

THE STRUCTURE OF CROSS-VALIDATION ERROR: STABILITY, COVARIANCE, AND MINIMAX LIMITS

IDO NACHUM^{*†}, RÜDIGER URBANKE^{*‡}, AND THOMAS WEINBERGER^{*‡}

ABSTRACT. Despite ongoing theoretical research on cross-validation (CV), many theoretical questions about CV remain widely open. This motivates our investigation into how properties of algorithm-distribution pairs can affect the choice for the number of folds in k -fold cross-validation.

Our results consist of a novel decomposition of the mean-squared error of cross-validation for risk estimation, which explicitly captures the correlations of error estimates across overlapping folds and includes a novel algorithmic stability notion, squared loss stability, that is considerably weaker than the typically required hypothesis stability in other comparable works.

Furthermore, we prove:

1. For every learning algorithm that minimizes empirical error, a minimax lower bound on the mean-squared error of k -fold CV estimating the population risk $L_{\mathcal{D}}$:

$$\min_{k|n} \max_{\mathcal{D}} \mathbb{E} \left[\left(\widehat{L}_{\text{CV}}^{(k)} - L_{\mathcal{D}} \right)^2 \right] = \Omega(\sqrt{k}/n),$$

where n is the sample size and k the number of folds. This shows that even under idealized conditions, for large values of k , CV cannot attain the optimum of order $1/n$ achievable by a validation set of size n , reflecting an inherent penalty caused by dependence between folds.

2. Complementing this, we exhibit learning rules for which

$$\max_{\mathcal{D}} \mathbb{E} \left[\left(\widehat{L}_{\text{CV}}^{(k)} - L_{\mathcal{D}} \right)^2 \right] = \Omega(k/n),$$

matching (up to constants) the accuracy of a hold-out estimator of a single fold of size n/k .

Together these results delineate the fundamental trade-off in resampling-based risk estimation: CV cannot fully exploit all n samples for unbiased risk evaluation, and its minimax performance is pinned between the k/n and $\sqrt{k}/n$ regimes.

1. INTRODUCTION

k -fold cross-validation (CV) is a popular model validation technique used in many settings in statistics, data science and machine learning, see [AC10] for a comprehensive introduction. Given the errors of the models obtained by training on subsets of the full sample and then validating on the remaining samples, the goal is usually of the following two: (Risk estimation) given a model chosen independently of the error computations, estimate its risk by averaging the errors; (Model selection) given the error estimates computed with CV, select the candidate model that looks best. In this work we mathematically analyze the accuracy of risk estimation under various statistical and algorithmic aspects by harnessing a novel error decomposition.

Although cross-validation is a workhorse for statistical validation in the empirical sciences, its theoretical properties remain surprisingly poorly understood. For instance, there is still no principled way to choose k , the number of folds. As noted by [AC10, Sec. 10.3]:

E-mail addresses: inachum@univ.haifa.ac.il, rudiger.urbanke@epfl.ch, thomas.weinberger@epfl.ch.

Key words and phrases. cross-validation, learning theory, algorithmic stability.

^{*} Alphabetical order.

[†] Department of Statistics, University of Haifa, Haifa, Israel.

[‡] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland.

“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.”

While well-known theoretical investigations such as [RW78; DW79b; BKL99; KR97; BE02; BGo4] are insightful and contributed significant progress towards our understanding of CV, they tend to suffer from at least one of the following limitations¹: (1) sufficient conditions for CV performance that can be arbitrarily loose (2) guarantees not in absolute terms but merely in relation to other error measures (e.g. empirical error or hold-out) (3) guarantees that are restricted to leave-one-out cross validation (4) results about certain statistical aspects of CV with no clear connection to the performance of CV.

The widespread confusion surrounding theoretical aspects of CV within the broader scientific community is perhaps best exemplified by recent informal online discussions [617; 280]. These discussions reveal the presence of numerous conflicting interpretations concerning the role of specific variance and covariance terms that emerge in the context of CV. Our results provide clear, theoretically grounded insights into these quantities, offering direction for selecting the number of folds based on properties of the algorithm and distribution at hand.

1.1. 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 $\mathcal{A} : \mathcal{Z}^\star \rightarrow \mathcal{H}$ that map a sample $S^n = (Z_1, \dots, Z_n) \in \mathcal{Z}^n$ 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 $S^n \sim \mathcal{D}^n$. As common in previous works, we assume throughout that $\mathcal{A}$ is permutation-invariant (symmetric): for any permutation π of $\{1, \dots, n\}$,

$$\mathcal{A}(S^n) = \mathcal{A}(S_\pi^n) \quad \text{a.s.}, \quad S_\pi^n := (Z_{\pi(1)}, \dots, Z_{\pi(n)}).$$

Fix $k \in \mathbb{N}$ with $k \mid n$. Partition the index set $\{1, \dots, n\}$ into k disjoint blocks $I_1, \dots, I_k$ of size $m := |I_i| = n/k$, and define $S_i = \{Z_j : j \in I_i\}$ and $S_{-i} = S^n \setminus S_i$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, the k -fold cross-validation estimator is

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

That is, $\widehat{L}_i^{(k)}$ is the average loss on the i th hold-out fold, and $\widehat{L}_{\text{CV}}^{(k)}$ averages these across folds.² We omit the subscript CV whenever it is clear from context.

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

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

where the population risk is

$$L(h) := \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(h(x), y)].$$

For the i th fold, we also write

$$L_i^{(k)}(S^n) := L(\mathcal{A}(S_{-i})),$$

i.e., the risk of the hypothesis trained on the complement S_{-i} of the i th hold-out block.

Finally, we denote the risks averaged over a sample of arbitrary size m (and the algorithm’s internal randomness) by

$$\bar{L}_m := \mathbb{E}_{S^m \sim \mathcal{D}^m, \mathcal{A}} [L(\mathcal{A}(S^m))],$$

when no ambiguity arises, we suppress explicit dependence on $\mathcal{A}$, $\mathcal{D}$, and the sample.

1.2. Background and Motivation. The main motivation for using k -fold cross-validation for risk estimation is that partitioning the data into non-overlapping subsets typically reduces statistical

¹We will address these limitations in more detail in our related works section.

²When $k = n$, the definition coincides with leave-one-out cross validation.

variability compared to relying on a single hold-out set [BKL99]. Moreover, compared to the empirical error, CV generally avoids overly optimistic error estimates caused by overfitting, which is a phenomenon prevalent when deploying overparametrized models such as neural networks.

While there are many schemes for partitioning the folds in CV, based for example on combinatorial partitioning [Sha93] or Monte-Carlo resampling [PC84], we focus on the variant where one partitions the sample into a non-overlapping partition of equi-sized folds. This includes for example the widely employed variants of 5- and 10-fold CV. In practice, guidelines for choosing the number of folds k are usually of heuristic nature. Typical lines of reasoning emphasize the importance of the following terms.

1. *Per-fold variance*: since each fold computes its own empirical estimate across $m := n/k$ i.i.d. test points independent of its training set, the variance per fold decreases as $1/m$.
2. *Inter-fold covariance*: a large number of small folds should intuitively lead to a higher correlation between the individual fold estimates. This is because for each fixed sample S^n , decreasing k means that the per-fold output hypotheses share a smaller fraction of the training set which should typically de-correlate the error estimates.
3. *Stability of the algorithm*: if $\mathcal{A}(S^n)$ typically behaves vastly differently than $\mathcal{A}(S^{n-m})$, the per-fold estimates can admit large (typically positive) biases w.r.t. L , causing a large bias of the aggregated estimate $\hat{L}_k$.

In other words, it is argued that choosing k is a balancing act that consists of regulating the overall variance $\text{Var}(\hat{L}_{\text{CV}}^{(k)}) = (1/k^2) [\sum_i \text{Var}(\hat{L}_i^{(k)}) + \sum_{i \neq j} \text{Cov}(\hat{L}_i^{(k)}, \hat{L}_j^{(k)})]$ (items 1 and 2) all while simultaneously not sacrificing too much stability (item 3).

Although the above reasoning is intuitively appealing, to the best of our knowledge there is no corresponding rigorous treatment in the literature. This motivates the following question.

Question 1. Which formal notion of algorithmic stability best captures the performance of cross-validation, and how does it quantitatively influence $\text{MSE}_{\text{CV}}^{(k)}$ relative to the overall variance across folds?

Moreover, as a minimal requirement, one may ask that CV perform at least as well as the empirical training error. This question, formalized in terms of so-called *sanity-check bounds*, was studied by [KR97], who proved that for loss-stable empirical risk minimizers 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 [AH98].

Another natural sanity-check is to require that CV perform no worse than a hold-out estimate over a single fold of size n/k . The work of [BKL99] confirms this property for a specific (non-standard) cross-validation setting, where the algorithm's final output on the full sample S^n is defined as the average of the hypotheses trained on each fold.

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

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

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

For illustration of the minimax CV risk, consider binary classification with the 0–1 loss function and a constant algorithm $\mathcal{A}_h$ that always outputs the same hypothesis h , independent of the input sample. If the population risk of h is p , then $\text{MSE}_{\text{CV}}^{(k)} = \frac{p(1-p)}{n}$.³ Therefore, $\mathfrak{R}_{\text{CV}}(\mathcal{A}_h) = 1/(4n)$.

³This follows because $L(\mathcal{A}_h, S^n) = p$ and $\hat{L}_{\text{CV}}^{(k)} \sim \text{Bin}(n, p)/n$, hence $\text{MSE}_{\text{CV}}^{(k)} = \text{Var}(\text{Bin}(n, p))/n^2$.

This simple example provides a natural baseline or reference point against which 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 instead on empirical risk minimization (ERM) algorithms, a standard assumption in statistical learning theory. This leads to the following question:

Question 2. *Can any ERM achieve an $O(1/n)$ minimax rate, and if not, how close can it get?*

We may also inquire about the opposite extreme: how far can an ERM deviate from the optimal minimax rate? Although previous work has shown that cross-validation outperforms a single hold-out estimator whenever $2 < k < n$ [BKL99], the magnitude of this improvement remains unquantified. Moreover, in the limiting cases $k = 2$ [BKL99] and $k = n$ [KR97], certain algorithms yield cross-validation estimates that coincide exactly with the corresponding hold-out estimates. This leads to the following open question:

Question 3. *For intermediate values $2 < k < n$, do there exist algorithms for which cross-validation performs no better than a hold-out estimator, up to a constant factor?*

1.3. Our Contributions. We now provide a high-level overview of our answers to Questions 1–3.

QUESTION 1. In Section 4.1, we derive a novel decomposition of the MSE of cross-validation in Lemma 4.1. This decomposition reveals two principal components: (i) a new notion of algorithmic stability, which we term *Squared Loss Stability* (SLS), and (ii) the covariance between loss estimates across different folds.

The decomposition also includes two additional correction terms, which under the practical assumption of low loss variance, are dominated by the squared loss stability and fold-covariance terms.

Overall, this result clarifies which formal quantities theoreticians and practitioners must consider when analyzing or applying cross-validation. Importantly, it shows that there is no universally optimal choice of k . We demonstrate this by analyzing two extreme cases that emphasize opposite regimes: On the one hand, a *linear function learner* exhibits poor squared loss stability, making stability the dominant term; in this case, using as many folds as possible (e.g., leave-one-out) is advantageous. On the other hand, the *majority algorithm* exhibits high squared-loss stability, making fold covariance the dominant term; thus, using fewer folds is preferable.

QUESTION 2. In Section 4.2, we show that ERM algorithms cannot achieve a minimax rate of $O(1/n)$. Specifically, every ERM algorithm $\mathcal{A}$ admits a minimax rate for the MSE that scales as

$$\mathfrak{R}_{\text{CV}}(\mathcal{A}) = \Omega\left(\frac{\sqrt{k}}{n}\right).$$

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

QUESTION 3. Although [BKL99] established that cross-validation outperforms a single hold-out estimator for all $2 < k < n$ (without quantifying the gap), we show in Section 4.3 that there are learning algorithms for which cross-validation achieves the same asymptotic rate as a single hold-out set. That is, for every k , there exists an algorithm $\mathcal{A}_k$ such that

$$\max_{\mathcal{D}} \text{MSE}_{\text{CV}}^{(k)}(\mathcal{A}_k, \mathcal{D}) = \Omega\left(\frac{k}{n}\right).$$

Thus, although cross-validation can outperform hold-out estimation in general, for certain algorithms this advantage is limited to at most a universal constant factor.

ON THE SIGNIFICANCE OF OUR STUDY OF THE MAJORITY ALGORITHM. While an algorithm that outputs a constant hypothesis is uninteresting from a theoretical standpoint (the MSE simply scales in accordance to a simple concentration of measure argument), the majority algorithm features a very rich behavior (as evidenced by the non-trivial proof found in Appendix B). This is despite majority being arguably the next simplest algorithm one could conceive: it can merely output two different hypotheses; and its decision rule is solely based on counting the occurrence of labels, while entirely disregarding the input features. Majority serves as a critical test case where our MSE bound demonstrably supersedes those of the foundational works [KR97; BKL99; KKV11; Kum+13], which highlights the importance of keeping the fold-covariance term intact (or carefully bounding it) when analyzing the MSE of CV.

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

To illustrate non-tightness of previous analyses, we can instantiate [Kum+13, Theorem 1] for the majority algorithm. Then, the contribution of the fold variance term is dominated by that of the loss stability parameter which is of order $1/\sqrt{n}$, yielding a variance upper bound of order $1/\sqrt{n}$ for any choice of k , the number of folds. By contrast, our analysis (Theorem 4.11) shows that choosing three folds achieves an MSE of order $1/n$ —the theoretical optimum.

2. 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 the two perhaps 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 $m > 1$ [Gas+24b].

Definition 2.1 (Hypothesis Stability). We call a pair $(\mathcal{A}, \mathcal{D})$ *hypothesis stable* with parameters (β_1, m) if

$$\mathbb{E}_{S^{n-m} \sim \mathcal{D}^{n-m}, S^m \sim \mathcal{D}^m, (x, y) \sim \mathcal{D}, \mathcal{A}} [\mathbf{1}_{\{\mathcal{A}(S^{n-m} \cup S^m)(x) \neq \mathcal{A}(S^{n-m})(x)\}}] < \beta_1.$$

Intuitively, hypothesis stability is a stronger assumption than necessary. 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 hypothesis-stable algorithm behaves almost like a constant algorithm—whose outputs, and hence fold predictions, are identical by definition. However, the key factor governing the accuracy of 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 are slightly perturbed.

Definition 2.2 (Loss Stability). We call a pair $(\mathcal{A}, \mathcal{D})$ *loss stable* with parameters (β_2, k) if

$$\mathbb{E}_{S^{n-k} \sim \mathcal{D}^{n-k}, S^k \sim \mathcal{D}^k, \mathcal{A}} [|L(\mathcal{A}(S^{n-k} \cup S^k)) - L(\mathcal{A}(S^{n-k}))|] < \beta_2.$$

One might think that loss stability on the other hand is necessary since each single validation in isolation is an unbiased estimator of the loss over $n - m$ samples, meaning that an aggregation of such estimates can only accurately predict $L(\mathcal{A}(S^n))$ if it is generally not too far from $L(\mathcal{A}(S^{n-m}))$.

This analogy is however not entirely rigorous since the individual estimates are correlated, and as we will see in Lemma 4.4, there exist pathological algorithms that do not admit low loss stability but whose loss can be accurately estimated with CV. This directly contradicts [KR97, Thm. 5.3], but this is because their result is erroneous (see Appendix E for clarification). In general, it is unclear in which cases loss stability is necessary for low MSE. Our form of loss stability below is provably necessary under low loss-variance, as we will show in Corollary 4.6.

Definition 2.3 (Squared Loss Stability). We call a pair $(\mathcal{A}, \mathcal{D})$ *squared loss stable* (SLS) with parameters (β, k) if

$$\mathbb{E}_{S^n \sim \mathcal{D}^n} [(\tilde{L}^{(k)}(\mathcal{A}, S^n) - L(\mathcal{A}(S^n)))^2] < \beta$$

where $\tilde{L}^{(k)}(\mathcal{A}, S^n) := \frac{1}{k} \sum_{i=1}^k L_i^{(k)}(\mathcal{A}, S^n)$ is the symmetrized leave- m loss.

Once can easily show that both β_2^2 and β are upper bounded by the same quantity $\mathbb{E}[(L_{n-k} - L_n)^2]$. For a further discussion of the role of squared loss stability, see the section after Lemma 4.1.

The following Lemma states that control over the first two moments of the risk allows us to bound the squared-loss stability

Lemma 2.4 (Bounds on the Squared Loss Stability). *Assume that the loss functional is bounded between 0 and 1 and that the risk has means $\mathbb{E}[L] = \bar{L}$ and $\mathbb{E}[L^{(k)}] = \bar{L}^{(k)}$ and denote the variances loss as $\sigma_n^2 := \text{Var}(L)$ and $\sigma_{n-m}^2 := \text{Var}(L_1^{(k)})$. Then, the squared loss stability $\mathbb{E}[(\tilde{L} - L)^2]$ can be bounded as*

$$(\bar{L}^{(k)} - \bar{L})^2 \leq \mathbb{E}[(\tilde{L}^{(k)} - L)^2] \leq (\sigma_{n-m} + \sigma)^2 + (\bar{L}^{(k)} - \bar{L})^2$$

This result will become useful later for controlling the stability of linear functions.

3. PRIOR WORK

The works [RW78] and [DW79b; DW79a] have been among the first to establish rigorous stability-based performance guarantees for classification problems using leave-one-out CV. Though in their works, they assume that the considered algorithms be 'local' (e.g. nearest neighbors) and the data distribution be 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 [BE02] 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. The principal aim of this field is the development of generalization bounds, typically in the form of high-probability upper bounds $L(\mathcal{A}(S)) < \hat{L}_{\text{emp}}(\mathcal{A}(S), S) + C$, where $\hat{L}_{\text{emp}}(\mathcal{A}(S), S)$ 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 [VC71; Blu+89] states it is sufficient and necessary 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 [Jia+19; Dzi+20] and theoretically [Gas+24a; Gas+24b].

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 algorithms stability, and CV can be computationally expensive).

In the 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 [KR97]. 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. [AH98] derived similar results for the more general case of k -fold CV.

Yet another valid *sanity-check* might be to require CV to do no worse than a single hold out set of corresponding size. The work of [BKL99] shows that this does indeed hold for a specific (non-standard) cross-validation setting.

Another influential line of work [BGo4; NB99] considers the limitations of unbiased estimation of the variance of CV.

A more recent line of work is [KKV11; Kum+13]. Therein, the authors devise upper bounds on the MSE based novel notions of loss stability. Unfortunately, the main Theorem in [KKV11] is erroneous (see Appendix F), which makes it difficult to assess the implications of these results. The follow-up work introduces a version of loss stability that leads to a stronger result [Kum+13, Theorem 1] since the related stability parameter is a lower bound on the one appearing in [KKV11, 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. Lastly, another key difference is that our Theorem 4.3 presents a characterization (i.e. two-sided bound) of the MSE of CV, not just an upper bound, and the gap between our lower and upper bound can approach zero (under low loss variance).

4. RESULTS

4.1. MSE Decomposition and the Role of Squared Loss Stability. We show that our new notion of SLS is one of the two principal components governing the MSE of CV which answers Question 1. The following lemma formalizes this relationship.

Lemma 4.1 (Decomposition of the MSE). *Denote the expected conditional variance of the risk as $\bar{\sigma}^2 := \mathbb{E} [\text{Var}_{(x,y) \sim \mathcal{D}} (\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n)]$. The MSE of k -fold cross-validation for a symmetric algorithm admits the exact decomposition:*

$$\begin{aligned} \text{MSE}_{\text{CV}}^{(k)} = & \underbrace{\mathbb{E}[(\tilde{L}^{(k)} - L)^2]}_{\text{Squared Loss Stability}} + \underbrace{\frac{k-1}{k} \cdot \text{Cov}(\tilde{L}_1^{(k)}, \tilde{L}_2^{(k)})}_{\text{Inter-fold Covariance}} + \underbrace{\frac{1}{k} \frac{\bar{\sigma}^2}{m}}_{\text{Single-Fold Estimation Variance}} \\ & + \underbrace{2 \text{Cov}(L, L_1^{(k)} - \tilde{L}_1^{(k)}) - \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)})}_{\text{Correction Terms}} \end{aligned}$$

Proof. See Appendix A.1. □

The individual terms in the decomposition have the following interpretations.

- The first term, the *squared loss stability*, is a measure of algorithmic stability that captures how quickly the averaged loss across the folds deteriorates as a function of n and k . Unlike other stability notions, SLS explicitly captures the inter-dependence of the different risks of the k cross-validated hypotheses that comes from the overlap of the training sets. Since we will show that being stable in the SLS sense is necessary (see Corollary 4.6), this suggests that MSE bounds based on stability notions that do not capture the fold-wise structure of CV must necessarily be loose in some settings.

- The second term is a constant $\in [1/2, 1)$ times the inter-fold covariances $\text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)})$. A large fold-covariance means that estimates from different cross-validated hypotheses tend to be small or large simultaneously, which degrades the $1/k$ variance reduction one would otherwise obtain from averaging k independent estimates.
- The per-fold variance term. This term (together with its $1/k$ pre-factor) typically contributes an irreducible $\Theta(1/n)$ error floor in case of a bounded loss function, see the Lemma below.
- The correction terms do not immediately offer a straightforward interpretation, but as we will see, it can easily be upper bounded under an additional assumption.

Lemma 4.2 (Expected Risk Variance for Bounded Loss). *For a loss functional ℓ bounded in $[0, M]$, the expected risk variance term $\bar{\sigma}^2 := \mathbb{E} [\text{Var}_{(x,y) \sim \mathcal{D}}(\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n)]$ is bounded as*

$$\bar{\sigma}^2 \leq \mathbb{E}[L_1^{(k)}(M - L_1^{(k)})] \leq \frac{M^2}{4}.$$

Proof. For a random variable X bounded in $[0, M]$ with mean μ , the variance is bounded by the Bhatia-Davis inequality as $\text{Var}(X) \leq (\mu - 0)(M - \mu) = \mu(M - \mu)$. Setting $X = [\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n]$ such that $\mu = L_1^{(k)}(S_{-1}^n)$ and taking the expectation gives $\mathbb{E}[\text{Var}(\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n)] \leq \mathbb{E}[L_1^{(k)}(M - L_1^{(k)})]$. Further, it is easy to see that this quantity is maximized when $L_1^{(k)}(S_{-1}^n) = M/2$ almost surely. $\square$

Our first result is that under the low variance assumption, we can bound the MSE from both sides.

Theorem 4.3 (Characterization of the MSE). *Denote the expected conditional variance of the risk as $\bar{\sigma}^2 := \mathbb{E} [\text{Var}_{(x,y) \sim \mathcal{D}}(\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n)]$. Assume $(\mathcal{A}, \mathcal{D})$ has risk variance $\sigma_n^2 := \text{Var}(L(\mathcal{A}, S^n))$. Then, the MSE of k -fold cross-validation (where $m = n/k$ is the fold size) is characterized by:*

$$\text{MSE}_{CV}^{(k)} = \mathbb{E}[(\tilde{L}^{(k)} - L)^2] + \frac{k-1}{k} \text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) + \frac{\bar{\sigma}^2}{n} + \mathcal{C}$$

where the magnitude of the correction term $\mathcal{C}$ is bounded by:

$$|\mathcal{C}| \leq \frac{k-1}{k} \sigma_{n-m}^2 + 2\sqrt{\frac{\sigma_n^2 \bar{\sigma}^2}{m}}$$

Proof. See Appendix A.2. $\square$

In order for the above characterization to become meaningful, $\mathcal{C}$ must be negligible compared to the largest of the main term which, necessitates small enough loss variances $\sigma_{n-m}^2, \sigma_n^2$. When comparing different algorithms or hyperparameter settings, we rely on their estimated performance (be it on a validation set, CV, or some other estimate) to make decisions. If these performance estimates have high variance due to the variability introduced by the training sample, it becomes difficult to confidently assert that one algorithm is truly better than another. Low variance of the population loss makes these comparisons more robust and increases our confidence in the selected model's expected performance in real-world scenarios. As part of the standard model selection process, practitioners typically evaluate multiple algorithms, during which those exhibiting high variability in validation error across samples are naturally excluded. For this reason, our low-variance assumption should be considered rather benign. In other words, for large enough samples size and low enough loss variance, the fold covariance and the loss stability essentially fully characterize the mean-squared error.

In general, loss stability is not necessary for ensuring small MSE of CV. This follows from the existence of pathological algorithm-distribution combinations that are quite squared loss stable but have arbitrarily small MSE.

Lemma 4.4. *There exists a algorithm-distribution combinations with squared loss stability $1/8$ and MSE 0.*

Proof. Consider the following setup. Let $\mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, with input distribution $\mathcal{D}_{\mathcal{X}} = \mathcal{U}(\mathcal{X})$, and conditional output distribution $f = \mathbf{1}_{\{x > 1/2\}}$ such that $\mathcal{D} = (\text{Id}, f) \circ \mathcal{D}_{\mathcal{X}}$. Consider the algorithm $\mathcal{A}(S^n) = \mathbf{1}_{\{1/2 - p/2 < \cdot < 1 - p/2\}}$ where $p = p(S) = \sum_i y_i/n$, $\mathcal{A}(S^{n-k}) = h_0$ and where h_0 is the constant zero hypothesis. Then, $\hat{L}^m = \sum_i \hat{L}_i^m/k = \sum_i y_i/n = p(S)$ and $L = p(S)$ so that the MSE is zero. Simultaneously, the squared loss stability is $\mathbb{E}[(L^m - L)^2] = \mathbb{E}_{p \sim \text{Bin}(k, 1/2)}[(1/2 - p)^2] = 1/(4n)$ which can be as large as $1/8$ for $n = 2, m = 1$. $\square$

Beyond the squared-loss stability, we can also generically bound the inter-fold covariance. Notably, the covariance can never be strongly negative and thereby balance out the influence of low loss stability. This lower bound follows from a geometric argument about the minimal pair-wise inner product of k vectors in euclidean space.

