# Technical Appendices and Supplementary Material

## A Further Discussion of Related Works

### A.1 Comparison to Gastpar et al. (2024)

The estimability setting studied in our paper was introduced by Gastpar, Nachum, Shafer, and Weinberger (2024). In Theorem 3 of their paper, they show a limitation on estimability (a learnability–estimability trade-off) for algorithm-dependent bounds that is fairly abstract and involves a total variation condition that might be hard to check in many cases. In contrast, Theorems 3.1 and 3.2 involve very concrete combinatorial and geometric conditions (VC dimension, orthogonal functions). Theorems 4 and 5 in their paper are more concrete, but they hold only for exactly orthogonal functions with strict algebraic structure (parity functions). In contrast, our Theorem 3.2 applies generally to any nearly-orthogonal function class (including classes that are exactly-orthogonal as a special case).

Unlike Gastpar et al. (2024), our work also presents positive results (Theorem 4.3 and Fact C.2), showing cases where generalization bounds for specific algorithms can be tight (even if, e.g., uniform convergence does not hold). The conceptual connections between estimability, stability and conditional variance appearing in those results was not present in Gastpar et al. (2024).

Finally, our techniques also differ from those of Gastpar et al. (2024). We use the Johnson–Lindenstrauss lemma, our technical lemma (Lemma H.1), and the duality of linear programming — expanding the arsenal of tools readily available for the study of estimability.

In summary, our work builds upon the foundation laid by Gastpar et al. (2024), but we make several important contributions that go beyond their results.

### A.2 Stability

In Definitions 4.1 and 4.2, we formalize simple stability conditions that guarantee the existence of tight generalization bounds, as we show in Theorem 4.3. There are many definitions of stability in the literature, and it is important to appreciate that Theorem 4.3 makes a nontrivial conceptual contribution by identifying the "correct" notion of stability for understanding estimability.

Definitions 4.1 and 4.2 are similar to the definition of hypothesis stability and loss stability in Kearns and Ron (1999), Elisseeff, Evgeniou, and Pontil (2005), and Rogers and Wagner (1978). Lei, Jin, and Ying (2022) use another similar definition for stability and utilize it to derive generalization bounds for GD and SGD.

In contrast, our definitions of stability are also reminiscent of the replace-one stability in Bousquet and Elisseeff (2002), but as we explain in Section 4, our definitions overcome an important limitation present in their definition. In particular, the memorization algorithm (Example 1.6), which is very estimable, is not stable according to the definition of stability of Bousquet and Elisseeff (2002), but it is stable according to our definitions.

### A.3 Neural Tangent Kernel and Mean-Field Theory

There are many works that study generalization using the neural tangent kernel (NTK) or mean-field theory (MFT) approach.[12] To the best of our knowledge, these works do not provide general necessary or sufficient conditions for generalization bounds to be tight, which is the focus of our work. Additionally, they study generalization bounds for fairly specific families of algorithms such as gradient descent (or idealized versions thereof), while our work applies to a broader and more general class of algorithms.[13]

[12] E.g., Aminian et al. (2023), Chen et al. (2020), Nishikawa et al. (2022), and Nitanda et al. (2021).

[13] We note that Theorems 3.1 and 3.2 apply to learning algorithms that achieve 0 training error. Because this property is satisfied by many contemporary learning algorithms (even if the labels are random), we do not view this as a significant limitation on the generality of our results. This assumption is not essential, and it could easily be relaxed in future work.

## B On Extending Our Results to Randomized Algorithms

For simplicity, in this paper we focus on deterministic learning rules. However, we recognize that the topic of randomized learning algorithms is very important, seeing as most algorithms used in practice today are randomized.

The estimability framework explored in this paper can be extended to handle randomized algorithms as well, and in fact the original work of Gastpar et al. (2024) already contains some initial treatment of randomized algorithms.

We expect that the results presented in this paper can be extended to randomized algorithms, and that the essence of the results remains mostly unchanged.

The first step in such an extension would be to clearly define what estimability means for randomized learning algorithms. A definition that one might initially consider is one where the estimator knows the randomness used by the algorithm, and must output a number that is with high probability close to the true population loss of the randomized algorithm. This definition is not very interesting, because a setting in which the estimator knows the randomness used by the randomized algorithm is equivalent to the setting of a deterministic algorithm, which is already covered by the results in this paper. Nonetheless, it is good to keep this definition in mind, because it means that our results for deterministic algorithms already apply as-is to randomized algorithms (like SGD) once the randomly chosen seed is fixed, which might be a simple and satisfactory approach for many purposes (SGD with a fixed random seed typically performs as well for most purposes as SGD with a fresh randomly-chosen seed).

Perhaps the more "correct" and interesting definition of estimability for randomized learning algorithms is one where the estimator knows the training set, but does not know the randomness used by the learning algorithm, and it is required to output a number that is close with high probability to the *expected* population loss of the randomized algorithm when executed with this training set (where the expectation is over the randomness of the algorithm). In this setting, we believe the essence of our results carries through, with an important conceptual difference: using randomness, one can always engineer a learning algorithm that is estimable, essentially by adding noise to the output of the algorithm. As the noise in the algorithm's output increases, the expected 0-1 loss of the algorithm becomes closer to $1/2$, and so the algorithm becomes estimable with a trivial estimator that simply always outputs the number $1/2$. (With intermediate amounts of noise, a number between 0 and $1/2$ will be optimal).

Consequently, for randomized algorithms, our lower bounds in Theorems 3.1 and 3.2 can no longer be stated as absolute limitations on estimability. Rather there is now a trade-off between the performance of the algorithm and its estimability. As one adds more noise, the algorithm becomes more estimable, but its performance degrades. Thus, the corresponding theorems for randomized algorithm would state that no algorithm can simultaneously make good predictions for some large set of labeling functions and also be estimable.

On the other hand, the upper bound in Theorem 4.3 that states that stable algorithms are estimable remains basically unchanged for randomized algorithms.

To summarize, under a suitable definition of estimability for randomized algorithms, we expect that our results would not change much, though the statement (and proof) of the lower bounds would be somewhat more complex. We leave this work to future research.

## C A Simple Characterization

The following definition is a variant of Definition 1.1. Such a variant allows us to have a simple characterization of estimability in Fact C.2. Namely, to understand whether an algorithm is estimable with respect to a set of distributions, one can examine the quantity $\mathbb{E}_{\mathcal{D}\sim\mathrm{U}(\mathbb{D}),S\sim\mathcal{D}^m}[\mathrm{var}(L_{\mathcal{D}}(A(S)) \mid S)]$.

**Definition C.1.** *Let $\mathbb{D}$ be a set of distributions and let $A$ be a learning algorithm. We say that $A$ is $(\varepsilon, m)$-estimable in $\ell_2$ with respect to $\mathbb{D}$, if there exists an estimator $\mathcal{E}$ such that*

$$\mathbb{E}_{\mathcal{D}\sim\mathrm{U}(\mathbb{D}),S\sim\mathcal{D}^m}\left[(\mathcal{E}(S) - L_{\mathcal{D}}(A(S)))^2\right] \leq \varepsilon$$

We remark that for bounded loss functions, one can move between Definition C.1 and Definition 1.1 using Markov's inequality. Furthermore, although the characterization in the following theorem is simple, it might provide a technical condition that will be useful for future work.

**Fact C.2.** *$A$ is $(\varepsilon, m)$-estimable in $\ell_2$ with respect to $\mathbb{D}$ if and only if*

$$\mathbb{E}_{\mathcal{D}\sim \mathrm{U}(\mathbb{D}), S\sim\mathcal{D}^m}[\mathrm{var}(L_{\mathcal{D}}(A(S)) \mid S)] \leq \varepsilon.$$

*Proof of Fact C.2.* The result that the minimum mean-square error (MMSE) estimator corresponds to the conditional expectation is a well-established theorem in probability theory (see, for instance, Section 7.9 in Grimmett and Stirzaker (2020)). For the sake of completeness, we present a proof of this result.

We will use the following simple claim.

**Claim C.3.** *Let $c_1, ..., c_k, p_1, ..., p_k \in \mathbb{R}$ such that $\sum_{i=1}^k p_i = 1$, then*

$$\mathrm{argmin}_{x\in\mathbb{R}} \sum_{i=1}^k p_i \cdot (x - c_i)^2 = \sum_{i=1}^k p_i \cdot c_i.$$

The claim follows by taking the derivative of $\sum_{i=1}^k p_i \cdot (x - c_i)^2$ with respect to $x$ which yields the equation:

$\sum_{i=1}^k 2p_i(x - c_i) = 0$ that implies $x = \sum_{i=1}^k p_i c_i$ since $\sum_{i=1}^k p_i = 1$.

The following shows that the estimator $\mathcal{E}^*(S) \coloneqq \mathbb{E}[L_{\mathcal{D}}(A(S)) \mid S]$ is optimal and the inequality follows from Claim C.3. Let $\mathcal{E}$ be any estimator for $A$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}\sim \mathrm{U}(\mathbb{D}), S\sim\mathcal{D}^m}\left[(\mathcal{E}(S) - L_{\mathcal{D}}(A(S)))^2\right] \\
&\quad= \sum_S \mathbb{P}(S) \sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S) \left(L_{\mathcal{D}}(A(S)) - \mathcal{E}(S)\right)^2 \\
&\quad= \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S) \left(L_{\mathcal{D}}(A(S)) - \mathcal{E}(S)\right)^2\right] \\
&\quad\geq \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S) \left(L_{\mathcal{D}}(A(S)) - \sum_{\mathcal{D}\in\mathbb{D}} [\mathbb{P}(\mathcal{D}|S) L_{\mathcal{D}}(A(S))]\right)^2\right] \\
&\quad= \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S) \left(L_{\mathcal{D}}(A(S)) - \mathcal{E}^*(S)\right)^2\right] \\
&\quad= \mathbb{E}_{\mathcal{D}\sim \mathrm{U}(\mathbb{D}), S\sim\mathcal{D}^m}\left[(\mathcal{E}^*(S) - L_{\mathcal{D}}(A(S)))^2\right].
\end{aligned}$$

This means that $A$ is square loss $(\varepsilon, m)$-estimable with respect to $\mathbb{D}$ if and only if $\mathcal{E}^*$ can achieve $\varepsilon$ accuracy. It achieves such accuracy if and only if $\mathbb{E}[\mathrm{var}(L_{\mathcal{D}}(A(S)) \mid S)] \leq \varepsilon$. This follows by the following equalities that complete the proof.

$$
\begin{aligned}
\mathbb{E}\left[\operatorname{var}(L_{\mathcal{D}}(A(S)) \mid S)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(L_{\mathcal{D}}(A(S)) - \mathbb{E}\left[L_{\mathcal{D}}(A(S))|S\right]\right)^2 \mid S\right]\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S)\left(L_{\mathcal{D}}(A(S)) - \mathbb{E}\left[L_{\mathcal{D}}(A(S)) \mid S\right]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S)\left(L_{\mathcal{D}}(A(S)) - \sum_{\mathcal{D}\in\mathbb{D}}\left[\mathbb{P}(\mathcal{D}|S)L_{\mathcal{D}}(A(S))\right]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{D}\in\mathbb{D}} \mathbb{P}(\mathcal{D}|S)\left(L_{\mathcal{D}}(A(S)) - \mathcal{E}^*(S)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}\sim \mathrm{U}(\mathbb{D}), S\sim\mathcal{D}^m}\left[\left(\mathcal{E}^*(S) - L_{\mathcal{D}}(A(S))\right)^2\right] \qquad \square
\end{aligned}
$$

# D Details for Example 1.8

For sample size $m \geq d + 10$, any ERM algorithm for $\mathcal{H}$ satisfies $p(m) \geq 0.999$, meaning it learns $\mathbb{D}$ well, and hence is $(0, 10^{-3}, d + 10)$-estimable on average. This holds because for an ERM to output the ground truth, it is clearly sufficient that only a single sample-consistent function exists in the concept class (the ground truth). Similarly, in the event that there are $t > 1$ sample-consistent functions, the success probability is given by $1/t$ due to the uniform prior over ground truth distributions. Parity functions are fully characterized by their coefficient vector $w = [w_1, \ldots, w_d]$. Since the labels $y$ are a bilinear function in the inputs $x$ and coefficients $w$, one can obtain $w$ from $m \geq d$ linearly independent samples $x_i$ by solving the linear system of equation $y = Xw$ with design matrix $X \in \{0, 1\}^{m\times d}$. More generally, $X$ having rank $d - k$ is equivalent to the event of having $t = 2^k$ sample-consistent functions (coefficient vectors) since every additional linearly independent row rules out half of all parity functions. Now assume $X$ consists of all i.i.d. Ber(½) entries and $y$ contains the labels of all samples. The probability of zero population loss can now be obtained from the law of total probability with the probabilities of rank deficiency computed according to Corollary 2.2 in Blake and Studholme (2006).

Similar calculations show that for smaller sample sizes, any ERM for $\mathcal{H}$ satisfies $p(d) \geq 0.61$, and $p(d - 1) \geq 0.38$. An application of Theorem 5 in Gastpar et al. (2024) shows that there exist ERM algorithms such that for any $6 \leq m \leq d$ there exists a collection $\mathbb{D}_m$ for which the algorithm is not $(0.25, 0.32, m)$-estimable on average. These algorithms have an inductive bias towards a subset $\mathcal{F} \subseteq \mathcal{H}$, such that they perform well for distributions labeled by a function from $\mathcal{F}$, and perform poorly for target functions from the complement of $\mathcal{F}$.

# E Proof of Theorem 3.1

Recall the definition of nearly-orthogonal functions (Definition 2.8). The proof of Theorem 3.1 uses a corollary of the Johnson–Lindenstrauss lemma (Theorem K.1), which states that random vectors in a high dimensional space are nearly orthogonal, as follows.[14]

**Claim E.1.** *Let $\varepsilon \in (0, 1/2)$, and let $d, n \in \mathbb{N}$ such that*

$$n \leq \exp\bigl(d\varepsilon^2/54\bigr).$$

*Let $\mathcal{U} = \mathrm{U}\bigl(\{\pm1\}^{[d]}\bigr)$ be the uniform distribution over functions $[d] \to \{\pm1\}$, and consider a random sequence $F$ of functions $F_1, \ldots, F_n$ sampled independently from $\mathcal{U}$. Then*

$$\mathbb{P}_{F\sim\mathcal{U}^n}\bigl[F \in \perp_{\varepsilon,[d]}\bigr] \geq 0.99.$$

*Proof of Claim E.1.* If $n = 1$ there is nothing to prove, so we assume $n \geq 2$. Let $R \sim \mathrm{U}\bigl(\{\pm1\}^{d\times n}\bigr)$ be a $d \times n$ matrix with entries in $\{\pm1\}$ chosen independently and uniformly at random. In particular, for each $i \in [n]$, the $i$-th column of $R$ is a vector of $d$ numbers in $\{\pm1\}$ chosen independently and uniformly at random. Hence, using $e_1, \ldots, e_n$ to denote the

[14] It is also possible to prove a similar claim by directly using concentration of measure (e.g., Hoeffding's inequality), without using the Johnson–Lindenstrauss lemma.

standard basis of $\mathbb{R}^n$, we identify the vector $Re_i$, which is the $i$-th column of $R$, with the random function $F_i : [d] \to \{\pm 1\}$.

Recall that for vectors $u, v \in \mathbb{R}^d$,

$$\|u - v\|_2^2 = \langle u - v, u - v\rangle = \|u\|_2^2 - 2\langle u, v\rangle + \|v\|_2^2,$$

so

$$\langle u, v\rangle = \frac{\|u\|_2^2 + \|v\|_2^2 - \|u - v\|_2^2}{2}. \tag{5}$$

Invoking Theorem K.1 with $s = n$, $\beta = 7$, $V = \{e_1, \dots, e_n\} \subseteq \mathbb{R}^n$, and $d, n, \varepsilon$ as in the claim statement implies that

$$\mathbb{P}_{R \sim \mathrm{U}(\{\pm 1\}^{d \times n})}\left[\begin{array}{l} \forall i, j \in [n], i \neq j : \\ \quad (1 - \varepsilon) \cdot 2 \leq \left\|\frac{1}{\sqrt{d}} Re_i - \frac{1}{\sqrt{d}} Re_j\right\|_2^2 \leq (1 + \varepsilon) \cdot 2 \end{array}\right] \geq 1 - \frac{1}{n^\beta}. \tag{6}$$

Hence, with probability at least $1 - 1/n^\beta \geq 1 - 1/2^7 \geq 0.99$ over the choice of $F$, every distinct $i, j \in [n]$ satisfy

$$\begin{aligned}
\left|\mathbb{E}_{x \sim \mathrm{U}([d])}[F_i(x) F_j(x)]\right| &= \left|\frac{1}{d} \sum_{x \in [d]} F_i(x) F_j(x)\right| = \left|\frac{1}{d}\langle Re_i, Re_j\rangle\right| && \text{(Identifying } F_i \text{ with } Re_i) \\
&= \left|\frac{\|Re_i\|_2^2 + \|Re_j\|_2^2 - \|Re_i - Re_i\|_2^2}{2d}\right| && \text{(By Eq. (5))} \\
&= \left|1 - \frac{1}{2}\left\|\frac{1}{\sqrt{d}} Re_i - \frac{1}{\sqrt{d}} Re_i\right\|_2^2\right| \\
&\leq \varepsilon, && \text{(By Eq. (6))}
\end{aligned}$$

as desired. $\square$

*Proof of Theorem 3.1.* Fix an $\mathcal{H}$-shattered set $\mathcal{X}_d \subseteq \mathcal{X}$ with cardinality $|\mathcal{X}_d| = d$, and for each $f : \mathcal{X}_d \to \{\pm 1\}$ let $\mathcal{D}_f = \mathrm{U}(\{(x, f(x)) : x \in \mathcal{X}_d\})$. Note that the distributions $\mathcal{D}_f$ are $\mathcal{H}$-realizable. We will show that there exists a collection $\mathbb{D} = \{\mathcal{D}_f : f \in \mathcal{F}\}$ that satisfies Eq. (4), where $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}_d}$ is a set of $k = 2^m + 1$ functions.

Consider the following experiment:

1. Sample a sequence of functions $G = (G_1, \dots, G_k)$ independently and uniformly at random from $\{\pm 1\}^{\mathcal{X}_d}$.
2. Sample a function $F$ uniformly from $G$.
3. Sample a sequence of points $X = (X_1, \dots, X_m)$ independently and uniformly at random from $\mathcal{X}_d$. ($X$ is sampled independently of $(G, F)$.)
4. For each $i \in [m]$, let $Y_i = F(X_i)$, let $Y = (Y_1, \dots, Y_m)$, and let $S = \Big((X_1, Y_1), \dots, (X_m, Y_m)\Big)$.

Let $\mathcal{P}$ be the joint distribution of $(G, F, X, Y, S)$. Consider the following events:

- $\mathcal{E}_1 = \{G \in \perp_{\varepsilon, \mathcal{X}_d}\}$ for $\varepsilon = 2/d^{1/4}$. By Claim E.1 and the choice of $k$, $\mathcal{P}(\mathcal{E}_1) \geq 0.99$ for $d$ large enough.[15]
- $\mathcal{E}_2 = \Big\{|\{X_1, \dots, X_m\}| = m\Big\}$. By Claim K.2 and the choice of $m$, $\mathcal{P}(\mathcal{E}_2) \geq 0.99$.

[15] We choose $d_0 \in \mathbb{N}$ to be the universal constant such that this inequality holds for all integers $d \geq d_0$ and all $m \leq \sqrt{d}/10$.

- $\mathcal{E}_3 = \{|G_S| = 2\}$. $\mathcal{P}(\mathcal{E}_3 \mid \mathcal{E}_2) \geq 1/e$. To see this, note that each function $G_i \in G \setminus \{F\}$ is chosen independently of $F$. Hence, the probability that a function $G_i$ agrees with $F$ on the $m$ distinct samples in $X$ (i.e., the probability that $G_i(X_j) = F(X_j)$ for all $j \in [m]$, given $\mathcal{E}_2$) is $p = 2^{-m}$. The functions in $G$ are chosen independently, so the number $T$ of functions in $G \setminus \{F\}$ that agree with $F$ on $m$ distinct samples has a binomial distribution $T \sim \mathrm{Bin}(k-1, p)$. So

$$\begin{aligned}
\mathbb{P}[T = 1] &= (k-1) \cdot p \cdot (1-p)^{k-2} = (1-p)^{k-2} \\
&\geq \left(e^{-\frac{p}{1-p}}\right)^{k-2} && (\forall p < 1:\ 1 - p \geq e^{-p/(1-p)}) \\
&= 1/e.
\end{aligned}$$

Let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_3$. Combining the above bounds yields

$$\begin{aligned}
\mathcal{P}(\mathcal{E}) &= \mathcal{P}(\mathcal{E}_1 \cap \mathcal{E}_3) \\
&\geq \mathcal{P}(\mathcal{E}_3) - \mathcal{P}\big(\mathcal{E}_1^C\big) \\
&\geq \mathcal{P}(\mathcal{E}_3 \mid \mathcal{E}_2) \cdot \mathcal{P}(\mathcal{E}_2) - \mathcal{P}\big(\mathcal{E}_1^C\big) \\
&\geq 0.99 \cdot 1/e - 0.01 > 1/3.
\end{aligned}$$

By an averaging argument, this implies that there exists $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}_d}$ such that $\mathcal{F} \in \perp_{\varepsilon, \mathcal{X}_d}$ for $\varepsilon = 2/d^{1/4}$ and

$$\mathcal{P}(|G_S| = 2 \mid G = \mathcal{F}) \geq 1/3. \tag{7}$$

Fix this $\mathcal{F}$, and let $A$ be an $\mathcal{F}$-interpolating learning rule. From the technical lemma (Lemma H.1), there exists a collection of $\mathcal{F}$-realizable distributions $\mathbb{D} \subseteq \Delta(\mathcal{X}_d \times \{\pm 1\})$ such that for any estimator $\mathcal{E} : (\mathcal{X} \times \{\pm 1\})^m \to [0, 1]$ that may depend on $\mathbb{D}$ and $A$,

$$\begin{aligned}
\mathbb{P}_{\substack{\mathcal{D} \sim \mathrm{U}(\mathbb{D}) \\ S \sim \mathcal{D}^m}}\left[\big|\mathcal{E}(S) - L_{\mathcal{D}}(A(S))\big| \geq \frac{1}{4} - \frac{\varepsilon}{4}\right] &\geq \frac{1}{2} \cdot \mathbb{P}_{\substack{\mathcal{D} \sim \mathrm{U}(\mathbb{D}) \\ S \sim \mathcal{D}^m}}[|\mathcal{F}_S| = 2] \\
&\geq \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}, && \text{(By Eq. (7))}
\end{aligned}$$

as desired. □

# F Proof of Theorem 3.2

*Proof of Theorem 3.2.* We take $\mathbb{D} = \{\mathcal{D}_f : f \in \mathcal{F}\}$ where $\mathcal{D}_f = \mathrm{U}(\{(x, f(x)) : x \in \mathcal{X}\})$. Fix a function $f^* \in \mathcal{F}$, let $S \sim (\mathcal{D}_{f^*})^m$, and consider the random variable $Z = |\mathcal{F}_S|$. We bound the expectation and variance of $Z$, and then show a lower bound on the probability that $Z \in \{2, 3\}$.

Let $S = \big((X_1, Y_1), \ldots, (X_m, Y_m)\big)$ and $X = \{X_1, \ldots, X_m\}$, and let $E$ denote the event in which $|X| = m$ (i.e., $S$ is collision-free). For each $f \in \mathcal{F}$, let $Z_f = \mathbb{1}(\forall i \in [m] : f(X_i) = Y_i)$, so that $Z = \sum_{f \in \mathcal{F}} Z_f$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S \sim (\mathcal{D}_{f^*})^m}[Z \mid E] &= \mathbb{E}\left[\sum_{f \in \mathcal{F}} Z_f \,\middle|\, E\right] \\
&= 1 + \sum_{\substack{f \in \mathcal{F} \\ f \neq f^*}} \mathbb{P}\big[\forall i \in [m] : f(X_i) = Y_i \,\big|\, E\big] && (Z_F = 1) \\
&\leq 1 + 2^m \cdot \left(\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{1000m}\right)^m && \text{(By Fact 2.9)} \\
&\leq 1 + e^{1/1000} < 2.002. && (8)
\end{aligned}$$

$$\mathop{\mathbb{E}}_{S \sim (\mathcal{D}_{f^*})^m}[Z \mid E] \geq 1 + 2^m \cdot \left(\frac{1}{2} - \frac{1}{2} \cdot \frac{1}{1000m}\right)^m \qquad \text{(By Fact 2.9)}$$

$$\geq 1 + e^{-1/500}. \qquad (1 - x \geq e^{-x/(1-x)}) \tag{9}$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{S \sim (\mathcal{D}_{f^*})^m}\left[Z^2 \mid E\right] &= \mathbb{E}\left[\left(\sum_{f \in \mathcal{F}} Z_f\right)\left(\sum_{g \in \mathcal{F}} Z_g\right) \;\middle|\; E\right] \\
&= \mathbb{E}\left[\left(1 + \sum_{\substack{f \in \mathcal{F} \\ f \neq f^*}} Z_f\right)\left(1 + \sum_{\substack{g \in \mathcal{F} \\ g \neq f^*}} Z_g\right) \;\middle|\; E\right] && (Z_{f^*} = 1) \\
&= \mathbb{E}\left[1 + 2\sum_{\substack{f \in \mathcal{F} \\ f \neq f^*}} Z_f + \sum_{\substack{f \in \mathcal{F} \\ f \neq f^*}} \sum_{\substack{g \in \mathcal{F} \\ g \neq f^*}} Z_f Z_g \;\middle|\; E\right] \\
&= \mathbb{E}\left[1 + 3\sum_{\substack{f \in \mathcal{F} \\ f \neq f^*}} Z_f + \sum_{\substack{f,g \in \mathcal{F} \setminus \{f^*\} \\ f \neq g}} Z_f Z_g \;\middle|\; E\right] \\
&= 1 + 3\left(\mathbb{E}[Z \mid E] - 1\right) + \sum_{\substack{f,g \in \mathcal{F} \setminus \{f^*\} \\ f \neq g}} \mathbb{E}\left[Z_f Z_g \;\middle|\; E\right].
\end{aligned} \tag{10}$$

$$\begin{aligned}
\sum_{\substack{f,g \in \mathcal{F} \setminus \{f^*\} \\ f \neq g}} \mathbb{E}\left[Z_f Z_g \;\middle|\; E\right] &= \sum_{\substack{f,g \in \mathcal{F} \setminus \{f^*\} \\ f \neq g}} \mathbb{P}\left[\forall i \in [m] :\; f(X_i) = g(X_i) = f^*(X_i) \;\middle|\; E\right] \\
&\leq 2^{2m} \cdot \left(\frac{1}{4} + \frac{3}{4} \cdot \frac{1}{1000m}\right)^m && \text{(By Claim J.1)} \\
&= \left(1 + \frac{3}{1000m}\right)^m \\
&\leq e^{3/1000}.
\end{aligned} \tag{11}$$

Combining Eqs. (8) to (11) yields

$$\begin{aligned}
\mathrm{Var}[Z \mid E] &= \mathbb{E}\left[Z^2 \mid E\right] - \left(\mathbb{E}[Z \mid E]\right)^2 \\
&\leq 1 + 3e^{1/1000} + e^{3/1000} - \left(1 + e^{-1/500}\right)^2 \\
&< 1.02.
\end{aligned}$$

By Lemma I.1,

$$\mathbb{P}[Z \in \{2,3\} \mid E] \geq 1 - \frac{\mathrm{Var}[Z \mid E]}{2} \geq 0.49.$$

Claim K.2 and $|\mathcal{X}'| \geq 100m^2$ imply that $\mathbb{P}[E] \geq 0.99$. Hence,

$$\mathbb{P}[Z \in \{2,3\}] \geq \mathbb{P}[E] \cdot \mathbb{P}[Z \in \{2,3\} \mid E] \geq 0.99 \cdot 0.49 \geq 0.48. \tag{12}$$

Finally, invoking our technical lemma (Lemma H.1) yields

$$\mathop{\mathbb{P}}_{\substack{F \sim \mathrm{U}(\mathcal{F}) \\ S \sim (\mathcal{D}_F)^m}}\left[\left|\mathcal{E}(S) - L_{\mathcal{D}_F}(A(S))\right| \geq \frac{1}{4} - \frac{1}{4000m}\right] \geq \frac{\mathbb{P}[Z \in \{2,3\}]}{3} \geq 0.16,$$

as desired. $\square$

## G Proof of Theorem 4.3

*Proof.* If $(A, \mathbb{D})$ is $(\alpha, \beta, m, k)$-hypothesis stable, then in particular $(A, \mathbb{D})$ is also $(\alpha, \beta, m, k)$-loss stable. Hence, it suffices to prove the claim for the case of loss stability. We construct a uniform estimator $\mathcal{E}$ as follows. Given a sample $S \in \mathcal{Z}^m$ for $\mathcal{Z} = (\mathcal{X} \times \{\pm 1\})$, let $S_1 \circ S_2 = S$ be the partition of $S$ such that $S_1 \in \mathcal{Z}^{m-k}$ and $S_2 \in \mathcal{Z}^k$. Take $\mathcal{E}(S) = L_{S_2}(A(S_1))$.

By the triangle inequality,

$$|\mathcal{E}(S) - L_{\mathcal{D}}(A(S))| \leq |\mathcal{E}(S) - L_{\mathcal{D}}(A(S_1))| + |L_{\mathcal{D}}(A(S_1)) - L_{\mathcal{D}}(A(S))|,$$

so

$$\begin{aligned}
\mathbb{P}_{S \sim \mathcal{D}^m}[|\mathcal{E}(S) - L_{\mathcal{D}}(A(S))| > \varepsilon] &\leq \mathbb{P}\left[\begin{array}{l} |L_{S_2}(A(S_1)) - L_{\mathcal{D}}(A(S_1))| > \alpha_0 \vee \\ |L_{\mathcal{D}}(A(S_1)) - L_{\mathcal{D}}(A(S))| > \alpha_1 \end{array}\right] \\
&\leq \mathbb{P}[|L_{S_2}(A(S_1)) - L_{\mathcal{D}}(A(S_1))| > \alpha_0] \\
&\quad + \mathbb{P}[|L_{\mathcal{D}}(A(S_1)) - L_{\mathcal{D}}(A(S))| > \alpha_1] \\
&\leq \beta_0 + \beta_1 = \delta,
\end{aligned}$$

where the final step follows from Hoeffding's inequality, the choice of $k$, and the stability of $A$. $\square$

## H Technical Lemma for Inestimability

**Lemma H.1.** *Let $m \in \mathbb{N}$, let $\varepsilon > 0$, let $\mathcal{X}$ be a finite set, let $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ such that $\mathcal{F} \in \perp_{\varepsilon, \mathcal{X}}$, and let $A\colon (\mathcal{X} \times \{\pm 1\})^m \to \{\pm 1\}^{\mathcal{X}}$ be an $\mathcal{F}$-interpolating learning rule. For each $f \in \mathcal{F}$ let $\mathcal{D}_f = \mathrm{U}(\{(x, f(x)) : x \in \mathcal{X}\})$, and for each $k \in \mathbb{N}$ let*

$$p_k = \mathbb{P}_{\substack{F \sim \mathrm{U}(\mathcal{F}) \\ S \sim (\mathcal{D}_F)^m}}[|\mathcal{F}_S| = k].$$

*Then for any estimator $\mathcal{E}\colon (\mathcal{X} \times \{\pm 1\})^m \to [0, 1]$ that may depend on $A$,*

$$\mathbb{P}_{\substack{F \sim \mathrm{U}(\mathcal{F}) \\ S \sim (\mathcal{D}_F)^m}}\left[\left|\mathcal{E}(S) - L_{\mathcal{D}_F}(A(S))\right| \geq \frac{1}{4} - \frac{\varepsilon}{4}\right] \geq \sum_{k \in \{2, \ldots, |\mathcal{F}|\}} \frac{p_k}{k}.$$

*Proof.* Consider the following experiment:

1. Sample a sequence of points $X = (X_1, \ldots, X_m)$ independently and uniformly at random from $\mathcal{X}$.
2. Sample a function $F$ uniformly from $\mathcal{F}$, independently of $X$.
3. For each $i \in [m]$, let $Y_i = F(X_i)$, let $Y = (Y_1, \ldots, Y_m)$, and let $S = \Big((X_1, Y_1), \ldots, (X_m, Y_m)\Big)$.

Let $\mathcal{P}$ be the joint distribution of $(X, F, Y, S)$. Fix $k \in \{2, \ldots, |\mathcal{F}|\}$, and let

$$s = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in (\mathcal{X} \times \{\pm 1\})^m$$

with $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ such that $|\mathcal{F}_s| = k$. Denote $\mathcal{F}_s = \{f_1, \ldots, f_k\}$. Then for any $i, j \in [k]$, $i \neq j$,

$$\begin{aligned}
\mathcal{P}(S = s \mid F = f_i) &= \mathcal{P}(X = x \mid F = f_i) \\
&= \mathcal{P}(X = x \mid F = f_j) && (X \perp F) \\
&= \mathcal{P}(S = s \mid F = f_j).
\end{aligned} \tag{13}$$

So,

$$\begin{aligned}
\mathcal{P}(F = f_i \mid S = s) &= \frac{\mathcal{P}(S = s \mid F = f_i) \cdot \mathcal{P}(F = f_i)}{\mathcal{P}(S = s)} \\
&= \frac{\mathcal{P}(S = s \mid F = f_j) \cdot \mathcal{P}(F = f_j)}{\mathcal{P}(S = s)} && (\text{By Eq. (13)}, F \sim \mathrm{U}(\mathcal{F})) \\
&= \mathcal{P}(F = f_j \mid S = s),
\end{aligned} \tag{14}$$

Seeing as $\mathcal{P}(F \in \mathcal{F}_s \mid S = s) = 1$, this implies that for all $i \in [k]$, $\mathcal{P}(F = f_i \mid S = s) = 1/k$.

Because $A$ is $\mathcal{F}$-interpolating, $A(s) \in \mathcal{F}_s$. Without loss of generality, denote $A(s) = f_1$. From $\mathcal{F} \in \perp_{\varepsilon,\mathcal{X}}$ and Fact 2.9, $L_{\mathcal{D}_{f_i}}(f_j) \geq \frac{1}{2} - \frac{\varepsilon}{2} := 2\alpha$ for all $i, j \in [k], i \neq j$. Hence,

$$\begin{aligned}
\mathcal{P}(L_{\mathcal{D}_F}(A(S)) = 0 \mid S = s) &= \mathcal{P}(F = A(S) \mid S = s) && (F, A(s) \in \mathcal{F}_s) \\
&= \mathcal{P}(F = f_1 \mid S = s) && (A(s) = f_1) \\
&= 1/k, && (15)
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{P}(L_{\mathcal{D}_F}(A(S)) \geq 2\alpha \mid S = s) &= \mathcal{P}(F \neq A(S) \mid S = s) \\
&= \mathcal{P}(F \in \{f_2, \dots, f_k\} \mid S = s) \\
&= (k-1)/k. && (16)
\end{aligned}$$

Hence, for any $\eta \in \mathbb{R}$,

$$\mathcal{P}\big(|L_{\mathcal{D}_F}(A(S)) - \eta| \geq \alpha \;\big|\; S = s\big) \geq \frac{1}{k}. \tag{17}$$

We conclude that for any estimator $\mathcal{E} : (\mathcal{X} \times \{\pm 1\})^m \to \mathbb{R}$,

$$\begin{aligned}
&\mathcal{P}(|L_{\mathcal{D}_F}(A(S)) - \mathcal{E}(S)| \geq \alpha) \\
&\quad\geq \sum_{k \in \{2,\dots,|\mathcal{F}|\}} \mathcal{P}\Big(|L_{\mathcal{D}_F}(A(S)) - \mathcal{E}(S)| \geq \alpha \bigwedge |\mathcal{F}_S| = k\Big) \\
&\quad= \sum_{k \in \{2,\dots,|\mathcal{F}|\}} \sum_{s:\, |\mathcal{F}_s| = k} \mathcal{P}\big(|L_{\mathcal{D}_F}(A(S)) - \mathcal{E}(S)| \geq \alpha \;\big|\; S = s\big) \cdot \mathcal{P}(S = s) \\
&\quad\geq \sum_{k \in \{2,\dots,|\mathcal{F}|\}} \sum_{s:\, |\mathcal{F}_s| = k} \inf_{\eta \in \mathbb{R}} \mathcal{P}\big(|L_{\mathcal{D}_F}(A(S)) - \eta| \geq \alpha \;\big|\; S = s\big) \cdot \mathcal{P}(S = s) \\
&\quad\geq \sum_{k \in \{2,\dots,|\mathcal{F}|\}} \sum_{s:\, |\mathcal{F}_s| = k} \frac{1}{k} \cdot \mathcal{P}(S = s) && \text{(By Eq. (17))} \\
&\quad= \sum_{k \in \{2,\dots,|\mathcal{F}|\}} \frac{1}{k} \cdot \mathcal{P}(|\mathcal{F}_S| = k)
\end{aligned}$$

as desired. □

# I Concentration Bound via Linear Programming

**Lemma I.1.** *Let $n \in \mathbb{N}$, $v_{\max} \in \mathbb{R}$. Let $Z$ be a random variable taking values in $[n]$ such that $\mu = \mathbb{E}[Z] \in [2, \sqrt{2} + 1]$ and $\mathrm{Var}[Z] \leq v_{\max}$. Then $\mathbb{P}[Z \in \{2, 3\}] \geq 1 - v_{\max}/2$.*

We prove this concentration of measure bound using the duality of linear programs (see Section 7.4.1 in Boyd and Vandenberghe, 2014 for an exposition of this approach).

