VOLUME HYPOTHESIS OR ALGORITHMIC BIAS: MARGIN CAN SEPARATE GRADIENT DESCENT FROM RANDOM SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Over-parameterized neural networks often generalize despite having the capacity to overfit their training data. Two proposed explanations are the inductive bias of optimization algorithms and the *volume hypothesis*, which posits that most interpolating solutions, measured under a fixed prior, generalize well. We investigate these explanations jointly by comparing Guess-and-Check (G&C), which fits the data through rejection sampling, and stochastic gradient descent (SGD). We focus on favorable settings under the *margin assumption*, where generalizing interpolators indeed carry significant volume, and optimization algorithms can utilize the structure through their respective bias.

For our investigation, we first provide a novel sample-dependent generalization bound for G&C. We then prove a sharp statistical separation between G&C and GD in a Gaussian model, and finally: We show empirically, on common datasets, that G&C generalizes by effectively searching a finite hypothesis class but achieves little margin, whereas SGD attains markedly larger input-space margins and better generalization. Our theory and experiments challenge the apparent equivalence of G&C and SGD: even when random sampling generalizes, optimization can exploit the landscape's margin structure to generalize better.

1 INTRODUCTION

Modern neural networks are commonly trained in an over-parameterized regime. Counter-intuitively, they carry capacity that enable to interpolate their training set, and yet reliably generalize to unseen data (Zhang et al., 2016; Belkin, 2021; Kaplan et al., 2020). Why training finds a generalizing interpolator, rather than one of the many that merely overfit, remains open, and the answer is generally attributed to some mix of the architecture and the optimization method.

Indeed, one approach explains generalization through the *inductive bias* of the learning procedure (Neyshabur et al., 2014; Belkin, 2021). Explicit regularization and implicit optimization bias can favor certain solutions among the many interpolators available in an over-parameterized model. This affectively restrict the capacity of the search space to be considered, and steer the algorithm towards *simple* or *structured* solutions.

A complementary line of work (Chiang et al., 2023; Huang et al., 2020) attributes generalization to the loss landscape geometry. It is assumed that under a fixed prior, conditioning on interpolating solutions, the volume of good classifiers is significant. This perspective is often depicted through the *Guess-and-Check* (G&C) algorithm that samples parameters from a fixed prior until it finds an interpolating solution. Rather than a practical training method, G&C serves as a thought algorithm that isolates the contribution of the landscape and prior from the bias of gradient-based optimization.

In this work, we compare and reconcile these approaches in arguably the most well-studied assumption: The *margin assumption*. For linear classification, the margin assumption assumes a separating hyperplane where example points lie well away from the decision boundary. Under this structural assumption, the generalizing solutions have significant volume under a fixed uniform prior over the sphere. On the other hand, the assumption has led to the design of explicitly margin-maximizing algorithms such as SVM (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1982), and it was observed that several gradient based methods are, implicitly, biased towards margin maximizing solutions (Soudry et al., 2018).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The margin assumption therefore provides a common setting for comparing these explanations: We ask specifically, when the landscape already supports generalization by random sampling, can optimization further exploit the same structure to generalize better?

Our Contribution Towards answering these questions, we first revisit G&C and provide novel generalization bounds that scale with the number of *distinct prediction patterns* produced. This improves over the existing analyses that are based on the *likelihood* of interpolation. We then study an overparameterized two-Gaussian model that admits poorly generalizing interpolators, yet enable learning. We obtain matching upper and lower bounds for G&C as well as a Gradient-based method. These bounds establish a sharp statistical separation in this model: G&C generalizes without necessarily finding large-margin solutions, whereas SGD, with a prescribed loss function, exploits the margin structure to achieve better generalization. Thus, a setting favorable to random interpolation can still admit a statistical advantage from optimization.

We then continue to an empirical study, and compare the solutions of G&C and SGD across tasks, architectures and sample sizes, on benchmark datasets as well as synthetic data. First, we reproduce existing experiments (Chiang et al., 2023) in which the two algorithms achieve comparable generalization performance. Rather than limiting the comparison to test accuracy, we extend these experiments by examining the properties of the returned solutions and the searches that produce them. This allows us to ask whether similar performance reflects similar generalization mechanisms.

Our experiments confirm and provide evidence for two distinct mechanisms for generalization as in our theoretical analysis. SGD returns solutions with substantially larger margins, and its generalization tracks those margins. G&C, in contrast, finds solutions with smaller margins, yet still generalizes. Its performance, instead, tracks the number of distinct prediction patterns encountered before interpolation. These observations support the distinction suggested by our theoretical analysis: Random search can already yield generalization through a union-bound argument, while gradient-based optimization contributes an additional bias toward larger-margin solutions. We further examine this distinction through stress tests and ablations, and find that the relationship between each algorithm's generalization and its respective measured quantity persists across these variations.

Related Work Generalization of overparameterized models is a highly active field of study. Several works take the approach of inductive bias (Hardt et al., 2016; Neyshabur et al., 2014), while others attempt to investigate the geometry of the loss landscape empirically (Chiang et al., 2023; Mingard et al., 2021; Ramanujan et al., 2020), as well as in theory (Alexander et al., 2025; Feder et al., 2025; Theisen et al., 2021; Buzaglo et al., 2024; Harel et al., 2024). The weak learnability of G&C algorithm under margin was studied by Blum (2005). We focus here on the Gaussian case and provide tight analysis for strong-learnability. Theisen et al. (2021) analyze G&C under an exact *equicorrelation* assumption on the data, deriving asymptotic expressions for the prediction accuracy. We study a Gaussian-mixture data without imposing equicorrelation, and establish non-asymptotic upper and lower bounds on the population error.

The bias of gradient-based methods toward large-margin solutions has been studied in several setups (Soudry et al., 2018; Lyu & Li, 2020; Jiang et al., 2019; Elsayed et al., 2018). Our work contributes, in the Gaussian setting, a novel direct finite-time analysis of GD and we obtain tight explicit and non-asymptotic generalization guarantees.

Chiang et al. (2023) observed the remarkable performance of G&C on small-scale data and compared it to GD. Peleg & Hein (2024) directly compare both algorithms and observe a performance gap as width increases. They also interpret SGD's width-dependent advantage in margin terms. We complement this analysis with estimates of input-space boundary distances, systematic comparisons of margins, search complexity, and a theoretical separation in a Gaussian model. They also track the likelihood of interpolation as sample size increases. They observe that this measure can be vacuous and identify the cause. In light of this, we propose the new measure of *number of distinct patterns*.

Beyond Linear Gaussian classification, a standard case-study for generalization is the more general model of Stochastic Convex Optimization (SCO). The implicit bias of GD has been extensively investigated (Dauber et al., 2020; Amir et al., 2021a;b). Several works suggest that SGD's generalization cannot be accounted for by its bias (Vansover-Hager et al., 2025). The affect of the loss landscape on generalization has also been studied (Schliserman et al., 2025). Particularly, Burla & Livni (2026) showed how some algorithms can learn while G&C and the volume hypothesis collapse.

2 SETUP

We consider a standard setting of supervised learning where *labeled examples* (x, y) drawn from a distribution $\mathcal{D}$ supported over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. In this setup, we also have parameter space Θ where each $\theta \in \Theta$ determines a classifier $h_\theta : \mathcal{X} \to \mathcal{Y}$. Throughout we focus on binary classification, i.e. $\mathcal{Y} = \{-1, +1\}$, and we assume that each parameter determines a real-valued function $f_\theta : \mathcal{X} \to \mathbb{R}$ that induces a predictor through its sign:

$$h_\theta = \text{sign}(f_\theta).$$

The *decision boundary* of θ is the zero-level set of f_θ: $\mathcal{B}_\theta = \{x : f_\theta(x) = 0\}$.

Given a fixed parameter space Θ, a learner is an algorithm that observes a finite sample $S = \{x_i, y_i\}_{i=1}^m$ of labeled examples and returns parameters θ. At a high level the objective of the learner is to minimize what is termed *The population loss*, defined for a given θ as:

$$L_{\mathcal{D}}(\theta) \;=\; \Pr_{(x,y)\sim\mathcal{D}}\big[\, h_\theta(x) \neq y \,\big]. \tag{1}$$

Our focus is on algorithms that are termed *interpolating* algorithms. These are algorithms that, given a sample S, return a solution from the interpolating set defined as:

$$\mathcal{I}(S) \;=\; \{\, \theta \in \Theta : h_\theta(x) = y,\ \forall (x, y) \in S \,\}. \tag{2}$$

For concreteness we will say that the algorithm fails if $\mathcal{I}(S) = \emptyset$. However, we mostly focus on the overparameterized regime where it is assumed that this set is nonempty.

2.1 GUESS-AND-CHECK (G&C)

Guess-and-Check is an interpolating algorithm that returns a solution by a sequence of draws. Given a parameter space Θ, G&C fixes a prior P over it, and draws $\theta \sim P$ independently and repetitively until the first interpolating draw. In more detail, let $\theta_1, \theta_2, \ldots$ be independent draws from a fixed prior P, independently of S, and define

$$\tau := \min\{t : \theta_t \in \mathcal{I}(S)\}, \qquad \widehat{\theta}_{\text{GC}} := \theta_\tau.$$

Alternatively, G&C can be defined as a sampler from the distribution:

$$P_S \;:=\; P\big(\,\cdot \mid \theta \in \mathcal{I}(S)\,\big). \tag{3}$$

An important quantity when analyzing G&C is the *interpolation probability*

$$p(S) := P(\mathcal{I}(S)).$$

One can show that the expected number of draws of $G\&C$ satisfy $\mathbb{E}[\tau|S] = 1/p(S)$. Classical methods (McAllester, 1999) provide guarantees w.p $1 - \delta$ on the population loss of order:

$$L_{\mathcal{D}}\left(\hat{\theta}_{GC}\right) = O\left(\frac{\log 1/p(S) + \log 1/\delta}{m}\right). \tag{4}$$

The above bound is very convenient in theory where we often assume a uniform, evenly distributed, prior over the hypothesis space. However, in practice, when the prior is often defined over the parameter space instead, we find that this bound may be too loose, and can even become vacuous. This limitation is already visible in Peleg & Hein (2024, Figure 3 and Section 3.3). There, G&C achieves above-chance accuracy in settings where the interpolation probability falls below the 2^{-m} baseline. As they note, this is caused by a bias toward constant predictions. Indeed, fix a prior P such that w.p $1 - 2^{-\Omega(m)}$ outputs a constant function, then the generalization bound above becomes vacous. This motivates counting *distinct prediction patterns* rather than raw draws. Therefore, in the experiment section we replace $1/p(S)$ from the classical bounds with a fine-grained measure $N(S)$ of distinct hypotheses. In section 3 we provide formal definition and justification for this replacement.

2.2 Stochastic Gradient Descent (SGD)

The family of Gradient Descent methods are optimization methods that are utilized to interpolate classification problems by a sequence of gradient-based updates. Given a parameter space Θ. SGD, without any momentum, weight decay or any regularization term, draws an initialization $\theta_0 \sim P_0$ and iterates

$$\theta_{t+1} \;=\; \theta_t \;-\; \eta\, \nabla_\theta \frac{1}{|B_t|} \sum_{(x,y)\in B_t} \ell\big(\theta_t; x, y\big), \tag{5}$$

over batches $B_t \subseteq S$ of size b, where ℓ is a differentiable surrogate for the zero-one loss, taken throughout as the cross-entropy.

Leave-One-Out SGD (SGD-LOO). Because SGD has hyperparameters, for fair comparison with the parameterless G&C we consider Leave-one-out (Stone, 1974) as a method to choose the learning rate. The Batch-size was set similar to the experiments conducted by Chiang et al. (2023).

Margin The margin of a classifier is a central concept when studying the generalization of optimization algorithms, and in particular Gradient based methods. In the homogeneous linear case, $f_\theta(x) = \langle \theta, x\rangle$, the margin, γ, is defined as the distance between the decision boundary and the closest point. If θ is normalized so that $\min_{x\in S} |f_\theta(x)| = 1$ on the sample, then the margin coincides with the inverse of the norm $\|\theta\| = 1/\gamma$. Advantageously, it is known (Cortes & Vapnik, 1995; Vapnik, 1998) that in the linear case, obtaining a classifier with large margin (as is performed by algorithms such as SVM) leads to superior generalization guarantees. Specifically, a large-margin solution can have generalization guarantee of order $O(1/\gamma^2 m)$, while the worst case generalization analysis for linear classifiers in dimension d scales linearly with the dimension d.

In non-linear classification, we will define the margin of a labeled point as its signed Euclidean distance to the decision boundary,

$$d_\theta(x, y) \;=\; \operatorname{sign}\big(y\, f_\theta(x)\big) \cdot \min\{\, \|x' - x\|_2 \;:\; x' \in \mathcal{B}_\theta \,\}, \tag{6}$$

For a nonlinear f_θ, there is no closed-form solution for the margin, and several estimators of it exist (Elsayed et al., 2018; Jiang et al., 2019). In our empirical investigations, we estimate it with the walk described by Moosavi-Dezfooli et al. (2016), which iterates the linearization of f_θ towards its zero-level set, starting from x_0:

$$x_{t+1} \;=\; x_t \;-\; \alpha\, \frac{f_\theta(x_t)}{\|\nabla_x f_\theta(x_t)\|_2^2}\, \nabla_x f_\theta(x_t), \tag{7}$$

with overshoot $\alpha > 1$, reporting $\|x_T - x_0\|_2$ with the sign of Equation (6). Per solution we report a low percentile of these distances over the training points rather than their minimum, which is brittle to a single outlying point. Remaining technical details are deferred to Section A.4.

3 Theoretical Analysis

In this section we provide two theoretical contributions: We begin with a more refined analysis of the generalization of G&C. We then proceed to analyze the sample complexity of the two learning methods in arguably the simplest model of linear classification under Gaussian data.

3.1 Generalizattion Bound for Rejection Sampling

Given a sample $S = ((x_i, y_i))_{i=1}^m \sim \mathcal{D}^m$, suppose that G&C produces the hypotheses $\theta_1, \ldots, \theta_\tau$ until interpolation. We define the random variable $N(S)$ to be:

$$N(S) = \left|\{(h_{\theta_t}(x_1), \ldots, h_{\theta_t}(x_m)) : t = 1, \ldots, \tau\}\right|.$$

We can state a generalization bound in terms of $N(S)$ instead of $1/p(S)$. A full proof as well as formal restatement are provided in section C.