Lemma 4.5. *For every algorithm-distribution pair, the covariance between the folds is bounded as follows.*

$$-1/(2n) \leq -1/[4(n - m)] \leq \text{Cov}(\hat{L}_1^k, \hat{L}_2^k) \leq \sigma_{n-m}^2 + 1/(4m).$$

Proof. See Appendix A.3. $\square$

Seeing how the covariance can never be strongly negative, it becomes clear that under the assumption of low loss variances, high squared loss stability is also a necessary condition for small MSE. This thereby rules out examples such as the one in Lemma 4.4.

Corollary 4.6. *It holds that*

$$\text{MSE}_{\text{CV}}^{(k)} \geq \mathbb{E}[(\tilde{L}^k - L)^2] - 1/(2n) - \sigma_{n-m}^2 - 2\sqrt{\frac{\sigma_n^2 \bar{\sigma}^2}{m}}$$

Proof. We simply combine the left-hand sides of Theorem 4.3 and Lemma 4.5. $\square$

To illustrate our decomposition and the role of SLS, we consider two algorithms. In the first, the SLS term dominates the MSE; in the second, the inter-fold covariance is the primary contributor. This demonstrates that there is no universally optimal choice of k in cross-validation: in the first case, performance improves with larger k , while in the second, smaller k is preferable.

4.1.1. *Linear Functions.* Let us consider multi-class classification with a randomized algorithm.

To set the stage, let us introduce the class of linear functionals $\mathbf{Lin}_q(d)$ over the vector space $\mathbb{F}_q^d$ where $\mathbb{F}_q$ is the finite field with q elements, with q prime.

$$\mathbf{Lin}_q(d) \equiv (\mathbb{F}_q^d)^* := \left\{ f_a : \mathbb{F}_q^d \rightarrow \mathbb{F}_q : a \in \mathbb{F}_q^d, f_a(x) = \sum_{i=1}^d a_i \cdot x_i \mod q \right\}$$

Note that for example, $\mathbf{Lin}_2(d)$ is the class of all parity functions of dimension d . We will consider throughout this section that the distribution is $\mathcal{D} = \mathcal{U}(\mathbb{F}_q^d)$, the uniform distribution over the space.

An elementary property of this class is that distinct pairs of linear functions agree on exactly a portion $1/q$ of the space. This means that for in-class learning, the risk is polarized between two dissimilar values, making this an interesting case-study for how CV performs under loss instability.

Lemma 4.7. *Each two distinct functions $f, h \in \mathbf{Lin}_q(d)$ agree on a fraction $1/q$ of the space and the 0 – 1 risk of the function h over samples from $\mathcal{D}_f = f \diamond \mathcal{U}(\mathbb{F}_q^d)$ is given by*

$$L_{\mathcal{D}_f}(h) = \begin{cases} 0 & h = f \\ 1 - 1/q & h \neq f \end{cases}$$

where $f \diamond \mathcal{U}(\mathbb{F}_q^d)$ denotes the distribution of the random variable $(X, f(X))$ where $X \sim \mathcal{U}(\mathbb{F}_q^d)$.

We will study the algorithm $\mathcal{A}_{lin} : \{\mathbb{F}_q^d\}^n \rightarrow \mathbf{Lin}_q(d)$ defined as the randomized empirical risk minimizer which outputs one of the sample-consistent linear functions uniformly at random. This algorithm is notably quite hypothesis unstable in the regime $n < d$. In that case, there exist at least q^{n-d} sample-consistent linear functions and $\mathcal{A}_{lin}$ picks one of them uniformly at random. At the same time, $\mathcal{A}_{lin}$ is quite hypothesis (and hence loss-) stable for $n \geq d$ since here $\mathcal{A}_{lin}$ will typically select the ground truth assuming that the number of linearly independent samples exceeds the number of linear constraints. One delicate detail that significantly complicates the analysis is the possibility that samples can be linearly dependent. For this reason, in every step of our analysis we need to condition on the set of samples being of a specific rank. This can be handled with random matrix theory results for finite fields [BS06]. In contrast to the majority algorithm, utilizing Theorem 4.3 now requires controlling the loss variances and the squared loss stability (utilizing Lemma 2.4), which further complicates the analysis.

Theorem 4.8 (MSE Bounds for CV on Linear Functions). *Let k be the number of folds, n be the total number of samples, and m be the size of each fold. Let d be the feature dimension and q be the finite field size.*

The Mean Squared Error (MSE) of k -fold cross-validation for $\mathcal{A}_{lin}$ is bounded as follows:

Case 1: $n < d$

$$MSE_{CV}^{(k)} = O\left(q^{-(d-n)}\right)$$

Case 2: $n \geq d$ and $n - m < d$

$$MSE_{CV}^{(k)} = 1 - O(1/q) = \Omega(1)$$

Case 3: $n - m \geq d$

$$MSE_{CV}^{(k)} = O\left(q^{-(n-m-d+1)}\right)$$

Proof. See Appendix C.1. □

We see that in this setting, in the case $n < d$ (where $\mathcal{A}_{lin}$ does typically not output the ground truth f), it is beneficial to choose m as large as possible since this can only decrease the MSE. In the case $n \geq d$ on the other hand it is beneficial to set $m = 1$ since the bound increases in m .

Lastly, we remark that here any MSE bound based on hypothesis stability must be loose since we are highly hypothesis unstable in the cases 1 and 2 above since $n - m \leq d$ implies the existence of multiple sample-consistent linear functions. Yet, our loss-stability based analysis correctly captures the MSE.

4.1.2. Majority Algorithm. In contrast to the algorithm $\mathcal{A}_{lin}$, which can produce many hypotheses with potentially large variations in their loss values, we now consider the opposite extreme—a setting in which the algorithm can output only two hypotheses with identical loss values.

Let the sample be $S^n = \{z_i\}_{i=1}^n \sim \mathcal{D}^n$ and define $Y := \sum_{i=1}^n y_i$.

The majority algorithm is defined as

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

where h_i denotes the hypothesis that outputs the constant value i .

We consider 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 $y_i \sim \text{Ber}(1/2)$. In this case, $Y \sim \text{Bin}(n, 1/2)$, and the population loss of $\mathcal{A}_{maj}$ equals $1/2$, independent of both the sample S and the sample size n . Consequently, analyzing its mean-squared error (MSE) reduces to controlling the covariance between folds.

Lemma 4.9. *The MSE of the majority algorithm equals $\frac{k-1}{k} \text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) + \frac{1}{4n}$.*

Proof. Since $L = \tilde{L}^{(k)} = 1/2$ and we have zero loss variance, this directly follows from Lemma 4.3. $\square$

Let us proceed with an informal analysis. First, $\mathcal{A}_{maj}$ is remarkably stable across many instances of S . In specific, whenever we know that Y is bounded away from $n/2$ by at least $m/2$, $\mathcal{A}_{maj}(S)$ and S are conditionally independent, and Hoeffding's inequality asserts that $\hat{L}^{(k)}$ concentrates around $1/2$, which is typically close in value to $1/2 \pm \Theta(1/\sqrt{n})$, leading to a conditional MSE on the order of $1/n$. By contrast, $\mathcal{A}_{maj}$ is conditionally highly hypothesis unstable in the regime $Y = n/2 \pm \theta(\sqrt{m})$. This follows from $\mathcal{A}_{maj}(S)$ having constant probability of changing from the all-ones to the all-zeros function (or vice-versa) upon removing a fold of size m so $\mathcal{A}_{maj}(S) \neq \mathcal{A}_{maj}(S_{-1})$. We fall into this unstable regime with probability proportional to $\sqrt{m}/\sqrt{n}$ due to Stirling's approximation of the central probability masses of Y . It follows that the algorithm becomes more hypothesis unstable as we decrease k (which makes sense because we removing a larger fold of size m). Yet, perhaps surprisingly, CV becomes *more accurate* as we decrease k (or increase m) as the following theorems suggest.

Theorem 4.10 (Fold-Covariance of Majority: Exact Combinatorial Form). *For $1 \leq m \leq n/2$, $m|n$, we have*

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

Proof. See Appendix B.5. $\square$

A more explicit version of the following result, including precise constants, is provided in Appendix B.

Theorem 4.11 (Fold-Covariance Asymptotics). *Throughout, let $n \geq 2$ and $m|n$.*

(A) *For all $m = \Omega(n^{1/5})$,*

$$\text{Cov}(n, m) = \Theta\left(\frac{1}{\sqrt{nm}}\right) = \Theta\left(\frac{\sqrt{k}}{n}\right).$$

(B) **Monotonicity and minimizer.** *For all sufficiently large n ,*

$$\text{Cov}(n, 1) > \text{Cov}(n, 2) > \dots > \text{Cov}(n, n/3) \text{ and } \text{Cov}(n, n/3) < \text{Cov}(n, n/2),$$

so consequently $k = 3$ minimizes fold covariance.

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, since the algorithm becomes less hypothesis stable. A similar lack of tightness, by more than constant factors, arises in existing analyses such as [KR97; BKL99; KKV11; Kum+13] when applied to the Majority algorithm.

For this reason, we identify *Majority as a natural benchmark*: achieving tightness for this instance should be regarded as a minimal requirement for any future theoretical bounds on the error of cross-validation.

4.2. A Minimax Lower Bound for Cross-Validation with ERM Algorithms. The answer to Question 2 follows as a corollary of the preceding analysis of the Majority algorithm. To establish this, we consider, for any ERM algorithm, a degenerate distribution supported on a single point x , where the labels are drawn uniformly from $\{0, 1\}$. In this case, an ERM must output a hypothesis whose label for x agrees with the majority label observed in the sample S . Thus, the behavior of any ERM under this distribution reduces directly to the analysis of the Majority algorithm.

Corollary 4.12. *For any ERM algorithm $\mathcal{A}$, it holds that*

$$\mathfrak{R}_{\text{CV}}(\mathcal{A}) = \Omega\left(\frac{\sqrt{k}}{n}\right).$$

This result shows that, in the distribution-free setting, no ERM algorithm can be designed to 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 $1/n$.

4.3. Algorithms Achieving the Hold-Out Rate. We conclude our results by showing that certain algorithms achieve, up to a constant factor, the same rate as a hold-out estimator, regardless of the number of folds. This result provides an affirmative answer to Question 3 by constructing an algorithm which attains (up to constant factors) the upper bound in Lemma 4.5.

To establish this result, we consider the setting of binary classification under the 0–1 loss and construct a simple family of algorithms that can output only the constant functions $h_0(x) = 0$ and $h_1(x) = 1$.

Definition 4.13 (*r*-Square-Wave Algorithm). An algorithm $\mathcal{A}$ is called an *r-square-wave algorithm* if, for a training sample $S^n = \{(x_i, y_i)\}_{i=1}^n$,

$$\mathcal{A}(S^n) = \begin{cases} h_0, & \text{if } \lfloor \frac{1}{\sqrt{r}} \sum_{i=1}^n y_i \rfloor \bmod 2 = 0, \\ h_1, & \text{if } \lfloor \frac{1}{\sqrt{r}} \sum_{i=1}^n y_i \rfloor \bmod 2 = 1 \end{cases}$$

Theorem 4.14 (Square-Wave Algorithm Fold-Covariance). Assume $k \geq 3$ and let $m|n$. Then, for sufficiently large m , the fold-covariance of the m -square-wave algorithm satisfies

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{c_0}{m} + E_L,$$

where c_0 is the main constant and E_L is an error term bounded by

$$|E_L| \leq \frac{c_R}{m} + O(m^{-3/2}).$$

where the above constants are given as

$$c_0 = \frac{1}{2} \sum_{j=0}^{\infty} e^{-\frac{\pi^2}{4}(2j+1)^2} \approx 0.0424, \quad c_R \leq 4 \times 10^{-4}.$$

In particular, since $c_0 > c_R$, we have $\text{Cov}(\hat{L}_1, \hat{L}_2) = \Theta(1/m)$ positive.

In other words, the fold-covariance of the square-wave algorithm is independent of n , no matter how large the shared training set (which is of size $n - 2m$) is, which is rather remarkable. The square-wave algorithm is carefully designed as to be robust to small changes in the training set, while simultaneously admitting large enough variation in the risk values it can achieve. Generally, these are two diametrically opposed algorithmic properties.

Theorem 4.14 gives the desired distribution-free result in the following corollary.

Corollary 4.15. For every k , there exists an algorithm $\mathcal{A}_k$ such that

$$\max_{\mathcal{D}} \text{MSE}_{\text{CV}}^{(k)}(\mathcal{A}_k, \mathcal{D}) = \Omega\left(\frac{1}{m}\right) = \Omega\left(\frac{k}{n}\right).$$

5. CONCLUSION

We presented a novel decomposition of the MSE for CV that illuminates the respective roles of squared-loss stability and fold covariance. In contrast, existing stability-based analyses are inherently limited, as they are not merely off by a constant factor—a fact underscored by our tight characterization of the Majority algorithm. Consequently, we propose that any future theoretical analysis of CV should benchmark its results against the Majority baseline introduced in Theorem 4.11.

While low loss variance is a fairly common property in practice, an interesting direction for future research is to identify alternative (and possibly weaker) conditions under which low squared-loss stability becomes necessary for achieving a small MSE.

Finally, as a natural extension of Corollary 4.12, 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.

ACKNOWLEDGEMENTS

todo

REFERENCES

- [AH98] M. Anthony and S. B. Holden. “Cross-validation for binary classification by real-valued functions: theoretical analysis”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pages 218–229.
- [AC10] S. Arlot and A. Celisse. “A survey of cross-validation procedures for model selection” (2010).
- [BG04] Y. Bengio and Y. Grandvalet. “No unbiased estimator of the variance of k-fold cross-validation”. *Journal of machine learning research* 5.Sep (2004), pages 1089–1105.
- [617] *Bias and variance in leave-one-out vs K-fold cross validation*. Cross Validated. URL:<https://stats.stackexchange.com/q/61783> (version: 2018-07-19). eprint: <https://stats.stackexchange.com/q/61783>.
- [BS06] I. F. Blake and C. Studholme. “Properties of random matrices and applications”. *Unpublished report available at https://www.cs.toronto.edu/~cvs/coding/random_report.pdf* (2006).
- [BKL99] A. Blum, A. Kalai, and J. Langford. “Beating the hold-out: Bounds for k-fold and progressive cross-validation”. In: *Proceedings of the twelfth annual conference on Computational learning theory*. 1999, pages 203–208.
- [Blu+89] A. Blumer et al. “Learnability and the Vapnik-Chervonenkis dimension”. *Journal of the ACM (JACM)* 36.4 (1989), pages 929–965.
- [BE02] O. Bousquet and A. Elisseeff. “Stability and generalization”. *Journal of machine learning research* 2.Mar (2002), pages 499–526.
- [DW79a] L. Devroye and T. Wagner. “Distribution-free inequalities for the deleted and holdout error estimates”. *IEEE Transactions on Information Theory* 25.2 (1979), pages 202–207.
- [DW79b] L. Devroye and T. Wagner. “Distribution-free performance bounds for potential function rules”. *IEEE Transactions on Information Theory* 25.5 (1979), pages 601–604.
- [Dzi+20] G. K. Dziugaite et al. “In search of robust measures of generalization”. *Advances in Neural Information Processing Systems* 33 (2020), pages 11723–11733.
- [Gas+24a] M. Gastpar et al. “Fantastic generalization measures are nowhere to be found”. In: *The Twelfth International Conference on Learning Representations*. 2024.
- [Gas+24b] M. Gastpar et al. “Which Algorithms Have Tight Generalization Bounds?” *arXiv preprint arXiv:2410.01969* (2024).
- [Jia+19] Y. Jiang et al. “Fantastic Generalization Measures and Where to Find Them”. In: *International Conference on Learning Representations*. 2019.
- [KKV11] S. Kale, R. Kumar, and S. Vassilvitskii. “Cross-validation and mean-square stability.” In: *ICS*. 2011, pages 487–495.
- [KR97] M. Kearns and D. Ron. “Algorithmic stability and sanity-check bounds for leave-one-out cross-validation”. In: *Proceedings of the tenth annual conference on Computational learning theory*. 1997, pages 152–162.
- [Kum+13] R. Kumar et al. “Near-optimal bounds for cross-validation via loss stability”. In: *International Conference on Machine Learning*. PMLR. 2013, pages 27–35.
- [NB99] C. Nadeau and Y. Bengio. “Inference for the generalization error”. *Advances in neural information processing systems* 12 (1999).
- [Pet12] V. V. Petrov. *Sums of independent random variables*. Springer Science & Business Media, 2012.
- [PC84] R. R. Picard and R. D. Cook. “Cross-validation of regression models”. *Journal of the American Statistical Association* 79.387 (1984), pages 575–583.
- [RW78] W. H. Rogers and T. J. Wagner. “A finite sample distribution-free performance bound for local discrimination rules”. *The Annals of Statistics* (1978), pages 506–514.
- [Sha93] J. Shao. “Linear model selection by cross-validation”. *Journal of the American statistical Association* 88.422 (1993), pages 486–494.

- [VC71] V. Vapnik and A. Y. Chervonenkis. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities”. *Theory of Probability & Its Applications* 16.2 (1971), pages 264–280.
- [280] *Variance of K -fold cross-validation estimates as $f(K)$: what is the role of stability?* Cross Validated. URL: <https://stats.stackexchange.com/q/280665>. (version: 2018-07-24). eprint: <https://stats.stackexchange.com/q/280665>.

APPENDIX A. MAIN PART

A.1. Proof of Lemma 4.1.

Proof. The proof begins by adding and subtracting $\tilde{L}^k$:

$$\begin{aligned}
\text{MSE}_{CV}^{(k)} &= \mathbb{E}[(\widehat{L}^{(k)} - L)^2] \\
&= \mathbb{E}[(\widehat{L}^{(k)} - \tilde{L}^{(k)} + (\tilde{L}^{(k)} - L))^2] \\
&= \mathbb{E}[(\widehat{L}^{(k)} - \tilde{L}^{(k)})^2] + \mathbb{E}[(\tilde{L}^{(k)} - L)^2] + 2\mathbb{E}[(\widehat{L}^{(k)} - \tilde{L}^{(k)})(\tilde{L}^{(k)} - L)] \\
&= \left(\text{Var}(\widehat{L}^{(k)}) + \text{Var}(\tilde{L}^{(k)}) - 2\text{Cov}(\widehat{L}^{(k)}, \tilde{L}^{(k)}) \right) + \mathbb{E}[(\tilde{L}^{(k)} - L)^2] + 2\text{Cov}(\widehat{L}^{(k)} - \tilde{L}^{(k)}, \tilde{L}^{(k)} - L) \\
&= \text{Var}(\widehat{L}^{(k)}) + \mathbb{E}[(\tilde{L}^{(k)} - L)^2] + \text{Var}(\tilde{L}^{(k)}) - 2\text{Var}(\tilde{L}^{(k)}) \\
&\quad - 2\text{Cov}(\widehat{L}^{(k)}, \tilde{L}^{(k)}) + 2\text{Cov}(\widehat{L}^{(k)}, \tilde{L}^{(k)}) - 2\text{Cov}(\widehat{L}^{(k)}, L) + 2\text{Cov}(\tilde{L}^{(k)}, L) \\
&= \text{Var}(\widehat{L}^{(k)}) + \mathbb{E}[(\tilde{L}^{(k)} - L)^2] - \text{Var}(\tilde{L}^{(k)}) + 2\text{Cov}(L, \tilde{L}^{(k)} - \widehat{L}^{(k)}). \tag{A.1}
\end{aligned}$$

We now substitute the fold-level decompositions for each term, leveraging symmetry. The key step is the Law of Total Variance for $\text{Var}(\widehat{L}_1^{(k)})$, via conditioning on S_{-1}^n :

$$\begin{aligned}
\text{Var}(\widehat{L}_1^{(k)}) &= \text{Var}(\mathbb{E}[\widehat{L}_1^{(k)} | S_{-1}^n]) + \mathbb{E}[\text{Var}(\widehat{L}_1^{(k)} | S_{-1}^n)] \\
&= \text{Var}(L_1^{(k)}) + \mathbb{E} \left[\text{Var}_{S_1 \sim \mathcal{D}^{\otimes m}} \left(\frac{1}{m} \sum_{(x,y) \in S_1} \ell(\mathcal{A}(S_{-1}^n)(x), y) \middle| S_{-1}^n \right) \right] \\
&= \text{Var}(L_1^{(k)}) + \frac{1}{m} \mathbb{E} [\text{Var}_{(x,y) \sim \mathcal{D}} (\ell(\mathcal{A}(S_{-1}^n)(x), y) | S_{-1}^n)] \quad (\text{since validation points are i.i.d.}) \\
&= \text{Var}(L_1^{(k)}) + \frac{\bar{\sigma}^2}{m}
\end{aligned}$$

The other terms decompose as:

- $\text{Var}(\widehat{L}^{(k)}) = \frac{1}{k} \text{Var}(\widehat{L}_1^{(k)}) + \frac{k-1}{k} \text{Cov}(\widehat{L}_1^{(k)}, \widehat{L}_2^{(k)})$
- $\text{Var}(\tilde{L}^{(k)}) = \frac{1}{k} \text{Var}(L_1^{(k)}) + \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)})$
- $\text{Cov}(L, \tilde{L}^{(k)} - \widehat{L}^{(k)}) = \text{Cov}(L, L_1^{(k)} - \widehat{L}_1^{(k)})$

Plugging these into the MSE expression of Eq. (A.1):

$$\begin{aligned}
\text{MSE}_{CV}^{(k)} &= \left(\frac{1}{k} \left(\text{Var}(L_1^{(k)}) + \frac{\bar{\sigma}^2}{m} \right) + \frac{k-1}{k} \text{Cov}(\widehat{L}_1^{(k)}, \widehat{L}_2^{(k)}) \right) + \mathbb{E}[(\tilde{L}^{(k)} - L)^2] \\
&\quad - \left(\frac{1}{k} \text{Var}(L_1^{(k)}) + \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)}) \right) + 2\text{Cov}(L, L_1^{(k)} - \widehat{L}_1^{(k)})
\end{aligned}$$

The variance terms cancel, yielding the final form. $\square$

A.2. Proof of Theorem 4.3.

Proof. The proof amounts to bounding the magnitude of the correction term $\mathcal{C} = 2\text{Cov}(L, L_1^{(k)} - \widehat{L}_1^{(k)}) - \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)})$. Using the triangle inequality:

$$|\mathcal{C}| \leq |2\text{Cov}(L, L_1^{(k)} - \widehat{L}_1^{(k)})| + \left| \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)}) \right|$$

We bound each term on the right-hand side separately. First, using the general identity for $\text{Var}(L_1^{(k)} - \widehat{L}_1^{(k)})$ and the law of total variance:

$$\begin{aligned}\text{Var}(L_1^{(k)} - \widehat{L}_1^{(k)}) &= \text{Var}(L_1^{(k)}) + \text{Var}(\widehat{L}_1^{(k)}) - 2\text{Cov}(L_1^{(k)}, \widehat{L}_1^{(k)}) \\ &= \text{Var}(L_1^{(k)}) + \left(\text{Var}(L_1^{(k)}) + \frac{\bar{\sigma}^2}{m} \right) - 2\text{Var}(L_1^{(k)}) = \frac{\bar{\sigma}^2}{m}\end{aligned}$$

Applying this to the covariance bound together with Cauchy-Schwarz:

$$\begin{aligned}|2\text{Cov}(L, L_1^{(k)} - \widehat{L}_1^{(k)})| &\leq 2\sqrt{\text{Var}(L) \cdot \text{Var}(L_1^{(k)} - \widehat{L}_1^{(k)})} \\ &\leq 2\sqrt{\sigma_n^2 \cdot \frac{\bar{\sigma}^2}{m}}\end{aligned}$$

Next, we bound

$$\left| \frac{k-1}{k} \text{Cov}(L_1^{(k)}, L_2^{(k)}) \right| \leq \frac{k-1}{k} \sqrt{\text{Var}(L_1^{(k)}) \text{Var}(L_2^{(k)})} \leq \frac{k-1}{k} \sigma_{n-m}^2$$

Combining the bounds for the two components gives the final result. $\square$

A.3. Proof of Lemma 4.5.

Proof. We can associate each random variable $F_i := (\widehat{L}_i - \bar{L}^k)$ with a vector in euclidean space given as $x_i := [F_i(S_1^n) \cdot \sqrt{\mathbb{P}(S_1^n)}, \dots, F_i(S_d^n) \cdot \sqrt{\mathbb{P}(S_d^n)}]$ where S_i^n are all the samples in the support of $\mathcal{D}^n$. With these associations, it is easy to verify that the standard inner product between the F_i equals the euclidean inner product of the corresponding vectors in euclidean space. We know that by symmetry that all of the pairwise inner products with $i \neq j$ are the same number, hence the problem boils down to determining how negative this number can at most become.

Given k d -dimensional vectors $x_1, \dots, x_k$ such that $\|x_i\|_2^2 = \|x_j\|_2^2 =: \|x\|_2^2$ for all i, j , and $x_i^T x_j =: \alpha$ for all $i \neq j$, we form the Gram matrix $G \in \mathbb{R}^{k \times k}$ with entries $G_{ij} = x_i^T x_j$:

$$G = \begin{pmatrix} \|x\|_2^2 & \alpha & \dots & \alpha \\ \alpha & \|x\|_2^2 & \dots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \dots & \|x\|_2^2 \end{pmatrix}$$

As G is a Gram matrix, it must be positive semidefinite. The eigenvalues of this specific matrix structure are $\lambda_1 = \|x\|_2^2 + (k-1)\alpha$ (with multiplicity 1) and $\lambda_2 = \|x\|_2^2 - \alpha$ (with multiplicity $k-1$). For G to be positive semidefinite, all eigenvalues must be non-negative:

$$\begin{aligned}\|x\|_2^2 - \alpha &\geq 0 \\ \|x\|_2^2 + (k-1)\alpha &\geq 0\end{aligned}$$

The first inequality yields $\alpha \leq \|x\|_2^2$ and assuming $k > 1$, the second inequality provides the lower bound $\alpha \geq -\frac{\|x\|_2^2}{k-1}$. The claim now follows from the fact that $\|x_i\|_2^2 = \text{Var}(\widehat{L}_i) \leq \sigma_{n-m}^2 + 1/(4m)$ and recalling that $k = n/m$. $\square$

A.4. Proof of Lemma 2.4.

Proof. Let $X_1, \dots, X_k$ be k identically distributed random variables, each with mean $\mathbb{E}[X_i] = \mu_X$ and variance $\text{Var}(X_i) = \sigma_X^2$. Let Y be a random variable with mean $\mathbb{E}[Y] = \mu_Y$ and variance $\text{Var}(Y) = \sigma_Y^2$. It is assumed that the variables X_i take values in the interval $[0, 1]$ for all $i \in \{1, \dots, k\}$, and so does Y .

Define the sample mean $\bar{X} = \frac{1}{k} \sum_{i=1}^k X_i$. Let $\sigma_{\bar{X}}^2 = \text{Var}(\bar{X})$ and define the squared difference $Z = (\bar{X} - Y)^2$.

We will prove the general statement that

$$(\mu_X - \mu_Y)^2 \leq \mathbb{E}[Z] \leq (\sigma_{\bar{X}} + \sigma_Y)^2 + (\mu_X - \mu_Y)^2$$

where the variance of the sample mean, $\sigma_{\bar{X}}^2$, is given by

$$\sigma_{\bar{X}}^2 = \frac{1}{k^2} \left(k \sigma_X^2 + \sum_{i \neq j} \text{Cov}(X_i, X_j) \right)$$

Lower Bound: Applying Jensen's inequality, we have

$$\mathbb{E}[Z] = \mathbb{E}[(\bar{X} - Y)^2] \geq (\mathbb{E}[\bar{X} - Y])^2 = (\mu_X - \mu_Y)^2$$

Upper Bound: We use the property that for any random variable D , $\mathbb{E}[D^2] = \text{Var}(D) + (\mathbb{E}[D])^2$. Applied to $D = \bar{X} - Y$, this gives

$$\mathbb{E}[Z] = \mathbb{E}[D^2] = \text{Var}(\bar{X} - Y) + (\mathbb{E}[\bar{X} - Y])^2 = \text{Var}(\bar{X} - Y) + (\mu_X - \mu_Y)^2$$

The variance of the difference, $\text{Var}(\bar{X} - Y)$, can be expanded as

$$\text{Var}(\bar{X} - Y) = \text{Var}(\bar{X}) + \text{Var}(Y) - 2 \text{Cov}(\bar{X}, Y)$$

So,

$$\text{Var}(\bar{X} - Y) = \sigma_{\bar{X}}^2 + \sigma_Y^2 - 2 \text{Cov}(\bar{X}, Y)$$

By the Cauchy-Schwarz inequality, $|\text{Cov}(\bar{X}, Y)| \leq \sigma_{\bar{X}} \sigma_Y$ hence $-2 \text{Cov}(\bar{X}, Y) \leq 2 \sigma_{\bar{X}} \sigma_Y$. Substituting this into the expression for $\text{Var}(\bar{X} - Y)$

$$\text{Var}(\bar{X} - Y) \leq \sigma_{\bar{X}}^2 + \sigma_Y^2 + 2 \sigma_{\bar{X}} \sigma_Y = (\sigma_{\bar{X}} + \sigma_Y)^2$$

Substituting this inequality back into the expression for $\mathbb{E}[Z]$:

$$\mathbb{E}[Z] = \text{Var}(\bar{X} - Y) + (\mu_X - \mu_Y)^2 \leq (\sigma_{\bar{X}} + \sigma_Y)^2 + (\mu_X - \mu_Y)^2$$

establishes the upper bound.

Finally, plugging in $X_i = L_i^k$, $Y = L$ and using $\sigma_{L^k}^2 \leq \sigma_{n-m}^2$, $\text{Cov}(X_i, X_j) \leq \sigma_{n-m}^2$ completes the proof. $\square$

APPENDIX B. MAJORITY ALGORITHM

Throughout this section, we consider the following setup.

B.1. Setup and Notation. For $1 \leq m \leq n/2$, $m|n$, 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 $\text{Cov}(\hat{L}_1, \hat{L}_2) = E(n, m)/4$ (see Thm. B.8 for details). We also denote $\text{Cov}(n, m) \equiv \text{Cov}(\hat{L}_1, \hat{L}_2)$ to highlight the roles of n, m .

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

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

and central binomial mass

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

B.2. Main Theorem.

Theorem B.1 (Fold-Covariance of the Majority Algorithm). *Throughout, let $n \geq 2$ and $m|n$.*

(A) Binomial form. *One has*

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

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

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

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

(C) Sublinear regime. *For all $\Omega(n^{1/5}) = m = o(n)$,*

$$\text{Cov}(n, m) = \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}} \left(1 - \frac{1}{8(m-1)}\right) + O(n^{-1}).$$

(D) Large m regime. *For all $\Omega(n^{2/3} \log^{1/3} n) = m \leq n/3$,*

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

(E) Exact expression for $m = n/2$. *It holds that*

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

(F) Monotonicity and minimizer. *For all sufficiently large n ,*

$$\text{Cov}(n, 1) > \text{Cov}(n, 2) > \dots > \text{Cov}(n, n/3) < \text{Cov}(n, n/2),$$

and consequently

$$\arg \min_{\substack{m|n \\ 1 \leq m \leq \lfloor n/2 \rfloor}} \text{Cov}(n, m) = \max\{d \mid n : d \leq n/3\}.$$

Proof. This is a consequence of collecting the results of Theorems B.9 to B.12. $\square$

B.3. Technical Lemmas. Let us first state a few technical results.

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

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

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 $\alpha := 4/(m-1)$.

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

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,

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

The product is

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

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 $a - b = (m-1)/2 - m/2 = -1/2$.

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

We also compute

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

Combining these results yields the displayed formula. $\square$

Lemma B.3 (Poisson summation for Gaussians). *Let $\gamma > 0$ and $\mu \in \mathbb{R}$. Define*

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

Then

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

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

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

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 (B.1). $\square$

Proposition B.4 (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

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

$$\Theta_{n,m} := \sum_{t \in \mathbb{Z}} \exp\left(-\pi^2 t^2 / (\alpha + \beta)\right) \exp(-2\pi i t \mu). \quad (\text{B.3})$$

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

$$\begin{aligned} & \sum_{j \in \mathbb{Z}} g_{m-1}(j)^2 g_N(\ell - j) \\ &= \frac{2}{\pi} \cdot \frac{1}{\sqrt{(m-1)(2n-3m-1)}} e^{-\frac{1}{2n-3m-1}} \sum_{t \in \mathbb{Z}} \exp\left(-\frac{\pi^2 t^2 (m-1)N}{2(2n-3m-1)}\right) e^{-2\pi i t \mu}. \end{aligned} \quad (\text{B.4})$$

Proof. By Lemma B.2, we have

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

where $C_{\text{prod}} = \frac{2}{\pi(m-1)} \sqrt{\frac{2}{\pi N}}$. Summing over $j \in \mathbb{Z}$ and applying Lemma B.3 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 B.2:

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

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

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

Lemma B.5 (Local Limit Theorem and Central binomial). *Let $r \geq 2$, $c := \lfloor r/2 \rfloor$ and $p_r(t) := 2^{-r} \binom{r}{t}$. Let $g_r(t) := \sqrt{\frac{2}{\pi r}} \exp(-(2t - r)^2 / (2r))$. There exists an absolute $C_0 > 0$ such that*

$$\sup_{t \in \mathbb{Z}} |p_r(t) - g_r(t)| \leq C_0 r^{-3/2}. \quad (\text{B.5})$$

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

$$\left| p_r(c) - g_r(c) \right| \leq C_0 r^{-3/2}, \quad g_r(c) = \begin{cases} \sqrt{\frac{2}{\pi r}}, & r \text{ even}, \\ \sqrt{\frac{2}{\pi r}} e^{-1/(2r)}, & r \text{ odd}. \end{cases} \quad (\text{B.6})$$

Hence, for all $r \geq 2$,

$$\sqrt{\frac{2}{\pi r}} e^{-1/(2r)} - C_0 r^{-3/2} \leq p_r(c) \leq \sqrt{\frac{2}{\pi r}} + C_0 r^{-3/2}. \quad (\text{B.7})$$

Proof. This is a classical uniform local limit theorem, see [Pet12, Chapter 7, Theorem 13] (with $p = q = \frac{1}{2}$). Evaluating at $t = c$ gives (B.6); the bounds (B.7) follow since $g_r(c)$ is as displayed. $\square$

B.4. Simplifying the Fold-Covariance.

Lemma B.6. *It holds that*

$$\text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) = \frac{4}{k^2} \mathbb{E}_Y \left[(C(k, Y))^2 \right].$$

where $C(k, Y) = \text{Cov}_{X_k}(X_k, \mathbf{1}_{X_k > m-Y})$, $m = (n - k)/2$, $X_k \sim \text{Bin}(k, 1/2)$ and $Y \sim \text{Bin}(n - 2k, 1/2)$ independent of each other.

Proof. Assume that $n - k$ is odd (to avoid ties) and that k divides n . Define $C(k, Y) = \text{Cov}_{X_k}(X_k, \mathbf{1}_{X_k > m-Y})$, as a covariance conditioned on Y and let $m = (n - k)/2$. Define $X_1, X_2 \sim \text{Bin}(k, 1/2)$ and $Y \sim \text{Bin}(n - 2k, 1/2)$ all independent were we interpret $X_1 = \{\text{Number of ones in the first fold}\}$, $X_2 = \{\text{Number of ones in the second fold}\}$, $Y = \{\text{Combined number of ones in the folds } 3, \dots, N\}$. Let p, q denote the probability mass functions corresponding to X_1, Y . By the law of total expectation it holds that

$$\mathbb{E}[\hat{L}_1^{(k)} \cdot \hat{L}_2^{(k)}] = \sum_{t=0}^{n-2k} \sum_{i,j=0}^k p(i)p(j)q(t) \cdot f(i, j, t)/k^2$$

where

$$f(i, j, t) = \begin{cases} (k-j)(k-i) & \text{if } t+i > (n-k)/2 \text{ and } t+j > (n-k)/2 \\ 2(k-j)i & \text{if } t+i > (n-k)/2 \text{ and } t+j < (n-k)/2 \\ ij & \text{if } t+i < (n-k)/2 \text{ and } t+j < (n-k)/2 \\ 0 & \text{else.} \end{cases}$$

The piece-wise defined function f can be explained as follows: when computing $\hat{L}_1^k$, we count the number of zeros in the first fold (i.e. $k - i$) as errors if the algorithm outputs the constant-one hypothesis which happens precisely when $t + j > (n - k)/2$, and else we count the number of ones (i). The same principle applies for the second fold with the roles of i and j reversed. The second case captures the case where exactly one of $\hat{L}_1, \hat{L}_2$ count zeros, and the other one counts ones.

Define $E = \mathbb{E}[\hat{L}_1^{(k)} \cdot \hat{L}_2^{(k)}] = \text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) + \mathbb{E}[\hat{L}_1^{(k)}]\mathbb{E}[\hat{L}_2^{(k)}] = \text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) + 1/4$.

With our definitions we can write $\mathbb{E}[\hat{L}_1^{(k)} \cdot \hat{L}_2^{(k)}] = \frac{1}{k^2} \mathbb{E}_Y[\mathbb{E}[f(X_1, X_2, Y)|Y]]$, where $\mathbb{E}[f(X_1, X_2, Y)|Y] = \sum_{i,j} p(i)p(j) \cdot f(i, j, Y)$. Let $c(Y) = m - Y$, $P_+(Y) = P(X_k > c(Y)|Y)$ and $P_-(Y) = P(X_k \leq c(Y)|Y)$, so $P_+(Y) + P_-(Y) = 1$. Now the conditional expectation $\mathbb{E}[f(X_1, X_2, Y)|Y]$ can be expressed using auxiliary functions $Q_s(Y)$ as

$$\begin{aligned} \mathbb{E}[f(X_1, X_2, Y)|Y] &= \mathbb{E} \left[(k - X_2)(k - X_1) \cdot \mathbf{1}_{\{Y+X_1 > (n-k)/2\}} \mathbf{1}_{\{Y+X_2 > (n-k)/2\}} \right. \\ &\quad + 2(k - X_2)X_1 \cdot \mathbf{1}_{\{Y+X_1 > (n-k)/2\}} \mathbf{1}_{\{Y+X_2 < (n-k)/2\}} \\ &\quad \left. + X_1X_2 \cdot \mathbf{1}_{\{Y+X_1 < (n-k)/2\}} \mathbf{1}_{\{Y+X_2 < (n-k)/2\}} | Y \right] \\ &= \mathbb{E}_{X_1} \left[(k - X_1) \mathbf{1}_{\{Y+X_1 > (n-k)/2\}} | Y \right] \mathbb{E}_{X_2} \left[(k - X_2) \mathbf{1}_{\{Y+X_2 > (n-k)/2\}} | Y \right] \\ &\quad + 2 \mathbb{E}_{X_1} \left[X_1 \mathbf{1}_{\{Y+X_1 > (n-k)/2\}} | Y \right] \mathbb{E}_{X_2} \left[(k - X_2) \mathbf{1}_{\{Y+X_2 < (n-k)/2\}} | Y \right] \\ &\quad + \mathbb{E}_{X_1} \left[X_1 \mathbf{1}_{\{Y+X_1 < (n-k)/2\}} | Y \right] \mathbb{E}_{X_2} \left[X_2 \mathbf{1}_{\{Y+X_2 < (n-k)/2\}} | Y \right] \\ &= Q_1(Y)^2 + 2Q_2(Y)Q_3(Y) + Q_4(Y)^2 \end{aligned}$$

where

- $Q_1(Y) = \mathbb{E}_{X_k}[(k - X_k) \mathbf{1}_{\{X_k > c(Y)\}} | Y]$
- $Q_2(Y) = \mathbb{E}_{X_k}[X_k \mathbf{1}_{\{X_k > c(Y)\}} | Y]$
- $Q_3(Y) = \mathbb{E}_{X_k}[(k - X_k) \mathbf{1}_{\{X_k \leq c(Y)\}} | Y]$
- $Q_4(Y) = \mathbb{E}_{X_k}[X_k \mathbf{1}_{\{X_k \leq c(Y)\}} | Y]$.

Let $C(Y)$ denote $C(k, Y)$ for brevity within this derivation. By definition of covariance, $C(Y) = Q_2(Y) - (k/2)P_+(Y)$. Thus, $Q_2(Y) = (k/2)P_+(Y) + C(Y)$. Similarly, we deduce: $Q_1(Y) = (k/2)P_+(Y) - C(Y)$, $Q_3(Y) = (k/2)P_-(Y) + C(Y)$ and $Q_4(Y) = (k/2)P_-(Y) - C(Y)$.

We can now compute the components necessary for $\mathbb{E}[f(X_1, X_2, Y)|Y]$

$$\begin{aligned} Q_1(Y)^2 &= \left(\frac{k}{2}P_+(Y) - C(Y)\right)^2 = \left(\frac{k}{2}\right)^2 P_+(Y)^2 - kP_+(Y)C(Y) + C(Y)^2 \\ 2Q_2(Y)Q_3(Y) &= 2\left(\frac{k}{2}P_+(Y) + C(Y)\right)\left(\frac{k}{2}P_-(Y) + C(Y)\right) \\ &= 2\left(\frac{k}{2}\right)^2 P_+(Y)P_-(Y) + kP_+(Y)C(Y) + kP_-(Y)C(Y) + 2C(Y)^2 \\ Q_4(Y)^2 &= \left(\frac{k}{2}P_-(Y) - C(Y)\right)^2 = \left(\frac{k}{2}\right)^2 P_-(Y)^2 - kP_-(Y)C(Y) + C(Y)^2 \end{aligned}$$

Summing these components, the terms linear in $C(Y)$ cancel out

$$-kP_+(Y)C(Y) + kP_+(Y)C(Y) + kP_-(Y)C(Y) - kP_-(Y)C(Y) = 0$$

and the remaining terms are

$$\begin{aligned} \mathbb{E}[f(X_1, X_2, Y)|Y] &= \left(\frac{k}{2}\right)^2 (P_+(Y) + P_-(Y))^2 + 4C(Y)^2 \\ &= \frac{k^2}{4} + 4C(Y)^2. \end{aligned}$$

Finally, we take the expectation with respect to Y

$$E = \frac{1}{k^2} \mathbb{E}_Y \left[\frac{k^2}{4} + 4C(k, Y)^2 \right] = \frac{1}{k^2} \left(\frac{k^2}{4} + 4\mathbb{E}_Y[C(k, Y)^2] \right).$$

Recalling that $E = \text{Cov}(\hat{L}_1^{(k)}, \hat{L}_2^{(k)}) + 1/4$, this completes the derivation. $\square$

Lemma B.7 (Simplification of Covariance Term). *Let $X_k \sim \text{Bin}(k, 1/2)$, $m = (n - k)/2$, and $a(Y) = \lfloor m - Y \rfloor$. Then,*

$$C(k, Y) = \text{Cov}_{X_k}(X_k, \mathbf{1}_{X_k > m - Y}) = \frac{k}{4}P(X_{k-1} = a(Y))$$

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

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

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

where $X_{k-1} \sim \text{Bin}(k - 1, 1/2)$. So, $C(k, Y) = \frac{k}{2}P(X_{k-1} \geq a(Y)) - \frac{k}{2}P(X_k \geq a(Y) + 1)$. To simplify $P(X_k \geq j + 1)$, let $X_k = X_{k-1} + B_k$, where $B_k \sim \text{Bernoulli}(1/2)$ is independent of X_{k-1} .

$$\begin{aligned} P(X_k \geq j + 1) &= P(X_{k-1} + B_k \geq j + 1 | B_k = 0)P(B_k = 0) + P(X_{k-1} + B_k \geq j + 1 | B_k = 1)P(B_k = 1) \\ &= \frac{1}{2}P(X_{k-1} \geq j + 1) + \frac{1}{2}P(X_{k-1} \geq j) \end{aligned}$$

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

$$\begin{aligned} C(k, Y) &= \frac{k}{2} \left[P(X_{k-1} \geq a(Y)) - \left(\frac{1}{2}P(X_{k-1} \geq a(Y) + 1) + \frac{1}{2}P(X_{k-1} \geq a(Y)) \right) \right] \\ &= \frac{k}{4} [P(X_{k-1} \geq a(Y)) - P(X_{k-1} \geq a(Y) + 1)] = \frac{k}{4} P(X_{k-1} = a(Y)) \end{aligned} \quad \square$$

B.5. Exact Combinatorial Form of the Fold-Covariance.

Theorem B.8. *It holds that*

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

Proof. We know from the previous Lemmas that $\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{1}{4} \mathbb{E}_Y [P(X = \lfloor m - Y \rfloor)^2]$.

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

□

B.6. Main Results for the Majority Algorithm Fold-Covariance.

Theorem B.9 (Sublinear m regime). *Fix integers n and $m|n$. Let $N := n - 2m$, and choose the parity-adjusted integer*

$$N_c \in \{\lfloor n - \frac{3}{2}m \rfloor, \lceil n - \frac{3}{2}m \rceil\} \quad \text{such that } N_c \equiv N \pmod{2},$$

so that $N_c/2$ is an admissible central index for $\text{Bin}(N_c, \frac{1}{2})$.

We bound $\text{Cov}(n, m)$ explicitly with error terms that are asymptotically negligible compared to the main term as long as $m = o(n)$.

(A) Precise binomial form. *One has*

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

uniformly for all $1 \leq m \leq n/3$.

(B) Explicit scalar form. *For all*

$$n^{1/5} \leq m \leq n/3$$

one has

$$\text{Cov}(n, m) = \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}} \left(1 - \frac{1}{8(m-1)} \right) + O\left(\frac{1}{n}\right).$$

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

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

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

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

where

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

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

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 (B.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

$$R_1 = \sum_{j=0}^{m-1} p_{m-1}(j)^2 \left(G_N(j) - G_{N_c} \right) + \underbrace{\sum_{j=0}^{m-1} p_{m-1}(j)^2 (\delta_N(j) - \delta_{N_c})}_{=: R_{\text{LLT}}}. \quad (\text{B.9})$$

STEP 3: BOUNDING THE PURE LLT REMAINDER. By $S_{m-1} \leq 1$ and the local limit theorem bound of B.5,

$$|R_{\text{LLT}}| \leq \sum_j p_{m-1}(j)^2 (|\delta_N(j)| + |\delta_{N_c}|) \leq C_{\text{LLT}} \left(\frac{1}{N^{3/2}} + \frac{1}{N_c^{3/2}} \right) = O\left(\frac{1}{n^{3/2}}\right).$$

STEP 4. N_c FROM THE FIRST-ORDER OPTIMAL GAUSSIAN CENTRAL TERM. Next, we bound $\sum_{j=0}^{m-1} p_{m-1}(j)^2 (G_N(j) - G_{N_c})$ with $N = n - 2m$, $N_c \equiv N \pmod{2}$, and define the discrete distribution w with probability weights

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

and corresponding expectation operator w : $\mathbb{E}_w[g] := \sum_j w_j g(j)$. Let $J \sim w$. We have $\mu_1 := \mathbb{E}_w[J] = \frac{m-1}{2}$ by symmetry. Our goal is to choose the parameter N_c as to “curvature match” the Gaussian prefactor G_{N_c} to the typical location of $m_0 - j$.

Since for a hypergeometric random variable $X \sim \text{Hypergeom}(N', K', n')$ it holds by definition that $P(\{X = k'\}) = \binom{K'}{k'} \binom{N'-K'}{n'-k'} / \binom{N'}{n'}$, we have

$$J \sim \text{Hypergeom}(N' = 2(m-1), K' = m-1, n' = m-1)$$

with variance

$$\text{Var}_w(J) = n' \frac{K'}{N'} \left(1 - \frac{K'}{N'}\right) \frac{N' - n'}{N' - 1} = \frac{(m-1)^2}{4(2m-3)} = \frac{m-1}{8} + O(1).$$

Since $\Delta_j = (\mu_1 - J) + (\frac{1}{2} - \theta)$ with $\theta \in [0, 1)$, is a shifted version of J we have $\mathbb{E}_w[\Delta_j] = \frac{1}{2} - \theta$ and

$$\mathbb{E}_w[\Delta_j^2] = \text{Var}_w(J) + (\frac{1}{2} - \theta)^2 = \frac{m-1}{8} + O(1).$$

Decomposition and expanding the error term. Set $c(t) := \sqrt{\frac{2}{\pi}} t^{-1/2}$. We can write

$$\begin{aligned} \sum_{j=0}^{m-1} p_{m-1}(j)^2 (G_N(j) - G_{N_c}) &= S_{m-1} \sum_{j=0}^{m-1} w_j (c(N) e^{-2\Delta_j^2/N} - c(N_c)) \\ &= S_{m-1} \left\{ c(N) \mathbb{E}_w [e^{-2\Delta_j^2/N} - 1] - (c(N_c) - c(N)) \right\}. \end{aligned}$$

Expand $c(N_c)$ around N (Taylor expansion with explicit remainder) and the exponential around 0:

$$\begin{aligned} c(N_c) &= c(N) + c'(N)(N_c - N) + \frac{1}{2} c''(\xi) (N_c - N)^2, \\ e^{-2\Delta_j^2/N} &= 1 - \frac{2}{N} \Delta_j^2 + R_j, \quad |R_j| \leq \frac{2}{N^2} \Delta_j^4, \end{aligned}$$

for some ξ between N and N_c , where $c'(t) = -\frac{1}{2} \sqrt{\frac{2}{\pi}} t^{-3/2}$, $c''(t) = \frac{3}{4} \sqrt{\frac{2}{\pi}} t^{-5/2}$. This yields

$$\begin{aligned} \frac{1}{S_{m-1}} \sum p(j)^2 (G_N(j) - G_{N_c}) &= \underbrace{c(N) \left(-\frac{2}{N} \mathbb{E}_w [\Delta_j^2] \right) + (-c'(N)(N_c - N))}_{\text{first-order terms}} \\ &\quad + \underbrace{c(N) \mathbb{E}_w [R_j] - \frac{1}{2} c''(\xi) (N_c - N)^2}_{\text{remainders}}. \end{aligned}$$

Choosing N_c to cancel the first order. Pick N_c so that the first-order bracket vanishes:

$$-c'(N)(N_c - N) = c(N) \frac{2}{N} \mathbb{E}_w [\Delta_j^2].$$

Since $c'(N) = -\frac{1}{2} \sqrt{\frac{2}{\pi}} N^{-3/2}$ and $c(N) = \sqrt{\frac{2}{\pi}} N^{-1/2}$, this equality is equivalent to

$$\frac{1}{2} \sqrt{\frac{2}{\pi}} N^{-3/2} (N_c - N) = \sqrt{\frac{2}{\pi}} N^{-1/2} \cdot \frac{2}{N} \mathbb{E}_w [\Delta_j^2] \iff N_c - N = 4 \mathbb{E}_w [\Delta_j^2].$$

Using $\mathbb{E}_w [\Delta_j^2] = \frac{(m-1)^2}{4(2m-3)} + O(1) = \frac{m-1}{8} + O(1)$, we get

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

and then we *parity-adjust* N_c to the nearest integer with $N_c \equiv N \pmod{2}$. This is exactly the choice in the theorem statement.

