or directly model algorithmic stability. Finally, it would be interesting to extend these ideas beyond classification to regression and other loss functions, where analogous finite-sample correlation effects may arise but interact differently with estimation error.

References

[1] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.

[2] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, Series B*, vol. 36, no. 2, pp. 111–147, 1974.

[3] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[4] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[5] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[6] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?" *Journal of the American Statistical Association*, vol. 119, no. 546, pp. 1434–1445, 2024.

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

[8] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[9] R. Kohavi and B. Becker, "Adult data set," 1996. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult