Proceedings of Machine Learning Research vol 336:1–38, 2026 39th Annual Conference on Learning Theory

Minimax Limits of k-Fold Cross-Validation via Majority

Ido Nachum INACHUM@UNIV.HAIFA.AC.IL
University of Haifa, Haifa, Israel.

Rüdiger Urbanke RUDIGER.URBANKE@EPFL.CH
EPFL, Lausanne, Switzerland.

Thomas Weinberger THOMAS.WEINBERGER@EPFL.CH
EPFL, Lausanne, Switzerland.

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the mean-squared error of k-fold cross-validation as a risk estimator, with particular emphasis on how its accuracy depends on the number of folds k. Despite the widespread use of cross-validation, principled guidance for choosing k is largely absent, mainly due to the complex dependence between fold-wise error estimates. To obtain sharp and interpretable results, we focus on the majority algorithm in binary classification, a minimal yet nontrivial empirical risk minimization procedure. We provide a fine-grained analysis of its cross-validation behavior, showing that even this simple algorithm exhibits subtle and delicate phenomena for which existing theory provides loose and even vacuous bounds. Leveraging this analysis, we introduce a minimax framework for cross-validation risk estimation and prove that no empirical risk minimization algorithm can achieve an $O(1/n)$ minimax mean-squared error when the number of folds grows with the number of samples n; instead, a lower bound of order $\Omega(\sqrt{k}/n)$ is unavoidable. Our results reveal fundamental limitations of cross-validation as a data-reuse strategy, clarify gaps and inaccuracies in prior theoretical work, and position the majority algorithm as a natural benchmark that any tight analysis of cross-validation should be able to explain.

Keywords: cross-validation, learning theory, algorithmic stability

1. Introduction

k-fold cross-validation (CV) is a widely used model validation technique in statistics, data science, and machine learning; see (Arlot and Celisse, 2010) for a comprehensive survey. Given a collection of errors obtained by training models on subsets of the data and validating them on the remaining observations, CV is typically used for one of two purposes: *(risk estimation)* estimating the risk of a model chosen independently of the error computations by averaging the validation errors; or *(model selection)* selecting, among a collection of candidate models, the one that minimizes the CV error.

The primary motivation for using k-fold CV for risk estimation is that partitioning the data into multiple non-overlapping subsets generally reduces statistical variability compared to relying on a single hold-out set (Blum et al., 1999). Moreover, unlike the empirical (training) error, CV typically mitigates the overly optimistic bias caused by overfitting, a phenomenon that is particularly pronounced for overparametrized models such as neural networks.

Many CV schemes have been proposed, including combinatorial partitioning (Shao, 1993) and Monte Carlo resampling (Picard and Cook, 1984). In this work, we focus on the standard variant in which the sample is partitioned into k non-overlapping folds of (approximately) equal size; 5-fold and 10-fold CV are especially common in practice. Despite its pervasive use as a validation tool in the empirical sciences, the theoretical properties of CV remain surprisingly poorly understood. In particular, there is still no principled method for choosing the number of folds k. As noted by Arlot and Celisse (2010, Sec. 10.3):

> *"VFCV [V-fold cross-validation] is certainly the most popular CV procedure, in particular because of its mild computational cost. Nevertheless, the question of choosing V remains widely open, even if indications can be given towards an appropriate choice."*

A central obstacle is that fold-wise errors are dependent, and this dependence is difficult to characterize sharply. As a result, relatively few works provide a precise mathematical analysis of how CV-based risk-estimation accuracy varies with k; see, for example, (Celisse, 2014; Arlot and Lerasle, 2016), which identify optimal choices of k in specific density estimation settings.

To ground the discussion, consider first an extreme baseline: binary classification with the 0–1 loss and a constant learning algorithm $\mathcal{A}_h$ that always outputs the same hypothesis h, independent of the sample. If the population risk of h is p, then the mean-squared error (MSE) of the k-fold CV risk estimator satisfies $\mathrm{MSE}_{\mathrm{CV}}^{(k)} = \frac{p(1-p)}{n}$ independent of k since $L(\mathcal{A}_h, \mathcal{D}) = p$ and $n \cdot \widehat{L}_{\mathrm{CV}}^{(k)} \sim \mathrm{Bin}(n,p)$, so $\mathrm{MSE}_{\mathrm{CV}}^{(k)} = \mathrm{Var}(\mathrm{Bin}(n,p))/n^2$.

A natural next step is to consider the simplest nontrivial extension of this baseline: the majority algorithm. Let $S^n = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be an i.i.d. sample, and define $Y := \sum_{i=1}^n y_i$. The majority algorithm is

$$\mathcal{A}_{\mathrm{maj}}(S^n) = \begin{cases} h_0 : x \mapsto 0, & \text{if } Y \le n/2, \\ h_1 : x \mapsto 1, & \text{if } Y > n/2, \end{cases}$$

where h_i denotes the constant hypothesis outputting i.

This is arguably the simplest empirical risk minimization (ERM) extension of the constant predictor: the hypothesis class grows from one constant function to two, and the decision rule depends only on the label counts (ignoring the features entirely). Yet, a precise analysis of its CV behavior is surprisingly delicate. Our study has two components:

1. k-fold CV is explicitly designed to reuse the same data for both training and validation. Accordingly, our aim is to isolate the effect of this data reuse on CV-based risk estimation. To this end, we benchmark CV-based risk estimation against an idealized baseline in which one trains on all n available samples and, in addition, has access to an independent validation set of size n.
2. We conduct a fine-grained analysis of the majority algorithm, which allows us to identify and rigorously analyze several gaps in the existing theoretical literature, including overly loose bounds.

1.1. Our Contributions

Fine-grained analysis of Majority. In Section 5.1, we carry out a fine-grained analysis of $\mathrm{MSE}_{\mathrm{CV}}^{(k)}$ for the Majority algorithm under a random-label model. The dominant contribution to the MSE arises from the covariance between validation errors across folds (see the notation in Section 2). In general, this covariance is difficult to analyze due to the complex dependencies induced by data reuse across folds.

We identify a structural property that the Majority algorithm satisfies that allows these dependencies to be handled explicitly. This property, which we call *Factorization* (Lemma 5), reduces the covariance computation to a simpler quantity depending only on the distribution of the sample within a single fold. Leveraging this result, we derive a closed-form expression for the covariance in Theorem 8, and hence for $\mathrm{MSE}_{\mathrm{CV}}^{(k)}$. In particular, we show in Theorem 9 that $\mathrm{MSE}_{\mathrm{CV}}^{(k)}$ scales as $\Theta(\sqrt{k}/n)$.

Minimax limits of cross-validation. Kearns and Ron (1997) proved that for loss-stable ERMs over VC classes, leave-one-out CV performs essentially no worse than the empirical error. Similar results for general k-fold CV were obtained by Anthony and Holden (1998).

Another natural requirement is that CV performs no worse than a hold-out estimate over a single fold of size n/k. The work of Blum et al. (1999) confirms this property for a specific (non-standard) cross-validation setting, where the algorithm's final output on the full sample S^n is the random classifier sampling uniformly at random from the k cross-validated hypotheses.

Importantly, these works do not *quantify* the advantage of CV. To make this notion precise, we define the *minimax cross-validation risk* for a given algorithm $\mathcal{A}$ as

$$\mathfrak{R}_{\mathrm{CV}}(\mathcal{A}) := \min_{k|n} \max_{\mathcal{D}} \mathrm{MSE}_{\mathrm{CV}}^{(k)}(\mathcal{A}, \mathcal{D}),$$

which represents the optimal achievable MSE over all choices of k in the absence of knowledge about the underlying distribution $\mathcal{D}$.

The constant algorithm discussed earlier attains its minimax for uniformly random labels, hence $\mathfrak{R}_{\mathrm{CV}}(\mathcal{A}_h) = 1/(4n)$. This provides a natural baseline or reference point to compare more sophisticated algorithms. Since overly simplistic procedures such as the constant algorithm are not practically useful—yet may achieve similar minimax rates—we focus on (non-constant) ERM algorithms, a standard assumption in statistical learning theory. This leads to the following question:

> **Question 1** *In classification with 0–1 loss, can any ERM achieve an $O(1/n)$ minimax rate for CV risk estimation, and if not, how close can it get?*

In Section 5.2, we show that ERM algorithms cannot achieve a minimax rate of $O(1/n)$ for $k = \omega_n(1)$ using a reduction to our in-depth study of the Majority algorithm: For any ERM algorithm and for every k it holds

$$\max_{\mathcal{D}} \mathrm{MSE}_{\mathrm{CV}}^{(k)}(\mathcal{A}, \mathcal{D}) = \Omega(\sqrt{k}/n).$$

Therefore, for any ERM algorithm $\mathcal{A}$ that achieves the minimax optimum with k^* folds, the MSE of cross-validation scales as

$$\mathfrak{R}_{\mathrm{CV}}(\mathcal{A}) = \Omega\left(\sqrt{k^*}/n\right).$$

Hence, even for the most carefully designed ERM, CV cannot exploit the entire dataset as if it were using a single hold-out set of size n; there remains a factor of $\sqrt{k^*}$ in the rate.

Limitations of existing theory and the Majority benchmark. A substantial body of work, most notably (Rogers and Wagner, 1978; Devroye and Wagner, 1979a; Blum et al., 1999; Kearns and Ron, 1997; Bousquet and Elisseeff, 2002; Kumar et al., 2013), has yielded important insights into the behavior of CV and significantly shaped the theoretical landscape. Nevertheless, when these results are evaluated against the Majority algorithm, they exhibit systematic shortcomings. In particular, existing guarantees typically suffer from at least one of the following limitations:

1. *Arbitrarily loose sufficient conditions.* Stability-based upper bounds, including those of Kumar et al. (2013), fail to capture the performance of even the majority algorithm. While the proof techniques underlying the results of Rogers and Wagner (1978); Devroye and Wagner (1979a); Kearns and Ron (1997); Bousquet and Elisseeff (2002) naturally extend to general k-fold cross-validation from leave-one-out CV, the resulting guarantees remain fundamentally misaligned with the behavior of Majority.
2. *Relative rather than absolute guarantees.* Several works provide guarantees only in comparison to other error measures rather than in absolute terms. For instance, Kearns and Ron (1997) compare leave-one-out CV to empirical error; however, our analysis of Majority shows that these bounds are loose across most regimes of their confidence parameter δ, particularly when considering the MSE of CV loss estimation.
3. *Mathematical inaccuracies in prior analyses.* Independent of our analysis of Majority, we identify errors in fundamental theorems in Kearns and Ron (1997) and Kale et al. (2011). Clarifying these issues is important because the affected results make claims about fundamental properties of CV estimation, and correcting them reveals several unresolved theoretical questions.

We discuss the Majority benchmark in Section 5.3 and the mathematical inaccuracies in Section 5.4. The benchmark analysis leads to the following conclusion:

> Majority is a natural benchmark and we advocate that demonstrating tightness for this instance should be a minimal requirement for any future bounds on the error of CV.

2. Setup and Notation

We start by establishing the framework for our investigation. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space, and set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We study (possibly randomized) learning rules $\mathcal{A} : \mathcal{Z}^\star \to \mathcal{H}$ that map a sample $S^n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ to a hypothesis $h = \mathcal{A}(S^n) \in \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The observations are i.i.d.: $Z_i \sim \mathcal{D}$, hence $S^n \sim \mathcal{D}^n$. The learning algorithm $\mathcal{A}$ may be randomized;

in that case, $\mathbb{E}_{\mathcal{A}}[\cdot]$ denotes expectation with respect to the internal randomness of $\mathcal{A}$. As is common in previous works, we assume throughout that $\mathcal{A}$ is permutation-invariant (symmetric): for any permutation π of $\{1, \ldots, n\}$, $\mathcal{A}(S^n) = \mathcal{A}(S^n_\pi)$ a.s., where $S^n_\pi := (Z_{\pi(1)}, \ldots, Z_{\pi(n)})$.

We assume throughout that the number of folds is an integer $k \geq 2$ with $k \mid n$. We partition the index set $\{1, \ldots, n\}$ into k disjoint blocks $I_1, \ldots, I_k$ of size $m := |I_i| = n/k$, and define $S_i = \{Z_j : j \in I_i\}$ and $S_{-i} = S^n \setminus S_i$. We denote the concatenation of two samples S_i and S_j by $S_i \circ S_j$.

Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the k-fold cross-validation estimator is

$$\widehat{L}^{(k)}_{\mathrm{CV}}(\mathcal{A}, S^n) = \frac{1}{k} \sum_{i=1}^{k} \widehat{L}^{(k)}_i, \quad \widehat{L}^{(k)}_i = \frac{1}{|S_i|} \sum_{(x,y) \in S_i} \ell\big(\mathcal{A}(S_{-i})(x),\, y\big) = \frac{k}{n} \sum_{(x,y) \in S_i} \ell\big(\mathcal{A}(S_{-i})(x),\, y\big).$$

That is, $\widehat{L}^{(k)}_i$ is the average loss on the ith hold-out fold, and $\widehat{L}^{(k)}_{\mathrm{CV}}$ averages these across folds.[1] We omit the subscript CV whenever it is clear from context. Since this work focuses on classification, we will often consider the 0–1 loss defined as $\ell(\hat{y}, y) = \mathbb{1}_{\{\hat{y} \neq y\}}$.

We assess the performance of cross–validation via the mean-squared error (MSE)

$$\mathrm{MSE}^{(k)}_{\mathrm{CV}}(\mathcal{A}, \mathcal{D}) := \mathbb{E}_{S^n \sim \mathcal{D}^n, \mathcal{A}}\left[\left(\widehat{L}^{(k)}_{\mathrm{CV}}(\mathcal{A}, S^n) - L(\mathcal{A}(S^n))\right)^2\right],$$

where the population risk is $L(h) := \mathbb{E}_{(x,y) \sim \mathcal{D}}\big[\ell(h(x), y)\big]$.

For the ith fold, we also write $L^{(k)}_i(S^n) := L(\mathcal{A}(S_{-i}))$, i.e., the risk of the hypothesis trained on the complement S_{-i} of the ith hold–out block.

3. Preliminaries on Algorithmic Stability

Before positioning our work within the context of previous works, it is instructive to familiarize oneself with commonly used notions of algorithmic stability. While there are many notions of algorithmic stability in the literature, we will focus on two of the most widely used variants. We also note that most classical works on the performance of leave-one-out CV consider the following notions for the special case where $m = 1$, while some newer works also consider leave-m notions with $m > 1$ (Gastpar et al., 2026).

Definition 1 (Hypothesis Stability) *We call a pair* $(\mathcal{A}, \mathcal{D})$ *hypothesis-stable with respect to some metric* $\mathrm{dist}(\cdot, \cdot)$ *with parameters* (β, m) *if*

$$\mathbb{E}_{S^{n-m} \sim \mathcal{D}^{n-m},\, S^m \sim \mathcal{D}^m,\, \mathcal{A}}\Big[\mathrm{dist}\big(\mathcal{A}(S^{n-m} \circ S^m), \mathcal{A}(S^{n-m})\big)\Big] \leq \beta.$$

Intuitively, hypothesis stability is a stronger assumption than necessary for error estimation. It provides a quantitative measure of how similar the hypotheses trained on different folds are to the one obtained from the full dataset. In this sense, a highly hypothesis-stable algorithm behaves almost like a constant algorithm. However, the key factor governing the accuracy of

1. When $k = n$, the definition coincides with leave-one-out cross-validation.

cross-validation (CV) error estimation is not the similarity of hypotheses themselves, but rather the stability of their loss values when a small subset of training samples is removed.

For this reason, it is more natural to require a weaker property, called *loss stability* (or *error stability*). This condition ensures that the per-fold loss estimates remain nearly unbiased, even when the training data is slightly perturbed.

Definition 2 (Loss Stability) *We call a pair $(\mathcal{A}, \mathcal{D})$ loss-stable with parameters (β, m) if*

$$\mathbb{E}_{S^{n-m}\sim\mathcal{D}^{n-m},\ S^m\sim\mathcal{D}^m,\ \mathcal{A}}\left[\left|L\big(\mathcal{A}(S^{n-m}\circ S^m)\big) - L\big(\mathcal{A}(S^{n-m})\big)\right|\right] \leq \beta.$$

Loss stability may appear necessary for accurate risk estimation: each fold estimate is unbiased for the risk of the algorithm trained on $n-m$ samples, so CV should estimate $L(\mathcal{A}(S^n))$ accurately only if this risk is close to $L(\mathcal{A}(S^{n-m}))$. However, this reasoning is incomplete because fold estimates are correlated. In fact, pathological algorithms can have poor loss stability while still admitting perfectly accurate CV risk estimates.

Lemma 3 *In the setting of classification with the 0–1 loss, there exist algorithm-distribution combinations whose loss-stability is $\beta = 1/4$, yet whose CV MSE is 0.*

Proof See Appendix B. ∎

This result contradicts Theorem 5.3 of Kearns and Ron (1997); as discussed in Appendix C, the discrepancy stems from an error in that theorem. In general, it remains unclear in which cases loss stability is strictly necessary for small MSE.

4. Prior Work

The works of Rogers and Wagner (1978); Devroye and Wagner (1979a,b) were among the first to establish rigorous stability-based performance guarantees for classification problems using leave-one-out CV. Although in their works, they assume that the considered algorithms are *local* (e.g. nearest neighbors) and the data distributions are arbitrary, their results directly generalize to the class of hypothesis-stable algorithms (in which case the bounds are no longer distribution-free).

The well-known work by Bousquet and Elisseeff (2002) provided a streamlined presentation of classical results and novel error bounds for leave-one-out CV and the empirical error under various strengthened assumptions on algorithmic stability and/or the loss functional.

Estimating the population loss in an algorithm-dependent manner is closely related to statistical learning theory (see, e.g., (Shalev-Shwartz and Ben-David, 2014)). The principal aim of this field is the development of generalization bounds, typically in the form of high-probability upper bounds $L(\mathcal{A}(S)) < \widehat{L}_{\text{emp}}(\mathcal{A}(S), S) + C$, where $\widehat{L}_{\text{emp}}(\mathcal{A}(S), S)$ denotes the empirical error over the training set and the generalization measure C accounts for the over-optimism of the empirical error induced by the complexity of the model. A classical result (Vapnik and Chervonenkis, 1971; Blumer et al., 1989) for binary classification with 0–1 loss states it is necessary and sufficient to let $C = \Theta(\sqrt{d/n})$ to ensure that the generalization bound holds in a tight manner even for the worst-case distribution, where d is the VC dimension, a combinatorial measure of the richness of the hypothesis class $\mathcal{H}$ associated with the algorithm. These bounds are often too pessimistic because they are not sensitive to the (possibly benign) characteristics of the specific data distribution at hand. Moreover, it can be shown that in overparametrized settings (which are ubiquitous in machine learning), generalization measures that are not distribution-dependent face limitations both empirically (Jiang et al., 2019; Dziugaite et al., 2020) and theoretically (Gastpar et al., 2024, 2026). With this in mind, CV becomes conceptually interesting as a flexible alternative to generalization measures for overparametrized settings, where the empirical error is typically uninformative, and a distribution-dependent measure is required—though admittedly CV is no silver bullet (theoretical bounds require estimating the algorithm's stability, and CV can be computationally expensive).

In light of this comparison, a sound minimal requirement is that CV performs at least as well as the empirical error. This question, formalized in terms of so-called *sanity-check bounds*, has been studied by Kearns and Ron (1997). One of their central results is that, for loss-stable empirical risk minimizers over VC classes, leave-one-out CV is guaranteed to perform essentially no worse than the empirical error. Anthony and Holden (1998) derived similar results for the more general case of k-fold CV.

Yet another natural sanity-check is to require CV to perform no worse than a single hold-out set of corresponding size. The work of Blum et al. (1999) shows that this does indeed hold for a specific (non-standard) cross-validation setting.

An influential line of work (Bengio and Grandvalet, 2004; Nadeau and Bengio, 1999) considers the limitations of unbiased estimation of the variance of CV.

A more recent line of work develops upper bounds on the MSE based on novel notions of loss stability (Kale et al., 2011; Kumar et al., 2013). As discussed in Appendix D, the main theorem of Kale et al. (2011) contains a technical error, which makes it difficult to assess the implications of the stated results. The follow-up work introduces a version of loss stability that leads to a stronger result (Kumar et al., 2013, Theorem 1) since the related stability parameter is a lower bound on the one appearing in (Kale et al., 2011, Theorem 2). In both works, the authors aim to bound the performance of the non-standard algorithm that at test time picks one of the cross-validated hypotheses uniformly at random, while we directly bound the MSE of the full-sample hypothesis.

Bayle et al. (2020) develop a central limit theorem for k-CV error around the k-fold test error under an abstract asymptotic linearity condition and propose consistent variance estimators that yield asymptotically exact confidence intervals and hypothesis tests for comparing two learning algorithms.

5. Results

We first present an exact analysis of the majority algorithm in binary classification with the 0–1 loss. Accordingly, the results in Sections 5.1–5.3 are specific to the binary setting $\mathcal{Y} = \{0, 1\}$ and to $\ell(\hat{y}, y) = \mathbb{1}_{\{\hat{y} \neq y\}}$, with the exception of Lemma 5. The subsequent minimax lower bound is proved by reducing arbitrary ERM algorithms to this binary majority problem. Since binary classification is a special case of multiclass classification, our lower bounds directly extend to multiclass classification with 0–1 loss.

Admissible fold sizes. Throughout this section we work with equal-size folds: whenever a fold count k is considered, we assume $k \mid n$ and write $m = n/k$. This convention is only meant to suppress inessential rounding effects. If $k \nmid n$, one could instead use approximately equal folds, or carry all arguments through with floor and ceiling corrections; for Majority, this only replaces the single exchangeable fold covariance by fold-size-dependent variants and does not change the covariance mechanism or its qualitative dependence on k. Thus, fixed-k asymptotics are understood along sequences with $k \mid n$; for example statements comparing $k = 2$ and $k = 3$ are understood along subsequences with $6 \mid n$.

5.1. MSE of Cross-Validation for Majority

We analyze the majority algorithm under a distribution $\mathcal{D}$ whose marginal over $\mathcal{X}$ is arbitrary, and whose labels y_i are i.i.d. draws from $\mathcal{Y} = \{0, 1\}$ with $y_i \sim \mathrm{Ber}(1/2)$. In this case, $Y \sim \mathrm{Bin}(n, 1/2)$, and the population loss of $\mathcal{A}_{\mathrm{maj}}$ equals $1/2$, independent of both the sample S and the sample size n. Consequently, analyzing its MSE reduces to controlling the covariance between folds.

Proposition 4 *The MSE of the majority algorithm with random labels equals*

$$\frac{k-1}{k}\,\mathrm{Cov}(\widehat{L}_1^{(k)}, \widehat{L}_2^{(k)}) + \frac{1}{4n}.$$

Proof Since we assume the labels are randomly distributed, $L(\mathcal{A}_{\mathrm{maj}}(S^n)) = 1/2$ almost surely and the estimator is unbiased, i.e., $\mathbb{E}[\widehat{L}_{\mathrm{CV}}^{(k)}] = 1/2$. Hence the MSE reduces to the variance, which can be decomposed as $\mathrm{MSE}_{\mathrm{CV}}^{(k)}(\mathcal{A}_{\mathrm{maj}}, \mathcal{D}) = \mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k}\widehat{L}_i\right) = \frac{1}{k}\mathrm{Var}(\widehat{L}_1) + \frac{k-1}{k}\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2)$ where $\mathrm{Var}(\widehat{L}_1) = \frac{1}{4m}$ since $\widehat{L}_1$ is independent of S_{-1} and $m \cdot \widehat{L}_1 \sim \mathrm{Bin}(m, 1/2)$. ■

5.1.1. Fold-Covariance and Loss Factorization

As we saw, for the simple class of algorithms with constant loss (such as Majority), the MSE reduces to the fold-covariance. In such settings, the fundamental challenge in analyzing cross-validation is the complex dependency structure introduced by the overlapping training sets. For any two distinct folds i and j, the training sets S_{-i} and S_{-j} share a large amount of data $S \setminus (S_i \circ S_j)$. This overlap creates a coupling between the cross-validated hypotheses $h_i = \mathcal{A}(S_{-i})$ and $h_j = \mathcal{A}(S_{-j})$ that is notoriously difficult to quantify.

We identify a structural property—*loss factorization*—that allows us to disentangle these dependencies. If the empirical loss can be separated into a local *test statistic* $\phi(S_i)$ and a global *decision function* $\psi(S_{-i})$, the complex interaction collapses into a simpler form.

Formally, the factorization $\widehat{L}_i = \mu + \phi(S_i)\psi(S_{-i})$ models a class of algorithms where the interaction between the folds is separable: the validation fold S_i influences the loss only through the scalar summary $\phi(S_i)$, and the training set S_{-i} influences the loss only through the response $\psi(S_{-i})$.

Lemma 5 (Factorization Lemma) *Let $S' := S\backslash(S_1 \circ S_2)$ denote the shared training set of the first two cross-validated hypotheses. Assume that the empirical losses of the cross-validated hypotheses have the form*

$$\widehat{L_i} = \mu + \phi(S_i)\psi(S_{-i}),$$

where $\mu = \mathbb{E}[\widehat{L_i}]$. Then,

$$\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = \mathbb{E}_{S'}\left[\left(\mathbb{E}_{S_1}[\phi(S_1)\psi(S' \circ S_1) \mid S']\right)^2\right].$$

Moreover, if $\phi(S_1)$ is centered ($\mathbb{E}[\phi(S_1)] = 0$), then

$$\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = \mathbb{E}_{S'}\left[\mathrm{Cov}(\phi(S_1), \psi(S' \circ S_1) \mid S')^2\right].$$

Proof Due to the form of the empirical fold loss, the covariance is equal to the inner product

$$\begin{aligned}\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) &= \mathbb{E}\left[\phi(S_1)\psi(S' \circ S_2) \cdot \phi(S_2)\psi(S' \circ S_1)\right] \\ &= \mathbb{E}\left[\phi(S_1)\psi(S' \circ S_1) \cdot \phi(S_2)\psi(S' \circ S_2)\right].\end{aligned}$$

Conditioning on S', the terms $T_1 := \phi(S_1)\psi(S' \circ S_1)$ and $T_2 := \phi(S_2)\psi(S' \circ S_2)$ become independent and identically distributed (since $S_1 \overset{d}{=} S_2$ and they are disjoint from S'):

$$\mathbb{E}[T_1 T_2 \mid S'] = \mathbb{E}_{S_1}[T_1 \mid S'] \cdot \mathbb{E}_{S_2}[T_2 \mid S'] = \left(\mathbb{E}_{S_1}[T_1 \mid S']\right)^2.$$

Taking the expectation over S' yields the result.

Centered case: by the definition of conditional covariance,

$$\mathbb{E}[\phi(S_1)\psi(S' \circ S_1) \mid S'] = \mathrm{Cov}(\phi(S_1), \psi(S' \circ S_1) \mid S') + \mathbb{E}[\phi(S_1) \mid S']\mathbb{E}[\psi(S' \circ S_1) \mid S'].$$

Since S_1 is independent of S' and $\mathbb{E}[\phi(S_1)] = 0$, the second term vanishes. ■

The term $\mathrm{Cov}(\phi(S_1), \psi(S' \circ S_1) \mid S')$ measures how strongly the decision function ψ (trained on the augmented set $S' \circ S_1$) correlates with the specific statistical fluctuations ϕ of the set S_1 that was added to it. In stable algorithms, adding S_1 to the training set should have negligible impact on the decision, leading to near-zero covariance. In unstable algorithms (like Majority near the decision boundary), the correlation is high, and the lemma precisely quantifies this instability.

Indeed, Majority is multiplicatively separable in the sense discussed above.

Corollary 6 *The majority algorithm with uniformly random labels admits the fold-covariance*

$$\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = 4 \cdot \mathbb{E}_Y\left[\mathrm{Cov}\left(\frac{X_m}{m}, \mathbb{1}_{\{X_m > \theta - Y\}} \middle| Y\right)^2\right]$$

where $\theta := (n-m)/2$, $X_t \sim \mathrm{Bin}(t, 1/2)$ and $Y \sim \mathrm{Bin}(n-2m, 1/2)$.

Proof Let $X_i := \sum_{(x,y)\in S_i} y$ and $Y_{-i} := \sum_{(x,y)\in S\setminus S_i} y$. It holds that

$$\begin{aligned}\widehat{L_i} &= \frac{X_i}{m}\mathbb{1}_{\{Y_{-i}\le\theta\}} + \frac{m-X_i}{m}\mathbb{1}_{\{Y_{-i}>\theta\}} = \frac{X_i}{m}\left(1-\mathbb{1}_{\{Y_{-i}>\theta\}}\right) + \left[1-\frac{X_i}{m}\right]\mathbb{1}_{\{Y_{-i}>\theta\}} \\ &= \left[1-\frac{2X_i}{m}\right](\mathbb{1}_{\{Y_{-i}>\theta\}} - 1/2) + \frac{1}{2}.\end{aligned}$$

This matches the form of Lemma 5 with:

- Test Statistic: $\phi(S_i) = 1 - \frac{2X_i}{m}$.
- Decision function: $\psi(S_{-i}) = \mathbb{1}_{\{Y_{-i}>\theta\}} - \frac{1}{2}$.
- For uniformly random labels, $\mathbb{E}[X_i] = m/2$, so $\mu = \mathbb{E}[\widehat{L_i}] = \frac{1}{2}$ and $\mathbb{E}[\phi(S_i)] = 0$.

Since shifting ϕ, ψ by constants does not change their covariance, the statement follows. ■

5.1.2. Precise Analysis of the MSE

For the remainder of this section, let us denote (with slight abuse of notation) $\mathrm{Cov}(n,m) \equiv \mathrm{Cov}(\widehat{L_1}, \widehat{L_2})$ to highlight the roles of n and m.

With the structural decomposition of the covariance established, we now turn to a precise quantitative analysis, transforming the probabilistic factorization into an exact combinatorial expression and, finally, deriving sharp non-asymptotic bounds on the MSE.

A key analytical step is the reduction of this covariance term to a point evaluation of the binomial probability mass function.

Lemma 7 *In the setting of Corollary 6, it holds that*

$$\mathrm{Cov}\left(\frac{X_m}{m}, \mathbb{1}_{\{X_m>\theta-Y\}} \,\middle|\, Y\right) = \frac{1}{4}\mathbb{P}\big(X_{m-1} = \lfloor\theta - Y\rfloor \mid Y\big).$$

Proof See Appendix A.4. ■

This relation is significant because it transforms the stability analysis from a moment-estimation problem into a counting problem.

Substituting this probability mass directly into the factorization formula converts the expectation over the shared training set count Y into a discrete convolution of binomial coefficients.

Theorem 8 (Exact Combinatorial Form of Fold-Covariance) *For $1 \le m \le n/2$, $m|n$, we have*

$$\mathrm{Cov}(n,m) = 2^{-n}\sum_{j=0}^{m-1}\binom{m-1}{j}^2\binom{n-2m}{\lfloor(n-m)/2\rfloor - j}.$$

Proof See Appendix A.4. ■

While Theorem 8 provides an exact description of the covariance, extracting explicit scaling laws requires a delicate non-asymptotic analysis. Indeed, up to normalization, the summand in Theorem 8 is a product of binomial probability masses, so the asymptotics depend on pointwise approximations to these masses across the relevant range of j. Consequently, our analysis relies fundamentally on a local limit theorem (LLT) to replace discrete binomial coefficients with Gaussian density functions, managing the approximation error in two distinct regimes.

The proof strategy bifurcates based on the relative size of the fold m:

1. **The Sublinear Regime ($m \ll n$):** In this setting, the summation range over the validation fold statistics is narrow relative to the scale of the training set fluctuations. Due to this scale separation, the probability mass function governing the training decision boundary is locally flat over the effective support of the validation statistic. This allows us to approximate the training term by a single *central coefficient*, decoupling the interaction.
2. **The Large m Regime:** When m is comparable to n, the scale separation vanishes and we cannot fix the training term. Instead, we approximate the entire summation as a *triple convolution* of Gaussian probability densities. Since the support is broad, the discrete sum converges rapidly to a Gaussian integral. The analysis here focuses on controlling the accumulated error from the Gaussian approximation across a large number of terms.

The above approach establishes the following result, a more explicit version of which (including precise constants) is provided in Appendix A.2.

Theorem 9 (Majority Asymptotics) *Throughout, let $n \geq 2$ and $m|n$. For Majority in the uniformly random-label classification model under 0–1 loss, the following statements hold:*

(A) *Uniformly over all $m = m_n$ with $m = \omega_n(1)$, and $m \leq n/2$,*

$$\mathrm{Cov}(n, m) = M_{n,m} + o_n(M_{n,m}) = \Theta\left(\frac{\sqrt{k}}{n}\right) \quad \textit{where} \quad M_{n,m} = \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}}.$$

(B) Monotonicity and minimizer of the Covariance.

Let $\mathcal{M}_n := \{m \in \mathbb{N} : m \mid n,\ 1 \leq m \leq n/2\}$. For all sufficiently large n, if $m_1, m_2 \in \mathcal{M}_n$ satisfy $m_1 < m_2 \leq n/3$, then

$$\mathrm{Cov}(n, m_1) > \mathrm{Cov}(n, m_2).$$

Moreover, along subsequences with $6 \mid n$,

$$\mathrm{Cov}(n, n/3) < \mathrm{Cov}(n, n/2).$$

Consequently, along subsequences with $3 \mid n$, the covariance-minimizing choice is $k = 3$.

(C) Minimizer of the MSE. *Along subsequences of even n, the MSE is minimized by $k = 2$ for all sufficiently large n. Moreover,*

$$\min_{k|n} \mathrm{MSE}_{\mathrm{CV}}^{(k)} = \mathrm{MSE}_{\mathrm{CV}}^{(2)} = \frac{1}{4n} + \frac{1}{2\pi n} + o(n^{-1}).$$

Proof *Proof of part (C).* By Proposition 4, $\mathrm{MSE}_{\mathrm{CV}}^{(k)} = \frac{k-1}{k}\,\mathrm{Cov}(n, n/k) + \frac{1}{4n}$. The second term is independent of k. On the range $m \le n/3$, both $\mathrm{Cov}(n,m)$ and the prefactor $(k-1)/k = 1 - m/n$ decrease with m, so their product is asymptotically bounded below by its continuous minimum at $m = n/3$. Hence, by part (B), it remains only to compare $m = n/3$ and $m = n/2$. Using part (A) and the endpoint asymptotic at $m = n/2$,

$$\frac{2}{3}\,\mathrm{Cov}(n, n/3) = \frac{1}{\sqrt{3}\pi n} + o(n^{-1}), \qquad \frac{1}{2}\,\mathrm{Cov}(n, n/2) = \frac{1}{2\pi n} + o(n^{-1}).$$

Since $1/(2\pi) < 1/(\sqrt{3}\pi)$, the full MSE is minimized by $k = 2$ for all sufficiently large n. ∎

We observe that the MSE scales as $\sqrt{k}/n$. In this setting, it is therefore advantageous to choose as few folds as possible. Notably, *hypothesis stability–based bounds* are not sufficiently fine-grained here: they incorrectly predict the MSE to increase when k decreases, see Section 5.3.

5.2. A Minimax Lower Bound for Cross-Validation with ERM Algorithms

The answer to Question 1 follows as a corollary of the preceding analysis of the Majority algorithm. Let $\mathcal{A}$ be an ERM algorithm over a nontrivial hypothesis class $\mathcal{H}$, meaning that there exist $h_0, h_1 \in \mathcal{H}$ and $x_0 \in \mathcal{X}$ such that $h_0(x_0) \neq h_1(x_0)$. Consider the distribution $\mathcal{D}$ supported on x_0, with the label drawn uniformly from $\{h_0(x_0), h_1(x_0)\}$. For this distribution and the 0–1 loss, empirical risk minimization reduces to choosing a classifier whose prediction at x_0 agrees with the majority label in the sample. Consequently, away from exact ties, the induced prediction at x_0 follows the binary Majority rule. Tie-breaking may depend on the ERM, but tie events affect only lower-order terms. Thus, the lower bound obtained above for Majority applies to any ERM over a nontrivial hypothesis class.

Corollary 10 *For any ERM algorithm $\mathcal{A}$ for classification with 0–1 loss such that $|\mathcal{H}| \ge 2$, it holds that*

$$\mathfrak{R}_{\mathrm{CV}}(\mathcal{A}) = \Omega\left(\frac{\sqrt{k^*}}{n}\right),$$

where k^* is the number of folds that achieves the minimax optimum.

This result shows that, in the distribution-free setting, no ERM algorithm with $k^*(n) = \omega_n(1)$ can utilize all n samples as efficiently as an independent validation set of the same size, whose mean-squared error decreases at the optimal rate of order $1/n$. Importantly, this result is agnostic to the specific value of k^*. While for Majority under random labels the optimal number of folds is $k = 2$, for other algorithms the minimax optimum may be attained at k^* growing with n, in which case the lower bound becomes even stronger.

Finally, as a straightforward corollary, our precise analysis of Majority allows us to make the gap between CV and using an independent hold-out set explicit.

Corollary 11 (Constant-factor gap for ERM) *Assume n is even. For any ERM algorithm $\mathcal{A}$ for classification with 0–1 loss such that $|\mathcal{H}| \ge 2$, there exists a data distribution $\mathcal{D}$ such that* *cross-validation incurs a multiplicative gap of*

$$\frac{\min_{k|n} MSE_{\mathrm{CV}}^{(k)}(\mathcal{A}, \mathcal{D})}{MSE_{\mathrm{val}}^{(n)}(\mathcal{A}, \mathcal{D})} = 1 + \frac{2}{\pi} + O(n^{-1})$$

where $MSE_{\mathrm{val}}^{(n)}$ denotes the MSE of an independent validation set of size n, and numerically, $1 + \frac{2}{\pi} \approx 1.637$.

Proof It suffices to consider the one-point random-label distribution used in the proof of Corollary 10. Under this distribution, any nontrivial ERM reduces to Majority, for which Theorem 9(C) shows that the MSE is minimized at $k = 2$, and $\mathrm{MSE}_{\mathrm{CV}}^{(2)} = \frac{1}{4n} + \frac{1}{2\pi n} + O(n^{-2})$. An independent validation set of size n has MSE $\mathbb{E}_{Y \sim \mathrm{Bin}(n,1/2)}[(1/2 - Y/n)^2] = 1/(4n)$. Taking the ratio yields the result. ■

5.3. Limitations of Existing Cross-Validation Theory: the Majority Benchmark

Arbitrarily loose sufficient conditions. We begin by examining the notion of hypothesis stability (Definition 1). Observe that the majority algorithm changes its output only when the sample is close to a tie. In the extreme case $k = n$ (leave-one-out CV), Majority flips its prediction only when the full sample has a tie or a one-vote margin, depending on parity and on the removed label. Since the probability of a tie is $\Theta(1/\sqrt{n})$, Majority is $(\Theta(1/\sqrt{n}), 1)$-hypothesis-stable.

For a general number of folds k, each fold has size $m = n/k$. By concentration of measure, the number of ones in a typical fold is $m/2 \pm O(\sqrt{m})$. For the output of Majority to change when a typical fold is removed, the full sample must therefore have a label count within $O(\sqrt{m})$ of $n/2$. The probability of this event is $\Theta(\sqrt{m}/\sqrt{n}) = \Theta(1/\sqrt{k})$. Consequently, Majority is $(\Theta(1/\sqrt{k}), m)$-hypothesis-stable. In particular, the stability of Majority deteriorates as the number of folds decreases.

If one extends the stability-based bounds of Rogers and Wagner (1978); Devroye and Wagner (1979a); Kearns and Ron (1997); Bousquet and Elisseeff (2002) from leave-one-out to general k-fold cross-validation, the resulting guarantees retain the same qualitative form,

$$L \leq \widehat{L}_{\mathrm{CV}}^{(k)} + O(\sqrt{\beta}),$$

where β denotes the hypothesis stability parameter. However, our analysis shows that CV provides more accurate risk estimates for Majority precisely when the algorithm is *less* stable. This mismatch is most pronounced at $k = 2$, where the stability-based bound differs from the true mean-squared error by a factor of $\Theta(n)$.

Similarly, Kumar et al. (2013) introduce a refined notion of stability, termed *loss stability* (Definition 2 in their paper, distinct from Definition 2 here), parameterized by γ. For the Majority algorithm, this notion effectively reduces to hypothesis stability with $m = 1$, except that the deviation inside the expectation is squared rather than taken in absolute value. By the discussion above, this yields $\gamma = \Theta(1/\sqrt{n})$ for Majority.

Instantiating Theorem 1 of Kumar et al. (2013) with the Majority algorithm, we find that the resulting upper bound on the MSE of the k-fold CV estimator is dominated by the term $(1-\frac{1}{k})\gamma$. Consequently, the bound becomes looser as k decreases. In the extreme case $k=2$, the predicted upper bound exceeds the true MSE of CV for Majority by a factor of $\Theta(\sqrt{n})$ ($\mathrm{MSE}_{\mathrm{CV}}^{(2)} = \Theta(1/n)$ by Theorem 9, and the bound predicts $\Theta(1/\sqrt{n})$).

Relative rather than absolute guarantees. Kearns and Ron (1997) compare leave-one-out cross-validation to empirical error and show that, for stable algorithms, the CV estimate is never much worse than the empirical error. However, this type of relative guarantee can be misleading when the quantity of interest is the MSE of risk estimation.

In particular, for Majority under leave-one-out CV we showed that $\mathrm{MSE}_{\mathrm{CV}}^{(n)} = \Theta(1/\sqrt{n})$. On the other hand, the empirical error is $\widehat{L}_{\mathrm{emp}} = \min(Y/n, 1-Y/n)$, and therefore $\left(\widehat{L}_{\mathrm{emp}} - \frac{1}{2}\right)^2 = \left(Y/n - \frac{1}{2}\right)^2$ which leads to an MSE of $\frac{1}{4n}$.

Thus, even though leave-one-out CV is guaranteed by Kearns and Ron (1997) to perform comparably to empirical error in a relative sense, its absolute performance in terms of MSE is worse by a factor of $\Theta(\sqrt{n})$. This gap is substantial, and it arises for an algorithm that is arguably among the most stable nontrivial learning rules.

5.4. Mathematical Inaccuracies in Prior Analyses

We also identify two concrete technical issues in prior analyses of cross-validation. These issues are qualitative rather than quantitative: the affected results make claims about fundamental properties of CV estimation that fail as stated.

First, Theorem 5.3 of Kearns and Ron (1997) claims a necessity of loss stability for accurate leave-one-out estimation. The proof overlooks possible cancellation between positive and negative deviations in the CV error. Lemma 3 gives a counterexample; see Appendix C.

Second, the main theorem of Kale et al. (2011) uses a conditional-expectation identity that is false in general: the empirical term uses a fixed validation fold, while the population term averages over an independent test point. Consequently, the advertised MSE bound does not follow from the proof; see Appendix D.

6. Conclusion

Our results show that the dependence induced by data reuse in cross-validation can impose a genuine statistical cost, even for very simple ERM procedures. As a natural extension of Corollary 10, it would be compelling to investigate which combined properties of algorithms and data distributions can yield improved minimax rates (or even attain the optimal $1/n$ rate) in settings beyond the distribution-free case.

References

Martin Anthony and Sean B Holden. Cross-validation for binary classification by real-valued functions: theoretical analysis. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 218–229, 1998.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 2010.

Sylvain Arlot and Matthieu Lerasle. Choice of v for v-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research*, 17(208):1–50, 2016.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350, 2020.

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

Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 203–208, 1999.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Alain Celisse. Optimal cross-validation in density estimation with the l^2-loss. *The Annals of Statistics*, 42(5):1879–1910, 2014.

Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979a.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979b.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic generalization measures are nowhere to be found. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Which algorithms have tight generalization bounds? *Advances in Neural Information Processing Systems*, 38: 36554–36589, 2026.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Innovations in Computer Science (ICS)*, pages 487–495, 2011.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 152–162, 1997.

Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR, 2013.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Advances in neural information processing systems*, 12, 1999.

Valentin V Petrov. *Sums of independent random variables*, volume 82. Springer Science & Business Media, 2012.

Richard R Picard and R Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

Vladimir N Vapnik and Alexey Y Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971.

Appendix A. Majority Algorithm

Throughout this section, we consider the following setup.

A.1. Setup and Notation

For $1 \le m \le n/2$, $m|n$, let $N := n - 2m$ and define

$$E(n,m) := 2^{-(n-2)} \sum_{j=0}^{m-1} \binom{m-1}{j}^2 \binom{n-2m}{\lfloor (n-m)/2 \rfloor - j}$$

such that $\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = E(n,m)/4$ (see Thm. 19 for details). We also denote $\mathrm{Cov}(n,m) \equiv \mathrm{Cov}(\widehat{L}_1, \widehat{L}_2)$ to highlight the roles of n, m.

Let $\mathsf{B}_r \sim \mathrm{Bin}(r, \frac{1}{2})$ with pmf $p_r(t) = 2^{-r}\binom{r}{t}$, and denote the Gaussian proxy

$$g_r(t) := \sqrt{\frac{2}{\pi r}} \exp\Big(-\frac{(2t-r)^2}{2r}\Big)$$

and central binomial mass

$$S_r := 2^{-2r}\binom{2r}{r}$$

A.2. Main Theorem

Theorem 12 (Fold-Covariance of the Majority Algorithm) *Throughout, let $n \ge 2$ and $m|n$.*

(A) Binomial form. *One has*

$$\mathrm{Cov}(n,m) = S_{m-1} \frac{1}{2\sqrt{\pi(2n-3m)}} + O\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right),$$

uniformly for all $1 \le m \le n/3$, where $S_{m-1} := 2^{-(2m-2)}\binom{2m-2}{m-1}$.

(B) Exact expression for $m = 1$. *It holds that*

$$\mathrm{Cov}(n,1) = 2^{-n}\binom{n-2}{\lfloor \frac{n-1}{2} \rfloor} = \sqrt{\frac{1}{8\pi(n-2)}} + O\Big(\frac{1}{n^{3/2}}\Big) = \sqrt{\frac{1}{8\pi n}} + O\Big(\frac{1}{n}\Big).$$

(C) Endpoint asymptotic for $m = n/2$. *It holds that*

$$\mathrm{Cov}\left(n, \tfrac{n}{2}\right) = \frac{1}{\pi(n-2)} + O\Big(\frac{1}{n^2}\Big) = \frac{1}{\pi n} + O\Big(\frac{1}{n^2}\Big).$$

(D) Large m regime. *Uniformly over all integer sequences $m = m_n$ satisfying $m = \omega_n(1)$ and $m \le n/3$,*

$$\mathrm{Cov}(n,m) = \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}} + O\left(\frac{1}{\sqrt{n}m^{3/2}}\right).$$

(E) Monotonicity and minimizer. *Let $\mathcal{M}_n := \{m \in \mathbb{N} : m \mid n,\ 1 \le m \le n/2\}$. For all sufficiently large n, if $m_1, m_2 \in \mathcal{M}_n$ satisfy $m_1 < m_2 \le n/3$, then*

$$\mathrm{Cov}(n,m_1) > \mathrm{Cov}(n,m_2).$$

Moreover, along subsequences with $6 \mid n$,

$$\mathrm{Cov}(n,n/3) < \mathrm{Cov}(n,n/2).$$

Consequently, along subsequences with $3 \mid n$, the covariance-minimizing choice is $k = 3$.

Proof This is a consequence of collecting the results of Theorems 20 to 23. ■

A.3. Technical Lemmas

Let us first state a few technical results.

Lemma 13 (Triple Gaussian Product) *Let $P(j) := g_{m-1}(j)^2\, g_N(\ell - j)$. With the parameters*

$$\alpha := \frac{4}{m-1}, \qquad \beta := \frac{2}{N}, \qquad \mu := \frac{\alpha \cdot \frac{m-1}{2} + \beta \cdot \frac{m}{2}}{\alpha+\beta} = \frac{(m-1)(2N+m)}{2(2N+m-1)},$$

the product $P(j)$ can be written as:

$$P(j) = \left(\frac{2}{\pi(m-1)}\sqrt{\frac{2}{\pi N}}\right)\exp\left(-\frac{1}{2N+m-1}\right)\exp\left(-(\alpha+\beta)(j-\mu)^2\right).$$

Furthermore, the sum of the rates is

$$\alpha + \beta = \frac{2(2N+m-1)}{(m-1)N}.$$

Proof Recall

$$g_r(t) := \sqrt{\frac{2}{\pi r}}\exp\left(-\frac{(2t-r)^2}{2r}\right).$$

Let $N := n - 2m$ and $\ell = (n-m)/2$.

We first write out the terms. Let $a := (m-1)/2$ and $\alpha := 4/(m-1)$.

$$g_{m-1}(j)^2 = \left(\sqrt{\frac{2}{\pi(m-1)}}\right)^2 \exp\left(-2\cdot\frac{2}{m-1}(j-\tfrac{m-1}{2})^2\right) = \frac{2}{\pi(m-1)}\, e^{-\alpha(j-a)^2}.$$

For the second term, let $b := m/2$ and $\beta := 2/N$. The exponent's center is $\ell - j - \frac{N}{2} = \frac{n-m}{2} - j - \frac{n-2m}{2} = \frac{m}{2} - j = -(j-b)$. Thus,

$$g_N(\ell - j) = \sqrt{\frac{2}{\pi N}}\,\exp\Big(-\frac{2}{N}(\ell - j - \tfrac{N}{2})^2\Big) = \sqrt{\frac{2}{\pi N}}\,e^{-\beta(j-b)^2}.$$

The product is

$$g_{m-1}(j)^2\,g_N(\ell - j) = \underbrace{\left(\frac{2}{\pi(m-1)}\sqrt{\frac{2}{\pi N}}\right)}_{:=C_{\mathrm{prod}}}\exp\{-\alpha(j-a)^2 - \beta(j-b)^2\}.$$

We complete the square for the exponential terms:

$$-\alpha(j-a)^2 - \beta(j-b)^2 = -(\alpha+\beta)(j-\mu)^2 - \frac{\alpha\beta}{\alpha+\beta}(a-b)^2,$$

where $\mu := (\alpha a + \beta b)/(\alpha+\beta)$ is as stated in the lemma. The constant term in the exponent depends on $a - b = (m-1)/2 - m/2 = -1/2$.

$$\frac{\alpha\beta}{\alpha+\beta}(a-b)^2 = \frac{1}{4}\cdot\frac{\frac{4}{m-1}\cdot\frac{2}{N}}{\frac{4}{m-1}+\frac{2}{N}} = \frac{1}{4}\cdot\frac{8/((m-1)N)}{(4N+2m-2)/((m-1)N)} = \frac{1}{2N+m-1}.$$

We also compute

$$\alpha+\beta = \frac{4}{m-1} + \frac{2}{N} = \frac{2(2N+m-1)}{(m-1)N}.$$

Combining these results yields the displayed formula. ∎

Lemma 14 (Poisson summation for Gaussians) *Let $\gamma > 0$ and $\mu \in \mathbb{R}$. Define*

$$f_{\gamma,\mu}(x) := e^{-\gamma(x-\mu)^2}.$$

Then

$$\sum_{j\in\mathbb{Z}} f_{\gamma,\mu}(j) = \sqrt{\frac{\pi}{\gamma}}\sum_{t\in\mathbb{Z}} e^{-\pi^2 t^2/\gamma}\,e^{-2\pi i t\mu}. \tag{1}$$

Proof Let $\mathcal{P}_{\gamma,\mu}(x) := \sum_{j\in\mathbb{Z}} f_{\gamma,\mu}(x+j)$ be the periodization (absolutely and uniformly convergent on $\mathbb{R}$). Then $\mathcal{P}_{\gamma,\mu}$ is 1–periodic and belongs to C^∞. Its complex Fourier series is

$$\mathcal{P}_{\gamma,\mu}(x) = \sum_{t\in\mathbb{Z}} c_t\,e^{2\pi i t x}, \qquad c_t = \int_0^1 \mathcal{P}_{\gamma,\mu}(x)\,e^{-2\pi i t x}\,dx.$$

By absolute convergence we may integrate termwise:

$$c_t = \sum_{j\in\mathbb{Z}} \int_0^1 e^{-\gamma(x+j-\mu)^2}\, e^{-2\pi i t x}\, dx = \int_{\mathbb{R}} e^{-\gamma(y-\mu)^2}\, e^{-2\pi i t y}\, dy =: \widehat{f}_{\gamma,\mu}(t),$$

after the change of variables $y = x + j$. The Gaussian Fourier transform is standard:

$$\widehat{f}_{\gamma,\mu}(t) = e^{-2\pi i t\mu} \int_{\mathbb{R}} e^{-\gamma z^2} e^{-2\pi i t z}\, dz = e^{-2\pi i t\mu} \sqrt{\frac{\pi}{\gamma}}\, e^{-\pi^2 t^2/\gamma}.$$

Thus

$$\mathcal{P}_{\gamma,\mu}(x) = \sqrt{\frac{\pi}{\gamma}} \sum_{t\in\mathbb{Z}} e^{-\pi^2 t^2/\gamma}\, e^{2\pi i t(x-\mu)}.$$

Evaluating at $x = 0$ gives

$$\sum_{j\in\mathbb{Z}} f_{\gamma,\mu}(j) = \mathcal{P}_{\gamma,\mu}(0) = \sqrt{\frac{\pi}{\gamma}} \sum_{t\in\mathbb{Z}} e^{-\pi^2 t^2/\gamma}\, e^{-2\pi i t\mu},$$

which is (1). ■

Lemma 15 (Lattice sum of the triple Gaussian) *With $N = n - 2m$, $\ell = (n-m)/2$, and the parameters*

$$\alpha = \frac{4}{m-1}, \quad \beta = \frac{2}{N}, \quad \mu = \frac{(m-1)(2N+m)}{2(2N+m-1)},$$

we have the exact identity

$$\sum_{j\in\mathbb{Z}} g_{m-1}(j)^2\, g_N(\ell - j) = \frac{2}{\pi} \cdot \frac{1}{\sqrt{(m-1)(2N+m-1)}}\, e^{-\frac{1}{2N+m-1}}\ \Theta_{n,m}, \tag{2}$$

$$\Theta_{n,m} := \sum_{t\in\mathbb{Z}} \exp\Big(-\pi^2 t^2/(\alpha+\beta)\Big)\ \exp\big(-2\pi i t\mu\big). \tag{3}$$

Equivalently, using $\alpha + \beta = \dfrac{2(2N+m-1)}{(m-1)N}$,

$$\sum_{j\in\mathbb{Z}} g_{m-1}(j)^2\, g_N(\ell - j)$$
$$= \frac{2}{\pi} \cdot \frac{1}{\sqrt{(m-1)(2n-3m-1)}}\, e^{-\frac{1}{2n-3m-1}} \sum_{t\in\mathbb{Z}} \exp\Big(-\frac{\pi^2 t^2 (m-1)N}{2(2n-3m-1)}\Big)\, e^{-2\pi i t\mu}. \tag{4}$$

Proof By Lemma 13, we have

$$g_{m-1}(j)^2\, g_N(\ell - j) = C_{\text{prod}} \cdot e^{-\frac{1}{2N+m-1}}\ \exp\Big(-(\alpha+\beta)(j-\mu)^2\Big),$$

where $C_{\text{prod}} = \frac{2}{\pi(m-1)}\sqrt{\frac{2}{\pi N}}$. Summing over $j \in \mathbb{Z}$ and applying Lemma 14 with $\gamma := \alpha + \beta$, we obtain

$$\begin{aligned}\sum_{j\in\mathbb{Z}} g_{m-1}(j)^2\, g_N(\ell - j) &= C_{\text{prod}} \cdot e^{-\frac{1}{2N+m-1}} \sum_{j\in\mathbb{Z}} e^{-(\alpha+\beta)(j-\mu)^2} \\ &= C_{\text{prod}} \cdot e^{-\frac{1}{2N+m-1}} \cdot \sqrt{\frac{\pi}{\alpha+\beta}} \sum_{t\in\mathbb{Z}} e^{-\pi^2 t^2/(\alpha+\beta)} e^{-2\pi i t\mu}.\end{aligned}$$

We now compute the combined prefactor. Using $\alpha + \beta = \frac{2(2N+m-1)}{(m-1)N}$ from Lemma 13:

$$\begin{aligned} C_{\text{prod}} \sqrt{\frac{\pi}{\alpha+\beta}} &= \left(\frac{2}{\pi(m-1)}\sqrt{\frac{2}{\pi N}}\right) \cdot \sqrt{\frac{\pi(m-1)N}{2(2N+m-1)}} \\ &= \left(\frac{2\sqrt{2}}{\pi^{3/2}(m-1)\sqrt{N}}\right) \cdot \left(\frac{\sqrt{\pi}\sqrt{m-1}\sqrt{N}}{\sqrt{2}\sqrt{2N+m-1}}\right) \\ &= \frac{2}{\pi\sqrt{m-1}\sqrt{2N+m-1}} = \frac{2}{\pi} \cdot \frac{1}{\sqrt{(m-1)(2N+m-1)}}. \end{aligned}$$

Substituting this prefactor back into the sum yields (2).

For (4), we substitute the expression for $\alpha + \beta$ into the exponent and use $N = n - 2m$ in the denominator, noting that $2N + m - 1 = 2(n - 2m) + m - 1 = 2n - 3m - 1$. ■

Lemma 16 (Local Limit Theorem and Central Binomial Mass) *Let $r \ge 2$, $c := \lfloor r/2 \rfloor$ and $p_r(t) := 2^{-r}\binom{r}{t}$. Let $g_r(t) := \sqrt{\frac{2}{\pi r}}\exp\big(-(2t-r)^2/(2r)\big)$. There exists an absolute $C_0 > 0$ such that*

$$\sup_{t\in\mathbb{Z}} |\, p_r(t) - g_r(t)\,| \;\le\; C_0\, r^{-3/2}. \tag{5}$$

In particular, at the center $t = c$,

$$\Big| p_r(c) - g_r(c) \Big| \;\le\; C_0\, r^{-3/2}, \qquad g_r(c) = \begin{cases} \sqrt{\frac{2}{\pi r}}, & r \text{ even}, \\ \sqrt{\frac{2}{\pi r}}\, e^{-1/(2r)}, & r \text{ odd}. \end{cases} \tag{6}$$

Hence, for all $r \ge 2$,

$$\sqrt{\frac{2}{\pi r}}\, e^{-1/(2r)} - C_0\, r^{-3/2} \;\le\; p_r(c) \;\le\; \sqrt{\frac{2}{\pi r}} + C_0\, r^{-3/2}. \tag{7}$$

Proof This is a classical uniform local limit theorem, see (Petrov, 2012, Chapter 7, Theorem 13) (with $p = q = \frac{1}{2}$). Evaluating at $t = c$ gives (6); the bounds (7) follow since $g_r(c)$ is as displayed. ■

Lemma 17 (Moments of the squared-binomial weights) *Let $r \geq 0$, and define*

$$p_r(j) := 2^{-r}\binom{r}{j}, \qquad S_r := \sum_{j=0}^{r} p_r(j)^2, \qquad w_j := \frac{p_r(j)^2}{S_r}, \quad j = 0, \ldots, r.$$

Let J have law $\mathbb{P}(J = j) = w_j$. Then

$$J \sim \mathrm{Hypergeom}(2r, r, r),$$

with the convention that $J = 0$ deterministically when $r = 0$. Moreover,

$$\mathbb{E}[J] = \frac{r}{2}, \qquad \mathrm{Var}(J) = \frac{r^2}{4(2r-1)} = \frac{r}{8} + O(1),$$

and, for every $|\eta| \leq 1/2$,

$$\mathbb{E}\left[\left(\frac{r}{2} - J + \eta\right)^2\right] = \frac{r}{8} + O(1), \qquad \mathbb{E}\left[\left(\frac{r}{2} - J + \eta\right)^4\right] = O\big((r+1)^2\big).$$

Proof By Vandermonde's identity,

$$S_r = 2^{-2r} \sum_{j=0}^{r} \binom{r}{j}^2 = 2^{-2r}\binom{2r}{r}.$$

Thus, for $r \geq 1$,

$$w_j = \frac{\binom{r}{j}^2}{\binom{2r}{r}} = \frac{\binom{r}{j}\binom{r}{r-j}}{\binom{2r}{r}}, \qquad j = 0, \ldots, r,$$

which is the probability mass function of $\mathrm{Hypergeom}(2r, r, r)$. Hence

$$\mathbb{E}[J] = r \cdot \frac{r}{2r} = \frac{r}{2}.$$

Using the standard variance formula for $X \sim \mathrm{Hypergeom}(N, K, n)$,

$$\mathrm{Var}(X) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1},$$

with $N = 2r$, $K = r$, and $n = r$, gives

$$\mathrm{Var}(J) = r \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{r}{2r-1} = \frac{r^2}{4(2r-1)} = \frac{r}{8} + O(1).$$

It remains to bound the fourth moment. View J as the number of marked elements in a sample of size r drawn without replacement from a population of $2r$ elements, exactly r of which are marked. By Hoeffding's inequality for sampling without replacement,

$$\mathbb{P}\left(\left|J-\frac{r}{2}\right|\geq t\right)\leq 2\exp\left(-\frac{2t^2}{r}\right),\qquad t\geq 0.$$

Therefore, using the tail-integral identity for nonnegative random variables,

$$\begin{aligned}\mathbb{E}\left[\left|J-\frac{r}{2}\right|^4\right] &= \int_0^\infty 4t^3\,\mathbb{P}\left(\left|J-\frac{r}{2}\right|\geq t\right)\,dt\\ &\leq 8\int_0^\infty t^3\exp\left(-\frac{2t^2}{r}\right)\,dt = r^2.\end{aligned}$$

Finally, since $\mathbb{E}[J-r/2]=0$, for $|\eta|\leq 1/2$,

$$\mathbb{E}\left[\left(\frac{r}{2}-J+\eta\right)^2\right]=\mathrm{Var}(J)+\eta^2=\frac{r}{8}+O(1),$$

and

$$|a+b|^4\leq 8|a|^4+8|b|^4$$

gives

$$\begin{aligned}\mathbb{E}\left[\left(\frac{r}{2}-J+\eta\right)^4\right] &\leq 8\,\mathbb{E}\left[\left|J-\frac{r}{2}\right|^4\right]+8|\eta|^4\\ &= O\big((r+1)^2\big).\end{aligned}$$

The case $r=0$ is immediate. ■

A.4. Exact Combinatorial Form of the Fold-Covariance

Lemma 18 (Simplification of Covariance Term) *Let $X_m\sim\mathrm{Bin}(m,1/2)$, $\theta=(n-m)/2$, and $a(y)=\lfloor\theta-y\rfloor$. Then, for every fixed y,*

$$\mathrm{Cov}_{X_m}(X_m,\mathbf{1}_{X_m>\theta-y})=\frac{m}{4}\mathbb{P}(X_{m-1}=a(y))$$

where $X_{m-1}\sim\mathrm{Bin}(m-1,1/2)$.