With this choice of N_c , the first-order terms vanish. We are left to bound the remainder terms from the Taylor expansion:

$$R_{\text{Taylor}} := S_{m-1} \cdot \left(c(N) \mathbb{E}_w [R_j] - \frac{1}{2} c''(\xi) (N_c - N)^2 \right).$$

We bound the two parts separately.

Second remainder term (from $c(N_c)$ expansion). We have $S_{m-1} \asymp m^{-1/2}$, $c''(\xi) = O(n^{-5/2})$ (since ξ is between $N, N_c \asymp n$), and $(N_c - N)^2 = (4 \mathbb{E}_w [\Delta_j^2])^2 = (O(m))^2 = O(m^2)$. Thus,

$$\left| S_{m-1} \cdot \frac{1}{2} c''(\xi) (N_c - N)^2 \right| \leq O(m^{-1/2}) \cdot O(n^{-5/2}) \cdot O(m^2) = O(m^{3/2} n^{-5/2}).$$

This term is dominated by $O(n^{-1})$ (since $m \leq n/2$) and is thus negligible.

First remainder term (from exponential expansion). Let $x_j := 2\Delta_j^2/N \geq 0$. The remainder is $R_j = e^{-x_j} - (1 - x_j)$. Since $e^{-x_j} \leq 1$ for $x_j \geq 0$, we have $R_j \leq x_j$ and

$$\mathbb{E}_w[R_j] \leq \mathbb{E}_w[x_j] = \mathbb{E}_w\left[\frac{2\Delta_j^2}{N}\right] = \frac{2}{N} \mathbb{E}_w[\Delta_j^2].$$

Using $\mathbb{E}_w[\Delta_j^2] = O(m)$, we have $\mathbb{E}_w[R_j] \leq O(m/N) = O(m/n)$. Now, we assemble the bound for this term, using $c(N) = O(N^{-1/2}) = O(n^{-1/2})$:

$$\left|S_{m-1} \cdot c(N) \cdot \mathbb{E}_w[R_j]\right| \leq O(S_{m-1}) \cdot O(N^{-1/2}) \cdot O(m/N) = O(m^{-1/2}) \cdot O(n^{-1/2}) \cdot O(m/n) = O(m^{1/2} n^{-3/2}).$$

Combined bound. The total error from this term is dominated by the exponential remainder:

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

STEP 5: COMPLETING (A).. Collecting Step 3 and Step 4 in (B.9), we have

$$|R_1| \leq \frac{C\sqrt{m}}{n^{3/2}}.$$

Since $m \leq O(n)$, this is $\leq C/n$. Furthermore, approximate P_{N_c} by its central Gaussian:

$$P_{N_c} = G_{N_c} + \delta_{N_c} = \sqrt{\frac{2}{\pi N_c}} + O\left(\frac{1}{N_c^{3/2}}\right) = \sqrt{\frac{2}{\pi N_c}} + O\left(\frac{1}{n^{3/2}}\right).$$

Thus from (B.8),

$$E(n, m) = S_{m-1} \sqrt{\frac{2}{\pi N_c}} + R_A(n, m), \quad R_A(n, m) = O\left(\frac{\sqrt{m}}{n^{3/2}}\right) + R_1 = O\left(\frac{\sqrt{m}}{n^{3/2}}\right),$$

uniformly for all m . This proves (A).

STEP 6: COMPLETING (B).. Use the two-correction Stirling expansion for the central mass (central binomial),

$$S_{m-1} = \frac{1}{\sqrt{\pi(m-1)}} \left(1 - \frac{1}{8(m-1)}\right) + \rho_m, \quad |\rho_m| \leq \frac{C}{(m-1)^{5/2}}.$$

Hence, from (A),

$$S_{m-1} \sqrt{\frac{2}{\pi N_c}} = \frac{2}{\pi \sqrt{(m-1)2N_c}} \left(1 - \frac{1}{8(m-1)}\right) + \rho_m \sqrt{\frac{2}{\pi N_c}}.$$

The extra m -side residual is bounded by

$$|\rho_m| \sqrt{\frac{2}{\pi N_c}} \leq \frac{C}{(m-1)^{5/2}} \cdot \frac{C'}{\sqrt{n}} = \frac{C''}{\sqrt{n} (m-1)^{5/2}}.$$

Combining with $R_A(n, m) = O(\sqrt{m}/n^{3/2})$ yields

$$|R_B(n, m)| \leq \frac{C_1 \sqrt{m}}{n^{3/2}} + \frac{C_2}{\sqrt{n} (m-1)^{5/2}} \leq \frac{C_3}{n} \quad \text{whenever } m \geq c n^{1/5}.$$

This proves (B). The threshold can be improved to $m \geq c n^{1/7}$, $m \geq c n^{1/9}$ etc. upon adding more explicit Stirling terms. For example, for one more term, the residual becomes $O(1/(\sqrt{n} (m-1)^{7/2}))$. $\square$

Theorem B.10 (MSE for small constant m and $m = n/2$). **(A) Strict decrease for fixed m -ranges.**

Let $K_0 \geq 2$ be fixed (independent of n). Then there exists $N_0 = N_0(K_0)$ such that for all $n \geq N_0$,

$$\text{Cov}(n, 1) > \text{Cov}(n, 2) > \cdots > \text{Cov}(n, K_0).$$

In fact, for each $m \in \{2, \dots, K_0\}$ and all $n \geq 8C(K_0 - 1)$,

$$\text{Cov}(n, m) \leq \text{Cov}(n, m-1) - \frac{1}{32(m-1)} S_{m-2} B_{m-1}.$$

where $S_{m-1} := 2^{-(2m-2)} \binom{2m-2}{m-1}$ and $b_m \in \{\lfloor n - \frac{3}{2}m \rfloor, \lceil n - \frac{3}{2}m \rceil\}$ be the parity-adjusted integer with $b_m \equiv n - 2m \pmod{2}$. Let

$$B_m := 2^{-b_m} \binom{b_m}{b_m/2} \quad (\text{central mass of } \text{Bin}(b_m, \frac{1}{2})).$$

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

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

Consequently, by the LLT (B.5),

$$\text{Cov}(n, 1) = \sqrt{\frac{1}{8\pi(n-2)}} + O\left(\frac{1}{n^{3/2}}\right) = \sqrt{\frac{1}{8\pi n}} + O\left(\frac{1}{n}\right).$$

(C) Exact expression for $m = n/2$. It holds that

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

Proof. **(A) Strict decrease for fixed m .** Write $A_m := S_{m-1}$ and $C_m := B_m$ so $E(n, m) = A_m C_m + R_m$. Exact ratio for A_m . Let $\ell = m-1$. Your slide gives for

$$p_\ell := 2^{-2\ell} \binom{2\ell}{\ell} = S_\ell, \quad \frac{p_{\ell+1}}{p_\ell} = \frac{2\ell+1}{2\ell+2} = 1 - \frac{1}{2(\ell+1)}.$$

Hence

$$\frac{A_m}{A_{m-1}} = \frac{S_{m-1}}{S_{m-2}} = \frac{2m-3}{2m-2} = 1 - \frac{1}{2(m-1)}. \quad (\star)$$

Monotonicity of C_m . The central mass $t \mapsto 2^{-t} \binom{t}{\lfloor t/2 \rfloor}$ is strictly decreasing in t (even/odd steps both go down), hence as m increases, b_m decreases and C_m increases. Applying Lemma B.5 yields

$$\frac{C_m}{C_{m-1}} = \sqrt{\frac{b_{m-1}}{b_m}} \left(1 + O\left(\frac{1}{n}\right)\right) = 1 + \frac{b_{m-1} - b_m}{2b_m} + O\left(\frac{1}{n}\right) = 1 + \frac{3}{4b_m} + O\left(\frac{1}{n}\right) = 1 + O\left(\frac{1}{n}\right),$$

uniformly (since $b_m \asymp n$ for fixed m).

Main-term ratio. Combine:

$$\frac{A_m C_m}{A_{m-1} C_{m-1}} = \left(1 - \frac{1}{2(m-1)}\right) \left(1 + O\left(\frac{1}{n}\right)\right) = 1 - \frac{1}{2(m-1)} + O\left(\frac{1}{n}\right).$$

Thus there exists $n_1(m)$ such that for all $n \geq n_1(m)$,

$$A_m C_m \leq \left(1 - \frac{1}{4(m-1)}\right) A_{m-1} C_{m-1}.$$

Now

$$E(n, m) - E(n, m-1) = (A_m C_m - A_{m-1} C_{m-1}) + (R_m - R_{m-1}).$$

Using $|R_m| \leq C/n$, for $n \geq 8C(m-1)$ we get

$$E(n, m) - E(n, m-1) \leq -\frac{1}{4(m-1)} A_{m-1} C_{m-1} + \frac{2C}{n} \leq -\frac{1}{8(m-1)} A_{m-1} C_{m-1} < 0,$$

which proves strict decrease at step $m - 1 \rightarrow m$. Taking $N_0(K_0) := \max_{2 \leq m \leq K_0} 8C(m - 1)$ yields the stated chain of inequalities.

(B) 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 $\text{Bin}(n - 2, \frac{1}{2})$; by Lemma B.5,

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

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

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

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

$$E\left(n, \frac{n}{2}\right) = \left(2^{-\ell} \binom{\ell}{r}\right)^2 =: q_{\ell}(r)^2,$$

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

Note that

$$\frac{\ell}{2} = \frac{n}{4} - \frac{1}{2}, \quad r - \frac{\ell}{2} = \begin{cases} +\frac{1}{2}, & n \equiv 0 \pmod{4}, \\ -\frac{1}{2}, & n \equiv 2 \pmod{4}, \end{cases} \Rightarrow \frac{2(r - \ell/2)^2}{\ell} = \frac{1}{2\ell}.$$

By Lemma B.5 (uniform for $\ell \geq 2$),

$$q_{\ell}(r) = \frac{1}{\sqrt{\pi\ell/2}} \exp\left(-\frac{1}{2\ell}\right) + O(\ell^{-3/2}).$$

Therefore

$$E\left(n, \frac{n}{2}\right) = q_{\ell}(r)^2 = \frac{2}{\pi\ell} \exp\left(-\frac{1}{\ell}\right) + O(\ell^{-2}) = \frac{4}{\pi(n-2)} \exp\left(-\frac{2}{n-2}\right) + O\left(\frac{1}{n^2}\right).$$

In particular,

$$E\left(n, \frac{n}{2}\right) = \frac{4}{\pi(n-2)} + O\left(\frac{1}{n^2}\right) = \frac{4}{\pi n} + O\left(\frac{1}{n^2}\right).$$

□

Theorem B.11 (MSE for large m). *Let $1 \leq m \leq n/3$ such that $m = \Omega\left(n^{2/3} \log^{1/3} n\right)$. Then,*

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

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

$$E(n, m) = \sum_{j=0}^{m-1} p_{m-1}(j)^2 p_N\left(\frac{N}{2} + \Delta_j\right), \quad \Delta_j := \ell - j - \frac{N}{2}.$$

We decompose

$$E(n, m) - \frac{2}{\pi\sqrt{(m-1)(2n-3m)}} = \underbrace{(E(n, m) - \sum_j p_{m-1}^2 g_N)}_{T_1} + \underbrace{(\sum_j p_{m-1}^2 g_N - \sum_j g_{m-1}^2 g_N)}_{T_2}$$

$$+ \underbrace{\left(\sum_J g_{m-1}^2 g_N - \sum_{\mathbb{Z}} g_{m-1}^2 g_N \right)}_{T_3} + \underbrace{\left(\sum_{\mathbb{Z}} g_{m-1}^2 g_N - \frac{2}{\pi \sqrt{(m-1)(2n-3m)}} \right)}_{T_4},$$

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

1) T_1 : replace p_N by g_N (uniform LLT). The LLT Lemma B.5 yields $\sup_t |p_N(t) - g_N(t)| \leq C_0 N^{-3/2}$, hence

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

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

$$|T_1| \leq \frac{C}{\sqrt{mn}^{3/2}}.$$

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

$$|T_2| = \left| \sum_{j=0}^{m-1} \delta_j (p_{m-1}(j) + g_{m-1}(j)) g_N\left(\frac{N}{2} + \Delta_j\right) \right| \leq \left(\sup_j |\delta_j| \right) \Sigma,$$

with

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

Uniform LLT (Lemma B.5) at scale $m-1$ gives $\sup_j |\delta_j| \leq C_0(m-1)^{-3/2}$. Moreover, upper bounding the average by the maximum,

$$\sum_{j=0}^{m-1} p_{m-1}(j) g_N\left(\frac{N}{2} + \Delta_j\right) \leq \sup_t g_N(t) \leq \sqrt{\frac{2}{\pi N}},$$

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

$$\sum_{j=0}^{m-1} g_{m-1}(j) g_N\left(\frac{N}{2} + \Delta_j\right) \leq \sum_{j \in \mathbb{Z}} g_{m-1}(j) g_N(\ell - j) \leq \frac{C'}{\sqrt{N}}.$$

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

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

3) T_3 : sum to integral.:

It holds that $T_3 = \sum_{i \notin [0, m-1]} g_{m-1}(i) \cdot g_{n-2m}(i)$. The product $g_C(i) := g_{m-1}(i) \cdot g_{n-2m}(i)$ is proportional to $N(\mu_C, \sigma_C^2)$ where:

$$\mu_C = \frac{(m-1)(n-2m)}{n-m-1} \quad \text{and} \quad \sigma_C^2 = \frac{(m-1)(n-2m)}{4(n-m-1)}$$

The tail boundary is at $i = m-1$ (since $m \leq n/3 \implies \mu_C \leq m-1$). The distance d (from mean to boundary) and the number of standard deviations z are:

$$\begin{aligned} d &= (m-1) - \mu_C = \frac{(m-1)^2}{n-m-1} \\ z &= \frac{d}{\sigma_C} = \frac{(m-1)^2}{n-m-1} \cdot \frac{2\sqrt{n-m-1}}{\sqrt{(m-1)(n-2m)}} = \frac{2(m-1)^{3/2}}{\sqrt{(n-m-1)(n-2m)}} \\ z^2 &= \frac{4(m-1)^3}{(n-m-1)(n-2m)} \end{aligned}$$

The tail sum $S_{tail} = T_L + T_R$ where $T_L = \sum_{i < 0} g_C(i)$ and $T_R = \sum_{i=m}^{\infty} g_C(i)$. The left tail T_L is exponentially small, since $\mu_C^2/(2\sigma_C^2) = 2\mu_C = \Theta(m)$:

$$\begin{aligned} T_L &\propto \sum_{i < 0} \exp\left(-\frac{(i - \mu_C)^2}{2\sigma_C^2}\right) = \sum_{j=1}^{\infty} \exp\left(-\frac{(-j - \mu_C)^2}{2\sigma_C^2}\right) \\ &\leq \sum_{j=1}^{\infty} \exp\left(-\frac{\mu_C^2 + 2j\mu_C}{2\sigma_C^2}\right) = \exp\left(-\frac{\mu_C^2}{2\sigma_C^2}\right) \sum_{j=1}^{\infty} \exp\left(-\frac{j\mu_C}{\sigma_C^2}\right) \\ &= O(\exp(-\Theta(m))) \quad (\text{since } \mu_C/\sigma_C^2 = O(1)) \end{aligned}$$

This is negligible, so S_{tail} is dominated by the right tail T_R . Since $g_C(i)$ is monotonically decreasing for $i > \mu_C$ (and $m > \mu_C$), we can use a standard integral bound:

$$\int_m^{\infty} g_C(x) dx \leq T_R \leq g_C(m) + \int_m^{\infty} g_C(x) dx$$

The integral $\int_m^{\infty} g_C(x) dx$ is the tail of a Gaussian, which has the asymptotic bound $O\left(\frac{1}{z} \exp(-z^2/2)\right)$, where $z = (m - \mu_C)/\sigma_C$. The term $g_C(m)$ is proportional to $O(\exp(-z^2/2))$. Since $1/z = \Theta(n/m^{3/2})$ is a large polynomial factor, the integral term dominates $g_C(m)$. Therefore, the sum T_R has the same asymptotic behavior as the integral:

$$S_{tail} = T_R = O\left(\int_m^{\infty} g_C(x) dx\right) = O\left(\frac{1}{z} \exp(-z^2/2)\right)$$

Now, let us find a lower bound on m such that T_3 is no larger than T_2 , or, equivalently,

$$S_{tail} \leq O\left(\frac{1}{\sqrt{n}m^{3/2}}\right)$$

The pre-factor is $1/z = \frac{\sqrt{(n-m-1)(n-2m)}}{2(m-1)^{3/2}} = \Theta(n/m^{3/2})$. The requirement hence becomes:

$$\begin{aligned} O\left(\frac{n}{m^{3/2}} \exp\left(-\frac{z^2}{2}\right)\right) &\leq O\left(\frac{1}{\sqrt{n}m^{3/2}}\right) \\ \implies \exp\left(-\frac{z^2}{2}\right) &\leq O(n^{-3/2}) \\ \implies z^2 &\geq 3 \log(n) - O(1) \end{aligned}$$

Thus, $z^2 = \Omega(\log n)$. The function $L(m) := z^2$ is monotonically increasing for $m \in [1, n/3]$. We find the lower bound m by solving $L(m) = \Omega(\log n)$ in the $m = o(n)$ regime, which gives the tightest constraint:

$$\frac{4(m-1)^3}{(n-m-1)(n-2m)} = \Omega(\log n)$$

The left side is $\Theta(m^3/n^2)$, so:

$$\Theta\left(\frac{m^3}{n^2}\right) = \Omega(\log n) \implies m^3 = \Omega(n^2 \log n)$$

This gives the final condition. Since $L(m)$ is increasing, any m satisfying this bound also satisfies the condition for all larger m up to $n/3$.

$$m = \Omega\left(n^{2/3} \log^{1/3} n\right)$$

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

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

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

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

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

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

Under the given conditions ($N \geq n/3$ and $m \leq n/3$), the exponent A simplifies to $A = \Theta(m)$. To obtain error $O(1/n)$, we require $e^{-\Theta(m)} \leq O(1/n)$, which is satisfied under our assumption $m = \Omega(\ln n)$. Hence $|\Theta_{n,m} - 1| = O(1/n)$ which gives overall error $A_{n,m} \cdot \Theta_{n,m} - 1 = (1 + O(1/n))(1 + O(e^{-cm})) - 1 = O(1/n) + O(e^{-cm})$ and

$$|T_4| = \frac{2}{\pi \sqrt{(m-1)(2n-3m)}} |A_{n,m} \Theta_{n,m} - 1| \leq \frac{2}{\pi \sqrt{(m-1)(2n-3m)}} (C/n + C'/n) \leq \frac{C}{\sqrt{mn}^{1.5}},$$

since $\frac{1}{\sqrt{(m-1)(2n-3m)}} \leq c/\sqrt{mn}$.

4) *Conclusion.* Adding the three bounds gives $|T_1| + |T_2| + |T_3| + |T_4| = O(\frac{1}{\sqrt{mn}^{3/2}}) + O(\frac{1}{\sqrt{nm}^{3/2}})$, uniformly for $1 \leq m \leq n/3$ provided that $m = \Omega(n^{2/3} \log^{1/3} n)$. The first error term is asymptotically dominated by the second error term and hence absorbed. The result is obtained by recalling $\text{Cov}(n, m) = E(n, m)/4$. $\square$

Theorem B.12 (Monotonicity and minimizer). *For all sufficiently large n , the function $\text{Cov}(n, m)$ over the divisors m of n satisfies:*

$$\text{Cov}(n, 1) > \text{Cov}(n, 2) > \cdots > \text{Cov}(n, m_0) < \text{Cov}(n, n/2),$$

where $m_0 = \max\{d \mid n : d \leq n/3\}$. Consequently, the minimizer of $\text{Cov}(n, m)$ is

$$\arg \min_{\substack{m \mid n \\ 1 \leq m \leq \lfloor n/2 \rfloor}} \text{Cov}(n, m) = \max\{d \mid n : d \leq n/3\}.$$

Proof. The proof consists of two main parts: first, proving the strictly decreasing behavior of $\text{Cov}(n, m)$ for divisors $m \leq n/3$, and second, proving the "uptick" at $m = n/2$.

PART I: MONOTONICITY FOR $m \leq n/3$. We show that $\text{Cov}(n, m)$ is a strictly decreasing function of m for $m \in \{d \mid n, d \leq n/3\}$.

First, for any fixed $K_0 \geq 2$, **Theorem B.10(A)** states that for all $n \geq N_0(K_0)$, we have $\text{Cov}(n, 1) > \text{Cov}(n, 2) > \cdots > \text{Cov}(n, K_0)$. This establishes the strict decrease for any fixed set of small divisors.

Second, for the range $m \geq \Omega\left(n^{2/3} \log^{1/3} n\right)$ up to $m \leq n/3$, we use **Theorem B.11** (MSE for large m). This theorem states

$$\text{Cov}(n, m) = f(n, m) + R(n, m),$$

$$\text{where } f(n, m) := \frac{1}{2\pi\sqrt{(m-1)(2n-3m)}} \text{ and } |R(n, m)| \leq \frac{C}{\sqrt{nm^{3/2}}}.$$

We analyze the monotonicity of the main term $f(n, m)$ by treating m as a continuous variable. The function is positive, so its monotonicity is the inverse of its denominator's. Let $D(m) := (m-1)(2n-3m)$. We find the derivative of the denominator with respect to m :

$$\frac{\partial D}{\partial m} = \frac{\partial}{\partial m}(2nm - 3m^2 - 2n + 3m) = 2n - 6m + 3.$$

In the specified range $1 \leq m \leq n/3$, this derivative is strictly positive:

$$\frac{\partial D}{\partial m} \geq 2n - 6(n/3) + 3 = 2n - 2n + 3 = 3 > 0.$$

Since the denominator $D(m)$ is strictly increasing for $m \in [1, n/3]$, its reciprocal is strictly decreasing, and thus $f(n, m)$ is strictly decreasing.

For sufficiently large n , the remainder $R(n, m)$ is of a smaller order than the main term. Specifically, from **Theorem B.9** (Sublinear m , $\mathbf{B}$), for $m \geq n^{1/5}$, the error is $O(1/n)$. The difference $f(n, m) - f(n, m+1) \approx -f'(n, m)$ is $\Omega(m^{-3/2}n^{-1/2})$, which is larger than $O(1/n)$ for $m \leq n^{1/3}$. The combination of **Theorem B.10(A)** (for small m) and the strict monotonicity of the asymptotic main term $f(n, m)$ (for larger m) is sufficient to conclude $\text{Cov}(n, m)$ is strictly decreasing over the entire range of divisors $m \leq n/3$.

PART II: THE UPTICK AT $m = n/2$. We now show that $\text{Cov}(n, m_0) < \text{Cov}(n, n/2)$ for $m_0 = \max\{d \mid n, d \leq n/3\}$.

First, from **Theorem B.10(C)**, we have the asymptotic value at $m = n/2$:

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

Second, we find the asymptotic value at m_0 . Since m_0 is the largest divisor $\leq n/3$, $m_0 = n/3 - \epsilon_n$, where $\epsilon_n = O(1)$. We use the main asymptotic term $f(n, m)$ from **Theorem B.11** (MSE for large m), as the error terms are of a lower order.

$$\begin{aligned} \text{Cov}(n, m_0) &= f(n, m_0) + O(m_0^{-3/2}n^{-1/2}) \\ &= \frac{1}{2\pi\sqrt{(m_0-1)(2n-3m_0)}} + O(n^{-2}) \\ &= \frac{1}{2\pi\sqrt{(\frac{n}{3} + O(1))(2n - 3(\frac{n}{3} + O(1)))}} + O(n^{-2}) \\ &= \frac{1}{2\pi\sqrt{(\frac{n}{3} + O(1))(n + O(1))}} + O(n^{-2}) \\ &= \frac{1}{2\pi\sqrt{n^2/3 + O(n)}} + O(n^{-2}) \\ &= \frac{1}{2\pi(n/\sqrt{3})} (1 + O(1/n))^{-1/2} + O(n^{-2}). \end{aligned}$$

We now compare the main terms. For all n large enough:

$$\text{Cov}(n, m_0) = \frac{\sqrt{3}}{2\pi n} + O(1/n^2) < \frac{1}{\pi n} + O(1/n^2) = \text{Cov}(n, n/2).$$

CONCLUSION. From Part I, $\text{Cov}(n, m)$ is strictly decreasing for all divisors $m \leq n/3$. This implies the minimum value in this range occurs at the largest divisor, $m_0 = \max\{d \mid n, d \leq n/3\}$. From Part II, we proved that $\text{Cov}(n, m_0) < \text{Cov}(n, n/2)$. This implies m_0 is the global minimizer over all divisors $m \leq n/2$.

Therefore,

$$\arg \min_{\substack{m \mid n \\ 1 \leq m \leq \lfloor n/2 \rfloor}} \text{Cov}(n, m) = \max\{d \mid n : d \leq n/3\}.$$

□

APPENDIX C. LINEAR FUNCTIONS

Lemma C.1. *The probability that an $n_1 \times n_2$ matrix with coefficients drawn i.i.d. from $\mathcal{U}(\{0, \dots, q-1\})$ has rank r is given by*

$$R_q(n_1, n_2, r) = \begin{bmatrix} n_2 \\ r \end{bmatrix}_q \sum_{l=0}^r (-1)^{r-l} \begin{bmatrix} r \\ l \end{bmatrix}_q q^{n_1(l-n_2) + \binom{r-l}{2}} \quad (\text{C.1})$$

where $\begin{bmatrix} n \\ k \end{bmatrix}_q := \prod_{i=0}^{k-1} \frac{q^{n-i}-1}{q^{k-i}-1}$ denote the so-called Gaussian coefficients.