Theorem 1 (Informal)**.** Suppose that G&C succeeds a.s. on a distribution $\mathcal{D}$. Then w.p. $1-\delta$ over the sample and runs of G&C:

$$L_{\mathcal{D}}(\widehat{\theta}_{\mathrm{GC}}) \le O\left(\frac{\log(N(S)/\delta)}{m}\right). \tag{8}$$

Comparing equation (8) to equation (4): Consider a prior that is biased towards a constant classifier. While $1/p(S)$ can be arbitrarily large, note that for the measure $N(S)$, given any randomly drawn output sequence of hypotheses, the constant hypothesis contributes only 1 to $N(S)$.

3.2 Gaussian Model

We next consider the model of labeled Gaussian mixture. Given $\mu \in \mathbb{R}^d$ with $\|\mu\| = \gamma > 0$, define the distribution $\mathcal{D}_\gamma$ by

$$x = y\mu + \frac{1}{\sqrt{d}}n, \qquad n \sim \mathcal{N}(0, I_d), \qquad y = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2, \end{cases} \tag{9}$$

where n and y are independent. We consider homogeneous linear classifiers $h_\theta(x) = \text{sign}(\langle \theta, x \rangle)$. For a fixed sample S, let $\widehat{\theta}_{\text{GC}}$ be the output of G&C with the uniform spherical prior:

$$\widehat{\theta}_{\text{GC}} \mid S \sim \text{Unif}\left\{\theta \in \mathbb{S}^{d-1} : y_i \langle \theta, x_i \rangle > 0 \text{ for every } i \leq m\right\}.$$

We next provide informal statements of our main results. Formal statements as well as proof are presented in section B. Our first result bound the performance of G&C.

Theorem 2 (Informal)**.** Let $\mathcal{D}_\gamma$ be the distribution in (9), for an arbitrary $\mu \in \mathbb{R}^d$ satisfying $\|\mu\| = \gamma$ and $1/\sqrt{m} \lesssim \gamma < 1$. For $\gamma\sqrt{d} = \Omega(\sqrt{\log m})$, and $m < d$, w.p $\Omega(1)$:

$$L_{\mathcal{D}_\gamma}\left(\widehat{\theta}_{\text{GC}}\right) = \Theta\left(\min\left\{1, \frac{1}{\gamma^2 m}\right\}\right).$$

In particular for $m\gamma^2 > 1$,

$$\mathbb{E}\left[L_{\mathcal{D}_\gamma}\left(\widehat{\theta}_{\text{GC}}\right)\right] = \Omega\left(\frac{1}{\gamma^2 m}\right)$$

The above result establishes a tight, up to constant factors, generalization guarantee for G&C. The assumption on $\gamma\sqrt{d}$ is necessary. Indeed, as discussed in equation (16), the classification error of any classifier is bounded by $\Omega(e^{-\gamma^2 d})$. In turn, when $\gamma^2 \ll \log m/d$, we start to obtain a generalization error above $\Omega(1/m)$. As such, any interpolating solution has large generalization gap.

Comparing the lower bound for the classification error of G&C to the sample complexity of max-margin under the Gaussian model (Wang & Thrampoulidis, 2021; Chatterji & Long, 2021; Cao et al., 2021) we immediately observe an exponential gap in performance. Moreover, Soudry et al. (2018), shows that the output of GD with logistic-loss, asymptotically converges to the max-margin classifier. Hence, in the limit one can relate its performance and obtain similar results. Here, we add further a non-asymptotic finite time analysis that bounds the generalization error directly at the same sample rate. More formally, given a sample S, define

$$F_S(\theta) := \frac{1}{m}\sum_{i=1}^{m} \log(1 + \exp(-y_i \theta^\top x_i)), \qquad \Rightarrow \qquad \theta_t = \theta_{t-1} - \eta \nabla F_S(\theta_{t-1}) \tag{10}$$

Theorem 3 (informal)**.** Let $\mathcal{D}_\gamma$ be the distribution in (9), for an arbitrary $\mu \in \mathbb{R}^d$ satisfying $\|\mu\| = \gamma$ and $0 < \gamma < 1$. For sufficiently large d, assume $\gamma\sqrt{d} = \Omega(\sqrt{\log m})$, and $m < d$. Run gradient descent as defined in equation (10), with a step size $0 < \eta \leq \frac{2}{5}$, initialized at any $\|\theta_0\| = O(1/\gamma)$. then for $\eta T = \Omega(1/\gamma^2)$, if $m\gamma^2 > 1$:

$$\mathbb{E}[L_{\mathcal{D}_\gamma}(\theta_T)] = e^{\Theta\left(-\gamma^2 d\right)}.$$

Formal statement and proof are provided in section B.2. Comparing Theorems 2 and 3, as we consider the overparameterized regime where $m \ll d$, we conclude that while G&C performs better than a worst case ERM, a well designed algorithm such as Gradient Descent can obtain an exponentially better performance.

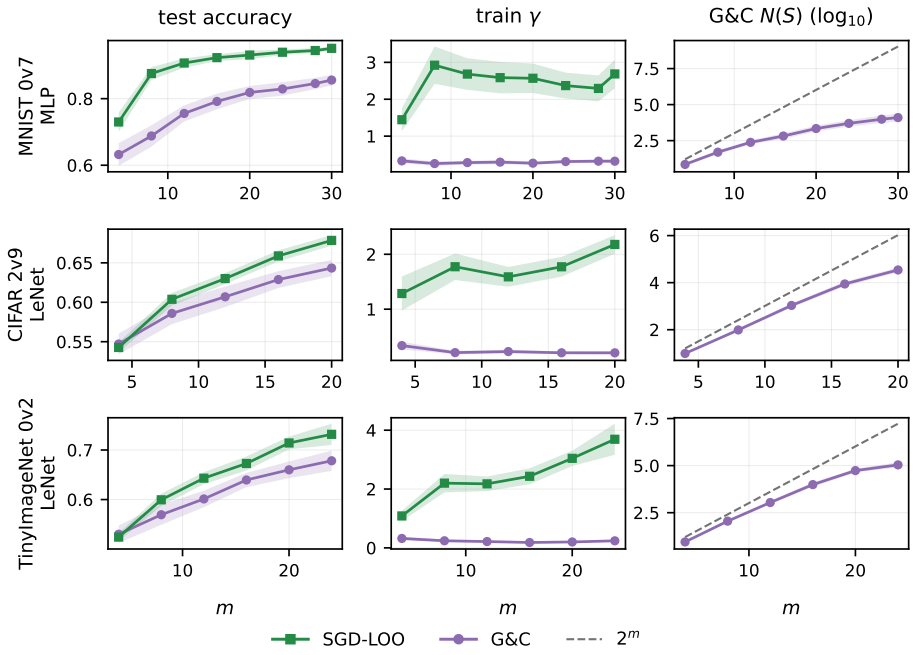


Figure 1: G&C and SGD-LOO across training-set sizes m, on three task-architecture pairs: test accuracy (left), train margin γ (middle), and G&C's $N(S)$ against the 2^m labelings of the sample, dashed (right). Bands are 95% intervals over 50 data seeds per point. Additional runs in Section A.5.

4 Experiments

Next, we provide empirical evidence that the algorithms are biased in different directions within $\mathcal{I}(S)$ and that their generalization performance are governed by their each respective measure. We first compare the two algorithms directly, across tasks, architectures and sample sizes, and measure how the properties above track generalization. We then add two further stress-test experiments to further measure the correlation between the success of the algorithm and its output behaviour.

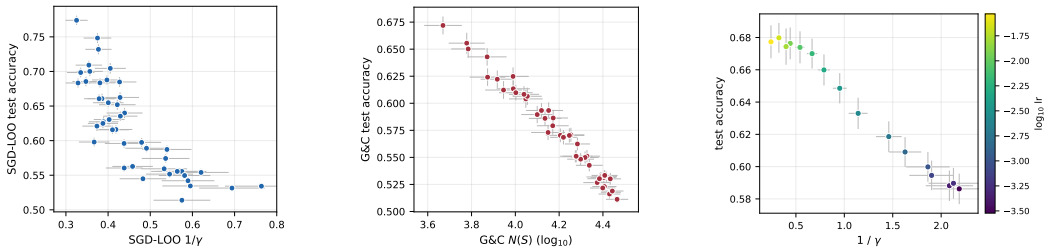
All experiments are binary classification tasks built from pairs of MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky, 2009), and TinyImageNet (Le & Yang, 2015) classes, on two architectures: LeNet (LeCun et al., 1998) and standard MLP, with further detail in Section A.1. In each of those, we measure test accuracy, an estimate of $L_{\mathcal{D}}$; the margin as defined in Equation (6), aggregated by a low percentile over the sample set;

For G&C, following equation (8) we measure $N(S)$ – the number of distinct hypotheses drawn until interpolation. Unlike the looser measure $1/p(S)$ used in previous work (Peleg & Hein, 2024), we indeed observe it being below the exponential-in-sample-size baseline.

Throughout the experiments, SGD returns the first iteration that interpolates $T = \min\{\, t : \theta_t \in \mathcal{I}(S) \,\}$, so that nothing is optimized past the point where the training error vanishes. This is consistent with Chiang et al. (2023). G&C draws its parameters from $\mathcal{U}[-1, 1]$ throughout the experiments below, following Chiang et al. (2023). We did not find this choice to matter substantially in our setting (Peleg & Hein (2024) report a similar pattern), and G&C in this regime seems to be relatively insensitive to the prior family across Kaiming, uniform and Gaussian choices.

4.1 Comparing G&C and SGD-LOO

We measure the quantities above under three separate variations: over the training-set size m, over the task, and over SGD's learning rate. Each is reported below with the tasks and architectures it uses. Throughout, G&C may exhaust its draw budget without finding an interpolating parameter, and every comparison is restricted to the samples where it did not (Section A.2).



(a) (b) (c)

Figure 2: Algorithm's test accuracy against either $1/\gamma$ or $N(S)$. a and b: each point is a binary CIFAR-10 task at m=16, on MLP, averaged over 200 data seeds, SGD-LOO against its $1/\gamma$ and G&C against its $N(S)$. figure 2c: each point is one learning rate on CIFAR-10 2v9 at m=16, on MLP, averaged over 200 data seeds, with colour giving the rate. G&C is outside the range shown and is given in the text. Additional runs in Sections A.6 and A.7.

Accuracy, margin and search cost against sample size. We sweep the training-set size m from 4 to 30 on MNIST, from 4 to 20 on CIFAR-10 and from 4 to 24 on TinyImageNet, running both algorithms on every data seed, reporting test accuracy, train margin and $N(S)$ (Figure 1). SGD-LOO reaches at least as high a test accuracy as G&C at every sample set size m, and strictly higher from m=8 upward. Additionally, we observe that G&C's margin is comparably lower. In contrast, though, the measure $\log(N(S))$ is strictly below the m threshold. Effectively, then, G&C samples from a relatively small $N(S)$, and reaches non-trivial generalization error; this is qualitatively consistent with the bound as described in Equation (8). Additional tasks are provided in Section A.5.

We further explore, across tasks, the correlation between each algorithm's performance and the relevant capacity measure. Fixing m=16 as a reference size (following Chiang et al. (2023) and Peleg & Hein (2024)), and taking all 45 binary pairs of CIFAR-10 classes, we run both algorithms on each pair, measuring the three quantities. We provide a scatter plot for each algorithm (Figures 2a and 2b). The scatter shows SGD-LOO's accuracy tracks its margin at a Spearman rank correlation of $\rho = -0.87$, and G&C's tracks its $N(S)$ at $\rho = -0.98$. Additional runs in Section A.6.

Learning-rate variance. In Section 2 we noted that SGD's performance varies widely with the learning rate, which is why we select it by leave-one-out rather than by hand. We find that this is necessary for proper comparison. In figure 2c we measure both test accuracy and train margin across different learning rates. We find that for different learning rates SGD can outperform or degrade with respect to G&C, so a single choice is enough to reproduce either of the conclusions already in the literature Chiang et al. (2023); Peleg & Hein (2024). Remarkably, though the performance and margin track each other closely across the grid ($\rho = -0.99$). While G&C could outperform SGD, its inverse-margin was 5.46 well outside the scope of the margin obtained by SGD. This is yet another evidence that its performance is not tracked by its margin, in contrast with SGD.

In Figure 3 we illustrate the decision boundary of both methods under such a projection: SGD's boundary is smooth and stands further from the training points in this projection, compared to G&C's more erratic behavior.

4.2 Stress Tests and Data Ablation

We further explore the margin advantage of SGD over G&C, and the correlation between the generalization of each algorithm and the proposed property, through controlled distributional variations on the data and on the prior.

4.2.1 Behavior under adversarial reweighting

As a first experiment, we construct a sequence of increasingly adversarial distributions to observe how this affects the algorithm's performance measures. In a nutshell, we begin with a benchmark distribution and reweigh the data points to make a new distribution which is, be design, harder or easier for the algorithm to learn. Then, when the algorithm's performance improves or deteriorates we track again the capacity measure in question.

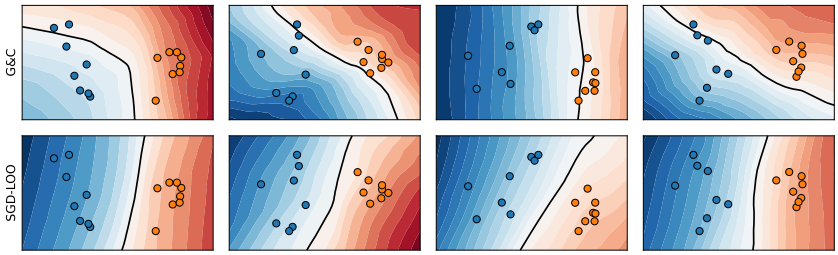


Figure 3: Decision boundaries of G&C (top) and SGD-LOO (bottom) on MNIST 0-vs-7, MLP width 64, m=16, four data seeds, drawn in PCA two-dimensional projection of the full 0-vs-7 training pool. Colour is f_θ, scaled independently in each panel, the black contour is $\mathcal{B}_\theta$, and the points are the training samples. The projection is described in Section A.8.