Proof Throughout, let $C(m,y):=\mathrm{Cov}_{X_m}(X_m,\mathbf{1}_{X_m>\theta-y})$. Let $a(y)=\lfloor\theta-y\rfloor$. The event $X_m>\theta-y$ is equivalent to $X_m\geq\lfloor\theta-y\rfloor+1=a(y)+1$. The covariance is $C(m,y)=\mathbb{E}_{X_m}[X_m\mathbf{1}_{X_m\geq a(y)+1}]-\mathbb{E}_{X_m}[X_m]\cdot\mathbb{P}(X_m\geq a(y)+1)$. Since $X_m\sim\mathrm{Bin}(m,1/2)$, its expectation is $\mathbb{E}_{X_m}[X_m]=m/2$. The first term is $\mathbb{E}_{X_m}[X_m\mathbf{1}_{X_m\geq a(y)+1}]=\sum_{j=a(y)+1}^{m}j\binom{m}{j}(1/2)^m$. Using $j\binom{m}{j}=m\binom{m-1}{j-1}$:

$$\sum_{j=a(y)+1}^{m}m\binom{m-1}{j-1}(1/2)^m=\frac{m}{2}\sum_{j'=a(y)}^{m-1}\binom{m-1}{j'}(1/2)^{m-1}=\frac{m}{2}\mathbb{P}(X_{m-1}\geq a(y))$$

where $X_{m-1} \sim \mathrm{Bin}(m-1, 1/2)$. So, $C(m,y) = \frac{m}{2}\mathbb{P}(X_{m-1} \geq a(y)) - \frac{m}{2}\mathbb{P}(X_m \geq a(y)+1)$. To simplify $\mathbb{P}(X_m \geq j+1)$, let $X_m = X_{m-1} + B$, where $B \sim \mathrm{Bernoulli}(1/2)$ is independent of X_{m-1}.

$$\begin{aligned}
\mathbb{P}(X_m \geq j+1) &= \mathbb{P}(X_{m-1} + B \geq j+1 | B=0)\mathbb{P}(B=0) + \mathbb{P}(X_{m-1} + B \geq j+1 | B=1)\mathbb{P}(B=1) \\
&= \frac{1}{2}\mathbb{P}(X_{m-1} \geq j+1) + \frac{1}{2}\mathbb{P}(X_{m-1} \geq j)
\end{aligned}$$

Substituting this with $j = a(y)$:

$$\begin{aligned}
C(m,y) &= \frac{m}{2}\left[\mathbb{P}(X_{m-1} \geq a(y)) - \left(\frac{1}{2}\mathbb{P}(X_{m-1} \geq a(y)+1) + \frac{1}{2}\mathbb{P}(X_{m-1} \geq a(y))\right)\right] \\
&= \frac{m}{4}\left[\mathbb{P}(X_{m-1} \geq a(y)) - \mathbb{P}(X_{m-1} \geq a(y)+1)\right] = \frac{m}{4}\mathbb{P}(X_{m-1} = a(y))
\end{aligned}$$

■

Theorem 19 *It holds that*

$$\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = 2^{-n} \sum_{j=0}^{m-1} \binom{m-1}{j}^2 \binom{n-2m}{\lfloor (n-m)/2 \rfloor - j}$$

Proof We know from combining Corollary 6 and Lemma 18 that $\mathrm{Cov}(\widehat{L}_1, \widehat{L}_2) = \frac{1}{4}\mathbb{E}_Y[\mathbb{P}(X = \lfloor \theta - Y \rfloor \mid Y)^2]$ where $X \sim \mathrm{Bin}(m-1, 1/2)$ is independent of Y. Furthermore,

$$\begin{aligned}
&\mathbb{E}_Y[\mathbb{P}(X = \lfloor \theta - Y \rfloor \mid Y)^2] \\
&\qquad = \mathbb{E}_Y[\mathbb{P}(X_1 = \lfloor \theta - Y \rfloor, X_2 = \lfloor \theta - Y \rfloor | Y)] \quad \text{(introducing } X_1, X_2 \text{ cond. indep. given } Y\text{)} \\
&\qquad = \mathbb{P}(X_1 = \lfloor \theta - Y \rfloor, X_2 = \lfloor \theta - Y \rfloor) \quad \text{(by Law of Total Expectation)} \\
&\qquad = \sum_{j=0}^{m-1} \mathbb{P}(X_1 = j, X_2 = j, \lfloor \theta - Y \rfloor = j) \quad \text{(summing over the support of } X_1, X_2\text{)} \\
&\qquad = \sum_{j=0}^{m-1} \mathbb{P}(X_1 = j, X_2 = j) \cdot \mathbb{P}(j = \lfloor \theta - Y \rfloor) \quad \text{(by independence of } (X_1, X_2) \text{ and } Y\text{)} \\
&\qquad = \sum_{j=0}^{m-1} \mathbb{P}(X_1 = j)^2 \mathbb{P}(\lfloor \theta - Y \rfloor = j) \quad \text{(by independence of } X_1, X_2\text{)} \\
&\qquad = \left(\frac{1}{2^{n-2}}\right) \sum_{j=0}^{m-1} \binom{m-1}{j}^2 \binom{n-2m}{\lfloor (n-m)/2 \rfloor - j} \quad \text{(writing out definition).}
\end{aligned}$$

Multiplying by the prefactor $1/4$ in the first display gives the claimed factor 2^{-n}. ■

A.5. Main Results for the Majority Algorithm Fold-Covariance

Theorem 20 (Precise binomial form) *Fix integers n and $m|n$. Let $N := n - 2m$, and choose N_C as the even integer closest to $n - \frac{3}{2}m$.*

Then, one has

$$\mathrm{Cov}(n, m) = S_{m-1} \frac{1}{2\sqrt{\pi(2n - 3m)}} + O\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right),$$

uniformly for all $1 \le m \le n/3$, where the error term is negligible once $m = o(n)$.

Proof Define

$$q_t(r) := 2^{-t}\binom{t}{r}.$$

Set

$$p(j) := 2^{-(m-1)}\binom{m-1}{j}, \quad m_0 := \left\lfloor \frac{n-m}{2} \right\rfloor, \quad P_N(r) := q_N(r), \quad P_{N_c} := q_{N_c}(N_c/2).$$

Then

$$E(n, m) = \sum_{j=0}^{m-1} p(j)^2\, P_N(m_0 - j) \;=\; S_{m-1}\, P_{N_c} \;+\; R_1, \quad R_1 := \sum_{j=0}^{m-1} p(j)^2 \big(P_N(m_0 - j) - P_{N_c}\big). \tag{8}$$

Step 1: LLT expansions. Apply (5) to $p_N(m_0 - j)$ and $S_{N_c/2}$.

$$p_N(m_0 - j) = G_N(j) + \delta_N(j), \qquad S_{N_c/2} = G_{N_c} + \delta_{N_c},$$

where

$$G_N(j) := \frac{1}{\sqrt{\pi N/2}} \exp\left(-\frac{2\Delta_j^2}{N}\right), \quad G_{N_c} := \frac{1}{\sqrt{\pi N_c/2}}, \quad \Delta_j := m_0 - j - \frac{N}{2} = \frac{m}{2} - j - \theta,\ \theta \in [0, 1),$$

and $|\delta_N(j)| \le C_{\mathrm{LLT}} N^{-3/2}$, $|\delta_{N_c}| \le C_{\mathrm{LLT}} N_c^{-3/2}$.

Rigorously, $P_{N_c} := p_{N_c}(c_{N_c})$ with $c_{N_c} := \lfloor N_c/2 \rfloor$, so $G_{N_c} := g_{N_c}(c_{N_c}) = \sqrt{\frac{2}{\pi N_c}} e^{-(2c_{N_c} - N_c)^2/(2N_c)}$. By (6) this is $\sqrt{\frac{2}{\pi N_c}}$ if N_c even or $\sqrt{\frac{2}{\pi N_c}} e^{-1/(2N_c)}$ if N_c odd. As $e^{-1/(2N_c)} = 1 + O(N_c^{-1})$, in both cases $G_{N_c} = \sqrt{\frac{2}{\pi N_c}} + O(N_c^{-3/2})$.

Step 2: Decomposition of R_1. Plugging in the Gaussian approximation, we get

$$R_1 = \sum_{j=0}^{m-1} p_{m-1}(j)^2 \Big(G_N(j) - G_{N_c}\Big) \;+\; \underbrace{\sum_{j=0}^{m-1} p_{m-1}(j)^2 \big(\delta_N(j) - \delta_{N_c}\big)}_{=: R_{\mathrm{LLT}}}. \tag{9}$$

Step 3: Bounding the LLT remainder. By $S_{m-1} \le 1$ and the local limit theorem bound of Lemma 16,

$$|R_{\mathrm{LLT}}| \;\le\; \sum_j p_{m-1}(j)^2\big(|\delta_N(j)| + |\delta_{N_c}|\big) \;\le\; C_{\mathrm{LLT}}\Big(\frac{1}{N^{3/2}} + \frac{1}{N_c^{3/2}}\Big) \;=\; O\Big(\frac{1}{n^{3/2}}\Big).$$

Step 4. N_c from the first–order optimal Gaussian central term. Next, we bound

$$\sum_{j=0}^{m-1} p(j)^2\big(G_N(j) - G_{N_c}\big),$$

where $N = n - 2m$. Define the probability weights

$$w_j := \frac{p(j)^2}{S_{m-1}}, \qquad j \in \{0, \dots, m-1\},$$

and let $\mathbb{E}_w$ denote expectation with respect to these weights. Let $J \sim w$. By symmetry,

$$\mathbb{E}_w[J] = \frac{m-1}{2}.$$

As before, write

$$\Delta_j := m_0 - j - \frac{N}{2} = \frac{m}{2} - j - \theta, \qquad \theta \in [0, 1).$$

Then

$$\Delta_J = \left(\frac{m-1}{2} - J\right) + \left(\frac{1}{2} - \theta\right).$$

By Lemma 17, applied with $r = m - 1$, we have

$$J \sim \mathrm{Hypergeom}\big(2(m-1), m-1, m-1\big),$$

and

$$\mathbb{E}_w[J] = \frac{m-1}{2}, \qquad \mathrm{Var}_w(J) = \frac{(m-1)^2}{4(2m-3)} = \frac{m-1}{8} + O(1).$$

Since

$$\Delta_J = \left(\frac{m-1}{2} - J\right) + \left(\frac{1}{2} - \theta\right), \qquad \theta \in [0, 1),$$

Lemma 17, with $\eta = 1/2 - \theta$, gives

$$\mathbb{E}_w[\Delta_J^2] = \frac{m-1}{8} + O(1), \qquad \mathbb{E}_w[\Delta_J^4] = O(m^2).$$

Set

$$c(t) := \sqrt{\frac{2}{\pi}}\, t^{-1/2}.$$

Then

$$\sum_{j=0}^{m-1} p(j)^2 \big(G_N(j) - G_{N_c}\big) = S_{m-1} \left\{ c(N)\mathbb{E}_w\left[\exp\left(-\frac{2\Delta_J^2}{N}\right) - 1\right] - \big(c(N_c) - c(N)\big) \right\}.$$

Let

$$x_J := \frac{2\Delta_J^2}{N}.$$

Since

$$0 \le e^{-x} - (1-x) \le \frac{x^2}{2}, \qquad x \ge 0,$$

we get

$$\begin{aligned} \mathbb{E}_w\left[\exp\left(-\frac{2\Delta_J^2}{N}\right) - 1\right] &= -\frac{2}{N}\mathbb{E}_w[\Delta_J^2] + O\left(\frac{\mathbb{E}_w[\Delta_J^4]}{N^2}\right) \\ &= -\frac{m-1}{4N} + O\left(\frac{1}{N} + \frac{m^2}{N^2}\right). \end{aligned}$$

On the other hand, since N_c is the even integer closest to $n - \frac{3}{2}m$, we have

$$N_c - N = \frac{m}{2} + O(1).$$

A Taylor expansion of $c(t)$ around N gives

$$\begin{aligned} c(N_c) - c(N) &= -\frac{c(N)}{2}\frac{N_c - N}{N} + O\left(c(N)\frac{(N_c - N)^2}{N^2}\right) \\ &= -c(N)\frac{m}{4N} + O\left(\frac{c(N)}{N} + c(N)\frac{m^2}{N^2}\right). \end{aligned}$$

Combining the last two displays,

$$\left|\sum_{j=0}^{m-1} p(j)^2 \big(G_N(j) - G_{N_c}\big)\right| \le C S_{m-1}\left(\frac{c(N)}{N} + c(N)\frac{m^2}{N^2}\right).$$

Since $m \le n/3$, we have $N \asymp n$ and $c(N) \asymp n^{-1/2}$. Also $S_{m-1} = O(m^{-1/2})$, with the convention that this bound is harmless at $m = 1$. Hence

$$\left|\sum_{j=0}^{m-1} p(j)^2 \big(G_N(j) - G_{N_c}\big)\right| = O\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right).$$

Step 5: Completing the proof Collecting Step 3 and Step 4, we have

$$|R_1| \le C\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right),$$

uniformly for $1 \le m \le n/3$. Furthermore,

$$P_{N_c} = G_{N_c} + \delta_{N_c} = \sqrt{\frac{2}{\pi N_c}} + O(N_c^{-3/2}).$$

Since

$$N_c = \frac{2n-3m}{2} + O(1),$$

we also have

$$\sqrt{\frac{2}{\pi N_c}} = \frac{2}{\sqrt{\pi(2n-3m)}} + O(n^{-3/2}).$$

Thus,

$$E(n,m) = S_{m-1}\frac{2}{\sqrt{\pi(2n-3m)}} + O\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right).$$

Recalling that $\mathrm{Cov}(n,m) = E(n,m)/4$, we obtain

$$\mathrm{Cov}(n,m) = S_{m-1}\frac{1}{2\sqrt{\pi(2n-3m)}} + O\left(n^{-3/2} + \frac{m^{3/2}}{n^{5/2}}\right),$$

uniformly for $1 \le m \le n/3$. This completes the proof.