Moreover, it holds that

$$R_q(n_1, n_2, r) = \begin{bmatrix} n_2 \\ r \end{bmatrix}_q q^{n_1(r-n_2)} \prod_{s=0}^{r-1} (1 - q^{s-n_1}) \quad (\text{C.2})$$

Proof. The first identity is a Corollary of [BS06, Corollary 2.2].

Let us continue by proving the second identity. The sum in (C.1), denoted by S_{sum} , can be rewritten by substituting $k = r - l$ (so $l = r - k$) as

$$\begin{aligned} S_{sum} &= \sum_{k=0}^r (-1)^k \begin{bmatrix} r \\ r-k \end{bmatrix}_q q^{n_1(r-k-n_2) + \binom{k}{2}} \\ &= \sum_{k=0}^r (-1)^k \begin{bmatrix} r \\ k \end{bmatrix}_q q^{n_1(r-n_2) - n_1 k + \binom{k}{2}} \quad (\text{since } \begin{bmatrix} r \\ r-k \end{bmatrix}_q = \begin{bmatrix} r \\ k \end{bmatrix}_q) \\ &= q^{n_1(r-n_2)} \sum_{k=0}^r (-1)^k \begin{bmatrix} r \\ k \end{bmatrix}_q (q^{-n_1})^k q^{\binom{k}{2}} \end{aligned}$$

Using the q -binomial theorem, which states $\sum_{k=0}^N \begin{bmatrix} N \\ k \end{bmatrix}_q x^k q^{\binom{k}{2}} = \prod_{s=0}^{N-1} (1 + xq^s)$, with $N = r$ and $x = (-1) \cdot q^{-n_1}$ yields

$$S_{sum} = q^{n_1(r-n_2)} \prod_{s=0}^{r-1} (1 - q^{-n_1} q^s) = q^{n_1(r-n_2)} \prod_{s=0}^{r-1} (1 - q^{s-n_1}).$$

□

Lemma C.2 (Rank Probability Asymptotics). *Assume that X is an $n_1 \times n_2$ matrix with coefficients drawn i.i.d. from $\mathcal{U}(\{0, \dots, q-1\})$. Denote $m_0 = \min(n_1, n_2)$ and $\Delta_0 = |n_1 - n_2|$.*

1. **Probability of Full Rank:** *The probability that the matrix X achieves its maximum possible rank m_0 is $R_q(n_1, n_2, m_0) = 1 - O(q^{-(\Delta_0+1)})$. This implies that for large q , random matrices are overwhelmingly likely to have full rank m_0 .*
2. **Probability of Specific Rank Deficiency:** *The probability of the rank being $m_0 - j$ for $j \geq 1$ (a rank deficiency of j) is $R_q(n_1, n_2, m_0 - j) = O(q^{-j(\Delta_0+j)})$. This shows that the probability of specific rank deficiencies decreases extremely rapidly with increasing deficiency j and with increasing q .*

Proof. **Probability of Full Rank**

Since the rank probability is symmetric in (n_1, n_2) , we can assume without loss of generality that $n_2 \leq n_1$. In this case, $m_0 = n_2$ and $\Delta_0 = n_1 - n_2$. Substitute $r = n_2$ into Eq. (C.2). Then, since $\begin{bmatrix} n_2 \\ n_2 \end{bmatrix}_q = 1$,

$$R_q(n_1, n_2, n_2) = \begin{bmatrix} n_2 \\ n_2 \end{bmatrix}_q q^{n_1(n_2-n_2)} \prod_{s=0}^{n_2-1} (1 - q^{s-n_1}) = \prod_{s=0}^{n_2-1} (1 - q^{s-n_1}).$$

Let $k = n_1 - s$. As s ranges from 0 to $n_2 - 1$, k ranges from n_1 down to $n_1 - n_2 + 1$:

$$R_q(n_1, n_2, n_2) = \prod_{k=n_1-n_2+1}^{n_1} (1 - q^{-k}) = (1 - q^{-(n_1-n_2+1)})(1 - q^{-(n_1-n_2+2)}) \cdots (1 - q^{-n_1}).$$

Hence for large q

$$R_q(n_1, n_2, n_2) = 1 - q^{-(n_1-n_2+1)} + O(q^{-(n_1-n_2+2)}).$$

This implies $R_q(n_1, n_2, n_2) = 1 - O(q^{-(n_1-n_2+1)}) = 1 - O(q^{-(\Delta_0+1)})$.

Probability of Specific Rank Deficiency

We want to show that $R_q(n_1, n_2, m_0 - j) = O(q^{-j(\Delta_0 + j)})$ for $j \geq 1$. Let $r = m_0 - j$.

First note that we can upper bound the Gaussian coefficient as

$$\begin{aligned} \begin{bmatrix} n_2 \\ r \end{bmatrix} &= \prod_{i=0}^{r-1} \frac{q^{n_2-i} - 1}{q^{r-i} - 1} \\ &= \prod_{i=0}^{r-1} \frac{q^{n_2-i} (1 - q^{-(n_2-i)})}{q^{r-i} (1 - q^{-(r-i)})} \\ &\leq (q^{n_2-r})^r \prod_{i=0}^{r-1} \frac{1}{(1 - q^{-(r-i)})} \end{aligned}$$

which is in $O(q^{(n_2-r)r})$ since every factor in the product can be expanded to a geometric series which is in $O(1 + q^{-1})$.

Assume $c_g > 0$ is a valid constant such that $\begin{bmatrix} n_2 \\ r \end{bmatrix} < c_g \cdot q^{(n_2-r)r}$. Putting things together we get

$$\begin{aligned} R_q(n_1, n_2, r) &\leq c_g q^{r(n_2-r)} q^{n_1(r-n_2)} \prod_{s=0}^{r-1} (1 - q^{s-n_1}) \\ &\leq c_g q^{-(n_1-r)(n_2-r)} \end{aligned}$$

which implies $R_q(n_1, n_2, r) = O(q^{-(n_1-r)(n_2-r)})$. Similar to before, we can assume without loss of generality that $n_1 \leq n_2$, in which case $m_0 = n_1$ and $\Delta_0 = n_2 - n_1$. Hence we can substitute $r = m_0 - j = n_1 - j$ which yields $R_q(n_1, n_2, r) = O(q^{j(n_2-n_1+j)}) = O(q^{-j(\Delta_0+j)})$ $\square$

Lemma C.3. Let X be an $n \times d$ matrix with entries in the finite field $\mathbb{F}_q$, and let $y \in \mathbb{F}_q^n$. Given the linear system $Xb = y$, where $b \in \mathbb{F}_q^d$ and the rank of X is r , the number of distinct solutions for b is q^{d-r} .

Proof. Since the system $Xb = y$ is consistent, there exists at least one particular solution $b_p \in \mathbb{F}_q^d$ such that $Xb_p = y$. Any other solution b can be expressed as $b = b_p + b_h$, where b_h is in the null space of X , denoted by $N(X)$. By the Rank-Nullity Theorem, the dimension of the null space is given by $\dim(N(X)) = d - \text{rank}(X)$ hence given that $\text{rank}(X) = r$, we have $\dim(N(X)) = d - r$. A vector space of dimension k over a finite field $\mathbb{F}_q$ contains q^k elements. Therefore, the null space $N(X)$ contains q^{d-r} distinct vectors b_h . Each distinct $b_h \in N(X)$ yields a distinct solution $b = b_p + b_h$. Thus, the number of distinct solutions for b is equal to the number of elements in $N(X)$, which is q^{d-r} . $\square$

Lemma C.4. Assume that we are given the ground truth linear function is f that labels the n feature vectors which are drawn uniformly at random from $\mathbb{F}_q$ and stacked in a matrix $X \in \mathbb{F}_q^{n \times d}$.

The population loss of any linear function $h \neq f$ is

$$1 - 1/q.$$

Moreover, the probability of the random linear solver to output the wrong concept given that the rank of X is r is given by

$$\mathbb{P}(\{A(S) \neq f\} | \text{Rank}(X) = r) = 1 - q^{r-d}.$$

Proof. A linear function from $\mathbb{F}_q^d$ to $\mathbb{F}_q$ can be written as $L(\mathbf{x}) = \mathbf{v} \cdot \mathbf{x}$ for a unique vector $\mathbf{v} \in \mathbb{F}_q^d$. Let $L_1(\mathbf{x}) = \mathbf{v}_1 \cdot \mathbf{x}$ and $L_2(\mathbf{x}) = \mathbf{v}_2 \cdot \mathbf{x}$. Since L_1 and L_2 are distinct, their corresponding vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ must be distinct, so $\mathbf{v}_1 \neq \mathbf{v}_2$.

The functions L_1 and L_2 agree at a point $\mathbf{x} \in \mathbb{F}_q^d$ if $L_1(\mathbf{x}) = L_2(\mathbf{x})$. This is equivalent to $\mathbf{v}_1 \cdot \mathbf{x} = \mathbf{v}_2 \cdot \mathbf{x}$, or $(\mathbf{v}_1 - \mathbf{v}_2) \cdot \mathbf{x} = 0$. Let $\mathbf{w} = \mathbf{v}_1 - \mathbf{v}_2$. Since $\mathbf{v}_1 \neq \mathbf{v}_2$, it follows that $\mathbf{w} \neq \mathbf{0}$. The set of points where L_1 and L_2 agree is the kernel of the linear functional $L_{\mathbf{w}} : \mathbb{F}_q^d \rightarrow \mathbb{F}_q$ defined by $L_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$. Since $\mathbf{w} \neq \mathbf{0}$, $L_{\mathbf{w}}$ is a non-zero linear functional.

The image of a non-zero linear functional $L_{\mathbf{w}} : \mathbb{F}_q^d \rightarrow \mathbb{F}_q$ is $\mathbb{F}_q$ itself. Thus, $\dim(\text{Im}(L_{\mathbf{w}})) = 1$. By the rank-nullity theorem, $\dim(\mathbb{F}_q^d) = \dim(\ker(L_{\mathbf{w}})) + \dim(\text{Im}(L_{\mathbf{w}}))$. So, $d = \dim(\ker(L_{\mathbf{w}})) + 1$, which implies $\dim(\ker(L_{\mathbf{w}})) = d - 1$.

The number of points in a subspace of dimension k over $\mathbb{F}_q$ is q^k . Therefore, the number of points $\mathbf{x}$ where $L_1(\mathbf{x}) = L_2(\mathbf{x})$ (i.e., the size of $\ker(L_{\mathbf{w}})$) is q^{d-1} . The total number of points in the space $\mathbb{F}_q^d$ is q^d . The fraction of points where L_1 and L_2 agree is $\frac{q^{d-1}}{q^d} = \frac{1}{q}$. Therefore, assuming that a test point z is drawn uniformly at random means that the population loss is $\mathbb{P}(\{h(z) = f(z)\}) = 1 - 1/q$. This proves the first statement.

The second statement follows immediately from Lemma C.3 by recalling that the random linear solver picks uniformly at random one of the linear functions which agree with the labeling of f across all n samples, and there are q^{d-r} such functions. $\square$

Lemma C.5 (Expected Loss of Random Linear Algorithm). *Assume we are given n' feature vector drawn independently and uniformly at random from $\mathcal{D} = \mathcal{U}(\mathbb{F}_q^d)$, with labels generated by an arbitrary linear function. We stack the feature vectors in a matrix $X \in \mathbb{F}_q^{n' \times d}$. Let $\Delta_0 := d - n'$ and $P_{\text{rank} < d} := \mathbb{P}(\{\text{Rank}(X) < d\})$ which is $O(q^{-(\Delta_0+1)})$ per Lemma C.2. Then, the expected population loss $\bar{L}_{n'} := \mathbb{E}_{S^{n'} \sim \mathcal{D}^{n'}, \mathcal{A}}[L(\mathcal{A}_{\text{lin}}, S^{n'})]$ of the random parity solver algorithm receiving n' samples can be bounded as follows.*

1. If $d > n'$:

$$(1 - 1/q) \left(1 - q^{-\Delta_0}\right) \leq \bar{L}_{n'} \leq (1 - 1/q) \left(1 - q^{-\Delta_0} + K_1 q^{-(2\Delta_0+1)}\right) \quad (\text{C.3})$$

Thus, $\bar{L}_{n'} \approx (1 - 1/q)(1 - q^{-\Delta_0})$.

2. If $d \leq n'$:

$$(1 - 1/q) \left((1 - 1/q) P_{\text{rank} < d} - K_2 q^{-(2\Delta_0+5)}\right) \leq \bar{L}_{n'} \leq (1 - 1/q)^2 P_{\text{rank} < d} \quad (\text{C.4})$$

Thus, $\bar{L}_{n'} \approx (1 - 1/q)^2 P_{\text{rank} < d}$.

Proof. Let $m_0 = \min(d, n')$ and assume f is the ground truth linear function that labeled the features stacked in X , so that we obtain samples $S^{n'}$. Using Lemma C.4, we can write (using the law of total expectation)

$$\begin{aligned} \bar{L}_{n'} &= \sum_{i=0}^{m_0} \mathbb{E}[L_{n'} | \{\text{Rank}(X) = i\}] \cdot \mathbb{P}(\{S^{n'} : \text{Rank}(X) = i\}) \\ &= \sum_{i=0}^{m_0} (1 - 1/q) \cdot \mathbb{E}_z[\mathbf{1}_{\{\mathcal{A}_{\text{lin}}(S^{n'})(z) \neq f(z)\}} | \{S^{n'} : \text{Rank}(X) = i\}] \cdot \mathbb{P}(\{S^{n'} : \text{Rank}(X) = i\}) \\ &= (1 - 1/q) \sum_{i=0}^{m_0} \mathbb{P}_z(\mathcal{A}_{\text{lin}}(S^{n'})(z) \neq f(z) | \{S^{n'} : \text{Rank}(X) = i\}) \cdot \mathbb{P}(\{S^{n'} : \text{Rank}(X) = i\}) \\ &= (1 - 1/q) \sum_{i=0}^{m_0} (1 - q^{i-d}) \cdot R_q(n', d, i) \end{aligned}$$

where R_q is defined as in Lemma C.1. This means that

$$\bar{L}_{n'} = \left(1 - \frac{1}{q}\right) S_0 \quad (\text{C.5})$$

where S_0 is defined as

$$S_0 = \sum_{i=0}^{\min\{d-1, n'\}} (1 - q^{i-d}) \cdot R_q(n', d, i). \quad (\text{C.6})$$

Based on the approximate rank probabilities of Lemma C.2, we directly obtain good bounds on S_0 for large enough q .

Bounds for S_0 **1. Case 1: $d > n'$**

Here the sum for S_0 runs up to $i = n'$. Let us first show the lower bound $S_0 \geq 1 - q^{-(d-n')}$.

Let $A = 1 - q^{-(d-n')}$. We have to show that $S_0 \geq A$. Consider the difference

$$\begin{aligned} S_0 - A &= \sum_{i=0}^{n'} (1 - q^{-(d-i)}) R_q(n', d, i) - \left(1 - q^{-(d-n')}\right) \\ &= \sum_{i=0}^{n'} (1 - q^{-(d-i)}) R_q(n', d, i) - \left(1 - q^{-(d-n')}\right) \sum_{i=0}^{n'} R_q(n', d, i) \quad (\text{since } \sum R_q(n', d, i) = 1) \\ &= \sum_{i=0}^{n'} [q^{-(d-n')} - q^{-(d-i)}] R_q(n', d, i) \\ &= \sum_{i=0}^{n'} q^{-(d-n')} (1 - q^{-(n'-i)}) R_q(n', d, i). \end{aligned}$$

It is easy to see that each summand is non-negative, hence the lower bound is proven.

For the upper bound, we can rewrite S_0 by isolating the contribution from $R_q(n', d, n')$ as

$$\begin{aligned} S_0 &= \sum_{i=0}^{n'} (1 - q^{-(d-i)}) R_q(n', d, i) \\ &= (1 - q^{-(d-n')}) R_q(n', d, n') + \sum_{i=0}^{n'-1} (1 - q^{-(d-i)}) R_q(n', d, i) \end{aligned}$$

Substituting $R_q(n', d, n') = 1 - \sum_{i=0}^{n'-1} R_q(n', d, i)$ yields

$$\begin{aligned} S_0 &= (1 - q^{-(d-n')}) \left(1 - \sum_{i=0}^{n'-1} R_q(n', d, i)\right) + \sum_{i=0}^{n'-1} (1 - q^{-(d-i)}) R_q(n', d, i) \\ &= (1 - q^{-(d-n')}) + \sum_{i=0}^{n'-1} [(1 - q^{-(d-i)}) - (1 - q^{-(d-n')})] R_q(n', d, i) \\ &= (1 - q^{-(d-n')}) + \sum_{i=0}^{n'-1} [q^{-(d-n')} - q^{-(d-i)}] R_q(n', d, i). \end{aligned}$$

The sum term in the last line is dominated by its $i = n' - 1$ term $q^{-(d-n')}(1 - q^{-1})R_q(n', d, n' - 1)$. Since by Lemma C.2 $R_q(n', d, n' - 1) = O(q^{-(d-n'+1)})$, this term is $O(q^{-(d-n')-(d-n'+1)}) = O(q^{-(2(d-n')+1)})$. Subsequent terms are of higher order in $1/q$. Letting $\Delta_0 = d - n'$, S_0 is hence bounded by

$$1 - q^{-\Delta_0} \leq S_0 \leq 1 - q^{-\Delta_0} + K_1 q^{-(2\Delta_0+1)}.$$

2. Case 2: $d \leq n'$

Let $\Delta_0 = n' - d$. Let $P_{\text{rank} < d} = \sum_{i=0}^{d-1} R_q(n', d, i)$. Note that by Lemma C.2 $P_{\text{rank} < d} = 1 - R_q(n', d, d) = O(q^{-(\Delta_0+1)})$.

Since $1 - q^{-j} \leq 1 - q^{-1}$ for $j \geq 1$, an immediate upper bound is $S_0 \leq (1 - q^{-1})P_{\text{rank} < d}$. To find a lower bound, consider the difference

$$\begin{aligned} (1 - q^{-1})P_{\text{rank} < d} - S_0 &= \sum_{i=0}^{d-1} [(1 - q^{-1}) - (1 - q^{-i+d})] R_q(n', d, i) \\ &= \sum_{i=0}^{d-1} (q^{-i+d} - q^{-1}) R_q(n', d, i) \end{aligned}$$

The term for $i = d - 1$ is zero. For $i \leq d - 2$, $q^{-i+d} - q^{-1}$ is negative and dominated by its first term (for $i = d - 2$): $(q^{-1} - q^{-2})R_q(n', d, d - 2)$. Since $R_q(n', d, d - 2) = O(q^{-2(\Delta_0+2)})$, that summand is of order $O(q^{-1} \cdot q^{-2(\Delta_0+2)}) = O(q^{-(2\Delta_0+5)})$.

So S_0 is bounded by

$$(1 - q^{-1})P_{\text{rank} < d} - K_2 q^{-(2\Delta_0+5)} \leq S_0 \leq (1 - q^{-1})P_{\text{rank} < d}$$

where $K_2 = O(1)$ is a positive constant.

□

Lemma C.6 (Loss Variance of the Random Linear Algorithm). *Let $\Delta_0 = d - n'$ and $P_{\text{rank} < d} = O(q^{-(\Delta_0+1)})$ (see Lemma C.2). The variance of the population loss $L_{n'}$ of the random parity algorithm can be bounded as follows.*

1. *If $d > n'$:*

$$(1 - 1/q)^2 \left(q^{-\Delta_0} (1 - q^{-\Delta_0}) - K'_1 q^{-(2\Delta_0+1)} \right) \leq \text{Var}(L_{n'}) \leq (1 - 1/q)^2 q^{-\Delta_0} (1 - q^{-\Delta_0})$$

Thus, $\text{Var}(L_{n'}) \approx (1 - 1/q)^2 q^{-\Delta_0} (1 - q^{-\Delta_0})$.

2. *If $d \leq n'$:*

$$\begin{aligned} \left(1 - \frac{1}{q}\right)^2 \left((1 - q^{-1}) P_{\text{rank} < d} (1 - (1 - q^{-1}) P_{\text{rank} < d}) - K'_2 q^{-(2\Delta_0+5)} \right) &\leq \text{Var}(L_{n'}) \\ &\leq \left(1 - \frac{1}{q}\right)^2 (1 - q^{-1}) P_{\text{rank} < d} (1 - (1 - q^{-1}) P_{\text{rank} < d}) \quad (\text{C.7}) \end{aligned}$$

Thus, $\text{Var}(L_{n'}) \approx (1 - 1/q)^3 P_{\text{rank} < d}$.

Proof. Recall that by Lemma C.4, the population loss can be written as $L_{n'} = (1 - 1/q) \cdot Z$ where $Z \sim \text{Ber}(\mathbb{P}(\mathcal{A}(S) \neq f))$. Hence the variance of the population loss $L_{n'}$ is given by

$$\text{Var}(L_{n'}) = \left(1 - \frac{1}{q}\right)^2 S_0(1 - S_0) \quad (\text{C.8})$$

where as before $S_0 = \mathbb{P}(\mathcal{A}(S) \neq f) = \sum_{i=0}^{m_0} (1 - q^{i-d}) \cdot R_q(n', d, i)$. To bound the variance, we therefore need bounds for $S_0(1 - S_0)$. For this we can reuse the bounds for S_0 derived in the proof of Lemma C.5.

Bounds for $S_0(1 - S_0)$

Let $\Delta_0 = |n' - d|$ and recall the bounds for S_0 derived in Lemma C.5. Let $f(x) = x(1 - x)$. This function is maximized at $x = 1/2$.

1. **Case 1:** $d > n'$ (so $\Delta_0 = d - n'$).

Here $S_0 = 1 - q^{-\Delta_0} + E_1$, where $0 \leq E_1 \leq K_1 q^{-(2\Delta_0+1)}$ where $K_1 = O(1)$. Since $q \geq 2$ and $\Delta_0 \geq 1$, $S_0 \geq 1 - q^{-1} \geq 1/2$. Thus, $f(S_0)$ is evaluated on the decreasing part of the parabola $f(x) = x(1 - x)$ (or at its maximum if $S_0 = 1/2$). The term $S_0(1 - S_0)$ is primarily determined by $1 - q^{-\Delta_0}$:

$$S_0(1 - S_0) = q^{-\Delta_0} (1 - q^{-\Delta_0}) - E_2 \quad (\text{C.9})$$

where $0 \leq E_2 \leq K'_1 q^{-(2\Delta_0+1)}$ for some $K'_1 = O(1)$. Hence the main term $q^{-\Delta_0} (1 - q^{-\Delta_0})$ serves as an upper bound. The error term E_2 contains $K_1 |1 - 2q^{-\Delta_0}| q^{-(2\Delta_0+1)}$ plus higher order terms.

2. **Case 2:** $d \leq n'$ (so $\Delta_0 = n' - d$).

Here $S_0 = (1 - q^{-1}) P_{\text{rank} < d} - E_2$, where $0 \leq E_2 \leq K_2 q^{-(2\Delta_0+5)}$ for $K_2 = O(1)$, and $P_{\text{rank} < d} = O(q^{-(\Delta_0+1)})$. Thus S_0 is small (i.e., $S_0 \ll 1/2$ for large q). The function $f(S_0)$ is evaluated on its increasing part. The term $S_0(1 - S_0)$ is primarily determined by $(1 - q^{-1}) P_{\text{rank} < d}$

$$S_0(1 - S_0) = (1 - q^{-1}) P_{\text{rank} < d} (1 - (1 - q^{-1}) P_{\text{rank} < d}) - E_3 \quad (\text{C.10})$$

where $0 \leq E_3 \leq K'_2 q^{-(2\Delta_0+5)}$ for some $K'_2 = O(1)$. The main term $(1 - q^{-1}) P_{\text{rank} < d} (1 - (1 - q^{-1}) P_{\text{rank} < d})$ serves as an upper bound. For large q , $S_0(1 - S_0) \approx (1 - q^{-1}) P_{\text{rank} < d}$.

Plugging the above bounds on $S_0(1 - S_0)$ into (C.8) concludes the proof. $\square$

Lemma C.7 (Expected Conditional Fold Variance). *For large q , the expected conditional variance of $\hat{L}_1$ given L_1 can be approximated as*

$$\mathbb{E}[\text{Var}(\hat{L}_1|L_1)] = O\left(\frac{1}{mq^{|n-d|+2} \cdot \mathbf{1}_{n \geq d}}\right)$$

Proof. First note that X having rank r implies that $\mathcal{A}_{lin}$ has loss $(1 - q^{r-d})(1 - 1/q)$ because according to Lemma C.3 it selects the ground truth w.p. q^{r-d} (incurring zero loss) and else it selects a linear function with loss $1 - 1/q$ (see Lemma 4.7).

Denoting $L_r = (1 - q^{r-d})(1 - 1/q)$, the expected conditional variance can be expressed as

$$\begin{aligned} \mathbb{E}[\text{Var}(\hat{L}_1|L_1)] &= \sum_{r=0}^{\min\{n,d\}} R_q(n, d, r) \frac{(1 - L_r)L_r}{m} \\ &\leq \frac{1}{m} \sum_{r=0}^{\min\{n,d\}} R_q(n, d, r) (q^{r-d} - q^{2(r-d)}) \end{aligned} \quad (\text{C.11})$$

where the first equality follows from the law of total expectation, since $m \cdot (\hat{L}_1|E_r) \sim \text{Bin}(m, L_r)$ conditioned on the event $E_r = \{\text{Rank}(X) = r\}$. Let $f(r) = q^{r-d} - q^{2(r-d)}$. We analyze $f(r)$:

- If $r = d$, then $f(d) = 0$.
- If $r < d$, let $s = d - r > 0$. Then $f(r) = q^{-s} - q^{-2s}$. Since $s \geq 1$ (as r and d are integers), $q^{-s} \geq q^{-2s}$ for $q \geq 1$. Thus, $f(r) = q^{-s}(1 - q^{-s})$. For large q , $1 - q^{-s}$ is close to 1. More formally, $f(r) = O(q^{-s}) = O(q^{-(d-r)})$.