The reweighting of the distribution is performed as follows: We begin with a uniformly weighted distribution $\mathcal{D}$ on restricted dataset. Then, in each iteration, we select a subset of size m from $\mathcal{D}_t$ through i.i.d sampling according to $\mathcal{D}_t$. We then search for a G&C solution on the selected subset to reach a predictor h_t, and evaluate the dataset $\mathcal{D}_t$ on it. The data is then reweighted, to give higher weight to "hard examples" as follows:

$$\mathcal{D}_{t+1}(x) \;\propto\; \mathcal{D}_t(x) \cdot \exp\big(-\alpha_t \,(2 \cdot \mathbb{1}[h_t(x) = y] - 1)\big), \qquad \alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right), \tag{11}$$

where ε_t is h_t's weighted error on $\mathcal{D}_t$, and w_{t+1}. The update of the distributions is as proposed by by AdaBoost (Freund & Schapire, 1997), with h_t playing the role of the round-t weak learner. Indeed, the framework of Adaboost generates adversarial distributions with the following guarantee: After fixed number of steps (that scale gracefully with the training set and performance of the weak learner) either the ensemble of output classifiers reaches a small error (i.e. strong learning) or we must reach an adversarial distribution on which the algorithm fails to produce even a weak learner, at which point G&C fails and we can observe if the capacity measure tracks this. We also apply a rank-reversal transformation on $\mathcal{D}_t$, swapping "hard" samples with "easy" ones, reversing the direction of the procedure so that rounds become progressively easier instead.

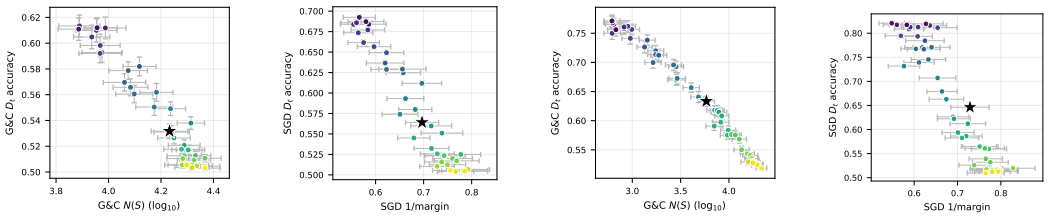
Results. On MNIST tasks, the ensemble inveed converges on $\mathcal{D}_0$. Thus, we report that G&C is strong enough to produce ensemble of weak learning, able to classify MNIST. CIFAR, though turns out to be more challenging. We indeed reach failure, where we defined failure as the round where ε_t consistently exceeds $\frac{1}{2}$ over multiple sample redraws. We plot the generalization scatter plots against the margins and against $N(S)$, across all the different reweighted distributions $\mathcal{D}_t$ in Figure 4. We find that the correlation theses of both algorithms hold throughout: as the rounds grow more adversarial, generalization decays, with G&C's $N(S)$ and SGD's $1/\gamma$ both growing; as they grow easier instead, generalization improves, and both shrink together (CIFAR 0v8: $\rho = -0.93$ for G&C and $\rho = -0.95$ for SGD-LOO; CIFAR 2v9: $\rho = -0.99$ and $\rho = -0.92$). Additional runs in Section A.9.

4.2.2 Behavior under Pretraining

Where boosting varies the task by reweighting the data, pretraining varies it by replacing the prior. In this experiment, we perform a small-scale pretraining to introduce the weights θ_0. We then swap SGD's initialization $P_0 \mapsto \theta_0$, and swap G&C's fixed parameter-space prior for perturbations around the pretrained weights, following a similar setup by Gan & Isola (2026).

$$\theta \;=\; \theta_0 \;+\; \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{12}$$

For θ_0, we use a deliberately simple held-out analogue. We select a downstream pair, and treat the rest of the classes as the pretraining pool. We split those into two groups, with the pretraining task being binary classification between those. This pretraining is done with standard sample size m and SGD. Importantly, this lets us keep a binary head that carries some information on the dataset, without being explicitly exposed to the downstream data. We vary the pretraining class split to generate a pool of checkpoints for per dataset.



(a) (b) (c) (d)

Figure 4: Adversarial Reweighting: Each point depicts a reweighted experiment, $\mathcal{D}_t$, towards harder (Equation (11)) or easier tasks. The star marks the initial distribution $t = 0$, so samples on the left are easier tasks and samples on the right are harder tasks. Error bars are 95% intervals on both axes, averaged from 200 data seeds. a and b: CIFAR-10 0v8, LeNet, over 38 total distributions, G&C against its $N(S)$ SGD-LOO against its $1/\gamma$. c and d: the same two quantities on CIFAR-10 2v9, LeNet, over 47 total distributions. Additional runs in Section A.9.

Table 1: Pretrained G&C and SGD solutions compared to default prior. Each row reports a downstream task, with 95% CI averaged across 12 different pretraining splits and 200 data seeds each.

task pool	G&C acc gain	$N(S)$ reduction ratio	SGD acc gain	SGD margin gain
CIFAR 2v9	0.102 ± 0.036	$0.06\times$ [0.03, 0.10]	0.035 ± 0.021	1.206 ± 0.226
CIFAR 0v8	0.031 ± 0.010	$0.12\times$ [0.08, 0.19]	0.019 ± 0.010	1.740 ± 0.217
TinyImageNet 0v1	0.098 ± 0.020	$0.25\times$ [0.19, 0.33]	0.086 ± 0.015	3.617 ± 0.578

Pretraining Results. Table 1 reports G&C's generalization gain and $N(S)$ reduction ratio, compared to an equivalent from-scratch floor, averaged across multiple data seeds and across pretraining splits. Both SGD and G&C achieve higher test accuracy; G&C draws from a significantly smaller $N(S)$, and SGD samples with higher margins. We note that in this section, SGD trains for a fixed T=50 epochs rather than stopping at interpolation, as explained in Section A.10.

5 Discussion

In this work, we examined algorithmic performance in relation to the *volume hypothesis* and *inductive bias*. Our experiments suggest a more intricate picture than attributing generalization to one or the other. Our theoretical model exhibits a setting where a substantial fraction of interpolating solutions generalize reasonably well, yet a carefully designed algorithm achieves significantly better performance. This motivates a more nuanced interpretation of the volume hypothesis: moderately good interpolators may be abundant, while reaching near-optimal performance may still require additional bias and algorithmic design.

Another important aspect of our work, is that the difference in margin suggests that the posterior over predictors induced by SGD and G&C can vary significantly. It has been conjectured that SGD behaves similar to a Bayesian sampler (Mingard et al., 2021). in our setting we observe distinct behavior with distinct performance. As such, the performance of GD cannot be naively connected to the success of G&C. Even in favorable setting for both algorithms, their behavior is distinctive.

A central limitation of our work and previous studies is the computational cost of scaling G&C. Rejection sampling can become prohibitive as the sample size and number of parameters increase, restricting direct empirical comparisons to small-scale settings. Our $N(S)$ bound highlights the distinction between statistical success and computational feasibility: meaningful generalization guarantees can coexist with even *super-exponential* rejection costs. Although the margin bounds we invoke have no explicit dimension dependence, this alone does not establish that the observed separation persists at larger scales. A concrete question is therefore how $N(S)$ and the margins evolve with sample size and model capacity. Predicting these quantities from the architecture, prior, and data distribution, without executing G&C, could clarify when random interpolation remains statistically effective and when optimization achieves a better rate.

AI USE STATEMENT

We used generative AI tools to assist with coding and running experiments, including developing and adapting code and monitoring experimental runs. All AI-assisted code was manually reviewed and tested by the authors. AI tools were also used to assist with LaTeX typesetting and to create figures, tables, plots, and visualizations from experimentally generated data. AI tools also assisted in finding critical ingredients for proving mathematical claims (e.g Šidák's inequality). All such claims were manually verified. During preparation of the final submission, we used AI tools to identify writing issues, verify and improve proofs, including typos, grammatical errors, potentially misleading statements, internal inconsistencies and loose bounds. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content, claims, and artifacts in this work.

REPRODUCIBILITY STATEMENT

Every experiment, table, and figure in this paper can be reproduced from the code assets provided with this submission. All non-synthetic tasks are built from public datasets, and every synthetic task and training pool is drawn by a fixed data seed found in the code assets.

REFERENCES

Yotam Alexander, Yonatan Slutzky, Yuval Ran-Milo, and Nadav Cohen. Do neural networks need gradient descent to generalize? a theoretical study. In *Advances in Neural Information Processing Systems*, 2025.

Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021a.

Idan Amir, Tomer Koren, and Roi Livni. SGD generalizes better than GD (and regularization doesn't help). In *Proceedings of the Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 63–92. PMLR, 2021b. URL https://proceedings.mlr.press/v134/amir21a.html.

Guillaume Aubrun, Justin Jenkinson, and Stanislaw J Szarek. Optimal constants in concentration inequalities on the sphere and in the gauss space. In *International Conference on High Dimensional Probability*, pp. 1–36. Springer, 2023.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Avrim Blum. Random projection, margins, kernels, and feature-selection. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pp. 52–68. Springer, 2005.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000. doi: 10.1002/(SICI)1098-2418(200005)16:3⟨277::AID-RSA4⟩3.0.CO;2-1. URL https://econ-papers.upf.edu/papers/376.pdf.

Joan Bruna, Loucas Pillaud-Vivien, and Aaron Zweig. On single index models beyond gaussian data. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4), 2015. doi: 10.1561/2200000050.

Tal Burla and Roi Livni. All erms can fail in stochastic convex optimization lower bounds in linear dimension. *arXiv preprint arXiv:2602.08350*, 2026.

Gon Buzaglo, Itamar Harel, Mor Shpigel Nacson, Alon Brutzkus, Nathan Srebro, and Daniel Soudry. How uniform random weights induce non-uniform bias: Typical interpolating neural networks generalize with narrow teachers. *arXiv preprint arXiv:2402.06323*, 2024.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QC10RmRbZy9.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85:45–70, 2019. doi: 10.1007/s10472-018-9613-y.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *Advances in Neural Information Processing Systems*, 33:7743–7753, 2020.

Lutz Dümbgen, Richard J. Samworth, and Jon A. Wellner. Bounding distributional errors via density ratios. *Bernoulli*, 27(2):818–852, 2021. doi: 10.3150/20-BEJ1256.

Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1803.05598.

Meir Feder, Ruediger Urbanke, and Yaniv Fogel. Information-theoretic framework for understanding modern machine-learning. *arXiv preprint arXiv:2506.07661*, 2025. doi: 10.48550/arXiv.2506.07661. URL https://arxiv.org/abs/2506.07661.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights, 2026. URL https://arxiv.org/abs/2603.12228.

Malay Ghosh. Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice*, 15:35, 2021. doi: 10.1007/s42519-020-00156-x.

R. D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. doi: 10.1214/aoms/1177731721.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Itamar Harel, William Hoza, Gal Vardi, Itay Evron, Nati Srebro, and Daniel Soudry. Provable tempered overfitting of minimal nets and typical nets. *Advances in Neural Information Processing Systems*, 37:53458–53524, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

W Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. 2020.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.00113.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. *CS 231N*, 7(7):3, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pp. 164–170, 1999.

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A Louis. Is sgd a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: Disentangling the effects of overparameterization of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11893–11902, 2020.

Matan Schliserman, Shira Vansover-Hager, and Tomer Koren. Flat minima and generalization: Insights from stochastic convex optimization. *arXiv preprint arXiv:2511.03548*, 2025.

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://www.jmlr.org/papers/v19/18-188.html.

Vladimir Spokoiny. Concentration of a high dimensional sub-Gaussian vector, 2023. Version 5, revised June 2024.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–147, 1974.

Ryan Theisen, Jason Klusowski, and Michael W. Mahoney. Good classifiers are abundant in the interpolating regime. In *International Conference on Artificial Intelligence and Statistics*, pp. 3376–3384. PMLR, 2021.

Shira Vansover-Hager, Tomer Koren, and Roi Livni. Rapid overfitting of multi-pass SGD in stochastic convex optimization. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/vansover-hager25a.html.

Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4030–4034. IEEE, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Appendix

A.1 Architectures

Both architectures are used at a single configuration throughout, with only the input dimension changing across datasets. All hidden units are ReLU.

MLP (width 64). Two hidden layers of 64 units: $d \to 64 \to 64 \to 2$, where d is the flattened input dimension.

LeNet (width factor 0.5). The LeNet-5 topology (LeCun et al., 1998) with every channel and feature count halved: 5×5 convolution to 3 channels, 2×2 max-pool, 5×5 convolution to 8 channels, 2×2 max-pool, then fully connected layers $\to 60 \to 42 \to 2$. All layers carry biases.

A.2 G&C implementation details

Supporting detail for the G&C procedure of Section 2.

Search budget. For most tasks, G&C is given a fixed budget of 5×10^6 draws per cell; if no draw satisfies $L_{\text{train}}(\theta) = 0$ within that budget, the run is recorded as a failure and discarded from computation. Importantly, one should be aware that this could shift the results slightly near the edge of feasibility - for harder tasks with larger sample sizes (e.g. CIFAR at $m > 20$), some data seeds become unsolvable within this budget.

Matched-seed comparison. To handle the above potential caveat, all of our comparisons are computed and averaged across the same data seeds in both SGD and G&C, and only on those where G&C succeeded. This ensures we do not implicitly leave out "hard" seeds that G&C fails to solve, while SGD takes them into account. This is done in every comparison between the algorithms across this work.

A.3 SGD-LOO Details

A more detailed explanation of the SGD-LOO algorithm.

For a pool of m training points and a candidate learning rate: run m leave-one-out folds (train on the $m-1$ remaining points), with R independent restarts per fold; pool all $m \times R$ outcomes into one score for that learning rate, with held-out accuracy as the target. Each fold trains on $m-1$ points of the pool and is scored on the one it left out, so no data outside S is ever consulted and the selection remains parameter-free rather than tuned. We perform this across a standard grid of learning rates, utilizing the standard Kaiming (He et al., 2015) initialization, and select the best-performing parameter. We then retrain fresh models on the full m-point pool. Only η is chosen this way; b and P_0 are held fixed throughout. Ultimately, we consider this as the core comparator with G&C throughout our experiments, as a method that absolves the bias found in arbitrary SGD.

A.4 MARGIN ESTIMATOR DETAILS

This section adds further technical detail on the margin estimator used throughout this work—specifically, the constants used in the walk defined in Section 2.2.

Stabilization and constants. We apply two stabilizations to Equation (7), both required once the walk encounters points sitting on a flat region of f_θ where dividing by $\|\nabla_x f_\theta\|^2 \approx 0$ produces an arbitrarily large and meaningless step. First, a trust region caps the per-step norm, so no single step can leave the neighborhood of x_0. Second, when $\|\nabla_x f_\theta\|$ falls below a threshold we replace the step by a small random perturbation, nudging the point off the flat spot rather than dividing by near-zero. The overshoot factor $\alpha > 1$ ensures the walk crosses the boundary rather than approaching it asymptotically. On MNIST and CIFAR we use $\alpha = 1.02$, tolerance $\tau = 10^{-4}$, at most $T_{\max} = 30$ iterations, and a per-point trust region of $\frac{1}{2}\|x_0\|_2$; the toy experiments use the same estimator with a fixed trust region appropriate to their input scale.

Convergence handling. The walk terminates when the boundary is crossed to within a set number of iterations; if it does not cross within $T_{\max} = 30$ iterations, the distance travelled so far is *not* a valid estimate of $\mathrm{dist}(x, \mathcal{B}_\theta)$. This is also supported in the findings of Moosavi-Dezfooli et al. (2016), who report the walk converging in far fewer iterations than our $T_{\max}$. We therefore exclude non-converged points from every reported aggregate.

A.5 ADDITIONAL MAIN SWEEP RESULTS

This section of the appendix provides additional sample size m sweeps for a variety of binary tasks, derived from MNIST, CIFAR and TinyImageNet, using both MLP and LeNet architectures, given in Figure 5.

A.6 ADDITIONAL ACROSS-TASK SCATTER PLOTS

This section of the appendix provides additional across-task scatter plots, in addition to Figure 2b and Figure 2a, which show CIFAR-10. The MNIST counterparts are given here, on both architectures (Figures 6 and 7).

A.7 ADDITIONAL ACROSS-LR SCATTER PLOTS

Supporting detail for Figure 2c. We sweep 19 learning rates spaced logarithmically from 3×10^{-4} to 7.4×10^{-2}, training one solution up to 200 data seeds at every rate, so each plotted point is a mean over the independent runs and the grid is matched across rates. We manually remove rates where SGD behaves erratically during its training, usually caused by large gradient steps at high LRs, leading to overflows.

We also provide additional learning-rate scatter plots on different tasks and architecture in Figure 8. G&C is not drawn in these panels. In the order shown it reaches $1/\gamma$ of 3.39, 6.29 and 3.22, at accuracies of 0.792, 0.624 and 0.722, in every case at a larger $1/\gamma$ than any rate on the grid.

A.8 DECISION-BOUNDARY VISUALISATIONS

To visualize the decision boundary, we project the full 784-dimensional MNIST into its top two principal components with standard PCA. PCA is performed on the entire 0-vs-7 training pool. This doesn't guarantee separability, but we find empirically that 0-vs-7 is a good candidate; it's a relatively easy task, and simple visualization of the full training set shows it's almost separable even in 2 dimensions. This enables us to find some visual information with this figure. It should be noted that this can cause "artifacts", such as seemingly incorrectly classified training points, as can be seen in Figure 3, but that is only due to the projection.

A.9 ADVERSARIAL REWEIGHTING

This section of the appendix provides additional data on the boosting experiment. For each task, we measure the test accuracy of both algorithms, SGD's train margins and G&C $N(S)$, along with the relevant across-round scatter plots (Figure 9).

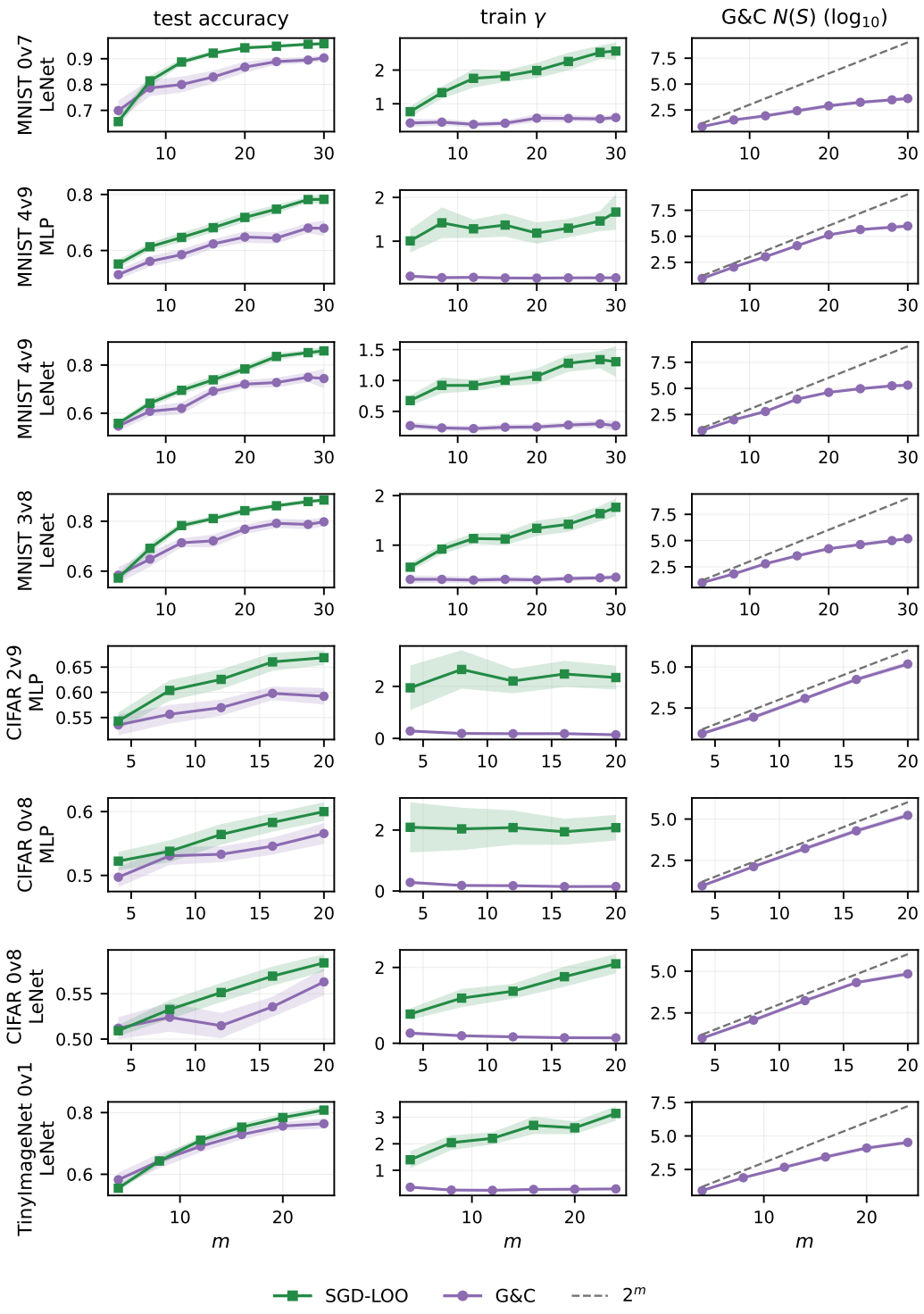


Figure 5: Additional sweep results: G&C and SGD-LOO across training-set sizes m on eight further task–architecture pairs, reporting test accuracy (left), train margin γ (middle) and $N(S)$ against the 2^m labelings of the sample, dashed (right). Bands are 95% intervals.

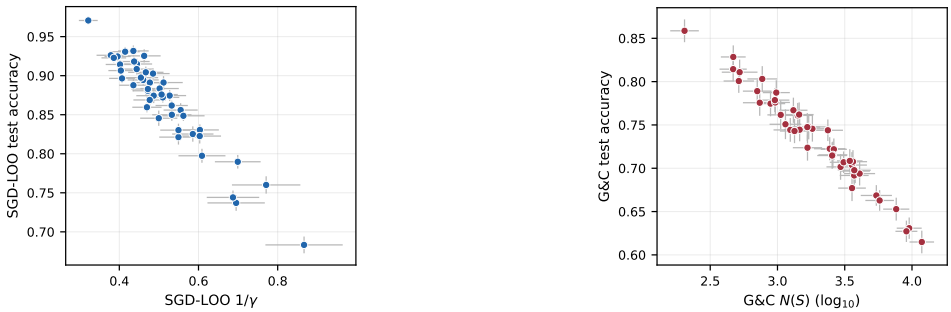


Figure 6: All 45 binary MNIST digit pairs at m=16, MLP width 64, 100 seeds per task: SGD-LOO against its $1/\gamma$ ($\rho = -0.92$, left) and G&C against its $N(S)$ ($\rho = -0.98$, right).

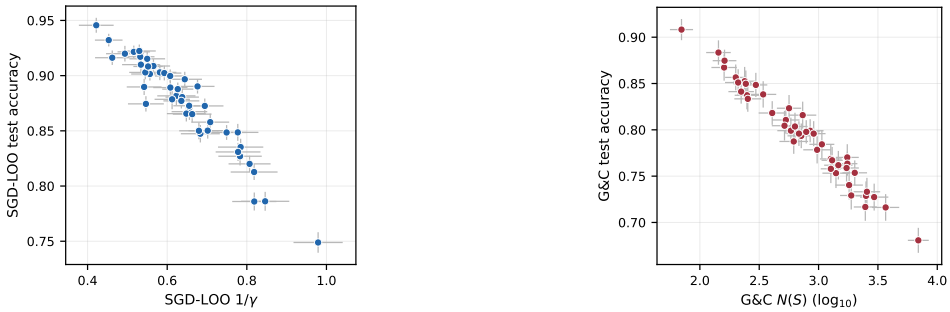


Figure 7: The same 45 MNIST pairs at m=16 on LeNet width 0.5: SGD-LOO against its $1/\gamma$ ($\rho = -0.95$, left) and G&C against its $N(S)$ ($\rho = -0.98$, right).

A.10 Pretraining

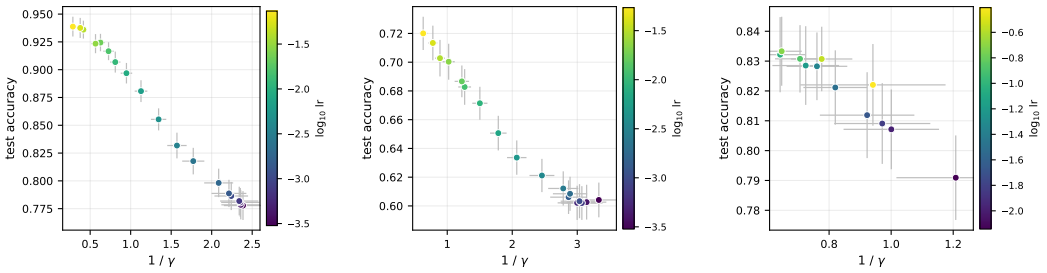
Supporting details and data for the pretraining experiment.

Fixed polarity. The downstream binary labeling is aligned to θ_0 once, at setup, and then frozen. Since pretraining in our case leads to unknown behavior, at θ_0, on the downstream binary task, we select the polarity of the task according to the zero-one error of θ_0 on the full training set. This enables us to review the impact on G&C on a range of "informative" and "uninformative" pretrained points.

Task pool. Both pools hold the downstream pair fixed (MNIST 0v7; CIFAR 2v9, bird vs. truck) and vary only the pretraining class split among the remaining classes.

σ calibration. Gan & Isola (2026) do not give a derivation, sweep, or validation procedure for their perturbation scale. To avoid picking an arbitrary number, we define the following simple procedure that is based only on the training set of the downstream task at the pretrained point. We sweep σ, and measure the solution density $\delta(\sigma) = P(\theta_0 + \sigma\varepsilon$ memorizes the m-point sample$)$. We perform this once per downstream task, not for every pretrained point. We then fix σ for the entire rest of the experiment. The main motivation for this is from Gan & Isola (2026) and their concept of thickets; we want to perturb to around where the solutions are the most dense. For MNIST, we used σ=0.05; For CIFAR σ=0.022; For TinyImageNet σ=0.008.

SGD training rule for this experiment. Unlike Section 4.1, SGD here trains for a fixed T=50 epochs rather than stopping at interpolation, keeping leave-one-out learning-rate selection. This is purely because pretraining lets SGD interpolate almost immediately, and the stop-at-interpolation rule does not transfer well to that regime; the margin-inducing properties of SGD are usually reported on the limit $T \to \infty$. From the pretrained point, SGD converged within 1-2 epochs, significantly lower than the rest of this work, which lead to unstable and noisy results.



(a) MNIST 0v7, MLP. (b) MNIST 4v9, MLP. (c) MNIST 3v8, LeNet.

Figure 8: The learning-rate comparison of Figure 2c on the cells not shown in the body. LeNet is swept over its own range of rates, since it does not reach zero training error at the rates used for the MLP.

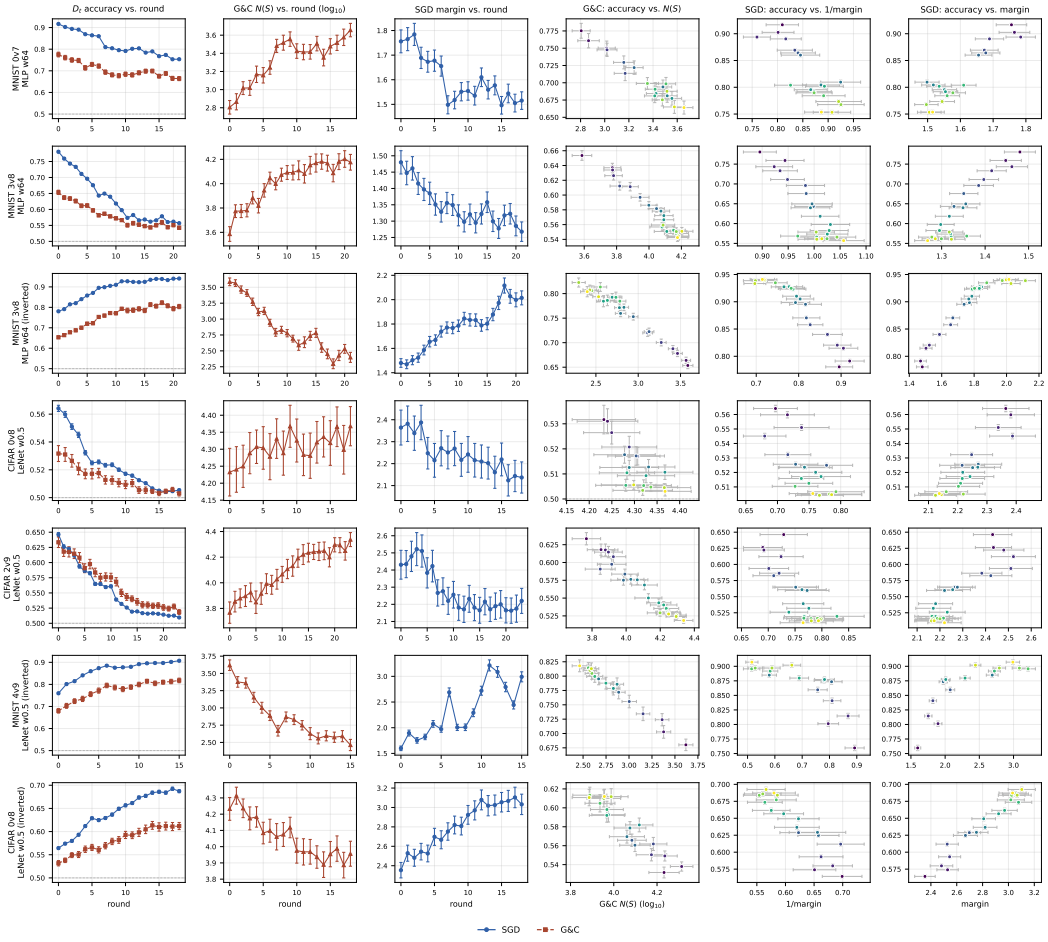


Figure 9: Full boosting-probe data, one row per task (SGD, 1500 data seeds per round; G&C, 200–400 data seeds per round). Columns: D_t accuracy vs. round, G&C $N(S)$ vs. round ($\log_{10}$), SGD margin vs. round, G&C accuracy against $N(S)$, SGD accuracy against $1/\text{margin}$, and SGD accuracy against margin (raw, not inverted). Scatter columns carry 95% CI error bars on both axes.