■

Theorem 21 (Endpoint cases)
(A) Exact expression for $m = 1$. *It holds that*

$$\mathrm{Cov}(n,1) = 2^{-n}\binom{n-2}{\lfloor \frac{n-1}{2} \rfloor}.$$

Consequently, by the LLT in Lemma 16,

$$\mathrm{Cov}(n,1) = \sqrt{\frac{1}{8\pi(n-2)}} \; + \; O\Big(\frac{1}{n^{3/2}}\Big) = \sqrt{\frac{1}{8\pi n}} \; + \; O\Big(\frac{1}{n}\Big).$$

(B) Endpoint asymptotic for $m = n/2$. *It holds that*

$$\mathrm{Cov}\left(n, \tfrac{n}{2}\right) = \frac{1}{\pi(n-2)} + O\Big(\frac{1}{n^2}\Big) = \frac{1}{\pi n} + O\Big(\frac{1}{n^2}\Big).$$

Proof

(A) The case $m = 1$. When $m = 1$, the sum has only $j = 0$ and $\binom{m-1}{0}^2 = 1$, so

$$E(n,1) = 2^{-(n-2)}\binom{n-2}{\lfloor \frac{n-1}{2} \rfloor}.$$

This is exactly the central (or near-central) mass of $\mathrm{Bin}(n-2, \frac{1}{2})$; by Lemma 16,

$$E(n,1) = \frac{1}{\sqrt{\pi(n-2)/2}} + O\Big(\frac{1}{(n-2)^{3/2}}\Big) = \sqrt{\frac{2}{\pi(n-2)}} + O\Big(\frac{1}{n^{3/2}}\Big) = \sqrt{\frac{2}{\pi n}} + O\Big(\frac{1}{n}\Big).$$

(B) The case $m = \frac{n}{2}$ (so n even). Set $\ell := \frac{n}{2} - 1$ and observe

$$E\Big(n, \tfrac{n}{2}\Big) = 2^{-(n-2)}\sum_{j=0}^{\ell}\binom{\ell}{j}^2\binom{0}{\lfloor n/4 \rfloor - j}.$$

Since $\binom{0}{r} = \mathbf{1}\{r = 0\}$, only the term $j = r := \lfloor n/4 \rfloor$ survives:

$$E\Big(n, \tfrac{n}{2}\Big) = \Big(2^{-\ell}\binom{\ell}{r}\Big)^2 =: q_\ell(r)^2,$$

i.e. it is the square of a symmetric binomial mass $q_\ell(r) = \Pr\{\mathrm{Bin}(\ell, \frac{1}{2}) = r\}$.

Note that

$$r - \frac{\ell}{2} = \begin{cases} 1/2, & n \equiv 0 \pmod 4, \\ 0, & n \equiv 2 \pmod 4. \end{cases}$$

In either case,

$$\frac{2(r - \ell/2)^2}{\ell} = O(\ell^{-1}),$$

and hence

$$q_\ell(r) = \sqrt{\frac{2}{\pi\ell}}\left(1 + O(\ell^{-1})\right).$$

Therefore

$$E(n, n/2) = q_\ell(r)^2 = \frac{2}{\pi\ell} + O(\ell^{-2}) = \frac{4}{\pi(n-2)} + O(n^{-2}),$$

and since $\mathrm{Cov}(n, n/2) = E(n, n/2)/4$,

$$\mathrm{Cov}(n, n/2) = \frac{1}{\pi(n-2)} + O(n^{-2}).$$

■

Theorem 22 (Covariance for large m) *Uniformly over all integer sequences $m = m_n$ satisfying $m = \omega_n(1)$ and $m \le n/3$,*

$$\mathrm{Cov}(n, m) = \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}} + O\left(\frac{1}{\sqrt{n}m^{3/2}}\right).$$

Proof Write $p_r(t) := 2^{-r}\binom{r}{t}$ and $g_r(t) := \sqrt{\frac{2}{\pi r}}\exp\big(-(2t-r)^2/(2r)\big)$. Set

$$E(n, m) = \sum_{j=0}^{m-1} p_{m-1}(j)^2\, p_N\Big(\tfrac{N}{2} + \Delta_j\Big), \qquad \Delta_j := \ell - j - \tfrac{N}{2}.$$

Let $N := n - 2m$, $\ell := \lfloor (n-m)/2 \rfloor$, and write

$$\varepsilon := \frac{n-m}{2} - \ell \in \{0, 1/2\}.$$

We decompose

$$E(n,m) - \frac{2}{\pi\sqrt{(m-1)(2n-3m)}} = \underbrace{\big(E(n,m) - \sum_J p_{m-1}^2 g_N\big)}_{T_1} + \underbrace{\big(\sum_J p_{m-1}^2 g_N - \sum_J g_{m-1}^2 g_N\big)}_{T_2}$$
$$+ \underbrace{\big(\sum_J g_{m-1}^2 g_N - \sum_{\mathbb{Z}} g_{m-1}^2 g_N\big)}_{T_3} + \underbrace{\big(\sum_{\mathbb{Z}} g_{m-1}^2 g_N - \frac{2}{\pi\sqrt{(m-1)(2n-3m)}}\big)}_{T_4},$$

where $J = \{0, \ldots, m-1\}$ in T_1, T_2.

1) T_1: replace p_N by g_N (uniform LLT). The LLT Lemma 16 yields $\sup_t |p_N(t) - g_N(t)| \le C_0 N^{-3/2}$, hence

$$|T_1| \;\le\; C_0 N^{-3/2} \sum_{j=0}^{m-1} p_{m-1}(j)^2 = C_0 N^{-3/2}\, p_{2m-2}(m-1).$$

By the same LLT at $r = 2m-2$, $p_{2m-2}(m-1) \le g_{2m-2}(m-1) + C_0(2m-2)^{-3/2} \le \frac{1}{\sqrt{\pi(m-1)}} + \frac{C_0}{2^{3/2}(m-1)^{3/2}}$. Since $N \ge n/3$,

$$|T_1| \;\le\; \frac{C}{\sqrt{m}n^{3/2}}.$$

2) T_2: replace p_{m-1} by g_{m-1}. Let $\delta_j := p_{m-1}(j) - g_{m-1}(j)$. Then

$$|T_2| = \Big|\sum_{j=0}^{m-1} \delta_j\,(p_{m-1}(j) + g_{m-1}(j))\, g_N\Big(\tfrac{N}{2} + \Delta_j\Big)\Big| \le \big(\sup_j |\delta_j|\big)\,\Sigma,$$

with

$$\Sigma := \sum_{j=0}^{m-1} (p_{m-1}(j) + g_{m-1}(j))\, g_N\Big(\tfrac{N}{2} + \Delta_j\Big).$$

Uniform LLT (Lemma 16) at scale $m-1$ gives $\sup_j |\delta_j| \le C_0(m-1)^{-3/2}$. Moreover, upper bounding the average by the maximum,

$$\sum_{j=0}^{m-1} p_{m-1}(j)\, g_N\Big(\tfrac{N}{2} + \Delta_j\Big) \le \sup_t g_N(t) \;\le\; \sqrt{\frac{2}{\pi N}},$$

and similarly, by extending to $\mathbb{Z}$ and using the lattice Gaussian–Gaussian convolution,

$$\sum_{j=0}^{m-1} g_{m-1}(j)\, g_N\Big(\tfrac{N}{2} + \Delta_j\Big) \le \sum_{j\in\mathbb{Z}} g_{m-1}(j)\, g_N(\ell - j) \;\le\; \frac{C'}{\sqrt{N}}.$$

Therefore $\Sigma \le C/\sqrt{N}$, and

$$|T_2| \;\le\; \frac{C_0}{(m-1)^{3/2}} \cdot \frac{C}{\sqrt{N}} \;\le\; \frac{C}{m^{3/2}\sqrt{n}}.$$

3) T_3: truncating the full-lattice Gaussian sum.

Here

$$T_3 = \sum_{j=0}^{m-1} g_{m-1}(j)^2 g_N(\ell - j) - \sum_{j\in\mathbb{Z}} g_{m-1}(j)^2 g_N(\ell - j),$$

where $N = n - 2m$ and $\ell = \lfloor (n-m)/2 \rfloor$. Put

$$\varepsilon := \frac{n-m}{2} - \ell \in \{0, 1/2\}, \qquad a := \frac{m-1}{2}, \qquad b := \frac{m}{2} - \varepsilon,$$

and

$$\alpha := \frac{4}{m-1}, \qquad \beta := \frac{2}{N}, \qquad \gamma := \alpha + \beta.$$

Then

$$g_{m-1}(j)^2 g_N(\ell - j) = C_{m,N} \exp\{-\alpha(j-a)^2 - \beta(j-b)^2\} = C'_{m,N} \exp\{-\gamma(j-\mu)^2\},$$

where

$$\mu = \frac{\alpha a + \beta b}{\gamma} = \frac{(m-1)(2N + m - 2\varepsilon)}{2(2N + m - 1)}$$

and

$$\sigma^2 := \frac{1}{2\gamma} = \frac{(m-1)N}{4(2N + m - 1)}.$$

Since $m \le n/3$, we have $N \ge m$. Hence, for all sufficiently large m,

$$\mu \ge cm, \qquad m-1-\mu \ge cm, \qquad \sigma^2 \le Cm$$

for universal constants $c, C > 0$. Therefore both tails outside $\{0, \dots, m-1\}$ are Gaussian tails at distance at least $c\sqrt{m}$ standard deviations. Consequently,

$$|T_3| \le C\frac{e^{-cm}}{\sqrt{mN}} \le C\frac{1}{\sqrt{n}\, m^{3/2}},$$

where the last inequality uses $N \asymp n$ for $m \le n/3$ and $\sup_{x\ge 1} xe^{-cx} < \infty$.

4) T_4: Gaussian triple product to the simple main term. The full-lattice Gaussian sum has the exact form

$$\sum_{j\in\mathbb{Z}} g_{m-1}(j)^2\, g_N(\ell-j) = \frac{2}{\pi\sqrt{(m-1)(2n-3m)}} \cdot A_{n,m}\,\Theta_{n,m},$$

with

$$A_{n,m} = \frac{\sqrt{2n-3m}}{\sqrt{2n-3m-1}}\, e^{-1/(2n-3m-1)} = 1 + O(1/n),$$

and

$$\Theta_{n,m} = 1 + 2\sum_{t\ge 1} \exp\Big(-\frac{\pi^2 t^2 (m-1)N}{2(2n-3m-1)}\Big)\cos(2\pi t\mu).$$

When $\varepsilon = 0$, this is exactly Lemma 15. When $\varepsilon = 1/2$, the same calculation applies with the second Gaussian center shifted from $m/2$ to $m/2 - \varepsilon$. This changes the constant factor

$$\exp\left\{-\frac{\alpha\beta}{\alpha+\beta}(a-b)^2\right\}$$

and the phase μ, but since $|a-b| \le 1/2$ and

$$\frac{\alpha\beta}{\alpha+\beta} = O(1/n),$$

the prefactor remains $1 + O(1/n)$. The bound on the theta correction is unchanged, as it uses only absolute values.

Let $A = \frac{\pi^2(m-1)N}{2(2n-3m-1)}$. Using the triangle inequality and $|\cos(\cdot)| \le 1$, we can bound the error term:

$$|\Theta_{n,m} - 1| \le 2\sum_{t\ge 1}\exp(-At^2)$$

Since $t^2 \geq t$ for $t \geq 1$, we can further bound this by a geometric series:

$$|\Theta_{n,m} - 1| \leq 2 \sum_{t \geq 1} \exp(-At) = 2 \frac{\exp(-A)}{1 - \exp(-A)}.$$

Since $m \leq n/3$, we have $N \asymp n$, and

$$A = \frac{\pi^2 (m-1) N}{2(2n - 3m - 1)} = \Theta(m).$$

Hence $|\Theta_{n,m} - 1| \leq Ce^{-cm}$. Consequently, $|A_{n,m}\Theta_{n,m} - 1| \leq C(n^{-1} + e^{-cm})$, and therefore

$$\begin{aligned} |T_4| &\leq \frac{C}{\sqrt{(m-1)(2n-3m)}} \left(n^{-1} + e^{-cm}\right) \\ &\leq C \left(\frac{1}{\sqrt{m}\, n^{3/2}} + \frac{1}{\sqrt{n}\, m^{3/2}} \right), \end{aligned}$$

where the last inequality uses $2n - 3m \asymp n$ and $e^{-cm} \leq C/m$.