*Proof.* Let $Z' = Z - \mu$. $Z'$ is a random variable with $\mathbb{E}[Z'] = 0$ and $\mathrm{Var}[Z'] = \mathrm{Var}[Z]$. Furthermore, $\mathbb{P}[Z \in \{2, 3\}] = \mathbb{P}[Z' \in \{2 - \mu, 3 - \mu\}]$. We show a lower bound on $\mathbb{P}[Z' \in \{2 - \mu, 3 - \mu\}]$ across all distribution of $Z'$ with the above moment constraints.

Indeed, let $X$ be a random variable taking values in $\{1 - \mu, 2 - \mu, \dots, n - \mu\}$ with $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] \leq v_{\max}$ such that $\mathbb{P}[X \in \{2 - \mu, 3 - \mu\}]$ is minimal. In particular, the distribution of $X$ is a solution to the following minimization problem.

$$\begin{aligned}
&\min_{\mathcal{D}_X} \mathbb{P}[X \in \{2 - \mu, 3 - \mu\}] \\
&\text{s.t.} \\
&\quad \mathbb{E}[X] = 0 \\
&\quad \mathrm{Var}[X] \leq v_{\max}
\end{aligned}$$

The minimization problem can be formulated as a linear program with variables $p_k = \mathbb{P}[X = k - \mu]$ for each $k \in [n]$.

$$\begin{aligned}
&\min_{\mathcal{D}_X} p_2 + p_3 \\
&\text{s.t.} \\
&\quad \sum_{k\in[n]} p_k \geq 1 \\
&\quad \sum_{k\in[n]} -p_k \geq -1 \\
&\quad \sum_{k\in[n]} p_k \cdot (k-\mu) \geq 0 \\
&\quad \sum_{k\in[n]} p_k \cdot (\mu-k) \geq 0 \\
&\quad \sum_{k\in[n]} -p_k \cdot (k-\mu)^2 \geq -v_{\max} \\
&\quad \forall k \in [n]:\ p_k \geq 0.
\end{aligned}$$

This linear program can be represented as

$$\begin{aligned}
&\min\, (0,1,1,0,\ldots,0)\cdot p \\
&\text{s.t.} \\
&\quad \begin{pmatrix} 1 & 1 & \ldots & 1 \\ -1 & -1 & \ldots & -1 \\ 1-\mu & 2-\mu & \ldots & n-\mu \\ \mu-1 & \mu-2 & \ldots & \mu-n \\ -(1-\mu)^2 & -(2-\mu)^2 & \ldots & -(n-\mu)^2 \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \geq \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ -v_{\max} \end{pmatrix} \\
&\quad p \geq 0.
\end{aligned}$$

Recall the symmetric duality

$$\begin{array}{lcl}
\min c^T x & & \max b^T y \\
\text{s.t.} & \leftrightsquigarrow & \text{s.t.} \\
\quad Ax \geq b & & \quad A^T y \leq c \\
\quad x \geq 0 & & \quad y \geq 0.
\end{array}$$

Hence, the dual linear program is

$$\begin{aligned}
&\max\, (1,-1,0,0,-v_{\max})\cdot y \\
&\text{s.t.} \\
&\quad \begin{pmatrix} 1 & -1 & 1-\mu & \mu-1 & -(1-\mu)^2 \\ 1 & -1 & 2-\mu & \mu-2 & -(2-\mu)^2 \\ 1 & -1 & 3-\mu & \mu-3 & -(3-\mu)^2 \\ & & & \vdots & \\ 1 & -1 & n-\mu & \mu-n & -(n-\mu)^2 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_5 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&\quad y \geq 0.
\end{aligned}$$

A direct calculation shows that the vector

$$y^* = \left(1, 0, \alpha, 0, \tfrac{1}{2}\right), \qquad \alpha = \frac{1}{\mu-1} - \frac{\mu-1}{2}$$

is a feasible solution for the dual program for any $\mu \in [2, \sqrt{2}+1]$. The value of the dual program at $y^*$ is $u = 1 - v_{\max}/2$. The weak duality theorem for linear programs implies that $u$ is a lower bound on the value of the primal problem. Hence,

$$\min \mathbb{P}[X \in \{2-\mu, 3-\mu\}] \geq u.$$

This implies that $\mathbb{P}[Z \in \{2,3\}] \geq u$, as desired. $\square$

## J Agreement Between Nearly-Orthogonal Functions

**Claim J.1.** *Let $\varepsilon > 0$, let $\mathcal{X}$ be a set, and let $f, g, h : \mathcal{X} \to \{\pm 1\}$ such that $\{f, g, h\} \in \perp_{\varepsilon,\mathcal{X}}$. Then $\mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) = g(x) = h(x)] \leq \frac{1}{4} + \frac{3\varepsilon}{4}$.*

*Proof.* Denote

$$\begin{aligned} a &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) = g(x) = h(x)] \\ b &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) \neq g(x) = h(x)] \\ c &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) = g(x) \neq h(x)] \\ d &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) \neq g(x) \neq h(x)] \end{aligned}$$

From $\{f, g, h\} \in \perp_{\varepsilon,\mathcal{X}}$ and Fact 2.9,

$$\begin{aligned} a + b &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[g(x) = h(x)] \leq \frac{1}{2} + \frac{\varepsilon}{2} \\ a + c &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) = g(x)] \leq \frac{1}{2} + \frac{\varepsilon}{2} \\ a + d &= \mathbb{P}_{x\sim \mathrm{U}(\mathcal{X})}[f(x) = h(x)] \leq \frac{1}{2} + \frac{\varepsilon}{2}. \end{aligned}$$

Adding these inequalities yields

$$3a + b + c + d \leq \frac{3}{2} + \frac{3\varepsilon}{2}.$$

From the identity $a + b + c + d = 1$,

$$2a \leq \frac{1}{2} + \frac{3\varepsilon}{2},$$

so $a \leq \frac{1}{4} + \frac{3\varepsilon}{4}$, as desired. □

## K Miscellaneous Lemmas

The following result from Achlioptas (2003) is a variant of a lemma of Johnson and Lindenstrauss (1984).

**Theorem K.1** (Johnson–Lindenstrauss)**.** *Let $n, s \in \mathbb{N}$, let $\varepsilon, \beta > 0$, and let $V \subseteq \mathbb{R}^s$ be a set with cardinality $|V| = n$. Let $d \in \mathbb{N}$ such that*

$$d \geq \frac{4 + 2\beta}{\varepsilon^2/2 - \varepsilon^3/3} \ln(n).$$

*Let $R$ be a $d \times s$ random matrix such that each entry is chosen independently and uniformly at random from $\{\pm 1\}$. Let $f_R : \mathbb{R}^s \to \mathbb{R}^d$ be given by $f_R(v) = (1/\sqrt{d}) \cdot Rv$. Then*

$$\mathbb{P}_{R\sim \mathrm{U}(\{\pm 1\}^{d\times s})}\left[\forall u, v \in V : \ (1 - \varepsilon)\|u - v\|_2^2 \leq \|f_R(u) - f_R(v)\|_2^2 \leq (1 + \varepsilon)\|u - v\|_2^2\right] \geq 1 - \frac{1}{n^\beta}.$$

**Claim K.2** (Converse to Birthday Paradox)**.** *Let $d, m \in \mathbb{N}$, and let $\beta \in (0, 1)$. If*

$$m \leq \min\left\{\sqrt{d \ln\left(\frac{1}{\beta}\right)}, \frac{d}{2}\right\}$$

*then $\mathbb{P}_{X\sim (\mathrm{U}([d]))^m}[|X| = m] \geq \beta$.*

*Proof.* We use the inequality $1 - x \geq e^{-x/(1-x)}$, which holds for $x < 1$.

$$\begin{aligned} \mathbb{P}_{X\sim (\mathrm{U}([d]))^m}[|X| = m] &= 1 \cdot \left(1 - \frac{1}{d}\right) \cdot \left(1 - \frac{2}{d}\right) \cdots \left(1 - \frac{m-1}{d}\right) \\ &\geq \prod_{k=0}^{m-1} \exp\left(-\frac{k}{d-k}\right) = \exp\left(-\sum_{k=0}^{m-1} \frac{k}{d-k}\right) \\ &\overset{(*)}{\geq} \exp\left(-\frac{2}{d}\sum_{k=0}^{m-1} k\right) \geq \exp\left(-\frac{m^2}{d}\right), \end{aligned}$$

where $(*)$ follows from $m \le d/2$. Solving $\exp\left(-\frac{m^2}{d}\right) \ge \beta$ yields the desired bound. $\square$

**Theorem K.3** (Hoeffding, 1963)**.** *Let $a, b, \mu \in \mathbb{R}$ and $m \in \mathbb{N}$. Let $Z_1, \ldots, Z_m$ be a sequence of i.i.d. real-valued random variables and let $Z = \frac{1}{m}\sum_{i=1}^{m} Z_i$. Assume that $\mathbb{E}[Z] = \mu$, and for every $i \in [m]$, $\mathbb{P}[a \le Z_i \le b] = 1$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}[|Z - \mu| > \varepsilon] \le 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right).$$

# L Experiments

## L.1 Motivation and Setup

Here, we examine if there are practical algorithms that admit loss stability or even hypothesis stability with substantial numerical values. To this end, we conduct experiments over a simple neural network architecture across four datasets: MNIST, FashionMNIST, CIFAR10, and CIFAR10 with random labels (figures 1-4, respectively). Throughout all experiments, we employ one-hidden-layer perceptrons with 512 hidden neurons. We train the models using stochastic gradient descent (SGD) with a momentum factor of 0.9 and a batch size of 1000, optimizing the cross-entropy loss. For every data set, we train the models across learning rates 0.1, 0.035,[16] and 0.01. We average all the curves over 10 random seeds (tied for the pairs of networks) and plot the standard deviation for all the curves.

The training procedure is as follows: we train two models in tandem, starting from the same random initialization. The first model is provided with the full training set, whereas the second model has $k = 100$ data points removed from its training set. These points are drawn uniformly at random before the beginning of the training, and fixed thereafter. After each epoch, we evaluate the training accuracy, test accuracy and hypothesis stability, i.e., the agreement between the two models (which we calculate across the test set).

We set our main focus on the agreement of the models since the most amenable way to show loss stability might be by way of proving hypothesis stability. The latter can perhaps be mathematically proven in the case of neural networks by analyzing the stability of the training dynamics under two slightly different training sets.

## L.2 Results

Across all experiments, the training and test accuracy of the model pairs are essentially identical throughout the training process. This suggests that at least simple models are loss stable across vision tasks. In order to reduce visual clutter, we hence only plot training and test accuracy of the first model (which has access to the full training set), respectively.

We observe higher agreement for simpler data sets and smaller learning rates. For example, the learning rate has a considerable effect on agreement for CIFAR10 ($\approx 0.65$ for learning rate 0.1 vs $\approx 0.8$ for learning rate 0.01).

The key takeaway from Figures 1 through 4 is that the agreement is consistently higher than the test accuracy. This relationship ensures that when applying the estimation procedure outlined in Theorem 4.3, we can avoid vacuous predictions of perfect accuracy. In the scenarios presented, the estimated accuracy will always be bounded away from 1, as it can be expressed as *test error + (1 - agreement)*. For instance, with a learning rate of 0.01, the maximum estimated accuracies are: 98% for MNIST (compared to 97.5% test accuracy), 90% for FashionMNIST (87% test accuracy), 72% for CIFAR10 (52% test accuracy), and 65% for CIFAR10 with random labels (10% test accuracy). These results illustrate a strong correlation between stability estimation and data complexity.

We repeat the same experiments, modifying the width of the hidden layer to investigate its impact on stability. The results, summarized in Table 1, reveal a strong positive correlation between network width and stability. This effect is particularly pronounced for more complex tasks, such as CIFAR10 and CIFAR10 with random labels. For instance, in the CIFAR10

[16] Except for CIFAR10, we present the results only for learning rate 0.1 and 0.01 to prevent clutter. The qualitative results are consistent across all datasets; that is, the curves of learning rate 0.035 lie between the curves of learning rate 0.1 and 0.01.

random labels setting with a learning rate of 0.01, increasing the width from 256 to 1024 neurons improves agreement from 32% to 50%, highlighting the stabilizing effect of greater network width.