B PROOFS FOR THE GAUSSIAN MODEL

Before delving into the proofs, we first discuss and analyze the performance of general algorithms in this model in relation with their correlation with μ. Denote:

$$Q(\theta) := \sqrt{d}\,\langle \mu, \theta \rangle. \tag{13}$$

For a unit parameter θ, the signed prediction on a new example satisfies $\sqrt{d}\, y\langle \theta, x\rangle = Q(\theta) + N$, where $N \sim \mathcal{N}(0,1)$. Writing Φ for the standard normal CDF and $\overline{\Phi} = 1 - \Phi$, we obtain that the classifier $x \mapsto \operatorname{sign}(\langle \theta, x\rangle)$ misclassifies one example with probability $\overline{\Phi}(Q(\theta))$:

$$L_{\mathcal{D}_\gamma}(\theta) = \overline{\Phi}(Q(\theta)). \tag{14}$$

By Cauchy Schwartz we obtain that μ is indeed the optimal classifier, where for each classifier:

$$L_{\mathcal{D}_\gamma}(\theta_T) \geq \overline{\Phi}\left(Q\left(\frac{\mu}{\|\mu\|}\right)\right) = \overline{\Phi}\left(\sqrt{d}\gamma\right),$$

we can bound the performance of any algorithm with the following inequality which we will use throughout the proofs (Gordon, 1941):

$$\frac{x}{1+x^2}\phi(x) \leq \overline{\Phi}(x) \leq \frac{\phi(x)}{x}, \qquad x > 0. \tag{15}$$

Indeed, as $\sqrt{d}\gamma > 1$, the lower bound in equation (15) provides

$$L_{D_\gamma}(\theta) > \frac{1}{2\sqrt{2\pi}} \exp(-\gamma^2 d). \tag{16}$$

and the performance of each classifier depends on its margin which is measured by its correlation with μ. Notice, however, that because $m < d$ there are several interpolating solutions with significantly small correlation with μ, hence interpolation does not guarantee generalization.

B.1 PROOF OF THEOREM 2

We begin by providing a formal restatement of theorem 2 with the exact bounds and constants:

Theorem 4. Fix a uniform prior P over the sphere, and assume

$$\sqrt{\frac{32\log(32m)}{d}} < \gamma \leq 1, \qquad \frac{1024}{\gamma^2} < m < d$$

Let $\mathcal{D}_\gamma$ be the distribution in (9), for an arbitrary $\mu \in \mathbb{R}^d$ satisfying $\|\mu\| = \gamma$. Then with probability at least $2/3$, jointly over $S \sim \mathcal{D}_\gamma^m$ and the randomness of G&C, the following bounds hold simultaneously:

$$\frac{e^{-40}}{128\sqrt{21}\gamma^2 m} \leq L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\mathrm{GC}}) \leq \frac{63}{\gamma^2 m}, \tag{17}$$

Our proof strategy is very simple. First, suppose we have a bound T_0 on the number of rounds until G&C interpolates the data. On the event $\tau \leq T_0$ G&C is equivalent to an algorithm that performs ERM on the first T_0 hypotheses picked independently of the data, breaking ties by the order in which they were drawn. The performance of such an algorithm can be estimated via the standard Occam's razor argument for a finite hypotheses class. On the other hand, to obtain a lower bound, we only need to argue that the probability to draw a "sufficiently-good" hypothesis is much smaller than $1/T_0$. That means that, within the class of hypotheses observed by G&C, no sufficiently-good hypothesis was likely drawn.

In more detail, Recall that P is uniform on $\mathbb{S}^{d-1}$. Since $m < d$, the signed sample vectors are linearly independent almost surely, so $p(S) > 0$ and $\tau < \infty$ almost surely. Suppose that we have a deterministic bound T_0 such that

$$\Pr(\tau > T_0) < \delta_0.$$

For every fixed hypothesis θ,

$$\Pr_S(\theta \in \mathcal{I}(S)) = \left(1 - L_{\mathcal{D}_\gamma}(\theta)\right)^m \le e^{-mL_{\mathcal{D}_\gamma}(\theta)}.$$

The union bound consequently gives, for $\epsilon > 0$,

$$\Pr\left[\exists t \le T_0 : \theta_t \in \mathcal{I}(S),\ L_{\mathcal{D}_\gamma}(\theta_t) > \epsilon\right] \le T_0 \mathop{\mathbb{E}}_{\theta\sim P}\left[e^{-mL_{\mathcal{D}_\gamma}(\theta)} \,\middle|\, L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right] \tag{18}$$

$$\le T_0 e^{-m\epsilon}. \tag{19}$$

Similarly, for every $r > 0$,

$$\Pr\left[\exists t \le T_0 : L_{\mathcal{D}_\gamma}(\theta_t) < r\right] \le T_0 P\left(L_{\mathcal{D}_\gamma}(\theta) < r\right). \tag{20}$$

And for every $r > 0$, the algorithm has population loss at least r with probability at least

$$1 - \delta_0 - T_0 P\left(L_{\mathcal{D}_\gamma}(\theta) < r\right).$$

The strategy is then to pick T_0 that leads to the tightest bound. In the next section, we provide a simplified case that already provides meaningful guarantees in some of the regimes and can also be seen as a general overview of our technique.

Brief overview with suboptimal rates

As a warm-up, we use a simple stopping-time estimate to obtain suboptimal rates. These bounds are already non-trivial, but the upper bound has an extra logarithmic factor, while the lower bound is missing the factor $1/\gamma^2$. Specifically, we show that

$$\Pr\left[\frac{e^{-12}}{80m} \le L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\mathrm{GC}}) \le \frac{5\log(128m)}{\gamma^2 m}\right] > \frac{3}{4}. \tag{21}$$

Following subsections can be, mostly, read independently of this overview and there we choose the stopping time more carefully and refine the analysis to obtain the sharper bounds. However, the purpose of this overview is to explain how the stopping-time estimate controls both bounds, and where the simple argument falls short before introducing the additional tools needed for the sharper rates.

For the analysis in this section we bound T_0 by considering the probability to draw a solution with sufficiently small population error. Specifically, let

$$p_{\mathrm{naive}} := P\left(L_{\mathcal{D}_\gamma}(\theta) \le \frac{1}{16m}\right),$$

and we take

$$T_{\mathrm{naive}} := \left\lceil \frac{3}{p_{\mathrm{naive}}} \right\rceil.$$

With probability at least $1 - e^{-3}$, these draws contain a hypothesis with population loss at most $1/(16m)$. The first such hypothesis is chosen independently of the sample, and the union bound shows that it fails to interpolate with probability at most $1/16$. Whenever it interpolates, G&C stops no later than that draw. Therefore,

$$\Pr(\tau > T_{\mathrm{naive}}) \le e^{-3} + \frac{1}{16}. \tag{22}$$

The naive upper bound. The upper bound follows once we estimate p_{naive}. Choose a_{naive} so that $\overline{\Phi}(a_{\mathrm{naive}}) = 1/(16m)$. The standard Gaussian tail bound $\overline{\Phi}(x) \le e^{-x^2/2}$ for $x \ge 0$ already gives

$$a_{\mathrm{naive}}^2 \le 2\log(16m).$$

By monotonicity, we have

$$p_{\mathrm{naive}} = P(Q(\theta) \ge a_{\mathrm{naive}}),$$

One can bound the probability of drawing a hypothesis with a given correlation (using, for example, Bruna et al. (2023, Lemma 2.1) and plugging in $\delta = 1$) and obtain a bound of the form

$$\begin{aligned} p_{\text{naive}} &\geq \frac{1}{4} e^{-(a_{\text{naive}}/\gamma+1)^2} \\ &\geq \frac{1}{4} e^{-\left(\sqrt{2\log(16m)}/\gamma+1\right)^2} \\ &\geq \frac{e^{-2}}{4} (16m)^{-4/\gamma^2} \qquad\qquad (x+1)^2 \leq 2x^2+2, \end{aligned} \tag{23}$$

The range condition $\max\{a_{\text{naive}}/\gamma, 1\} \leq \sqrt{d}/4$ in the cited lemma follows from our assumptions. By (19), except on an additional event of probability $1/32$, every interpolating hypothesis among these draws has population loss at most $\log(32T_{\text{naive}})/m$. Since $T_{\text{naive}} \leq 4/p_{\text{naive}}$, (23) gives

$$\log(32T_{\text{naive}}) \leq \log 512 + 2 + \frac{4\log(16m)}{\gamma^2} \leq \frac{5\log(128m)}{\gamma^2},$$

where the last inequality uses $\gamma \leq 1$. We conclude that

$$\Pr\left[L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\text{GC}}) \leq \frac{5\log(128m)}{\gamma^2 m}\right] \geq 1 - e^{-3} - \frac{1}{16} - \frac{1}{32}. \tag{24}$$

The naive lower bound. For the lower bound, we use the same sequence of T_{naive} hypotheses. We want to choose c_{naive} so that, having chosen our stopping time, the probability to obtain a classifier with Q-score at least c_{naive} becomes negligible. Define the set

$$A_\alpha = \left\{\theta \in \mathbb{S}^{d-1} : Q(\theta) \geq \alpha\right\}$$

It forms a spherical cap with angular radius

$$\arccos\left(\frac{\alpha}{\gamma\sqrt{d}}\right), \qquad 0 \leq \alpha < \gamma\sqrt{d},$$

we exploit the following monotonicity statement (see Aubrun et al. (2023, Proposition 6)), and one can show that, for $0 \leq a < c < \gamma\sqrt{d}$,

$$\begin{aligned} \frac{P(Q(\theta) \geq c)}{P(Q(\theta) \geq a)} &\leq \left(\frac{\sin\left(\arccos(c/(\gamma\sqrt{d}))\right)}{\sin\left(\arccos(a/(\gamma\sqrt{d}))\right)}\right)^{d-1} \\ &= \left(\frac{1 - c^2/(\gamma^2 d)}{1 - a^2/(\gamma^2 d)}\right)^{(d-1)/2} \\ &= \left(1 - \frac{c^2 - a^2}{\gamma^2 d - a^2}\right)^{(d-1)/2} \\ &\leq \exp\left(-\frac{(c^2 - a^2)(d-1)}{2\gamma^2 d}\right). \end{aligned}$$

Set

$$c_{\text{naive}} := \sqrt{a_{\text{naive}}^2 + 24\gamma^2}.$$

Our assumptions ensure that $c_{\text{naive}} < \gamma\sqrt{d}$. Applying the comparison with $a = a_{\text{naive}}$ and $c = c_{\text{naive}}$, we obtain

$$P(Q(\theta) \geq c_{\text{naive}}) \leq e^{-6} p_{\text{naive}}. \tag{25}$$

Thus hypotheses above c_{naive} are much rarer than the hypotheses used to bound the stopping time. Since $T_{\text{naive}} \leq 4/p_{\text{naive}}$, the union bound now gives

$$\Pr\left[\exists t \leq T_{\text{naive}} : Q(\theta_t) \geq c_{\text{naive}}\right] \leq 4e^{-6}. \tag{26}$$

It remains to translate the score c_{naive} back into population loss. Writing ϕ for the standard normal density. For $x, A > 0$, the substitution $z = \sqrt{t^2 + A}$ gives

$$\overline{\Phi}(\sqrt{x^2 + A}) = e^{-A/2} \int_x^\infty \frac{t}{\sqrt{t^2 + A}} \phi(t)\, dt.$$

The fraction in the integral lies between $x/\sqrt{x^2 + A}$ and 1. Hence

$$e^{-A/2} \frac{x}{\sqrt{x^2 + A}} \overline{\Phi}(x) \le \overline{\Phi}(\sqrt{x^2 + A}) \le e^{-A/2} \overline{\Phi}(x). \tag{27}$$

Applying the lower bound with $x = a_{\text{naive}}$ and $A = 24\gamma^2$,

$$\overline{\Phi}(c_{\text{naive}}) \ge e^{-12\gamma^2} \frac{a_{\text{naive}}}{c_{\text{naive}}} \frac{1}{16m}.$$

Since $\overline{\Phi}(a_{\text{naive}}) \le 1/16 < \overline{\Phi}(1)$, we have $a_{\text{naive}} > 1$. Together with $\gamma \le 1$, this gives

$$\frac{a_{\text{naive}}}{c_{\text{naive}}} = \frac{1}{\sqrt{1 + 24\gamma^2/a_{\text{naive}}^2}} \ge \frac{1}{5},$$

and hence

$$\overline{\Phi}(c_{\text{naive}}) \ge \frac{e^{-12}}{80m}. \tag{28}$$

If G&C stops within T_{naive} draws and none of those draws has score at least c_{naive}, its population loss is at least $\overline{\Phi}(c_{\text{naive}})$. Combining this observation with (22), (18), and (26), we obtain equation (21).