The approximation of the sum relies on the asymptotic behavior of $R_q(n, d, r)$ for large q , see Lemma C.2. Let $m_0 = \min\{n, d\}$ be the maximum possible rank of the $n \times d$ matrix, and let $\Delta_0 = |n - d|$ be the absolute difference of its dimensions. Recall that

1. **Probability of Full Rank:** $R_q(n, d, m_0) = 1 - O(q^{-(\Delta_0+1)})$
2. **Probability of Specific Rank Deficiency:** for $j \geq 1$ (a rank deficiency of j), $R_q(n, d, m_0 - j) = O(q^{-j(\Delta_0+j)})$

We now analyze the sum by considering two cases for the relationship between n and d .

Case 1: $n \geq d$. In this scenario, the maximum rank is $m_0 = d$, and $\Delta_0 = n - d$. The sum runs from $r = 0$ to d . The term in the sum for $r = d$ is $R_q(n, d, d) \cdot f(d) = R_q(n, d, d) \cdot 0 = 0$. Thus, the sum is effectively over $r \leq d - 1$. Define $S_j = R_q(n, d, d - j) \cdot f(d - j)$. We have $f(d - j) = q^{-j} - q^{-2j} = O(q^{-j})$. Using Property 2 for R_q yields $R_q(n, d, d - j) = O(q^{-j(\Delta_0+j)}) = O(q^{-j(n-d+j)})$ hence $S_j = O(q^{-j(n-d+j)}) \cdot q^{-j} = O(q^{-j(n-d+j+1)})$.

It is easy to see that S_1 is the dominant term in the sum over $r < d$ and hence the sum $\sum_{r=0}^{d-1} R_q(n, d, r) f(r)$ is $O(q^{-(n'-d+2)})$. Consequently, by (C.11), $\mathbb{E}[\text{Var}(\hat{L}_1|L_1)] \leq \frac{1}{m} \cdot O(q^{-(n-d+2)}) = O\left(\frac{1}{mq^{n-d+2}}\right)$.

Case 2: $n < d$. In this scenario, the maximum rank is $m_0 = n$, and $\Delta_0 = d - n$. The sum runs from $r = 0$ to n with terms $S_j = R_q(n, d, n - j) \cdot g(n - j)$ where $g(n - j) = q^{(n-j)-d} - q^{2((n-j)-d)}$.

Since both $g(n - j)$ and $R_q(n, d, n - j)$ are decreasing in j , S_0 (the term for $r = n$) is the dominant term and the sum $\sum_{r=0}^n R_q(n, d, r) g(r)$ is $O(q^{-(d-n)})$. Consequently, $\mathbb{E}[\text{Var}(\hat{L}_1|L_1)] = \frac{1}{m} \cdot O(q^{-(d-n)}) = O\left(\frac{1}{mq^{d-n}}\right)$.

Combined Result: If $n \geq d$, the exponent of q in the denominator is $n - d + 2 = |n - d| + 2$. If $n < d$, the exponent of q in the denominator is $d - n = |n - d|$. Combining hence yields

$$\mathbb{E}[\text{Var}(\hat{L}_1|L_1)] = O\left(\frac{1}{mq^{|n-d|+2} \cdot \mathbf{1}_{n \geq d}}\right).$$

□

C.1. Proof for Theorem 4.8.

Proof. By Theorem 4.3 and Lemmas 2.4 it holds that

$$\begin{aligned} \text{MSE} &\leq (\bar{L}_n - \bar{L}_{n-m})^2 + (\sigma_{n-m} + \sigma_m)^2 + \frac{k-1}{k} \text{Cov}(\hat{L}_1, \hat{L}_2) + \frac{\bar{\sigma}^2}{n} + \frac{k-1}{k} \sigma_{n-m}^2 + 2\sigma_n \sqrt{\frac{\bar{\sigma}^2}{m}} \\ &= O\left((\bar{L}_n - \bar{L}_{n-m})^2 + \max(\{\sigma_{n-m}^2, \sigma_n^2\}) + \mathbb{E}[\text{Var}(\hat{L}_1|L_1)] + \sigma_n \sqrt{\mathbb{E}[\text{Var}(\hat{L}_1|L_1)]}\right) \end{aligned}$$

where we used the facts that $\frac{k-1}{k} \in (1/2, 1)$, $\bar{\sigma}^2 = m \cdot \mathbb{E}[\text{Var}(\hat{L}_1|L_1)]$, $\text{Cov}(\hat{L}_1, \hat{L}_2) \leq \text{Var}(\hat{L}_1) = \mathbb{E}[\text{Var}(\hat{L}_1|L_1)] + \text{Var}(L_1)$.

Similarly, for the lower bound,

$$\text{MSE} \geq (\bar{L}_n - \bar{L}_{n-m})^2 - \frac{k-1}{k} \sigma_{n-m}^2 - 2\sigma_n \sqrt{\mathbb{E}[\text{Var}(\hat{L}_1|L_1)]} = (\bar{L}_n - \bar{L}_{n-m})^2 + O(\sigma_{n-m}^2) + O\left(\sigma_n \sqrt{\mathbb{E}[\text{Var}(\hat{L}_1|L_1)]}\right)$$

where for the first inequality we simply ignored some of the positive terms of Theorem 4.3.

Let $\beta_{low} := (\bar{L}_{n_t} - \bar{L}_n)^2$. Recall $P(x, d, q) = C_P q^{-(x-d+1)}$ for $d \leq x$ [called $P_{rank < d}$ earlier].

Expected losses ($\bar{L}_x$) and loss variances (σ_x):

We recall from Lemma C.5 and Lemma C.6:

- If $d > x$: $\bar{L}_x = \Theta((1 - q^{-1})(1 - q^{-(d-x)}))$, $\sigma_x^2 = \Theta((1 - q^{-1})^2 q^{-(d-x)}(1 - q^{-(d-x)}))$. So $\bar{L}_x = O(1)$, $\bar{L}_x = 1 - O(1/q)$, $\sigma_x = O(q^{-(d-x)/2})$.
- If $d \leq x$: $\bar{L}_x = \Theta((1 - q^{-1})^2 P(x, d, q))$, $\sigma_x^2 = \Theta((1 - q^{-1})^3 P(x, d, q))$. So, $\bar{L}_x = O(q^{-(x-d+1)})$, $\sigma_x = O(q^{-(x-d+1)/2})$.

Now we are ready to put all pieces together. **MSE Bounds: Case 1: $n < d$ (Upper Bound)**

- Loss variances: we have $\sigma_n^2 = O(q^{-(d-n)})$, $\sigma_{n_t}^2 = O(q^{-(d-n_t)})$. So $\max(\{\sigma_{n-m}^2, \sigma_n^2\}) = O(q^{-(d-n)})$.
- Expected losses: $\bar{L}_n = \Theta((1 - q^{-1})(1 - q^{-(d-n)}))$, $\bar{L}_{n_t} = \Theta((1 - q^{-1})(1 - q^{-(d-n_t)}))$. Hence $\Delta \bar{L} = \Theta((1 - q^{-1})(q^{-(d-n)} - q^{-(d-n_t)})) = O(q^{-(d-n)}(1 - q^{-m}))$. Since $m \geq 1$, $q \geq 2$ it holds $(1 - q^{-m}) = O(1)$. So $(\Delta \bar{L})^2 = O(q^{-2(d-n)})$.
- Combined bound:

$$\begin{aligned} \text{MSE} &= O\left((\bar{L}_n - \bar{L}_{n-m})^2 + \max(\{\sigma_{n-m}^2, \sigma_n^2\}) + \mathbb{E}[\text{Var}(\hat{L}_1|L_1)] + \sigma_n \sqrt{\mathbb{E}[\text{Var}(\hat{L}_1|L_1)]}\right) \\ &= O(q^{-2(d-n)} + q^{-(d-n)} + m^{-1} q^{-(d-n)} + q^{-(d-n)} (m^{-1/2} q^{-\frac{d-n}{2}})) \\ &= O(q^{-(d-n)}) \end{aligned}$$

Case 2: $n \geq d$ and $n_t < d$ (Lower Bound)

- Loss variances: $\sigma_n^2 = O(q^{-(n-d+1)})$, $\sigma_{n_t}^2 = O(q^{-(d-n_t)})$.
- Expected losses: $\bar{L}_n = O(q^{-(n-d+1)})$. $\bar{L}_{n_t} = O(1)$. $\Delta \bar{L} = \bar{L}_{n_t} - \bar{L}_n = 1 - O(1/q) - O(q^{-(n-d+1)}) = 1 - O(1/q)$ and hence $(\Delta \bar{L})^2 = 1 - O(1/q)$.
- Combined bound: The loss stability term dominates all other terms, hence $\text{MSE} \geq 1 - O(1/q)$ which is $\Omega(1)$ since $q \geq 2$.

Case 3: $n_t \geq d$ (Upper Bound)

- Loss variances: $\sigma_n^2 = O(q^{-(n-d+1)})$, $\sigma_{n_t}^2 = O(q^{-(n_t-d+1)})$. So $\max(\{\sigma_{n-m}^2, \sigma_n^2\}) = O(q^{-(n_t-d+1)})$.

- Expected losses: $\bar{L}_n = \Theta(q^{-(n-d+1)})$, $\bar{L}_{n_t} = \Theta(q^{-(n_t-d+1)})$. Hence, $\Delta\bar{L} = O(q^{-(n_t-d+1)})$ and $(\Delta\bar{L})^2 = O(q^{-2(n_t-d+1)})$.
- Combined bound:

$$\text{MSE} = O(q^{-2(n_t-d+1)} + q^{-(n_t-d+1)} + m^{-1}q^{-(n-d+1)} + q^{-(n-d+1)}m^{-1/2}q^{-(n-d+1)/2}) = O(q^{-(n_t-d+1)}).$$

□

APPENDIX D. FOLD COVARIANCE OF THE SQUARE-WAVE ALGORITHM

D.1. Setup and Definitions. Throughout, let $N = n - 2m$. Let $W \sim \text{Bin}(m, 1/2)$, $a = s/\sqrt{m}$, and $p_w = P(W = w) = \binom{m}{w} 2^{-m}$. The function $f(a)$ is defined as

$$f(a) = \mathbb{E}_W \left[\left(\frac{W - m/2}{m} \right) \epsilon \left(a + \frac{W}{\sqrt{m}} \right) \right] = \sum_{w=0}^m p_w \left(\frac{w - m/2}{m} \right) (-1)^{\lfloor a + w/\sqrt{m} \rfloor} \quad (\text{D.1})$$

Let $\mu = m/2$ and $\sigma = \sqrt{m}/2$ be the mean and standard deviation of W . Let $g(w)$ be the PDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable:

$$g(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w-\mu)^2}{2\sigma^2}} = \sqrt{\frac{2}{\pi m}} e^{-\frac{2(w-m/2)^2}{m}} \quad (\text{D.2})$$

Let $h_g(w)$ be the Gaussian-weighted term:

$$h_g(w) = g(w) \left(\frac{w - m/2}{m} \right) \quad (\text{D.3})$$

Let $\psi(w) = \epsilon(a + w/\sqrt{m}) = (-1)^{\lfloor a + w/\sqrt{m} \rfloor}$. We seek a lower bound for $|f(a)|$.

Our first step is to greatly simplify the fold-covariance via a factorization.

D.2. Factorization of the Fold-Covariance.

Theorem D.1 (Factorization Identity). *It holds that*

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \mathbb{E}[f(S/\sqrt{m})^2]$$

where

$$f(a) := \mathbb{E}_W \left[\left(\frac{W}{m} - \frac{1}{2} \right) \epsilon \left(a + \frac{W}{\sqrt{m}} \right) \right].$$

Proof. It follows from the definitions of $\hat{L}_1, \hat{L}_2$ that

$$\begin{aligned} \hat{L}_1 - \frac{1}{2} &= \left(\frac{W_1}{m} - \frac{1}{2} \right) (-1)^{Y_1} = \left(\frac{W_1}{m} - \frac{1}{2} \right) \epsilon \left(\frac{S + W_2}{\sqrt{m}} \right), \\ \hat{L}_2 - \frac{1}{2} &= \left(\frac{W_2}{m} - \frac{1}{2} \right) \epsilon \left(\frac{S + W_1}{\sqrt{m}} \right). \end{aligned}$$

Define $W \sim \text{Bin}(m, \frac{1}{2})$ independent of S . Then, by conditioning on S and using independence of W_1, W_2 ,

$$\begin{aligned} \text{Cov}(\hat{L}_1, \hat{L}_2) &= \mathbb{E} \left[\left\{ \left(\frac{W_1}{m} - \frac{1}{2} \right) \epsilon \left(\frac{S + W_1}{\sqrt{m}} \right) \right\} \left\{ \left(\frac{W_2}{m} - \frac{1}{2} \right) \epsilon \left(\frac{S + W_2}{\sqrt{m}} \right) \right\} \right] \\ &= \mathbb{E}_S \left[\left(\mathbb{E}_W \left[\left(\frac{W}{m} - \frac{1}{2} \right) \epsilon \left(\frac{S+W}{\sqrt{m}} \right) \right] \right)^2 \right] \\ &= \mathbb{E}[f(S/\sqrt{m})^2], \end{aligned}$$

where

$$f(a) := \mathbb{E}_W \left[\left(\frac{W}{m} - \frac{1}{2} \right) \epsilon \left(a + \frac{W}{\sqrt{m}} \right) \right].$$

□

Proof Sketch. The proof in this section aims to find an asymptotic value for $\text{Cov}(\hat{L}_1, \hat{L}_2) = \mathbb{E}_S[f(a)^2]$.

Step 1: Simplify the Problem (Sum $\rightarrow$ Integral). We start with $f(a)$, a difficult discrete sum over a Binomial distribution. The first step is to get rid of the discrete sum and approximate it with a continuous integral, which is easier to manipulate.

$$f(a) = \underbrace{\sum_{w=0}^m p_w \left(\frac{w - m/2}{m} \right) \psi(w)}_{\text{Discrete, hard}} \xrightarrow{\text{Approximation}} \underbrace{\int_{-\infty}^{\infty} h_g(t) \psi(t) dt}_{\text{Continuous, easier}}$$

This is a standard analysis step. We replace the Binomial PMF p_w with a Gaussian PDF $g(w)$, and the sum with an integral.

Result: $f(a) = I_g + O(m^{-1})$, where I_g is the integral.

Step 2: Evaluate the Integral (Integral $\rightarrow$ New Sum). Now we must solve the integral I_g which is an integral of a smooth function $h_g(t)$ against a high-frequency square wave $\psi(t)$. Our function $h_g(t)$ is special: it's related to the derivative of a Gaussian ($h_g(t) \propto ue^{-u^2/2}$). The integral $\int ue^{-u^2/2} du$ is trivial. We split the integral at the jump points of $\psi(t)$. This turns the integral into a sum:

$$I_g = \sum_{r \in \mathbb{Z}} (-1)^r \int_{u_r}^{u_{r+1}} u \varphi(u) du = \sum_{r \in \mathbb{Z}} (-1)^r \left[-\varphi(u) \right]_{u_r}^{u_{r+1}}$$

This is a **telescoping sum** and simplifies the integral I_g into the much cleaner discrete sum.

Result: $I_g = \frac{1}{\sqrt{m}} \sum_{r \in \mathbb{Z}} (-1)^r e^{-2(r-C_m)^2}$.

Step 3: Analyze the New Sum (The First Fourier Tool: PSF). We have successfully simplified $f(a)$, but now we have a new problem: an alternating, shifted sum of a sampled Gaussian. The **Poisson Summation Formula (PSF)** is the precise tool for relating a sum of samples of a function to a sum of samples of its Fourier transform. The PSF converts our complicated, slowly-converging sum $\sum (-1)^r f(r - \delta)$ into a different sum that converges *extremely* fast. We apply the PSF to $f(x) = e^{-2x^2}$. The Fourier transform $\hat{f}(s)$ is also a Gaussian, $e^{-\pi^2 s^2/2}$, which decays very rapidly. The resulting sum in the frequency domain is:

$$\Theta(\delta) = \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi\delta)$$

This sum is dominated by its first term ($j = 0$). This is our main analytic expression for $f(a)$.

Result: $f(a) = \frac{1}{\sqrt{m}} \Theta(\delta(a)) + O(m^{-1})$.

Step 4: Analyze the Expectation (The Second Fourier Tool: Series). We are finally ready to tackle the main goal, $\text{Cov}(\hat{L}_1, \hat{L}_2) = \mathbb{E}_S[f(a)^2]$.

$$\text{Cov}(\hat{L}_1, \hat{L}_2) \approx \mathbb{E}_S \left[\left(\frac{1}{\sqrt{m}} \Theta(\delta_S) \right)^2 \right] = \frac{1}{m} \mathbb{E}_S[g(\delta_S)] \quad \text{where} \quad g(x) = \Theta(x)^2.$$

We now need to find the expectation of a periodic function $g(x)$ where its phase δ_S is a random variable. The most natural way to analyze a periodic function is to decompose it into its average value and its oscillations. This is the definition of a **Fourier Series**. We write $g(\delta_S) = c_0 + \sum_{l \neq 0} c_l e^{2\pi i l \delta_S}$. By linearity of expectation:

$$\mathbb{E}_S[g(\delta_S)] = c_0 + \sum_{l \neq 0} c_l \mathbb{E}_S[e^{2\pi i l \delta_S}]$$

The expectation on the r.h.s. is the characteristic function of S . We show that this term is very small for $l \neq 0$.

Result: $\mathbb{E}_S[g(\delta_S)] = c_0 + E_{\text{fourier}}$, where E_{fourier} is a small bias.

Step 5: Connecting the Two Fourier Tools. We know that if $g = \Theta \cdot \Theta$, then the Fourier coefficients of g (the c_l 's) are the *discrete convolution* of the Fourier coefficients of Θ (the $\hat{\Theta}_r$'s).

$$c_l = (\hat{\Theta} * \hat{\Theta})(l) = \sum_{r \in \mathbb{Z}} \hat{\Theta}(r) \hat{\Theta}(2l - r)$$

We use the coefficients $\hat{\Theta}_r$ we found in Step 3 (from the PSF) to compute the c_l we need for Step 4. This calculation gives us the final numerical values for our main term (c_0) and our bias terms (c_1 , etc.).

D.3. Technical Lemmas.

Lemma D.2 (Euler-Maclaurin Summation Formula). *Let $a, b \in \mathbb{R}$ such that $b - a \in \mathbb{N}^+$. Let $p \geq 2$ be an integer. Let f be a function with p continuous derivatives on $[a, b]$. Then,*

$$\begin{aligned} \sum_{i=a}^b f(i) &= \int_a^b f(t) dt + \frac{f(a) + f(b)}{2} + \sum_{j=2}^p \frac{b_j}{j!} \left(f^{(j-1)}(b) - f^{(j-1)}(a) \right) \\ &\quad - \int_a^b \frac{\bar{B}_p(1-t)}{p!} f^{(p)}(t) dt \end{aligned}$$

where b_j are the Bernoulli numbers and $\bar{B}_p(x) = B_p(\{x\})$ is the p -th periodic Bernoulli polynomial.

Proof. This is a standard result from numerical analysis. We also use the property $\bar{B}_p(1-t) = \bar{B}_p(-t) = (-1)^p \bar{B}_p(t)$. $\square$

We now apply this lemma to the sum $S_g = \sum_{w \in I_c} h_g(w) \psi(w)$, where $h_g(w)$ is smooth but $\psi(w) = (-1)^{\lfloor a+w/\sqrt{m} \rfloor}$ is a step function.

Lemma D.3 (Tail Bound). *Let $I_c = \{w \in \mathbb{Z} : |w - \mu| \leq \sqrt{2 \log m} \cdot \sigma\} = \{w : |w - m/2| \leq \sqrt{m \log m}\}$. Let $I_{tail} = \{0, \dots, m\} \setminus I_c$. Then $f(a) = \sum_{w \in I_c} p_w \left(\frac{w-\mu}{m} \right) \psi(w) + E_1$, where $|E_1| = O(m^{-2})$.*

Proof. The error E_1 is the sum over the tails:

$$\begin{aligned} |E_1| &= \left| \sum_{w \in I_{tail}} p_w \left(\frac{w-\mu}{m} \right) \psi(w) \right| \leq \sum_{w \in I_{tail}} p_w \left| \frac{w - m/2}{m} \right| \\ &\leq \sum_{w \in I_{tail}} p_w \left(\frac{m/2}{m} \right) = \frac{1}{2} P(W \in I_{tail}) = \frac{1}{2} P(|W - \mu| > \sqrt{2 \log m} \cdot \sigma) \end{aligned}$$

By Hoeffding's inequality, $P(|W - \mu| > t) \leq 2e^{-2t^2/m}$. Setting $t = \sqrt{m \log m}$, we have:

$$|E_1| \leq \frac{1}{2} \left(2e^{-2(m \log m)/m} \right) = e^{-2 \log m} = m^{-2}.$$

Thus $f(a) = \sum_{w \in I_c} p_w \left(\frac{w-\mu}{m} \right) \psi(w) + O(m^{-2})$. $\square$

Lemma D.4 (Gaussian Approximation). *Let $S_c = \sum_{w \in I_c} p_w \left(\frac{w-\mu}{m} \right) \psi(w)$. Then $S_c = \sum_{w \in I_c} h_g(w) \psi(w) + E_2$, where $|E_2| = O(m^{-3/2} \log m)$.*

Proof. We expand $S_c = \sum_{w \in I_c} (p_w - g(w)) \left(\frac{w-\mu}{m} \right) \psi(w) + \sum_{w \in I_c} h_g(w) \psi(w)$. By the local limit theorem of Lemma B.5 (for $W \sim \text{Bin}(m, 1/2)$), for $w \in I_c$:

$$p_w = g(w) + E_{LLT}$$

where $E_{LLT}(w) = O(m^{-3/2})$. The error E_2 is therefore bounded as

$$|E_2| = \left| \sum_{w \in I_c} (p_w - g(w)) \left(\frac{w-\mu}{m} \right) \psi(w) \right| \leq \sum_{w \in I_c} |E_{LLT}(w)| \left| \frac{w-\mu}{m} \right|$$

$$\leq \sum_{w \in I_c} O(m^{-3/2}) \left(\frac{\sqrt{m \log m}}{m} \right) = \sum_{w \in I_c} O(m^{-2} \sqrt{\log m})$$

The number of terms $|I_c|$ is $2\sqrt{m \log m} + 1 = O(\sqrt{m \log m})$. Hence,

$$|E_2| \leq O(\sqrt{m \log m}) \cdot O(m^{-2} \sqrt{\log m}) = O(m^{-3/2} \log m).$$

□

Lemma D.5 (Sum-to-Integral Approximation). *Let $m \in \mathbb{P}$, $\mu = m/2$, and define*

$$g(t) = \sqrt{\frac{2}{\pi m}} \exp\left(-\frac{2(t-\mu)^2}{m}\right), \quad h_g(t) = g(t) \frac{t-\mu}{m}.$$

Fix $a \in \mathbb{R}$ and let $\psi(t) = (-1)^{\lfloor a+t/\sqrt{m} \rfloor}$. Let

$$W_L = \left\lceil \mu - \sqrt{m \log m} \right\rceil, \quad W_R = \left\lfloor \mu + \sqrt{m \log m} \right\rfloor,$$

and define

$$S_g := \sum_{w=W_L}^{W_R} h_g(w) \psi(w), \quad I_g := \int_{W_L}^{W_R} h_g(t) \psi(t) dt.$$

Then

$$S_g = I_g + E_3, \quad |E_3| \leq C_m m^{-1} + C^* m^{-3/2} \sqrt{\log m},$$

for constants C_m, C^ . In particular, $|E_3| = O(1/m)$.*

Proof. Let the real jump points be $t_j := \sqrt{m}(j - a)$, so $\psi(t) = (-1)^j$ on $[t_j, t_{j+1})$. Define the integer blocks

$$A_j := \lceil t_j \rceil, \quad B_j := \lceil t_{j+1} \rceil - 1, \quad R_j := \{w \in \mathbb{Z} : A_j \leq w \leq B_j\},$$

so that $\psi(w) = (-1)^j$ on R_j and, crucially,

$$B_j + 1 = \lceil t_{j+1} \rceil = A_{j+1}.$$

Step 1 (Decomposition). Set

$$I'_g := \sum_j (-1)^j \int_{A_j}^{B_j} h_g(t) dt, \quad E_3 = S_g - I_g = (S_g - I'_g) + (I'_g - I_g) =: E_{\text{EM}} + E_{\text{BM}}.$$

(The sum is over all j such that R_j has a non-empty intersection with $[W_L, W_R]$).

Step 2 (Euler–Maclaurin on each block). The Euler–Maclaurin formula in Lemma D.2 (with $p = 2$) gives the error for a single block R_j as:

$$\text{Error}_j = \frac{h_g(A_j) + h_g(B_j)}{2} + \frac{B_2}{2} (h'_g(B_j) - h'_g(A_j)) - \int_{A_j}^{B_j} \frac{\bar{B}_2(1-t)}{2} h''_g(t) dt.$$

The total error E_{EM} is the alternating sum $E_{\text{EM}} = \sum_j (-1)^j \text{Error}_j$, where the sum runs over j from j_L to j_R . Let $P_j = j_R - j_L + 1 = \Theta(\sqrt{\log m})$ be the number of blocks. We analyze the three parts of this sum separately.