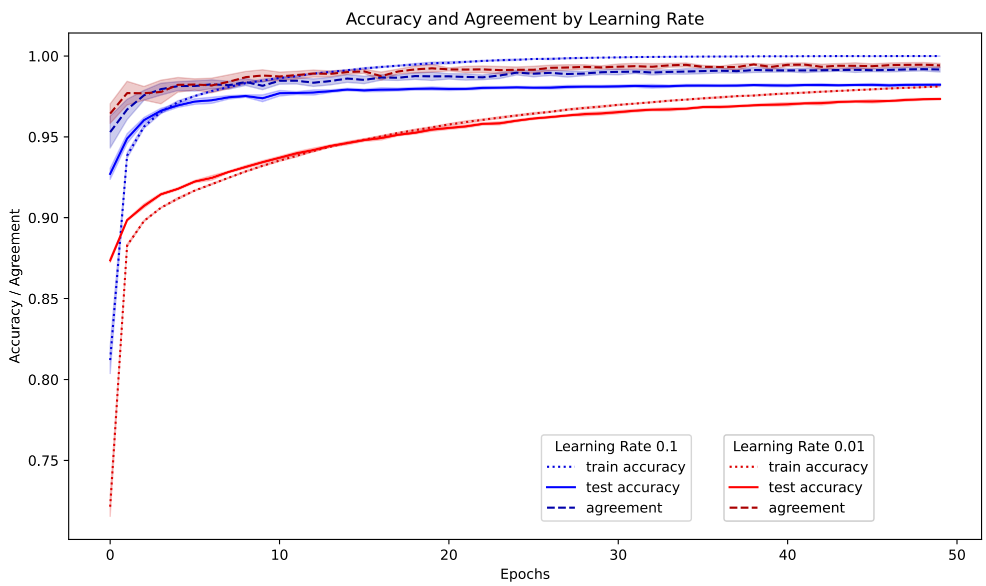


Figure 1: MNIST

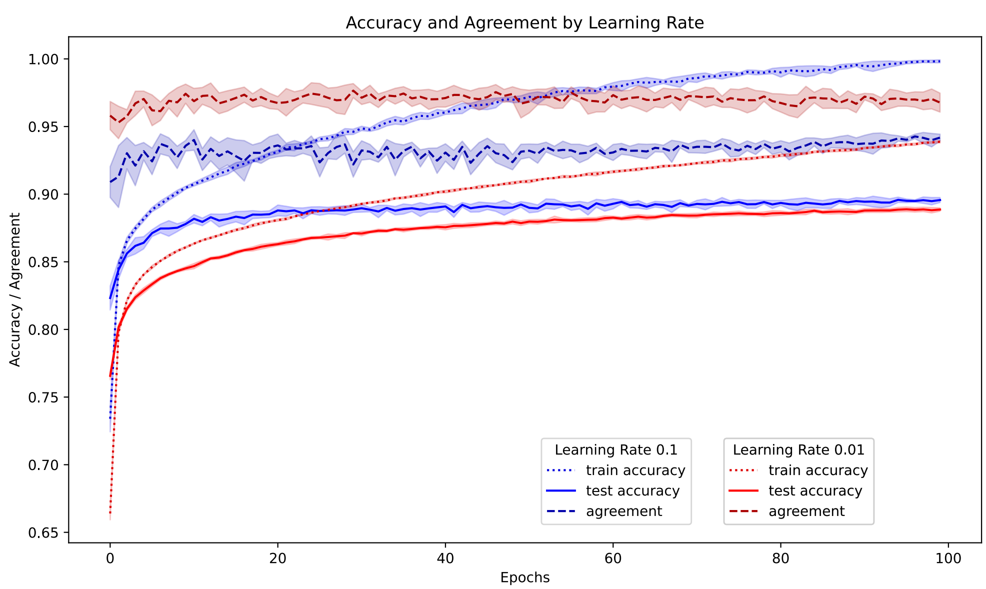


Figure 2: FashionMNIST

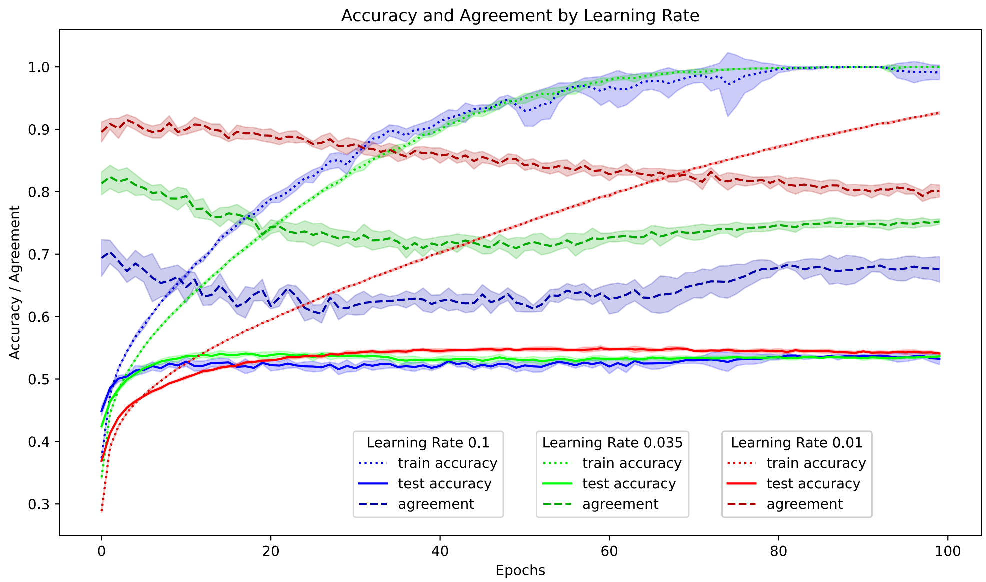


Figure 3: CIFAR10

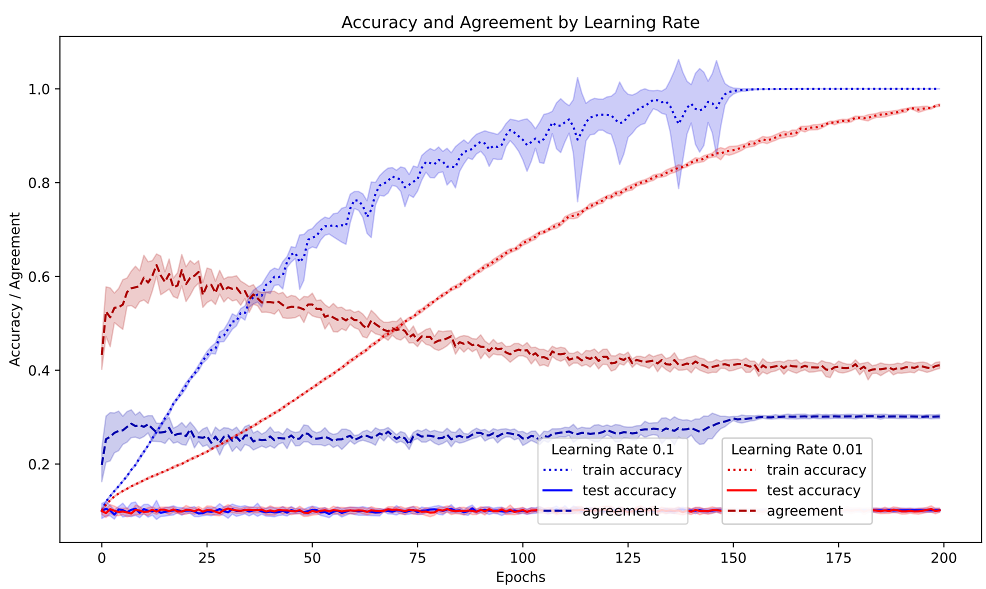


Figure 4: CIFAR10 with random labels

| MNIST | | | FMNIST | | | CIFAR10 | | | CIFAR10 - RAND | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **#N** | **lr** | **Agree** | **#N** | **lr** | **Agree** | **#N** | **lr** | **Agree** | **#N** | **lr** | **Agree** |
| 256 | 0.1 | 99% | 256 | 0.1 | 92% | 256 | 0.1 | 62% | 256 | 0.1 | 21% |
| 256 | 0.01 | 99.5% | 256 | 0.01 | 97% | 256 | 0.01 | 71% | 256 | 0.01 | 32% |
| 512 | 0.1 | 99% | 512 | 0.1 | 94% | 512 | 0.1 | 67% | 512 | 0.1 | 30% |
| 512 | 0.01 | 99.5% | 512 | 0.01 | 97% | 512 | 0.01 | 80% | 512 | 0.01 | 41% |
| 1024 | 0.1 | 99% | 1024 | 0.1 | 95% | 1024 | 0.1 | 76% | 1024 | 0.1 | 39% |
| 1024 | 0.01 | 99.5% | 1024 | 0.01 | 98% | 1024 | 0.01 | 85% | 1024 | 0.01 | 50% |

Table 1: Agreement percentages across datasets with varying number of neurons in the hidden layer (#N) and learning rates (lr). The setup is the same as in L.1 except for the number of training epochs, which is $\{50, 150, 150, 300\}$ for {MNIST, FMNIST, CIFAR10, CIFAR10 random}, respectively. In scenarios where agreement has not yet reached saturation, agreement is positively correlated with the width of the network.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This is a theory paper and we present all our results informally for better readability before the formal results in later sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: no limitations.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All proofs are in the appendix

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We run simple experiments and we explicitly mention all hyperparameters to reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: the experiments are elementary; nevertheless, if reviewers wish to see the code, we can provide it.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the experimental details appear in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: There are error bars in all our figures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No] .

   Justification: Very simple experiments executed on a laptop, no special resources needed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This is a theory paper so there are no ethical concerns for potential harmful consequences.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: Purely theoretical work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: Purely theoretical work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: No external resources used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: No new assets introduced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This is a theory paper with no crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: This is a theory paper with no crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: No use of LLMs as mentioned above.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.