UPPER BOUND

To remove the logarithm in the upper bound, we make two corrections: First we bound the stopping time more carefully, and second, we estimate the conditional expectation in (18), rather than simply bounding it by $e^{-m\epsilon}$ as appears in equation (19).

To bound the stopping time, we bound the mass of the interpolating set directly, rather than only counting hypotheses with small population loss. The following lemma gives this bound using Šidák's inequality. We first use it to obtain the upper and lower bounds, and provide its proof in section B.1.1.

Lemma 1. Suppose $0 < \gamma \le 1$, $m \le d$, and $b \ge 2$ satisfies

$$m\overline{\Phi}(b) \le \frac{1}{128\gamma^2}, \qquad \gamma^2 d \ge 32 \log(32m).$$

Then

$$\Pr_S \left[p(S) \ge e^{-32/\gamma^2} \overline{\Phi}(b/\gamma) \right] \ge \frac{13}{16}.$$

With lemma 1 at hand, we choose the stopping time bound as follows: choose b, p_0, and T_0 by

$$\overline{\Phi}(b) = \frac{1}{128\gamma^2 m}, \qquad p_0 := e^{-32/\gamma^2} \overline{\Phi}(b/\gamma), \qquad T_0 := \left\lceil \frac{3}{p_0} \right\rceil. \tag{29}$$

Our assumptions in theorem 2 imply those of lemma 1. Indeed, since $\gamma^2 m > 1$, the lower bound at 2 gives

$$\overline{\Phi}(b) < \frac{1}{128} < \frac{2}{5}\phi(2) \le \overline{\Phi}(2),$$

so by monotonicity, our choice of b guarantees $b \ge 2$, by assumption $\gamma^2 d > 32 \log(32m)$, and by construction $m\overline{\Phi}(b) = 1/(128\gamma^2)$. Therefore:

$$\begin{aligned}
\Pr(\tau > T_0) &\le \Pr\left(p(S) < p_0\right) + \Pr\left(\tau > T_0 | p(S) \ge p_0\right) \\
&\le \frac{3}{16} + \mathbb{E}\left[(1 - p(S))^{T_0} \middle| p(S) \ge p_0\right] && \text{lemma 1} \\
&\le \frac{3}{16} + e^{-3} && T_0 \ge 3/p_0
\end{aligned} \tag{30}$$

The distribution of the score. To bound the conditional expectation, we use the following consequence of Dümbgen et al. (2021, Corollary 5).

Lemma 2 (Score density)**.** Suppose $d \geq 4$, let $\|\mu\| = \gamma > 0$, and draw $\theta \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. The score $Q(\theta) = \sqrt{d}\,\langle \mu, \theta \rangle$ has a density $f_{d,\gamma}$ satisfying

$$f_{d,\gamma}(q) \leq \sqrt{\frac{d}{d-3}}\,\frac{1}{\gamma}\phi(q/\gamma) \leq \frac{2}{\gamma}\phi(q/\gamma), \qquad q \in \mathbb{R}. \tag{31}$$

Proof. By rotational symmetry, $Q(\theta)/\gamma$ has the same distribution as $\sqrt{d}\,\theta_1$. Apply Dümbgen et al. (2021, Corollary 5) with $n = d$, $k = 1$, and $r_n = \sqrt{d}$, using their equation (1) to express the result as a bound on the density ratio. Rescaling by γ gives the first inequality. The second follows from $d/(d-3) \leq 4$ for $d \geq 4$. □

Bounding the conditional expectation. Set

$$\epsilon := \frac{63}{\gamma^2 m}. \tag{32}$$

Since $\frac{63}{\gamma^2 m} < \frac{63}{1024} < \frac{1}{16}$, we have that $\epsilon < \overline{\Phi}(1)$, and we can choose $\ell > 1$ so that

$$\overline{\Phi}(\ell) = \epsilon = \frac{63}{\gamma^2 m}. \tag{33}$$

By (18), it is enough to show

$$T_0 \mathop{\mathbb{E}}_{\theta \sim P}\left[e^{-mL_{\mathcal{D}_\gamma}(\theta)} \,\middle|\, L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right] \leq e^{-5}. \tag{34}$$

Then $1 < \ell < b$, and $L_{\mathcal{D}_\gamma}(\theta) > \epsilon$ is equivalent to $Q(\theta) < \ell$. By symmetry,

$$P\left(L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right) \geq P(Q(\theta) < 0) = \frac{1}{2}.$$

Using (31), we obtain

$$\begin{aligned} &\mathop{\mathbb{E}}_{\theta \sim P}\left[e^{-mL_{\mathcal{D}_\gamma}(\theta)} \,\middle|\, L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right] \\ &\qquad = \frac{\displaystyle\int_{-\infty}^{\ell} f_{d,\gamma}(q) e^{-m\overline{\Phi}(q)}\,dq}{P\left(L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right)} \\ &\qquad \leq 4\int_{-\infty}^{\ell} \frac{1}{\gamma}\phi(q/\gamma) e^{-m\overline{\Phi}(q)}\,dq. \end{aligned} \tag{35}$$

Define

$$g(q) := \frac{1}{\gamma}\phi(q/\gamma)e^{-m\overline{\Phi}(q)}, \qquad I := \int_{-\infty}^{\ell} g(q)\,dq.$$

For $0 < q \leq \ell$, the upper bound in equation (15) gives:

$$\phi(q) \geq q\overline{\Phi}(q) \geq q\epsilon.$$

Thus $q/\phi(q) \leq 1/\epsilon$ for every $q \leq \ell$, with the inequality immediate when $q \leq 0$.

Since

$$\frac{d}{dq}e^{-m\overline{\Phi}(q)} = m\phi(q)e^{-m\overline{\Phi}(q)},$$

integration by parts gives

$$\begin{aligned} I &\leq \frac{g(\ell)}{m\phi(\ell)} + \frac{1}{m}\int_{-\infty}^{\ell}\left(\frac{1}{\gamma^2} - 1\right)\frac{q}{\phi(q)}g(q)\,dq \\ &\leq \frac{g(\ell)}{m\phi(\ell)} + \frac{I}{m\gamma^2\epsilon} \\ &= \frac{g(\ell)}{m\phi(\ell)} + \frac{I}{63} && m\gamma^2\epsilon = 63. \end{aligned}$$

Rearranging,

$$I \le \frac{63}{62m}\frac{g(\ell)}{\phi(\ell)}.$$

Finally, applying equation (15) at ℓ and ℓ/γ gives

$$\phi(\ell) \ge \ell\epsilon, \qquad \phi(\ell/\gamma) \le \frac{2\ell}{\gamma}\overline{\Phi}(\ell/\gamma),$$

where we used $\ell/\gamma \ge 1$. Consequently,

$$\begin{aligned}\int_{-\infty}^{\ell} g(q)\,dq &\le \frac{63}{62m\gamma}\frac{\phi(\ell/\gamma)}{\phi(\ell)}e^{-m\epsilon} \\ &\le \frac{126}{62m\gamma^2\epsilon}\overline{\Phi}(\ell/\gamma)e^{-m\epsilon} \\ &= \frac{1}{31}\overline{\Phi}(\ell/\gamma)e^{-m\epsilon}.\end{aligned}$$

Substituting into (35),

$$\mathbb{E}_{\theta\sim P}\left[e^{-mL_{\mathcal{D}_\gamma}(\theta)}\,\middle|\,L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right] \le \frac{4}{31}\overline{\Phi}(\ell/\gamma)e^{-m\epsilon}. \tag{36}$$

Applying the upper comparison in (27) with $x = \ell$ and $A = b^2 - \ell^2$ shows that

$$b^2 - \ell^2 \le 2\log\frac{\overline{\Phi}(\ell)}{\overline{\Phi}(b)} = 2\log 8064 < 18,$$

and

$$\frac{b}{\ell} = \sqrt{1 + \frac{b^2-\ell^2}{\ell^2}} < 5.$$

The lower comparison, now with $x = \ell/\gamma$ and $A = (b^2-\ell^2)/\gamma^2$, gives

$$\frac{\overline{\Phi}(\ell/\gamma)}{\overline{\Phi}(b/\gamma)} \le \frac{b}{\ell}e^{(b^2-\ell^2)/(2\gamma^2)} \le 5e^{9/\gamma^2}.$$

Combining equations (36) and (30):

$$\begin{aligned}&T_0\,\mathbb{E}_{\theta\sim P}\left[e^{-mL_{\mathcal{D}_\gamma}(\theta)}\,\middle|\,L_{\mathcal{D}_\gamma}(\theta) > \epsilon\right] \\ &\quad\le \frac{16}{31}e^{32/\gamma^2}\frac{\overline{\Phi}(\ell/\gamma)}{\overline{\Phi}(b/\gamma)}e^{-m\epsilon} \\ &\quad\le \frac{80}{31}e^{-22/\gamma^2} \\ &\quad< e^{-5}.\end{aligned} \tag{37}$$

where we used $m\epsilon = 63/\gamma^2$ and $\gamma \le 1$. This proves (34). Combining (34) with (18) and (30), we conclude that

$$\Pr_{S,\mathrm{GC}}\left[L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\mathrm{GC}}) \le \frac{63}{\gamma^2 m}\right] \ge 1 - \frac{3}{16} - e^{-3} - e^{-5}. \tag{38}$$

LOWER BOUND

We use the same b, p_0, and T_0 from (29). To obtain the lower bound, we show that a hypothesis with substantially smaller population loss is unlikely to appear within these T_0 draws. Set

$$c := \sqrt{b^2 + 80}.$$

Integrating the density bound in (31) gives

$$P(Q(\theta) \ge c) \le 2\overline{\Phi}(c/\gamma).$$

Apply the upper bound in equation (27) with $x = b/\gamma$ and $A = 80/\gamma^2$:

$$\begin{aligned}
\Pr\left[\exists t \le T_0 : Q(\theta_t) \ge c\right] &\le T_0 P(Q(\theta) \ge c) \\
&\le 2T_0 \overline{\Phi}(c/\gamma) \\
&\le 2T_0 e^{-40/\gamma^2} \overline{\Phi}(b/\gamma) && \text{equation (27)} \\
&\le 8e^{-8/\gamma^2} && T_0 \le \frac{4}{p_0} = \frac{4e^{32/\gamma^2}}{\overline{\Phi}(b/\gamma)} \\
&\le 8e^{-8} && \gamma \le 1, \qquad (39)
\end{aligned}$$

If G&C stops within T_0 draws and none of those hypotheses has score at least c, its population loss is at least $\overline{\Phi}(c)$. The lower comparison in (27), applied with $x = b$ and $A = 80$, gives

$$\begin{aligned}
L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\mathrm{GC}}) &\ge \overline{\Phi}(c) \\
&\ge e^{-40} \frac{b}{c} \overline{\Phi}(b) \\
&\ge \frac{e^{-40}}{128\sqrt{21}} \frac{1}{\gamma^2 m}, \qquad (40)
\end{aligned}$$

where the last inequality uses

$$\overline{\Phi}(b) = \frac{1}{128\gamma^2 m}, \qquad \frac{b}{c} = \frac{1}{\sqrt{1 + 80/b^2}} \ge \frac{1}{\sqrt{21}},$$

since $b \ge 2$.

Consequently, this lower bound holds with probability at least

$$1 - \frac{3}{16} - e^{-3} - 8e^{-8}.$$

Unlike the naive bound, it retains the factor $1/\gamma^2$.

Both bounds use the same T_0 and the same stopping event. Combining (30), (34), and (39), together with the union bound in (18), we obtain

$$\begin{aligned}
&\Pr_{S,\mathrm{GC}}\left[\frac{e^{-40}}{128\sqrt{21}} \frac{1}{\gamma^2 m} \le L_{\mathcal{D}_\gamma}(\widehat{\theta}_{\mathrm{GC}}) \le \frac{63}{\gamma^2 m}\right] \\
&\qquad \ge 1 - \frac{3}{16} - e^{-3} - e^{-5} - 8e^{-8} > \frac{2}{3}. \qquad (41)
\end{aligned}$$

The concludes the proof of theorem 4

B.1.1 Proof of lemma 1

Let us treat the parameter θ, whose distributed uniformly on the sphere, as $\theta = H/\|H\|$ where $H \sim \mathcal{N}(0, I_d)$. Notice that normalization does not affect interpolation. Therefore, $y_i\langle x_i, \theta\rangle > 0$ if and only if $y_i\langle x_i, H\rangle > 0$. As such, we can bound the probability of H being an interpolating solution instead of θ.

By symmetry of the noise, write the signed examples as

$$z_i := y_i x_i = \mu + \frac{1}{\sqrt{d}} n_i, \qquad n_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d).$$

Rotate coordinates so that $\mu = \gamma e_1$, and write

$$n_i = (a_i, v_i), \qquad a_i \sim \mathcal{N}(0, 1), \qquad v_i \sim \mathcal{N}(0, I_{d-1}),$$

where a_i and v_i are independent.

Conditional on $\langle \mu, H\rangle = q$, we have $H_1 = q/\gamma$ and $H_\perp \sim \mathcal{N}(0, I_{d-1})$, where $H = (H_1, H_\perp)$. Hence

$$\langle z_i, H\rangle = q\left(1 + \frac{a_i}{\gamma\sqrt{d}}\right) + \frac{\langle v_i, H_\perp\rangle}{\sqrt{d}}. \qquad (42)$$

In turn, for a fixed sample, let p_i be the probability of misclassifying the i-th example conditional on $\langle \mu, H \rangle = b$, then:

$$p_i := \overline{\Phi}\left(\frac{b\left(1 + a_i/(\gamma\sqrt{d})\right)}{\|v_i\|/\sqrt{d}} \right). \tag{43}$$

We will show that, with probability at least $13/16$ over S,

$$\sum_{i=1}^{m} p_i \le \frac{4}{\gamma^2}, \qquad p_i \le \frac{3}{8} \text{ for every } i, \qquad 1 + \frac{a_i}{\gamma\sqrt{d}} > 0 \text{ for every } i. \tag{44}$$

First, we explain why these conditions suffice.