- **Remainder Integral Term (h''_g):** Let $E_{\text{EM},3} = -\sum_j (-1)^j \int_{A_j}^{B_j} \frac{\bar{B}_2(1-t)}{2} h''_g(t) dt$. We bound its magnitude by the integral over the entire window $I_c = [W_L, W_R]$. Since $|\bar{B}_2(x)|$ is bounded by a constant C_B :

$$|E_{\text{EM},3}| \leq \sum_j \int_{A_j}^{B_j} \left| \frac{\bar{B}_2(1-t)}{2} \right| |h''_g(t)| dt \leq C_B \int_{W_L}^{W_R} |h''_g(t)| dt$$

$$\begin{aligned}
&\leq C_B \cdot (\text{Length of } I_c) \cdot \|h_g''\|_\infty \\
&\leq C_B \cdot O(\sqrt{m \log m}) \cdot O(m^{-2}) = O(m^{-3/2} \sqrt{\log m}).
\end{aligned}$$

- **Boundary Term (h_g):** Let $E_{EM,1} = \frac{1}{2} \sum_j (-1)^j (h_g(A_j) + h_g(B_j))$. We rewrite this sum by grouping terms adjacent to the block boundaries, using $A_{j+1} = B_j + 1$.

$$\begin{aligned}
E_{EM,1} &= \frac{1}{2} \left[\sum_{j=j_L}^{j_R} (-1)^j h_g(A_j) + \sum_{j=j_L}^{j_R} (-1)^j h_g(B_j) \right] \\
&= \frac{1}{2} \left[(-1)^{j_L} h_g(A_{j_L}) + \sum_{j=j_L+1}^{j_R} (-1)^j h_g(A_j) \right. \\
&\quad \left. + \sum_{j=j_L}^{j_R-1} (-1)^j h_g(B_j) + (-1)^{j_R} h_g(B_{j_R}) \right]
\end{aligned}$$

We group the two inner sums by re-indexing the first one ($j \rightarrow j+1$):

$$\sum_{j=j_L}^{j_R-1} (-1)^{j+1} h_g(A_{j+1}) + \sum_{j=j_L}^{j_R-1} (-1)^j h_g(B_j) = \sum_{j=j_L}^{j_R-1} (-1)^j (h_g(B_j) - h_g(A_{j+1}))$$

Using $A_{j+1} = B_j + 1$, the inner sum becomes:

$$\sum_{j=j_L}^{j_R-1} (-1)^j (h_g(B_j) - h_g(B_j + 1)) = \sum_{j=j_L}^{j_R-1} (-1)^{j+1} (h_g(B_j + 1) - h_g(B_j))$$

By the Mean Value Theorem, $h_g(B_j + 1) - h_g(B_j) = h'_g(\xi_j)$ for some $\xi_j \in (B_j, B_j + 1)$. Thus, the exact expression is:

$$E_{EM,1} = \frac{1}{2} \left[(-1)^{j_L} h_g(A_{j_L}) + (-1)^{j_R} h_g(B_{j_R}) + \sum_{j=j_L}^{j_R-1} (-1)^{j+1} h'_g(\xi_j) \right]$$

The two endpoint terms $h_g(A_{j_L})$ and $h_g(B_{j_R})$ are located near W_L and W_R , where $|h_g(t)| = O(m^{-3} \sqrt{\log m})$ and are negligible. For the main sum, we use the triangle inequality:

$$\left| \frac{1}{2} \sum_{j=j_L}^{j_R-1} (-1)^{j+1} h'_g(\xi_j) \right| \leq \frac{1}{2} \sum_{j=j_L}^{j_R-1} |h'_g(\xi_j)| \leq \frac{1}{2} (P_j - 1) \|h'_g\|_\infty$$

Since $P_j = O(\sqrt{\log m})$ and $\|h'_g\|_\infty = O(m^{-3/2})$, this sum is $O(m^{-3/2} \sqrt{\log m})$. The total contribution is $|E_{EM,1}| = O(m^{-3/2} \sqrt{\log m})$.

- **Boundary Term (h'_g):** Let $E_{EM,2} = \frac{B_2}{2} \sum_j (-1)^j (h'_g(B_j) - h'_g(A_j))$. A similar telescoping argument applies, replacing h_g with h'_g and h'_g with h''_g .

$$E_{EM,2} = \frac{B_2}{2} \left[(\text{Endpoints}) + \sum_{j=j_L}^{j_R-1} (-1)^{j+1} h''_g(\zeta_j) \right]$$

The endpoint terms are negligible. The main sum is bounded by:

$$\left| \frac{B_2}{2} \sum_{j=j_L}^{j_R-1} (-1)^{j+1} h''_g(\zeta_j) \right| \leq \frac{|B_2|}{2} (P_j - 1) \|h''_g\|_\infty$$

This is $O(\sqrt{\log m}) \cdot O(m^{-2}) = O(m^{-2} \sqrt{\log m})$.

Combining these, the total Euler-Maclaurin error E_{EM} is dominated by the remainder integral and h_g boundary terms:

$$|E_{EM}| \leq |E_{EM,1}| + |E_{EM,2}| + |E_{EM,3}| \leq O(m^{-3/2} \sqrt{\log m}).$$

Step 3 (Boundary mismatch). Write

$$E_{BM} = \sum_j (-1)^j \left[\int_{A_j}^{B_j} h_g - \int_{t_j}^{t_{j+1}} h_g \right] = \sum_j (-1)^j \left[- \int_{t_j}^{A_j} h_g - \int_{B_j}^{t_{j+1}} h_g \right].$$

We pair the "tail" of block j with the "head" of block $j + 1$.

- Tail of j : $T_j = (-1)^j \left(- \int_{B_j}^{t_{j+1}} h_g \right)$
- Head of $j + 1$: $H_{j+1} = (-1)^{j+1} \left(- \int_{t_{j+1}}^{A_{j+1}} h_g \right)$

Their sum is:

$$T_j + H_{j+1} = (-1)^{j+1} \left(\int_{B_j}^{t_{j+1}} h_g - \int_{t_{j+1}}^{A_{j+1}} h_g \right) = (-1)^{j+1} \int_{B_j}^{A_{j+1}} h_g$$

Using $A_{j+1} = B_j + 1$, this becomes:

$$T_j + H_{j+1} = (-1)^{j+1} \int_{B_j}^{B_j+1} h_g(t) dt.$$

The full sum E_{BM} thus collapses to a sum over these unit intervals, plus the two un-paired residuals at the global endpoints (W_L, W_R):

$$E_{\text{BM}} = \sum_j (-1)^{j+1} \int_{B_j}^{B_j+1} h_g(t) dt + (\text{Endpoint residuals})$$

Let $s_j := \int_{B_j}^{B_j+1} h_g(t) dt$. The sum E_{BM} collapses to $S = \sum_j (-1)^{j+1} s_j$ plus the two un-paired endpoint residuals. These residuals are integrals of length < 1 at $t \approx \mu \pm \sqrt{m \log m}$, where $|h_g(t)| = O(m^{-3} \sqrt{\log m})$. Their contribution is thus negligible relative to $O(1/m)$.

We now bound the alternating sum S . Let $s_j := \int_{B_j}^{B_j+1} h_g(t) dt$. We need to bound the alternating sum $S = \sum_j (-1)^{j+1} s_j$.

The total variation of the sequence $\{s_j\}$ is $\sum_j |s_{j+1} - s_j|$. This sequence samples the smooth function $h_g(t)$, so its total variation is bounded by the total variation of $h_g(t)$ itself, $\int |h'_g(t)| dt$.

$$\begin{aligned} \int_{-\infty}^{\infty} |h'_g(t)| dt &= \int_{-\infty}^{\infty} \left| \frac{2}{\sqrt{2\pi} m^{3/2}} (1 - u^2) e^{-u^2/2} \right| \frac{dt}{du} du \quad (u = \frac{2(t - \mu)}{\sqrt{m}}) \\ &= \int_{-\infty}^{\infty} \frac{2}{\sqrt{2\pi} m^{3/2}} |1 - u^2| e^{-u^2/2} \left(\frac{\sqrt{m}}{2} \right) du \\ &= \frac{1}{\sqrt{2\pi} m} \int_{-\infty}^{\infty} |1 - u^2| e^{-u^2/2} du = \frac{C_{\text{int}}}{m} = O(1/m). \end{aligned}$$

For an alternating sum $S = \sum (-1)^j s_j$, its magnitude is bounded by the sum of its variations, $|S| \leq \sum |s_{j+1} - s_j| + |s_{\text{start}}| + |s_{\text{end}}|$. Since the total variation is $O(1/m)$ and the endpoint terms s_j are $O(m^{-M})$, we have $|S| = O(1/m)$.

Hence, $|E_{\text{BM}}| = O(1/m)$.

Conclusion. The total error is

$$|E_3| \leq |E_{\text{EM}}| + |E_{\text{BM}}| \leq O(m^{-3/2} \sqrt{\log m}) + O(1/m) = O(1/m).$$

□

Lemma D.6 (Alternating Poisson Summation with Shift). *Let $f : \mathbb{R} \rightarrow \mathbb{C}$ be an **even function** ($f(x) = f(-x)$) that is continuous, integrable, and decays sufficiently fast (e.g., $f \in \mathcal{S}(\mathbb{R})$, the Schwartz space). Let $\delta \in \mathbb{R}$ be a shift.*

Let the Fourier transform be defined as $\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i \xi x} dx$. Then the following identity holds:

$$\sum_{N \in \mathbb{Z}} (-1)^N f(N - \delta) = \sum_{N \in \mathbb{Z}} \hat{f}(N + 1/2) e^{2\pi i (N + 1/2) \delta}$$

Proof. We begin with the standard Poisson Summation Formula (PSF), which states that for a suitable function $g(x)$:

$$\sum_{N \in \mathbb{Z}} g(N) = \sum_{m \in \mathbb{Z}} \hat{g}(m)$$

To evaluate the sum $S = \sum_{N \in \mathbb{Z}} (-1)^N f(N - \delta)$, we define an auxiliary function $g(x)$. Using the identity $(-1)^N = e^{i\pi N}$, we set:

$$g(x) = e^{i\pi x} f(x - \delta)$$

Next, we compute the Fourier transform $\hat{g}(m)$ of $g(x)$:

$$\begin{aligned} \hat{g}(m) &= \int_{-\infty}^{\infty} g(x) e^{-2\pi i m x} dx \\ &= \int_{-\infty}^{\infty} e^{i\pi x} f(x - \delta) e^{-2\pi i m x} dx \\ &= \int_{-\infty}^{\infty} f(x - \delta) e^{-2\pi i x(m - 1/2)} dx \end{aligned}$$

We apply the substitution $u = x - \delta$, which implies $x = u + \delta$ and $du = dx$.

$$\begin{aligned} \hat{g}(m) &= \int_{-\infty}^{\infty} f(u) e^{-2\pi i (u + \delta)(m - 1/2)} du \\ &= e^{-2\pi i \delta(m - 1/2)} \int_{-\infty}^{\infty} f(u) e^{-2\pi i u(m - 1/2)} du \\ &= e^{-2\pi i \delta(m - 1/2)} \cdot \hat{f}(m - 1/2) \end{aligned}$$

Substituting this result back into the standard PSF, we have:

$$\sum_{N \in \mathbb{Z}} (-1)^N f(N - \delta) = \sum_{m \in \mathbb{Z}} \hat{f}(m - 1/2) e^{-2\pi i \delta(m - 1/2)}$$

This identity holds for any suitable function f . To arrive at the form stated in the lemma, we now apply the assumption that f is an even function.

If $f(x)$ is even, its Fourier transform $\hat{f}(\xi)$ is also even, i.e., $\hat{f}(\xi) = \hat{f}(-\xi)$.

We re-index the sum on the right-hand side. Let $m = -p$, where $p \in \mathbb{Z}$.

$$\begin{aligned} \sum_{m \in \mathbb{Z}} \hat{f}(m - 1/2) e^{-2\pi i \delta(m - 1/2)} &= \sum_{p \in \mathbb{Z}} \hat{f}(-p - 1/2) e^{-2\pi i \delta(-p - 1/2)} \\ &= \sum_{p \in \mathbb{Z}} \hat{f}(-(p + 1/2)) e^{2\pi i \delta(p + 1/2)} \end{aligned}$$

Applying the even property $\hat{f}(-(p + 1/2)) = \hat{f}(p + 1/2)$, the sum becomes:

$$\sum_{p \in \mathbb{Z}} \hat{f}(p + 1/2) e^{2\pi i \delta(p + 1/2)}$$

Finally, relabeling the summation index p to N yields the desired result:

$$\sum_{N \in \mathbb{Z}} (-1)^N f(N - \delta) = \sum_{N \in \mathbb{Z}} \hat{f}(N + 1/2) e^{2\pi i (N + 1/2) \delta}$$

□

Lemma D.7 (Periodicity of Θ and Θ^2). *The function $\Theta(x) = \sum_{j=0}^{\infty} C_j \cos((2j + 1)\pi x)$ has period 2 and $\Theta(x)^2$ has period 1.*

Proof.

$$\begin{aligned}
\Theta(x+2) &= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi(x+2)) \\
&= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x + (2j+1)2\pi) \\
&\text{Since } \cos(y+2\pi J) = \cos(y) \text{ for any integer } J = (2j+1) : \\
&= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x) = \Theta(x)
\end{aligned}$$

$$\begin{aligned}
\Theta(x+1) &= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi(x+1)) \\
&= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x + (2j+1)\pi)
\end{aligned}$$

Using $\cos(y + \pi J) = (-1)^J \cos(y)$, and $J = (2j+1)$ is always odd:

$$\begin{aligned}
&= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x + \pi(2j+1)) \\
&= \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x + \pi) \\
&= \sum_{j=0}^{\infty} -C_j \cos((2j+1)\pi x) = -\Theta(x)
\end{aligned}$$

Therefore, $g(x+1) = \Theta(x+1)^2 = (-\Theta(x))^2 = \Theta(x)^2 = g(x)$ and $g(x)$ has a period 1 Fourier series $g(x) = \sum_{l \in \mathbb{Z}} c_l e^{2\pi i l x}$. $\square$

Lemma D.8 (Derivation of Fourier Coefficients $\hat{\Theta}(p)$).

$$\Theta(x) = \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x)$$

has fourier coefficients

$$\hat{\Theta}(p) = \begin{cases} \frac{1}{2} C_{(|p|-1)/2} & \text{if } p \text{ is odd} \\ 0 & \text{if } p \text{ is even} \end{cases}$$

Proof. We want to find the coefficients $\hat{\Theta}(p)$ for the complex Fourier series of $\Theta(x)$ with period 2 (by Lemma D.7).

$$\Theta(x) = \sum_{p \in \mathbb{Z}} \hat{\Theta}(p) e^{i\pi p x}$$

We start with the definition of $\Theta(x)$:

$$\Theta(x) = \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi x)$$

Using Euler's formula, $\cos(\theta) = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$:

$$\begin{aligned}
\Theta(x) &= \sum_{j=0}^{\infty} C_j \left[\frac{1}{2} (e^{i(2j+1)\pi x} + e^{-i(2j+1)\pi x}) \right] \\
&= \sum_{j=0}^{\infty} \frac{C_j}{2} e^{i(2j+1)\pi x} + \sum_{j=0}^{\infty} \frac{C_j}{2} e^{-i(2j+1)\pi x}
\end{aligned}$$

We now compare this expression, term by term, to the target series $\sum_{p \in \mathbb{Z}} \hat{\Theta}(p) e^{i\pi p x}$.

Case 1: p is even. The terms in our expanded sum only involve exponents $i(2j+1)\pi x$ and $-i(2j+1)\pi x$. Since $(2j+1)$ is always odd, there are no even values for p . Therefore,

$$\hat{\Theta}(p) = 0 \quad (\text{if } p \text{ is even})$$

Case 2: p is odd and positive. An odd, positive p must be of the form $p = 2j+1$ for some $j \geq 0$. We look at the first sum: $\sum_{j=0}^{\infty} \frac{C_j}{2} e^{i(2j+1)\pi x}$. By matching the exponent $p = 2j+1$, we find the coefficient:

$$\hat{\Theta}(p) = \frac{C_j}{2}$$

Since $j = (p-1)/2$, and $p = |p|$ for $p > 0$, we can write this as:

$$\hat{\Theta}(p) = \frac{1}{2} C_{(|p|-1)/2} \quad (\text{if } p \text{ is odd, positive})$$

Case 3: p is odd and negative. An odd, negative p must be of the form $p = -(2j+1)$ for some $j \geq 0$. We look at the second sum: $\sum_{j=0}^{\infty} \frac{C_j}{2} e^{-i(2j+1)\pi x}$. By matching the exponent $p = -(2j+1)$, we find the coefficient:

$$\hat{\Theta}(p) = \frac{C_j}{2}$$

Since $j = (-p-1)/2 = (|p|-1)/2$, we can write this as:

$$\hat{\Theta}(p) = \frac{1}{2} C_{(|p|-1)/2} \quad (\text{if } p \text{ is odd, negative})$$

Conclusion: Combining all cases, the coefficient is non-zero only if p is odd, and in that case, it is $\frac{1}{2} C_{(|p|-1)/2}$. This gives:

$$\hat{\Theta}(p) = \begin{cases} \frac{1}{2} C_{(|p|-1)/2} & \text{if } p \text{ is odd} \\ 0 & \text{if } p \text{ is even} \end{cases}$$

□

D.4. Main Results.

Theorem D.9 (Integral Evaluation and Final Bound). *Let $\delta = \{a + \sqrt{m}/2\}$ be the fractional part. Then*

$$f(a) = \frac{(-1)^{\lfloor a + \sqrt{m}/2 \rfloor}}{\sqrt{m}} \sum_{j=0}^{\infty} e^{-\frac{\pi^2(2j+1)^2}{8}} \cos((2j+1)\pi\delta) + E_{total} \quad (\text{D.4})$$

where $|E_{total}| = O(m^{-1})$.

Proof. Combining Lemmas Theorems D.3 to D.5, we have

$$f(a) = \int_{I_c} h_g(t) \psi(t) dt + E_{total}$$

where $I_c = [m/2 - \sqrt{m \log m}, m/2 + \sqrt{m \log m}]$ and the total error from the sum approximation is $|E_{total}| = O(m^{-1})$.

We now show that the integral over I_c can be extended to all of $\mathbb{R}$, incurring a negligible error. Let $I_{tail} = \mathbb{R} \setminus I_c$. The error from extending the integral is E_{tail_int} :

$$E_{tail_int} := \int_{I_{tail}} h_g(t) \psi(t) dt$$

We can bound its magnitude:

$$\begin{aligned} |E_{tail_int}| &\leq \int_{I_{tail}} |h_g(t)| |\psi(t)| dt = \int_{I_{tail}} g(t) \frac{|t - m/2|}{m} dt \\ &= \int_{|t - m/2| > \sqrt{m \log m}} \sqrt{\frac{2}{\pi m}} e^{-\frac{2(t - m/2)^2}{m}} \frac{|t - m/2|}{m} dt \end{aligned}$$

We substitute $u = \frac{2(t - m/2)}{\sqrt{m}}$, so $t - m/2 = u\sqrt{m}/2$ and $dt = (\sqrt{m}/2)du$. The integration region $|t - m/2| > \sqrt{m \log m}$ becomes $|u| > 2\sqrt{\log m}$.

$$\begin{aligned} |E_{tail_int}| &\leq \int_{|u| > 2\sqrt{\log m}} \sqrt{\frac{2}{\pi m}} e^{-u^2/2} \frac{|u\sqrt{m}/2|}{m} \left(\frac{\sqrt{m}}{2} du \right) \\ &= \int_{|u| > 2\sqrt{\log m}} \sqrt{\frac{2}{\pi m}} e^{-u^2/2} \frac{|u|m}{4m} du = \frac{1}{2\sqrt{m}} \int_{|u| > 2\sqrt{\log m}} |u| \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \\ &= \frac{1}{2\sqrt{m}} \int_{|u| > 2\sqrt{\log m}} |u| \varphi(u) du = \frac{1}{\sqrt{m}} \int_{2\sqrt{\log m}}^{\infty} u \varphi(u) du \\ &= \frac{1}{\sqrt{m}} [-\varphi(u)]_{2\sqrt{\log m}}^{\infty} = \frac{1}{\sqrt{m}} \left(0 - \left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{(2\sqrt{\log m})^2}{2}} \right) \right) \\ &= \frac{1}{\sqrt{2\pi m}} e^{-2 \log m} = \frac{1}{\sqrt{2\pi m}} m^{-2} = O(m^{-5/2}) \end{aligned}$$

This error $E_{tail_int} = O(m^{-5/2})$ is asymptotically smaller than E_{total} and is therefore absorbed by the latter.

We have

- $w = m/2 + u\sqrt{m}/2 \implies dw = (\sqrt{m}/2)du$
- $g(w)dw = \varphi(u)du$, where $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
- $\frac{w - m/2}{m} = \frac{u\sqrt{m}/2}{m} = \frac{u}{2\sqrt{m}}$
- $\psi(w) = \epsilon(a + \frac{m/2 + u\sqrt{m}/2}{\sqrt{m}}) = \epsilon(a + \sqrt{m}/2 + u/2)$

Let $C_m = a + \sqrt{m}/2$.

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \varphi(u) \left(\frac{u}{2\sqrt{m}} \right) \epsilon(C_m + u/2) du \\ &= \frac{1}{2\sqrt{m}} \int_{-\infty}^{\infty} u \varphi(u) (-1)^{\lfloor C_m + u/2 \rfloor} du \end{aligned}$$

The sign changes at $u_p = 2(p - C_m)$ for $p \in \mathbb{Z}$. Together with using $\int u \varphi(u) du = -\varphi(u) + \text{constant}$:

$$\begin{aligned} I &= \frac{1}{2\sqrt{m}} \sum_{p=-\infty}^{\infty} (-1)^p \int_{u_p}^{u_{p+1}} u \varphi(u) du \\ &= \frac{1}{2\sqrt{m}} \sum_{p=-\infty}^{\infty} (-1)^p [-\varphi(u)]_{u_p}^{u_{p+1}} = \frac{1}{2\sqrt{m}} \sum_{p=-\infty}^{\infty} (-1)^p (\varphi(u_p) - \varphi(u_{p+1})) \end{aligned}$$

This is a telescoping sum which simplifies to:

$$I = \frac{1}{2\sqrt{m}} \left(2 \sum_{p=-\infty}^{\infty} (-1)^p \varphi(u_p) \right) = \frac{1}{\sqrt{m}} \sum_{p=-\infty}^{\infty} (-1)^p \varphi(2(p - C_m))$$

$$= \frac{1}{\sqrt{2\pi m}} \sum_{p=-\infty}^{\infty} (-1)^p e^{-\frac{(2(p-C_m))^2}{2}} = \frac{1}{\sqrt{2\pi m}} \sum_{p=-\infty}^{\infty} (-1)^p e^{-2(p-C_m)^2}$$

Let $p_0 = \lfloor C_m \rfloor$ and $\delta = \{C_m\} = C_m - p_0$. Let $j = p - p_0$.

$$I = \frac{1}{\sqrt{2\pi m}} \sum_{j=-\infty}^{\infty} (-1)^{j+p_0} e^{-2(j-\delta)^2} = \frac{(-1)^{p_0}}{\sqrt{2\pi m}} \sum_{j=-\infty}^{\infty} (-1)^j e^{-2(j-\delta)^2}$$

Let $\Theta(\delta) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2(j-\delta)^2}$. By Lemma D.6, $\sum_{N \in \mathbb{Z}} (-1)^N f(N-\delta) = \sum_{N \in \mathbb{Z}} \hat{f}(N+1/2) e^{2\pi i(N+1/2)\delta}$. Here $f(x) = e^{-2x^2}$. Its Fourier transform is $\hat{f}(s) = \int_{\mathbb{R}} e^{-2x^2} e^{-2\pi i s x} dx = \sqrt{\frac{\pi}{2}} e^{-\pi^2 s^2/2}$. We get that

$$\begin{aligned} \Theta(\delta) &= \sum_{N=-\infty}^{\infty} \sqrt{\frac{\pi}{2}} e^{-\frac{\pi^2(N+1/2)^2}{2}} e^{i\pi(2N+1)\delta} \\ &= \sqrt{\frac{\pi}{2}} \sum_{N=-\infty}^{\infty} e^{-\frac{\pi^2(2N+1)^2}{8}} (\cos(\pi(2N+1)\delta) + i \sin(\pi(2N+1)\delta)) \end{aligned}$$

The sine terms cancel (e.g., $N = 0$ and $N = -1$). The cosine terms are even, so we sum over $j \geq 0$:

$$\Theta(\delta) = \sqrt{\frac{\pi}{2}} \cdot 2 \sum_{j=0}^{\infty} e^{-\frac{\pi^2(2j+1)^2}{8}} \cos(\pi(2j+1)\delta) = \sqrt{2\pi} \sum_{j=0}^{\infty} e^{-\frac{\pi^2(2j+1)^2}{8}} \cos(\pi(2j+1)\delta)$$

Substituting $\Theta(\delta)$ back into the expression for I :

$$\begin{aligned} I &= \frac{(-1)^{p_0}}{\sqrt{2\pi m}} \left[\sqrt{2\pi} \sum_{j=0}^{\infty} e^{-\frac{\pi^2(2j+1)^2}{8}} \cos(\pi(2j+1)\delta) \right] \\ &= \frac{(-1)^{p_0}}{\sqrt{m}} \sum_{j=0}^{\infty} e^{-\frac{\pi^2(2j+1)^2}{8}} \cos(\pi(2j+1)\delta) \end{aligned}$$

This proves the theorem. $\square$

Lemma D.10 (Asymptotic for the Magnitude of $f(a)$). Let $C_j = e^{-\pi^2(2j+1)^2/8}$ and $\delta = \{a + \sqrt{m}/2\}$. The magnitude of $f(a)$ has the asymptotic form

$$|f(a)| = \frac{C_0 |\cos(\pi\delta)|}{\sqrt{m}} + E_{mag}(a)$$

where the error term $E_{mag}(a)$ is bounded by

$$|E_{mag}(a)| \leq \frac{C_{tail}}{\sqrt{m}} + O(m^{-1}).$$

Here, $C_{tail} = \sum_{j=1}^{\infty} C_j$ is a small constant ($C_0 \approx 0.2917$ and $C_{tail} \approx 1.5 \times 10^{-5}$).