5) Conclusion. Adding the four bounds gives $|T_1|+|T_2|+|T_3|+|T_4| = O(\frac{1}{\sqrt{m} n^{3/2}}) + O(\frac{1}{\sqrt{n} m^{3/2}})$, uniformly for $m = \omega_n(1)$ and $m \leq n/3$. The first error term is asymptotically dominated by the second error term and hence absorbed. The result is obtained by recalling $\mathrm{Cov}(n, m) = E(n, m)/4$. ■

Theorem 23 (Monotonicity and minimizer) *Let*

$$\mathcal{M}_n := \{m \in \mathbb{N} : m \mid n,\ 1 \leq m \leq n/2\}.$$

For all sufficiently large n, the covariance is strictly decreasing over admissible fold sizes up to $n/3$: if $m_1, m_2 \in \mathcal{M}_n$ and

$$m_1 < m_2 \leq n/3,$$

then

$$\mathrm{Cov}(n, m_1) > \mathrm{Cov}(n, m_2).$$

Moreover, along subsequences with $6 \mid n$,

$$\mathrm{Cov}(n, n/3) < \mathrm{Cov}(n, n/2).$$

Consequently, along subsequences with $3 \mid n$, the fold covariance is minimized at $k = 3$, among all admissible equal-fold choices.

Proof We write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for a universal constant C. We first prove the monotonicity for admissible $m_1 < m_2 \leq n/3$.

Suppose first that $m_2 = o(n)$. We use the combinatorial estimate from Theorem 20, which is precise in the sublinear regime. Define

$$S_{m-1} := 2^{-2m+2}\binom{2m-2}{m-1}, \qquad \Phi(n,m) := S_{m-1}\frac{1}{2\sqrt{\pi(2n-3m)}}.$$

Then, uniformly for $1 \le m \le n/3$,

$$\operatorname{Cov}(n,m) = \Phi(n,m) + O\Big(n^{-3/2} + m^{3/2}n^{-5/2}\Big). \tag{10}$$

Let

$$d := m_2 - m_1 > 0.$$

Since $m_1, m_2 \mid n$, we have

$$\operatorname{lcm}(m_1, m_2) \mid n.$$

Moreover, $\gcd(m_1, m_2) \mid d$, hence $\gcd(m_1, m_2) \le d$. Therefore

$$n \ge \operatorname{lcm}(m_1, m_2) = \frac{m_1 m_2}{\gcd(m_1, m_2)} \ge \frac{m_1 m_2}{d},$$

and so $d \ge \frac{m_1 m_2}{n} \ge \frac{m_1^2}{n}$. Of course also $d \ge 1$. Consequently,

$$\frac{d}{m_1} \ge \max\left\{\frac{1}{m_1}, \frac{m_1}{n}\right\}. \tag{11}$$

We now compare the leading terms $\Phi(n, m_1)$ and $\Phi(n, m_2)$. Using the exact ratio

$$\frac{S_t}{S_{t-1}} = \frac{2t-1}{2t} = 1 - \frac{1}{2t},$$

we obtain

$$\frac{S_{m_2-1}}{S_{m_1-1}} = \prod_{r=m_1}^{m_2-1}\left(1 - \frac{1}{2r}\right) \le \exp\left(-\frac{1}{2}\sum_{r=m_1}^{m_2-1}\frac{1}{r}\right).$$

Hence, for $m_2 = o(n)$,

$$\begin{aligned}
\log\frac{\Phi(n,m_2)}{\Phi(n,m_1)} &= \log\frac{S_{m_2-1}}{S_{m_1-1}} + \frac{1}{2}\log\frac{2n-3m_1}{2n-3m_2} \\
&\le -c_1 \min\left\{1, \frac{d}{m_1}\right\} + C_1\frac{d}{n} \\
&\le -c_2 \min\left\{1, \frac{d}{m_1}\right\},
\end{aligned}$$

for all sufficiently large n, since $m_2 = o(n)$. Therefore

$$\Phi(n, m_1) - \Phi(n, m_2) \geq c_3\, \Phi(n, m_1) \min\left\{1, \frac{d}{m_1}\right\}. \tag{12}$$

It remains to check that the error in (10) is negligible relative to the right-hand side of (12). Since

$$\Phi(n, m_1) \asymp \frac{1}{\sqrt{m_1 n}},$$

and since $\frac{d}{m_1} \geq \max\left\{\frac{1}{m_1}, \frac{m_1}{n}\right\}$ due to (11), we have, in the case $d \leq m_1$,

$$\frac{n^{-3/2}}{\Phi(n, m_1)(d/m_1)} \lesssim \frac{\sqrt{m_1}/n}{\max\{1/m_1, m_1/n\}} = o(1).$$

Also $d \leq m_1$ implies $m_2 \leq 2m_1$, and therefore

$$\frac{m_2^{3/2} n^{-5/2}}{\Phi(n, m_1)(d/m_1)} \lesssim \frac{m_1^2/n^2}{\max\{1/m_1, m_1/n\}} = o(1).$$

If $d > m_1$, then $\min\{1, d/m_1\} = 1$, and

$$\frac{n^{-3/2}}{\Phi(n, m_1)} \lesssim \frac{\sqrt{m_1}}{n} = o(1),$$

while, using $m_1 \leq m_2 = o(n)$,

$$\frac{m_2^{3/2} n^{-5/2}}{\Phi(n, m_1)} \lesssim \frac{m_2^{3/2}\sqrt{m_1}}{n^2} \leq \left(\frac{m_2}{n}\right)^2 = o(1).$$

Combining these bounds with (10) and (12) gives

$$\mathrm{Cov}(n, m_1) > \mathrm{Cov}(n, m_2)$$

throughout the sublinear regime $m_2 = o(n)$.

Next suppose $m_1 = o(n)$ but m_2 is not $o(n)$. Then the preceding estimates give

$$\mathrm{Cov}(n, m_1) \asymp \frac{1}{\sqrt{m_1 n}},$$

whereas Theorem 22 gives $\mathrm{Cov}(n, m_2) = O(n^{-1})$. Since $m_1 = o(n)$, we have $1/\sqrt{m_1 n} \gg 1/n$, and the desired inequality follows.

It remains only to consider the regime in which m_1 is of order n. Write

$$m_i = \frac{n}{k_i}, \qquad i = 1, 2.$$

Since $m_1 < m_2 \le n/3$, we have integers $k_1 > k_2 \ge 3$. Along any such linear-regime subsequence, the possible k_i's are bounded, so after passing to a subsequence we may regard k_1, k_2 as fixed. Theorem 22 yields

$$\mathrm{Cov}\left(n, \frac{n}{k}\right) = \frac{k}{2\pi n\sqrt{2k-3}} + O(n^{-2})$$

for every fixed integer $k \ge 3$. The function

$$h(k) := \frac{k}{\sqrt{2k-3}}$$

is strictly increasing on the integers $k \ge 3$: indeed $h(4) > h(3)$, and for real $k > 3$,

$$h'(k) = \frac{k-3}{(2k-3)^{3/2}} > 0.$$

Thus $k_1 > k_2$ implies

$$\mathrm{Cov}\left(n, \frac{n}{k_1}\right) > \mathrm{Cov}\left(n, \frac{n}{k_2}\right)$$

for all sufficiently large n. This proves the claimed monotonicity over admissible $m \le n/3$.

Finally assume $6 \mid n$. Then Theorem 22 at $m = n/3$ and Theorem 21(B) at $m = n/2$ give

$$\mathrm{Cov}(n, n/3) = \frac{\sqrt{3}}{2\pi n} + O(n^{-2}),$$

and

$$\mathrm{Cov}(n, n/2) = \frac{1}{\pi n} + O(n^{-2}).$$

Since $\sqrt{3}/2 < 1$, it follows that

$$\mathrm{Cov}(n, n/3) < \mathrm{Cov}(n, n/2)$$

for all sufficiently large n with $6 \mid n$.

If $3 \mid n$, then $m = n/3$ is admissible. The preceding monotonicity shows that it minimizes the covariance among all admissible $m \le n/3$. The only admissible fold size in $(n/3, n/2]$, if it exists, is $m = n/2$, corresponding to $k = 2$; when $6 \mid n$ the comparison above rules it out, and when n is odd it is not admissible. Hence $m = n/3$, equivalently $k = 3$, is the admissible covariance minimizer along $3 \mid n$. ■

Appendix B. Proof of Lemma 3

Proof Consider the following setup. Let $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, with input distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where the marginal over $\mathcal{X}$ is uniform and $y = f(x)$ where $f = \mathbf{1}_{\{x>1/2\}}$. Consider the algorithm that outputs $\mathcal{A}(S^n) = \mathbf{1}_{\{1/2-p/2< \cdot <1-p/2\}}$ with $p = p(S) = \sum_i y_i/n$, and $\mathcal{A}(S^{n-m}) = h_0$, where h_0 is the constant zero hypothesis. Then, $\widehat{L}_{\mathrm{CV}}^{(k)} = \sum_i \widehat{L}_i^{(k)}/k = \sum_i y_i/n = p(S)$ and $L(A(S^n)) = p(S)$, so that the MSE is zero. For the loss-stability term, note that $L(A(S^{n-m})) = 1/2$. Hence, for $n = 2$ and $m = 1$,

$$\mathbb{E}\left[|L(A(S^n)) - L(A(S^{n-m}))|\right] = \mathbb{E}_{Y\sim\mathrm{Bin}(2,1/2)}\left|\frac{Y}{2} - \frac{1}{2}\right| = \frac{1}{4}.$$

■

Appendix C. Error in Theorem 5.3 of Kearns and Ron (1997)

Let us first recall their notion of stability in our notation. We say that a deterministic algorithm $\mathcal{A}$ has error stability (β_1, β_2) if $\mathbb{P}_{S^{n-1},(x,y)}[|L(\mathcal{A}(S^n)) - L(\mathcal{A}(S^{n-1}))| \geq \beta_2] \leq \beta_1$ where $S^n = S^{n-1} \circ (x,y)$, and both β_1 and β_2 may be functions of n.

Consider the proof of their Theorem 5.3. There, they define the random variable $\chi(S^n) = \widehat{L}_{\mathrm{CV}}^{(n)} - L(\mathcal{A}(S^n))$ and assume without loss of generality that with probability at least $\beta_1/2$, $L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) \geq \beta_2$.

Next, their Lemma 4.1 correctly asserts that the expected cross-validation estimate equals the expected estimate of a single hold-out set, i.e., $\mathbb{E}_{S^n}[\chi(S^n)] = \mathbb{E}_{S^n}[L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n))]$. Combined with the fact that with probability at least $\beta_1/2$, $L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) \geq \beta_2$, the authors then claim that $\mathbf{E}_{S^n}[\chi(S^n)] \geq \frac{\beta_1}{2} \cdot \beta_2$.

This is incorrect, since $L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) \geq \beta_2$ on part of the sample space does not rule out that this quantity can also be negative at other times. To illustrate this, let us consider an extreme case where $\beta_1 = \beta_2 = 1$ by assuming that $\mathbb{P}(\{L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) = 1\}) = \beta_1/2 = 1/2$. This assumption does not rule out the possibility that $\mathbb{P}(\{L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) = -1\}) = 1/2$. In that case, $\mathbb{E}_{S^n}[\chi(S^n)] = 0$, violating the alleged lower bound $\frac{\beta_1}{2} \cdot \beta_2 = 1/2$.

This directly contradicts our Lemma 3 because non-zero loss stability according to our Definition 2 implies a lower bound on their high-probability error stability parameters, yet we prove in Lemma 3 that one can have non-zero loss stability and simultaneously zero MSE (which necessitates $\mathbf{E}_{S^n}[\chi(S^n)] = 0$).

Appendix D. Error in Theorem 2 of Kale et al. (2011)

The key ingredient for deriving the main result of Kale et al. (2011, Theorem 2) is to obtain an upper bound on $\mathrm{Cov}_{S^n}(\widehat{L}_1^{(k)} - L_1^{(k)}, \widehat{L}_2^{(k)} - L_2^{(k)})$ (in their notation $\mathrm{Cov}_U(\mathrm{gen}_1, \mathrm{gen}_2)$) that scales linearly with a parameter measuring a certain notion of algorithmic stability (*"mean square stability"*). To do so, the supposed identity $\mathbb{E}_{S_2}[\widehat{L}_1^{(k)} - L_1^{(k)} \mid S_1, S_3, \ldots, S_k] = 0$ (in their notation $\mathbb{E}_{T'}[\mathrm{gen}_1 \mid S, T] = 0$) is used twice. Define $S' := S^n \setminus S_2$ and $S'' := S^n \setminus (S_1 \circ S_2)$.

One can see that

$$\begin{aligned}
\mathbb{E}_{S_2}[\widehat{L}_1^{(k)} - L_1^{(k)} \mid S_1, S_3, \ldots, S_k] &= \mathbb{E}_{S_2}\left[\frac{1}{m}\sum_{z' \in S_1} \ell(\mathcal{A}(S_{-1}^n), z') - \mathbb{E}_z[\ell(\mathcal{A}(S_{-1}^n), z)] \,\middle|\, S'\right] \\
&= \mathbb{E}_{S_2}\left[\frac{1}{m}\sum_{z' \in S_1} \ell(\mathcal{A}(S_{-1}^n), z') \,\middle|\, S'\right] - \mathbb{E}_{S_2,z}[\ell(\mathcal{A}(S_{-1}^n), z) \mid S'']
\end{aligned}$$

where the two terms in the last line are functions of S' and S'', respectively, and their difference is non-zero in general.