Fix a sample satisfying (44). By Šidák's inequality (Šidák, 1967, Corollary 1), we have for centered jointly Gaussian variables $U_1, \dots, U_m$ and positive thresholds $t_1, \dots, t_m$,

$$\Pr\left(|U_i| < t_i \text{ for every } i\right) \ge \prod_i \Pr(|U_i| < t_i). \tag{45}$$

Apply this inequality to

$$U_i := \frac{\langle v_i, H_\perp \rangle}{\sqrt{d}}, \qquad t_i := b\left(1 + \frac{a_i}{\gamma\sqrt{d}}\right).$$

If $|U_i| < t_i$ for every i, then $\langle z_i, H \rangle = t_i + U_i > 0$ for every i, so the hypothesis interpolates. One can show that, $\Pr(|U_i| < t_i \mid S) = 1 - 2p_i$. Consequently,

$$\Pr_H\left(\theta \in \mathcal{I}(S) \mid \langle \mu, H \rangle = b,\ S\right) \ge \prod_i (1 - 2p_i).$$

Using $\log(1-x) \ge -x/(1-x)$ for $0 \le x < 1$ and $p_i \le 3/8$,

$$\prod_i (1 - 2p_i) \ge \exp\left(-8\sum_i p_i\right) \ge e^{-32/\gamma^2}.$$

Finally, by (42), notice that conditioning on $q > b$ only increases the probability of interpolation and the same lower bound holds for every $q \ge b$. Integrating over $\langle \mu, H \rangle \ge b$ gives

$$p(S) \ge e^{-32/\gamma^2}\overline{\Phi}(b/\gamma).$$

Verifying the conditions on the sample. We first bound the individual probabilities p_i. The Gaussian and chi-square tail bounds give, for each i,

$$\Pr\left(a_i < -\frac{\gamma\sqrt{d}}{2}\right) \le e^{-\gamma^2 d/8}, \qquad \Pr\left(\|v_i\| > 2\sqrt{d}\right) \le e^{-d/2}.$$

For the second bound, see Spokoiny (2023, Theorem B.1). Since $\gamma \le 1$, both probabilities are at most $e^{-\gamma^2 d/8}$. A union bound therefore gives

$$\Pr_S\left[\exists i : a_i < -\frac{\gamma\sqrt{d}}{2} \text{ or } \|v_i\| > 2\sqrt{d}\right] \le 2me^{-\gamma^2 d/8} \le \frac{1}{16}, \tag{46}$$

where the last inequality uses $\gamma^2 d \ge 8\log(32m)$. Outside this event we have $1 + a_i/(\gamma\sqrt{d}) \ge 1/2$, and by equation (43) and $b > 2$:

$$p_i \le \overline{\Phi}(b/4) \le \overline{\Phi}(1/2) < \frac{3}{8}.$$

To control the sum, we bound it directly using a second-order expansion. Averaging over the training example, conditional on $H_\perp$, gives

$$\mathbb{E}_S[p_i] = \mathbb{E}\left[\overline{\Phi}\left(\frac{b}{\sqrt{V}}\right)\right], \qquad V := \frac{Y + b^2/\gamma^2}{d}, \qquad Y \sim \chi^2_{d-1}.$$

For $v > 0$, define

$$F(v) := \overline{\Phi}\left(\frac{b}{\sqrt{v}}\right).$$

Differentiating,

$$F'(v) = \frac{b}{2v^{3/2}}\phi\left(\frac{b}{\sqrt{v}}\right), \qquad F''(v) = \frac{b(b^2-3v)}{4v^{7/2}}\phi\left(\frac{b}{\sqrt{v}}\right).$$

Thus

$$F''(v) \le \frac{b^3}{4\sqrt{2\pi}}\, v^{-7/2} e^{-b^2/(2v)}.$$

The expression $v^{-7/2}e^{-b^2/(2v)}$ is maximized at $v = b^2/7$. Consequently,

$$F''(v) \le \frac{7^{7/2}e^{-7/2}}{4\sqrt{2\pi}}\frac{1}{b^4} < \frac{3}{b^4}, \qquad v > 0.$$

Taylor's theorem around $v = 1$ therefore gives:

$$F(v) \le \overline{\Phi}(b) + \frac{b\phi(b)}{2}(v-1) + \frac{3}{2b^4}(v-1)^2. \tag{47}$$

Since Y has mean $d-1$ and variance $2(d-1)$,

$$\mathbb{E}[V-1] = \frac{b^2/\gamma^2 - 1}{d} \le \frac{b^2}{\gamma^2 d},$$

and

$$\mathbb{E}\left[(V-1)^2\right] = \frac{2(d-1) + (b^2/\gamma^2-1)^2}{d^2} \le \frac{2}{d} + \frac{b^4}{\gamma^4 d^2},$$

where we used $b \ge 2$ and $\gamma \le 1$. Taking expectations in (47) and summing over the examples,

$$\begin{aligned}\mathbb{E}_S\left[\sum_{i=1}^m p_i\right] &\le m\overline{\Phi}(b) + \frac{mb^3\phi(b)}{2\gamma^2 d} + \frac{3m}{b^4 d} + \frac{3m}{2\gamma^4 d^2} \\ &\le m\overline{\Phi}(b) + \frac{b^3\phi(b)}{2\gamma^2} + \frac{3}{b^4} + \frac{3}{2\gamma^4 d},\end{aligned}$$

where the second inequality uses $m \le d$.

For $b \ge 2$, the function $b^3\phi(b)$ is decreasing, so

$$b^3\phi(b) \le 8\phi(2) < \frac{1}{2}.$$

Our assumptions also give $\gamma^2 d \ge 32$ and $m\overline{\Phi}(b) \le 1/(128\gamma^2)$. Using $\gamma \le 1$, we conclude that

$$\begin{aligned}\mathbb{E}_S\left[\sum_{i=1}^m p_i\right] &\le \frac{1}{128\gamma^2} + \frac{1}{4\gamma^2} + \frac{3}{16\gamma^2} + \frac{3}{64\gamma^2} \\ &= \frac{63}{128\gamma^2} < \frac{1}{2\gamma^2}.\end{aligned} \tag{48}$$

Markov's inequality therefore gives

$$\Pr_S\left(\sum_{i=1}^m p_i > \frac{4}{\gamma^2}\right) \le \frac{1}{8}. \tag{49}$$

Combining this with (46), with probability at least $1 - 1/16 - 1/8 = 13/16$, we have

$$\sum_{i=1}^m p_i \le \frac{4}{\gamma^2}, \qquad p_i \le \frac{3}{8} \text{ for every } i, \qquad 1 + \frac{a_i}{\gamma\sqrt{d}} > 0 \text{ for every } i.$$

B.2 PROOF OF THEOREM 3

We now state a formal restatement of theorem 3 with the exact bounds and constants:

Theorem 5. Consider the Gaussian model in equation (9), with $\|\mu\| = \gamma$, and assume

$$d \geq 4, \qquad 0 < \gamma \leq 1, \qquad \frac{1}{\gamma^2} < m < d, \qquad \gamma^2 d \geq 16 \log(4m).$$

Run gradient descent as defined in equation (10), initialized at $\|\theta_0\| \leq R$, with a fixed step size $0 < \eta \leq 2/5$. Assume

$$\eta T \geq \frac{e^{900\gamma R}}{9\gamma^2},$$

then,

$$\frac{1}{2\sqrt{2\pi}} \exp(-d\gamma^2) \leq \mathbb{E}\left[L_{\mathcal{D}_\gamma}(\theta_T)\right] \leq 4 \exp\left(-\frac{d\gamma^2}{1152}\right). \tag{50}$$

Notice that the lower bound was already established in equation (16). Next, recall that $z_i = y_i x_i$ and let us denote by $\bar{z} = \frac{1}{m}\sum z_i$. We claim that with probability at least

$$1 - me^{-\gamma^2 d/8} - 2e^{-3d/4}$$

on the sample we have that:

$$\min_{i \leq m} \langle \mu, z_i \rangle \geq \frac{\gamma^2}{2}, \qquad \|\bar{z}\| \leq 3\gamma, \qquad \frac{1}{m}\sum_{i=1}^{m} \|z_i\|^2 \leq 10, \tag{51}$$

By taking the second derivative, one can show that F_S is β smooth where

$$\beta = \frac{1}{4m}\sum_{i=1}^{m} \|z_i\|^2,$$

hence by our assumption in equation (51) we have that it is $5/2$-smooth. We defer the proof to the end of this section. Our starting point is the special case of $\theta_0 = 0$.

Lemma 3. Consider the Gaussian model in equation (9), with $\|\mu\| = \gamma$, and assume

$$d \geq 4, \qquad 0 < \gamma \leq 1, \qquad \frac{1}{\gamma^2} < m < d, \qquad \gamma^2 d \geq 16 \log(4m).$$

Assume the inequalities in equation (51) hold. and run gradient descent as defined in equation (10), initialized at $\theta_0 = 0$, with a fixed step size $0 < \eta \leq 2/5$. for $T \geq 1$ iterations Then, for every integer $1 \leq t \leq T$:

$$\frac{\langle \theta_t, \mu \rangle}{\|\theta_t\|} \geq \frac{\gamma}{12}. \tag{52}$$

Notice that equations (14) and (52) can be used to obtain a generalization bound immediately for the case $\theta_0 = 0$. We next prove lemma 3 and then continue to the general case.

Proof. Assume $\theta_0 = 0$. W.l.o.g assume $t = T$ (running the algorithm for $T \geq t$ steps will not change the result). We claim that for every v such that $F_S(v) \leq F_S(\theta_T)$:

$$\|\theta_T - v\|^2 \leq \|v\|^2 \tag{53}$$

Indeed, the standard descent lemma (Bubeck, 2015, eq. 3.5) gives for $\eta \leq 1/\beta$:

$$\eta \|\nabla F_S(\theta_t)\|^2 \leq 2(F_S(\theta_t) - F_S(\theta_{t+1})). \tag{54}$$

Hence

$$\begin{aligned}
\|\theta_{t+1} - v\|^2 &= \|\theta_t - v\|^2 - 2\eta\langle \nabla F_S(\theta_t), \theta_t - v\rangle + \eta^2 \|\nabla F_S(\theta_t)\|^2 \\
&\leq \|\theta_t - v\|^2 - 2\eta\langle \nabla F_S(\theta_t), \theta_t - v\rangle + 2\eta(F_S(\theta_t) - F_S(\theta_{t+1})) \quad \text{equation (54)} \\
&\leq \|\theta_t - v\|^2 - 2\eta(F_S(\theta_t) - F_S(v)) + 2\eta(F_S(\theta_t) - F_S(\theta_{t+1})) \qquad \text{convexity} \\
&= \|\theta_t - v\|^2 - 2\eta(F_S(\theta_{t+1}) - F_S(v))
\end{aligned}$$

summing over $t = 0, \ldots, T-1$. we obtain that

$$\begin{aligned}
\|\theta_T - v\|^2 &\leq \|v\|^2 - 2\eta \sum_{t=1}^{T} (F_S(\theta_t) - F_S(v)) \\
&\leq \|v\|^2 - 2\eta T (F_S(\theta_T) - F_S(v)) && \text{equation (54)} \\
&\leq \|v\|^2
\end{aligned}$$

This last inequality can be rewritten as:

$$2\langle \theta_T, v \rangle \geq \|\theta_T\|^2.$$

Now choose v such that:

$$v = \frac{\|\theta_T\| \|\bar{z}\|}{\min_i \langle \mu, z_i \rangle} \mu,$$

Set $\ell(x) = \log(1 + \exp(-x))$ and recall that ℓ is monotone:

$$\begin{aligned}
F_S(v) &= \frac{1}{m} \sum_{i=1}^{m} \ell \left(\|\theta_T\| \|\bar{z}\| \frac{\langle \mu, z_i \rangle}{\min_i \langle \mu, z_i \rangle} \right) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \ell \left(\|\theta_T\| \|\bar{z}\| \right) && \frac{\langle \mu, z_i \rangle}{\min_i \langle \mu, z_i \rangle} \geq 1 \\
&\leq \ell(\langle \theta_T, \bar{z} \rangle) && \text{C.S} \\
&\leq F_S(\theta_T) && \text{convexity}
\end{aligned}$$

In turn, we have that:

$$2 \frac{\langle \theta_T, v \rangle}{\|\theta_T\|} \geq \|\theta_T\|, \tag{55}$$

By definition of v and assumption on $\|\bar{z}\|$ and $\langle \mu, z_i \rangle$:

$$2 \frac{\langle \theta_T, \mu \rangle}{\|\theta_T\|} \geq \frac{\min \langle \mu, z_i \rangle}{\|\bar{z}\|} \geq \frac{\gamma}{6}.$$

□

We continue with the proof of the general case. We will prove that in the general case, if equation (51) hold then

$$\frac{\langle \theta_T, \mu \rangle}{\|\theta_T\|} \geq \frac{\gamma}{24} \tag{56}$$

Thus, with probability at least

$$1 - m e^{-\gamma^2 d/8} - 2e^{-3d/4}$$

$$\begin{aligned}
L_{\mathcal{D}_\gamma}(\theta_T) &= \overline{\Phi} \left(\sqrt{d} \, \frac{\langle \theta_T, \mu \rangle}{\|\theta_T\|} \right) \\
&\leq \overline{\Phi} \left(\frac{\gamma \sqrt{d}}{24} \right) \\
&\leq \exp \left(-\frac{d\gamma^2}{1152} \right).
\end{aligned}$$

Taken together

$$\begin{aligned}
\mathbb{E}[L_{\mathcal{D}_\gamma}(\theta_T)] &\leq \exp \left(-\frac{d\gamma^2}{1152} \right) + m e^{-\gamma^2 d/8} + 2e^{-3d/4} \\
&\leq \exp \left(-\frac{d\gamma^2}{1152} \right) + e^{-\gamma^2 d/16} + 2e^{-3\gamma^2 d/4} \\
&\leq 4 \exp \left(-\frac{d\gamma^2}{1152} \right)
\end{aligned}$$

To see equation (56) holds, let us denote by $\tilde{\theta}_t$ the trajectory of GD on the same sample as before but initialized at zero, i.e. $\tilde{\theta}_0 = 0$ and $\tilde{\theta}_t$ satisfies equation (10). Now, let θ_T be the output when intializing at $\|\theta_0\| < R$. The non-expansiveness of gradient update step (see (Hardt et al., 2016, Lemma 3.6) provides:

$$\|\tilde{\theta}_T - \theta_T\| \leq \|\theta_0 - 0\| \leq R.$$

Next, assume $\|\tilde{\theta}_T\| \geq 25R$. In particular,

$$\begin{aligned}\frac{\langle \theta_T, \mu\rangle}{\|\theta_T\|} &\geq \frac{\langle \tilde{\theta}_T, \mu\rangle + \langle \theta_T - \tilde{\theta}_T, \mu\rangle}{\|\theta_T\|} \\ &\geq \frac{\frac{\gamma}{12}\|\tilde{\theta}_T\| - R\gamma}{\|\theta_T\|} && \text{equation (55)}\end{aligned}$$