Proof. From Theorem D.9, we have the asymptotic equality:

$$f(a) = I_0(a) + I_{tail}(a) + E_{total}$$

where

$$\begin{aligned} I_0(a) &= \frac{(-1)^{\lfloor a+\sqrt{m}/2 \rfloor}}{\sqrt{m}} C_0 \cos(\pi\delta) \\ I_{tail}(a) &= \frac{(-1)^{\lfloor a+\sqrt{m}/2 \rfloor}}{\sqrt{m}} \sum_{j=1}^{\infty} C_j \cos((2j+1)\pi\delta) \\ |E_{total}| &= O(m^{-1}) \end{aligned}$$

We want to find the error $E_{mag}(a) = |f(a)| - |I_0(a)|$. By the reverse triangle inequality, $||X+Y| - |X|| \leq |Y|$, we can set $X = I_0(a)$ and $Y = I_{tail}(a) + E_{total}$.

$$\begin{aligned} |E_{mag}(a)| &= ||I_0(a) + (I_{tail}(a) + E_{total})| - |I_0(a)|| \\ &\leq |I_{tail}(a) + E_{total}| \\ &\leq |I_{tail}(a)| + |E_{total}| \end{aligned}$$

We now bound the two error components:

1. $|I_{tail}(a)| \leq \frac{1}{\sqrt{m}} \left| \sum_{j=1}^{\infty} C_j \cos((2j+1)\pi\delta) \right| \leq \frac{1}{\sqrt{m}} \sum_{j=1}^{\infty} C_j =: \frac{C_{tail}}{\sqrt{m}}$
2. $|E_{total}| = O(m^{-1})$

□

So far we have obtained an approximation for $a = S/\sqrt{m}$:

$$f(a) = \frac{1}{\sqrt{m}} \Theta(\delta(a)) + E(a) \quad (\text{D.5})$$

where $\delta(a) = \{a + \sqrt{m}/2\}$ and $E(a)$ is an error term such that $|E(a)| = O(m^{-1})$ uniformly. Next, we want to find a lower bound for $\text{Cov}(\hat{L}_1, \hat{L}_2) = \mathbb{E}_S[f(S/\sqrt{m})^2]$, where $S \sim \text{Bin}(N, 1/2)$ and N is a positive integer multiple of m .

Lemma D.11 (Expansion of the Expectation). *The expected value $\text{Cov}(\hat{L}_1, \hat{L}_2)$ is given by*

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{1}{m} \mathbb{E}_S [\Theta(\delta_S)^2] + O(m^{-3/2})$$

where $\delta_S = \{S/\sqrt{m} + \sqrt{m}/2\}$.

Proof. We square the expression for $f(a)$ and take the expectation over S :

$$\begin{aligned} \text{Cov}(\hat{L}_1, \hat{L}_2) &= \mathbb{E}_S \left[\left(\frac{1}{\sqrt{m}} \Theta(\delta_S) + E_S \right)^2 \right] \\ &= \mathbb{E}_S \left[\frac{1}{m} \Theta(\delta_S)^2 + \frac{2}{\sqrt{m}} \Theta(\delta_S) E_S + E_S^2 \right] \\ &= \frac{1}{m} \mathbb{E}_S [\Theta(\delta_S)^2] + \frac{2}{\sqrt{m}} \mathbb{E}_S [\Theta(\delta_S) E_S] + \mathbb{E}_S [E_S^2] \end{aligned}$$

We bound the error terms using the uniform bounds $|\Theta(\delta_S)| \leq C_\Theta$ and $|E_S| \leq C m^{-1}$ for some $C > 0$.

Cross Term:

$$\begin{aligned} \left| \frac{2}{\sqrt{m}} \mathbb{E}_S [\Theta(\delta_S) E_S] \right| &\leq \frac{2}{\sqrt{m}} \mathbb{E}_S [|\Theta(\delta_S)| |E_S|] \\ &\leq \frac{2}{\sqrt{m}} \mathbb{E}_S [C_\Theta \cdot (C m^{-1})] \\ &= O(m^{-3/2}) \end{aligned}$$

Squared Error Term:

$$\begin{aligned} |\mathbb{E}_S [E_S^2]| &\leq \mathbb{E}_S [|E_S|^2] \leq \mathbb{E}_S [(C m^{-1})^2] \\ &= C^2 m^{-2} = O(m^{-2}) \end{aligned}$$

Since $O(m^{-2})$ is asymptotically smaller than $O(m^{-3/2})$, the dominant error is $O(m^{-3/2})$. □

It remains to bound $\mathbb{E}_S[\Theta(\delta_S)^2]$, the main term. Recall

$$\Theta(\delta) := \sum_{j=0}^{\infty} C_j \cos((2j+1)\pi\delta) \quad \text{where} \quad C_j = e^{-\frac{\pi^2(2j+1)^2}{8}}$$

Lemma D.12 (Fourier Expansion). *The expectation $\mathbb{E}_S[\Theta(\delta_S)^2]$ (with $g = \Theta^2$ and c_l as its Fourier coefficients) is:*

$$\mathbb{E}_S[\Theta(\delta_S)^2] = \sum_{l \in \mathbb{Z}} c_l e^{\pi i l (\sqrt{m} + N/\sqrt{m})} \cos(\pi l / \sqrt{m})^N = c_0 + E_{\text{fourier}}$$

Proof. Let $g(x) = \Theta(x)^2$. Since $g(x)$ is 1-periodic by Lemma D.7, we can write it as a standard Fourier series:

$$g(x) = \sum_{l \in \mathbb{Z}} c_l e^{2\pi i l x}$$

where $\delta_S = S/\sqrt{m} + \sqrt{m}/2$. Since g is 1-periodic, $g(\{y\}) = g(y)$.

$$\begin{aligned} \mathbb{E}_S[g(\delta_S)] &= \mathbb{E}_S \left[g \left(\left\{ \frac{S}{\sqrt{m}} \right\} + \frac{\sqrt{m}}{2} \right) \right] \\ &= \mathbb{E}_S \left[\sum_{l \in \mathbb{Z}} c_l e^{2\pi i l (S/\sqrt{m} + \sqrt{m}/2)} \right] \quad (\text{Substitute Fourier series}) \\ &= \sum_{l \in \mathbb{Z}} c_l e^{\pi i l \sqrt{m}} \mathbb{E}_S \left[e^{i(2\pi l / \sqrt{m})S} \right] \quad (\text{Linearity of } \mathbb{E}_S) \end{aligned}$$

The term $\mathbb{E}_S[e^{itS}]$ is the characteristic function $\Phi_S(t)$ of $S \sim \text{Bin}(N, 1/2)$, evaluated at $t_l = 2\pi l / \sqrt{m}$. The characteristic function for $\text{Bin}(N, p)$ is $\Phi(t) = (1 - p + pe^{it})^N$. For $p = 1/2$:

$$\Phi_S(t) = \left(\frac{1}{2} + \frac{1}{2} e^{it} \right)^N = \left(\frac{e^{it/2}}{2} (e^{-it/2} + e^{it/2}) \right)^N = \cos(t/2)^N e^{iNt/2}$$

Now, substitute $t = t_l = 2\pi l / \sqrt{m}$:

$$\Phi_S(t_l) = \cos \left(\frac{2\pi l / \sqrt{m}}{2} \right)^N e^{iN(2\pi l / \sqrt{m})/2} = \cos(\pi l / \sqrt{m})^N e^{iN\pi l / \sqrt{m}}$$

Substitute this back into the sum:

$$\begin{aligned} \mathbb{E}_S[g(\delta_S)] &= \sum_{l \in \mathbb{Z}} c_l e^{\pi i l \sqrt{m}} \left(\cos(\pi l / \sqrt{m})^N e^{iN\pi l / \sqrt{m}} \right) \\ &= \sum_{l \in \mathbb{Z}} c_l e^{i\pi l (\sqrt{m} + N/\sqrt{m})} \cos(\pi l / \sqrt{m})^N \end{aligned}$$

□

Lemma D.13 (Bound on Off-Center Contributions). *Let $R = N/m \geq 0$ be an integer. Further, let*

$$\mathbb{E}_S[g(\delta_S)] = \sum_{l \in \mathbb{Z}} c_l e^{i\pi l (\sqrt{m} + N/\sqrt{m})} \cos(\pi l / \sqrt{m})^N$$

and

$$E_{\text{fourier}} = \sum_{l \neq 0} c_l e^{\pi i l (\sqrt{m} + N/\sqrt{m})} \cos(\pi l / \sqrt{m})^N.$$

It holds for the $l \neq 0$ contributions that

$$E_{\text{fourier}} \leq 2c_1 e^{-(\pi^2/2)R} + \frac{4C}{\pi^2(1+2R)} e^{-\pi^2(1+2R)} + O(m^{-1/2}e^{-m})$$

where (with $C_j := e^{-\frac{\pi^2}{8}(2j+1)^2}$)

$$c_1 = \frac{1}{4}C_0^2 + \frac{1}{2} \sum_{j=0}^{\infty} C_j C_{j+1}, \quad C = (1/4) \cdot \sum_{p \in \mathbb{Z}} e^{-(\pi^2/4)p^2}.$$

Proof. We provide a quantitative bound for $|E_{\text{fourier}}| \leq \sum_{l \neq 0} |c_l| |\cos(\pi l / \sqrt{m})|^N$. Let $x_l = \pi l / \sqrt{m}$. We split the sum into $l \in L_{\text{small}}$ and $l \in L_{\text{large}}$. Let $\alpha = \pi^2/8$. Let $C_\alpha = \sum_{j \in \mathbb{Z}} e^{-2\alpha j^2} = \sum_{j \in \mathbb{Z}} e^{-(\pi^2/4)j^2}$. This is a constant (related to the Jacobi theta function, $\vartheta_3(0, e^{-\pi^2/4})$). Let $C = C_\alpha/4$. **Part 1: Small l (The Constant Bias Term)**

Let $L_{\text{small}} = \{l \in \mathbb{Z} : 0 < |l| \leq \sqrt{m}/2\}$. For $l \in L_{\text{small}}$, the argument $x_l = \pi l / \sqrt{m}$ is in the interval $[-\pi/2, \pi/2]$ (excluding 0). In this interval, the inequality $\cos(x) \leq e^{-x^2/2}$ holds.

$$\begin{aligned} |\cos(x_l)|^N &= |\cos(\pi l / \sqrt{m})|^N \\ &\leq \left(e^{-(\pi l / \sqrt{m})^2/2} \right)^N = e^{-N\pi^2 l^2 / (2m)} \\ &= e^{-(N/m)\pi^2 l^2 / 2} = e^{-R\pi^2 l^2 / 2} \end{aligned}$$

This bound is a constant that depends only on l and R . The contribution from this part of the sum, S_1 , is:

$$\begin{aligned} S_1 &= \sum_{l \in L_{\text{small}}} |c_l| |\cos(x_l)|^N \\ &\leq \sum_{0 < |l| \leq \sqrt{m}/2} |c_l| e^{-R\pi^2 l^2 / 2} \\ &\leq \sum_{l \neq 0} |c_l| e^{-R\pi^2 l^2 / 2} \end{aligned}$$

Using $\cos x \leq e^{-x^2/2}$ for $|x| \leq \pi/2$ and $x_l = \pi l / \sqrt{m}$, for $0 < |l| \leq \sqrt{m}/2$ we obtain

$$|\cos(x_l)|^N \leq \exp\left(-\frac{\pi^2}{2} R l^2\right), \quad R = \frac{N}{m}.$$

Hence

$$S_1 = \sum_{0 < |l| \leq \sqrt{m}/2} |c_l| |\cos(x_l)|^N \leq 2 \sum_{l \geq 1} c_l e^{-(\pi^2/2)R l^2}.$$

Isolating the $l = \pm 1$ mode gives the exact leading term

$$S_1 = 2c_1 e^{-(\pi^2/2)R} + \text{Tail}(R), \quad \text{Tail}(R) := 2 \sum_{l \geq 2} c_l e^{-(\pi^2/2)R l^2}.$$

Here c_1 admits the explicit convergent series (with $C_j := e^{-\frac{\pi^2}{8}(2j+1)^2}$)

$$c_1 = \frac{1}{4}C_0^2 + \frac{1}{2} \sum_{j=0}^{\infty} C_j C_{j+1}.$$

Moreover, using the Gaussian bound on Fourier coefficients from Part 2, $|c_l| \leq C e^{-(\pi^2/4)l^2}$ with $C = C_\alpha/4$ and $C_\alpha = \sum_{p \in \mathbb{Z}} e^{-(\pi^2/4)p^2}$, the tail is uniformly bounded by

$$0 \leq \text{Tail}(R) \leq 2C \sum_{l \geq 2} e^{-(\pi^2/4)(1+2R)l^2} \leq \frac{4C}{\pi^2(1+2R)} e^{-\frac{\pi^2}{4}(1+2R) \cdot 4} = \frac{4C}{\pi^2(1+2R)} e^{-\pi^2(1+2R)}.$$

In particular,

$$\text{Tail}(R) = o_{R \rightarrow \infty}(1) \quad \text{and} \quad S_1 = 2c_1 e^{-(\pi^2/2)R} + o_{R \rightarrow \infty}(1),$$

uniformly in m .

Part 2: Large l (The Vanishing Error Term) Let $L_{\text{large}} = \{l \in \mathbb{Z} : |l| > \sqrt{m}/2\}$. The contribution S_2 is $S_2 \leq \sum_{|l| > \sqrt{m}/2} |c_l|$. We must now quantitatively bound the tail of the Fourier coefficients c_l for $g(x) = \Theta(x)^2$.

Bound on Fourier Coefficients c_l .

To find c_l , we first write $\Theta(x)$ in its complex (period 2) series $\Theta(x) = \sum_{p \in \mathbb{Z}} \hat{\Theta}(p) e^{i\pi p x}$. By inspection of $\cos(A\pi x) = \frac{1}{2}(e^{iA\pi x} + e^{-iA\pi x})$:

$$\Theta(x) = \sum_{j=0}^{\infty} C_j \left(\frac{e^{i(2j+1)\pi x} + e^{-i(2j+1)\pi x}}{2} \right)$$

This is a sum over odd integers $p = \pm(2j+1)$. By Lemma D.8, the coefficients are:

$$\hat{\Theta}(p) = \begin{cases} \frac{1}{2} C_{(|p|-1)/2} & \text{if } p \text{ is odd} \\ 0 & \text{if } p \text{ is even} \end{cases}$$

Let $\alpha = \pi^2/8$, so $C_j = e^{-\alpha(2j+1)^2}$. This means $C_{(|p|-1)/2} = e^{-\alpha p^2}$. This gives the bound: $|\hat{\Theta}(p)| \leq \frac{1}{2} e^{-\alpha p^2}$ for all $p \in \mathbb{Z}$.

Now, we find $c_l = \int_0^1 g(x) e^{-2\pi i l x} dx = \int_0^1 \Theta(x)^2 e^{-2\pi i l x} dx$:

$$\begin{aligned} c_l &= \int_0^1 \left(\sum_{p \in \mathbb{Z}} \hat{\Theta}(p) e^{i\pi p x} \right) \left(\sum_{q \in \mathbb{Z}} \hat{\Theta}(q) e^{i\pi q x} \right) e^{-2\pi i l x} dx \\ &= \sum_{p, q \in \mathbb{Z}} \hat{\Theta}(p) \hat{\Theta}(q) \int_0^1 e^{i\pi(p+q-2l)x} dx \end{aligned}$$

Since $\hat{\Theta}$ is non-zero only for p, q odd, $p+q$ is even. Thus $p+q-2l$ is always an even integer. The integral $\int_0^1 e^{i\pi(2J)x} dx = \int_0^1 e^{2\pi i J x} dx$ is 1 if $J = 0$ and 0 if $J \neq 0$. So, the integral is 1 only if $p+q-2l = 0$, i.e., $q = 2l - p$.

$$c_l = \sum_{p \in \mathbb{Z}} \hat{\Theta}(p) \hat{\Theta}(2l - p) = \sum_{p \in \text{Odd}} \hat{\Theta}(p) \hat{\Theta}(2l - p)$$

We bound $|c_l|$ using our bound for $|\hat{\Theta}(p)|$:

$$\begin{aligned} |c_l| &\leq \sum_{p \in \text{Odd}} |\hat{\Theta}(p)| |\hat{\Theta}(2l - p)| \leq \sum_{p \in \mathbb{Z}} \left(\frac{1}{2} e^{-\alpha p^2} \right) \left(\frac{1}{2} e^{-\alpha(2l-p)^2} \right) \\ &= \frac{1}{4} \sum_{p \in \mathbb{Z}} e^{-\alpha(p^2 + (2l-p)^2)} = \frac{1}{4} \sum_{p \in \mathbb{Z}} e^{-\alpha(2p^2 - 4lp + 4l^2)} \\ &= \frac{1}{4} \sum_{p \in \mathbb{Z}} e^{-\alpha(2(p-l)^2 + 2l^2)} = \frac{1}{4} e^{-2\alpha l^2} \sum_{p \in \mathbb{Z}} e^{-2\alpha(p-l)^2} \\ &= \frac{1}{4} e^{-2\alpha l^2} \sum_{j \in \mathbb{Z}} e^{-2\alpha j^2} \quad (\text{Let } j = p - l) \end{aligned}$$

The sum $C_\alpha = \sum_{j \in \mathbb{Z}} e^{-2\alpha j^2}$ is a constant. Thus, we have a rigorous Gaussian bound:

$$|c_l| \leq C e^{-\beta l^2} \quad \text{where } \beta = 2\alpha = \pi^2/4 \text{ and } C = C_\alpha/4$$

Now we bound the tail sum S_2 :

$$\begin{aligned} S_2 &\leq \sum_{|l| > \sqrt{m}/2} C e^{-\beta l^2} = 2C \sum_{l=\lfloor \sqrt{m}/2 \rfloor + 1}^{\infty} e^{-\beta l^2} \\ &\leq 2C \int_{\lfloor \sqrt{m}/2 \rfloor}^{\infty} e^{-\beta x^2} dx \leq 2C \int_{\sqrt{m}/2 - 1}^{\infty} e^{-\beta x^2} dx \end{aligned}$$

We use the standard Gaussian tail bound $\int_t^{\infty} e^{-\beta x^2} dx \leq \frac{1}{2\beta t} e^{-\beta t^2}$ for $t > 0$. Let $t = \sqrt{m}/2 - 1$. For $m \geq 16$, $t \geq \sqrt{m}/4$.

$$S_2 \leq 2C \left[\frac{1}{2\beta(\sqrt{m}/2 - 1)} e^{-\beta(\sqrt{m}/2 - 1)^2} \right]$$

$$\begin{aligned}
&\leq \frac{C}{\beta(\sqrt{m}/4)} e^{-\beta(m/4 - \sqrt{m} + 1)} \\
&= O(m^{-1/2} e^{-\beta m/4}) = O(m^{-1/2} e^{-\pi^2 m/16})
\end{aligned}$$

This error term S_2 vanishes exponentially in m .

□

Theorem D.14 (Main Result). *Let $R := N/m \geq 0$ be an integer. Then, for sufficiently large m , the quantity $\text{Cov}(\hat{L}_1, \hat{L}_2) = \mathbb{E}_S[f(S/\sqrt{m})^2]$ satisfies*

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{c_0}{m} + E_L,$$

where c_0 is the main constant and E_L is an error term bounded by

$$|E_L| \leq \frac{\Delta(R)}{m} + O(m^{-3/2}).$$

Here, $\Delta(R)$ is a positive bias constant, exponentially small in R :

$$\Delta(R) := 2c_1 e^{-(\pi^2/2)R} + \frac{4C}{\pi^2(1+2R)} e^{-\frac{\pi^2}{4}(1+2R)},$$

with $C = 1/4 \sum_{p \in \mathbb{Z}} e^{-(\pi^2/4)p^2}$. The above constants

$$c_0 = \frac{1}{2} \sum_{j=0}^{\infty} e^{-\frac{\pi^2}{4}(2j+1)^2}, \quad c_1 = \frac{1}{4} C_0^2 + \frac{1}{2} \sum_{j=0}^{\infty} C_j C_{j+1}, \quad C_j = e^{-\frac{\pi^2}{8}(2j+1)^2},$$

are absolute (numerically $c_0 \approx 0.0424$, $c_1 \approx 0.0212$, $C_\alpha \approx 1.17$).

In particular, since $c_0 > \Delta(R)$ for $R \geq 1$, we have $\text{Cov}(\hat{L}_1, \hat{L}_2) = \Theta(1/m)$ positive.

Proof. By Theorems D.11 to D.13, we have the exact asymptotic:

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{1}{m} \mathbb{E}_S[\Theta(\delta_S)^2] + O(m^{-3/2}) = \frac{1}{m} (c_0 + E_{\text{fourier}}) + O(m^{-3/2}),$$

where, by Lemma D.13,

$$|E_{\text{fourier}}| \leq 2c_1 e^{-(\pi^2/2)R} + \frac{4C}{\pi^2(1+2R)} e^{-\frac{\pi^2}{4}(1+2R)} + O(m^{-1/2} e^{-m}) =: \Delta(R) + O(m^{-1/2} e^{-m}).$$

Substituting this back into the expression for $\text{Cov}(\hat{L}_1, \hat{L}_2)$:

$$\text{Cov}(\hat{L}_1, \hat{L}_2) = \frac{1}{m} (c_0 + E_{\text{fourier}}) + O(m^{-3/2})$$

or, equivalently,

$$\frac{1}{m} (c_0 - \Delta(R) - S_2) \leq \text{Cov}(\hat{L}_1, \hat{L}_2) \leq \frac{1}{m} (c_0 + \Delta(R) + S_2) + O(m^{-3/2})$$

Since S_2/m is absorbed by the $O(m^{-3/2})$ error, this simplifies to the claimed two-sided bound.

To show $\text{Cov}(\hat{L}_1, \hat{L}_2) = \Omega(1/m)$, we must ensure the lower bound is positive. Since $c_0 \approx 0.0424$ and $\Delta(R)$ is exponentially small in R , the constant $C_L(R) := c_0 - \Delta(R)$ is strictly positive for $R \geq 1$. Numerically, for $R = 1$:

$$\Delta(1) \approx 2(0.0212)e^{-\pi^2/2} + \dots \approx 0.000305 + 0.000024 = 0.000329$$

So $C_L(1) \approx 0.042402 - 0.000329 \approx 0.04207 > 0$. Because $C_L(R)$ is positive and bounded away from zero for all $R \geq 1$, we have $\text{Cov}(\hat{L}_1, \hat{L}_2) = \Omega(1/m)$.

On the other hand, for $R = 0$ (i.e., $N = 0$), the $l \neq 0$ Fourier mass does not decay with R : $|E_{\text{fourier}}| = \left| \sum_{l \neq 0} c_l e^{i\pi l \sqrt{m}} \right| \leq \sum_{l \neq 0} |c_l|$. This sum is a constant, which can be bounded by

$\sum |c_l| \leq C \sum e^{-(\pi^2/4)l^2} \approx 0.33$. Since this is much larger than $c_0 \approx 0.0424$, the lower bound $c_0 - |E_{\text{fourier}}|$ becomes negative. Thus, no uniform positive lower bound can be ensured in that case, and the $R \geq 1$ (i.e., $N \geq m$) condition is necessary for a meaningful bound. $\square$

APPENDIX E. ERROR IN THEOREM 5.3 IN [KR97]

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 $S^n = S^{n-1} \cup (x, y)$, and both β_1 and β_2 may be functions of n .

Let us proceed with the proof of their Theorem 5.3. There, they define the random variable $\chi(S^n) = \hat{L}^k - L(\mathcal{A}(S^n))$ 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 asserts that the expected cross-validation estimate equals the expected estimate of a single hold-out set, i.e., $\mathbb{E}_{S^n}[\chi(S^n)] = L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n))$. By this Lemma and the fact that with probability at least $\beta_1/2$, $L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) \geq \beta_2$, they claim that $\mathbb{E}_{S^n}[\chi(S^n)] \geq \frac{\beta_1}{2} \cdot \beta_2$.

This is incorrect, since $L(\mathcal{A}(S^{n-1})) - L(\mathcal{A}(S^n)) \geq \beta_2$ for some of the time 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 a non-zero squared loss stability implies a lower bound on their error stability parameters, yet we prove in Lemma 3 that one can have non-zero squared loss stability and simultaneously zero MSE (which necessitates $\mathbb{E}_{S^n}[\chi(S^n)] = 0$).

APPENDIX F. ERROR IN THEOREM 2 IN [KKV11]

The key ingredient for deriving their main result [KKV11, Theorem 2] is to obtain an upper bound on $\text{Cov}_{S^n}(\hat{L}_1^{(k)} - L_1^{(k)}, \hat{L}_2^{(k)} - L_2^{(k)})$ (in their notation $\text{cov}_U(\text{gen}_1, \text{gen}_2)$) that scales linearly with a parameter measuring a certain notion of algorithmic stability (mean square stability). To do so, the supposed identity $\mathbb{E}_{S_2}[\hat{L}_1^{(k)} - L_1^{(k)} \mid S_1, S_3, \dots, S_N] = 0$ (in their notation $\mathbb{E}_{T'}[\text{gen}_1 \mid S, T] = 0$) is used twice. Define $S' := S^n \setminus S_2$ and $S'' := S^n \setminus (S_1 \cup S_2)$. We see that

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

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