By our assumption the nominator is positive and we have that:

$$\begin{aligned}\frac{\langle \theta_T, \mu\rangle}{\|\theta_T\|} &\geq \frac{\frac{\gamma}{12}\|\tilde{\theta}_T\| - R\gamma}{\|\tilde{\theta}_T\| + R} \\ &= \gamma\left(\frac{1}{24} + \frac{\|\tilde{\theta}_T\| - 25R}{24(\|\tilde{\theta}_T\| + R)}\right) \\ &\geq \frac{\gamma}{24} && \|\tilde{\theta}\| \geq 25R\end{aligned}$$

And equation (56) holds. It remains to show that $\|\tilde{\theta}_T\| \geq 25R$. To see that, we first show that $e^{36\langle \tilde{\theta}_t, \mu\rangle}$ increases by at least $9\eta\gamma^2$ at every iteration. From this we conclude that:

$$e^{36\|\tilde{\theta}_T\|\gamma} \geq e^{36\langle \tilde{\theta}_T, \mu\rangle} \geq 9\eta T\gamma^2.$$

And the result follows for our choice

$$\frac{\log\left(9\eta T\gamma^2\right)}{36\gamma} > 25R.$$

Write

$$a_t := \langle \widetilde{\theta}_t, \mu\rangle, \qquad s_{i,t} := \langle \widetilde{\theta}_t, z_i\rangle, \qquad \bar{z} := \frac{1}{m}\sum_{i=1}^m z_i.$$

Again, by monotinicity (equation (54)) we have that:

$$F_S(\widetilde{\theta}_t) \leq F_S(0) = \log 2,$$

Consequently,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \log(1 + e^{s_{i,t}}) &= F_S(\widetilde{\theta}_t) + \langle \widetilde{\theta}_t, \bar{z}\rangle \\ &\leq \log 2 + 3\gamma\|\widetilde{\theta}_t\| && \|\bar{z}\| \leq 3\gamma \\ &\leq \log 2 + 36a_t.\end{aligned}$$

The GD update then yields:

$$\begin{aligned}a_{t+1} - a_t &= \frac{\eta}{m}\sum_{i=1}^m \frac{\langle \mu, z_i\rangle}{1 + e^{s_{i,t}}} \\ &\geq \frac{\eta\gamma^2}{2m}\sum_{i=1}^m \frac{1}{1 + e^{s_{i,t}}} && \langle \mu, z_i\rangle \geq \frac{\gamma^2}{2} \\ &\geq \frac{\eta\gamma^2}{2}\exp\left(-\frac{1}{m}\sum_{i=1}^m \log(1 + e^{s_{i,t}})\right) && \text{Jensen's inequality} \\ &\geq \frac{\eta\gamma^2}{4}e^{-36a_t}.\end{aligned}$$

Using $e^u - 1 \geq u$, we therefore obtain

$$\begin{aligned} e^{36a_{t+1}} - e^{36a_t} &= e^{36a_t}\left(e^{36(a_{t+1}-a_t)} - 1\right) \\ &\geq 36e^{36a_t}(a_{t+1} - a_t) \\ &\geq 9\eta\gamma^2. \end{aligned}$$

Summing over $t = 0, \ldots, T-1$ and using $a_0 = 0$ gives

$$e^{36a_T} \geq 1 + 9\eta\gamma^2 T.$$

Hence,

$$\|\tilde{\theta}_t\| \geq \frac{a_t}{\gamma} \geq \frac{\log\left(1 + 9\eta\gamma^2 T\right)}{36\gamma} \geq 25R.$$

And this proves the general case.

Finally, We conclude the proof by demonstrating that equation (51) holds with probability at least $1 - me^{-\gamma^2 d/8} - 2e^{-3d/4}$.

Proof of equation (51) Indeed, the Gaussian tail bound and the union bound give

$$\Pr_S\left[\min_{i \leq m}\langle \mu, z_i\rangle < \frac{\gamma^2}{2}\right] \leq me^{-\gamma^2 d/8}. \tag{57}$$

To obtain the two remaining inequalities we use the following upper tail-bound for Chi-square distribution (see (Ghosh, 2021, Theorem 1)): If $Y \sim \chi^2_k$ then

$$\Pr(Y > 4k) \leq \exp\left(-\frac{3 - \log 4}{2}k\right) \leq e^{-3k/4}. \tag{58}$$

Write $\bar{z} = m^{-1}\sum_i z_i$. Since

$$md\|\bar{z} - \mu\|^2 \sim \chi^2_d, \qquad \sum_{i=1}^m \|n_i\|^2 \sim \chi^2_{md},$$

equation (58) gives

$$\Pr_S\left(\|\bar{z} - \mu\| > \frac{2}{\sqrt{m}}\right) \leq e^{-3d/4},$$

and

$$\Pr_S\left(\sum_{i=1}^m \|n_i\|^2 > 4md\right) \leq e^{-3md/4}.$$

$m\gamma^2 > 1$ implies then,

$$\|\bar{z}\| \leq \gamma + \frac{2}{\sqrt{m}} \leq 3\gamma,$$

and

$$\frac{1}{m}\sum_{i=1}^m \|z_i\|^2 \leq 2\gamma^2 + \frac{2}{md}\sum_{i=1}^m \|n_i\|^2 \leq 2\gamma^2 + 8 \leq 10.$$

Consequently, the sample satisfies equation (51) with probability at least

$$1 - me^{-\gamma^2 d/8} - 2e^{-3d/4}$$

C GENERALIZATION BOUND FOR G&C

We next give the formal statement for theorem 1 together with a formal proof. Let $S = ((x_i, y_i))_{i=1}^m \sim \mathcal{D}^m$ be a sample for binary classification, for a realization of a sequence $\theta_1, \theta_2, \ldots$ by the G&C algorithm write

$$N(S) := \left|\left\{\left(h_{\theta_t}(x_1), \ldots, h_{\theta_t}(x_m)\right) : 1 \leq t \leq \tau\right\}\right|.$$

Thus $N(S)$ counts distinct prediction patterns on the sample, observed until termination by G&C.

Theorem 6. Assume that G&C terminates almost surely. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$ jointly over S and G&C,

$$L_{\mathcal{D}}(\widehat{\theta}_{\mathrm{GC}}) \le \frac{48}{m} \log\left(\frac{40N(S)^2}{\delta}\right). \tag{59}$$

To prove the theorem, We first establish a generalization bound for a fixed hypothesis class in terms of its expected log-pattern count. In more detail, let $\mathcal{H}$ be a fixed nonempty hypothesis class. Denote by $N_{\mathcal{H}}(S)$ the number of distinct prediction patterns it realizes on the inputs of S:

$$N_{\mathcal{H}}(S) = |\{(h(x_1),\dots,h(x_m)) : h \in \mathcal{H}\}|.$$

and write

$$a_{\mathcal{H}} := \mathbb{E}_S \log N_{\mathcal{H}}(S).$$

Lemma 4. For every $\epsilon > 0$:

$$\Pr_S\left[\exists \theta \in \mathcal{H} : L_S(\theta) = 0,\ L_{\mathcal{D}}(\theta) > \epsilon\right] \le 4\exp\left(4a_{\mathcal{H}} - \frac{m\epsilon}{4}\right). \tag{60}$$

Proof. We begin by providing a generalization bound for a fixed hypothesis class w.r.t the number of trace patters observed on a sample. This result is due to (Cortes et al., 2019, Corollary 4) that states in our notation:

$$\Pr_S\left[\exists \theta \in \mathcal{H} : L_S(\theta) = 0,\ L_{\mathcal{D}}(\theta) > \epsilon\right] \le 4 \mathop{\mathbb{E}}_{\mathbf{S}\sim\mathcal{D}^{2m}}[N_{\mathcal{H}}(\mathbf{S})]\exp\left(-\frac{m\epsilon}{4}\right). \tag{61}$$

Notice that $\mathbf{S}$ is a sample of $2m$ i.i.d examples and $N(\mathbf{S})$ is the number of distinct hypotheses restricted to $\mathbf{S}$. Now, by a simple counting argument we have that:

$$\mathop{\mathbb{E}}_{\mathbf{S}\sim\mathcal{D}^{2m}}[N_{\mathcal{H}}(\mathbf{S})] \le \mathop{\mathbb{E}}_{(S_1,S_2)\sim\mathcal{D}^m\times\mathcal{D}^m}[N_{\mathcal{H}}(S_1)\cdot N_{\mathcal{H}}(S_2)] = \left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[N_{\mathcal{H}}(S)]\right)^2.$$

Where the last equality follows from independence between the samples. Now we use the following inequality due to (Boucheron et al., 2000, Corollary 1), after converting to natural logarithms and weakening the constants:

$$a_{\mathcal{H}} = \mathbb{E}\log N_{\mathcal{H}}(S) \ge \frac{1}{2}\log \mathbb{E}[N_{\mathcal{H}}(S)] = \frac{1}{4}\log\left(\mathbb{E}[N_{\mathcal{H}}(S)]\right)^2, \tag{62}$$

Taken together,

$$\mathbb{E}[N_{\mathcal{H}}(\mathbf{S})] = \exp\left(\log \mathbb{E}[N_{\mathcal{H}}(\mathbf{S})]\right) \le e^{4a_{\mathcal{H}}}.$$

□

The preceding lemma concerns a fixed class. We next apply it to a prefix of the search.

Lemma 5. Fix a deterministic sequence $\theta_1, \theta_2, \dots$. Given a sample S, let τ be its first interpolating index, and let $N(S)$ count the distinct prediction patterns of $\theta_1,\dots,\theta_\tau$ on S. Suppose $\tau < \infty$ almost surely. Then, for every $n \ge 1$ and $\epsilon > 0$,

$$\Pr_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\theta_\tau) > \epsilon,\ N(S) \le n\right] \le 5ne^{-m\epsilon/48}. \tag{63}$$

Proof. For each $t \ge 1$, define

$$\mathcal{H}_t := \{\theta_1,\dots,\theta_t\}, \qquad N_t(S) := N_{\mathcal{H}_t}(S), \qquad a_t := \mathbb{E}_S \log N_t(S).$$

Both $N_t(S)$ and a_t are nondecreasing in t, and $N(S) = N_\tau(S)$. Choose the deterministic cutoff

$$t_* := \min\left\{t \ge 1 : a_t > \frac{m\epsilon}{24}\right\},$$

with $t_* = \infty$ if the set is empty. Notice that t_* depends only the distribution and the fixed sequence, and not on the sample.

We next consider two events.

First consider the event that there exists $t < t_*$ such that $L_S(\theta_t) = 0$ and $L_{\mathcal{D}}(\theta_t) > \epsilon$. If $t_\star < \infty$, apply lemma 4 to $\mathcal{H}_{t_*-1}$ and we obtain:

$$\begin{aligned}\Pr_S\left[\tau < t_*,\ L_{\mathcal{D}}(\theta_\tau) > \epsilon\right] &\le 4\exp\left(4\frac{m\epsilon}{24} - \frac{m\epsilon}{4}\right) \\ &= 4e^{-m\epsilon/12}. \end{aligned} \tag{64}$$

If $t_\star = \infty$ a similar argument follows by applying lemma 4 to $\mathcal{H} = \cup\mathcal{H}_t$

Next, suppose $t_* < \infty$ and consider the event that $\tau \ge t_*$ and that $N(S) \le n$ (notice that if $t_* = \infty$ this event cannot occur and we have only the first event to consider). This implies:

$$N_{t_*}(S) \le n.$$

Write $a = a_{t_*} > m\epsilon/24$ and $r = \log n$. We now apply the concentration inequality in (Boucheron et al., 2000, Corollary 1) (again converting to natural logarithms) and we have that: for $0 \le u \le a$,

$$\Pr_S\left[\log N_{t_*}(S) \le a - u\right] \le \exp\left(-\frac{u^2}{2a}\right). \tag{65}$$

If $r < a$, apply this with $u = a - r$ to obtain

$$\begin{aligned}\Pr_S\left[\log N_{t_*}(S) \le r\right] &\le \exp\left(-\frac{(a-r)^2}{2a}\right) \\ &\le e^{r-a/2} \le ne^{-m\epsilon/48}.\end{aligned}$$

If $r \ge a$, the final bound exceeds 1 and holds trivially. Consequently,

$$\Pr_S\left[\tau \ge t_*,\ N(S) \le n\right] \le ne^{-m\epsilon/48}. \tag{66}$$

Combining (64) and (66) gives

$$\begin{aligned}\Pr_S\left[L_{\mathcal{D}}(\theta_\tau) > \epsilon,\ N(S) \le n\right] &\le 4e^{-m\epsilon/12} + ne^{-m\epsilon/48} \\ &\le 5ne^{-m\epsilon/48},\end{aligned}$$

as claimed. □

Completing the proof of theorem 6. Condition on the entire sequence of draws, which is independent of S. Joint almost-sure termination implies that, for almost every such sequence, the search terminates almost surely over S. Fix one of these sequences. Since the sequence is independent of S, the sample remains distributed as $\mathcal{D}^m$. Recall that, by our assumption, for almost every sequence, the search terminates almost surely over S.

For $j = 0, 1, \ldots$, set

$$n_j := 2^j, \qquad \epsilon_j := \frac{48}{m}\log\left(\frac{10n_j^2}{\delta}\right).$$

By lemma 5,

$$\Pr_S\left[L_{\mathcal{D}}(\theta_\tau) > \epsilon_j,\ N(S) \le n_j\right] \le 5n_j e^{-m\epsilon_j/48} = \frac{\delta}{2^{j+1}}.$$

A union bound over $j \ge 0$ shows that, with probability at least $1 - \delta$, all these events are avoided. On this event, choose $j = \lceil \log_2 N(S) \rceil$. Since $N(S) \le n_j \le 2N(S)$,

$$L_{\mathcal{D}}(\theta_\tau) \le \epsilon_j = \frac{48}{m}\log\left(\frac{10n_j^2}{\delta}\right) \le \frac{48}{m}\log\left(\frac{40N(S)^2}{\delta}\right).$$

This holds for almost every fixed sequence of draws of hypotheses. Averaging over that sequence proves the joint-probability statement. □