EFFICIENT AND NOISE-TOLERANT PAC LEARNING OF MULTICLASS LINEAR CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise-tolerant PAC learning of linear models has long been a central problem in machine learning. In this paper, we study the multiclass linear classifiers $\{h_w : x \mapsto \arg\max_{y\in[k]} w_y \cdot x, x \in \mathbb{R}^d, w \in \mathbb{R}^{kd}\}$ under the challenging nasty noise. We give the first computationally efficient algorithm for this problem, assuming that the marginal distribution is a mixture of distributions with bounded covariance and that the uncorrupted distribution satisfies a margin condition. Our algorithm consists of two main ingredients: a convex optimization program that prunes corrupted samples by bounding the pairwise squared distance, and a standard multiclass hinge loss minimization. We provide a multiclass gradient analysis and show that after the pruning, the hinge loss minimization naturally enjoys a strong robustness. Specially, we establish a deterministic condition that for any corrupted sample surviving the pruning, there exists a close-by reference point that enjoys the margin condition like a clean sample. As a result, we show that efficient PAC learning is possible with a sample complexity of $\tilde{O}(k^2d)$, where $\tilde{O}$ hides logarithmic terms, for the class of multiclass linear classifiers under a significant amount of nasty noise.

1 INTRODUCTION

This paper revisits the problem of learning multiclass linear classifiers (MLCs) under noise corruptions. Robust learning problems have been extensively studied in binary learning scenarios. However, our understanding of when and how we can achieve algorithmic robustness in multiclass learning settings, i.e. $k > 2$, falls short. Notably, it was shown recently in Diakonikolas et al. (2025b) that distribution-free learning of multiclass linear classifiers under random classification noise suffers from a super-polynomial statistical query lower bound, whereas the binary case $k = 2$ is learnable in polynomial time (Blum et al., 1998). This contrast shows that efficient robust learning in multiclass settings differs drastically from that in binary settings. Hence, it is natural to ask:

> Can the robustness guarantees for binary classification be extended to multiclass classification under natural assumptions while preserving computational efficiency?

We answer this question in the affirmative under both concentration and margin conditions, building on recent progress in robust learning of halfspaces (Talwar, 2020; Shen, 2025). The motivation for combining these conditions comes from the limitations of less structured settings. In distribution-free halfspace learning, the tolerable malicious noise rate is bounded above by $\epsilon/(1 + \epsilon)$, where ϵ is the target error (Kearns & Li, 1988). Under isotropic logconcave marginals, efficient algorithms achieve malicious noise tolerance proportional to ϵ (Awasthi et al., 2017), but these guarantees still require the noise rate to decrease with the desired error. Combining suitable marginal distributional assumptions with a margin condition can change this picture substantially. Under their respective structural assumptions, Talwar (2020) shows that hinge loss minimization tolerates a constant fraction of adversarial label corruptions, while Shen (2025) obtains constant malicious noise tolerance through reweighted hinge loss minimization. These results demonstrate that appropriate structure can allow small error on the uncorrupted distribution even when the noise rate does not vanish with ϵ.

Extending this approach to multiclass learning addresses a gap that recent multiclass results under Gaussian marginals do not directly resolve. While previous results provide improper agnostic learning guarantees for multiclass linear classifiers (Diakonikolas et al., 2025a; 2026) under label corruptions,

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.

our objective is proper PAC learning with vanishing error rate ϵ on the uncorrupted distribution, even in the presence of a significant amount of feature corruption. We show that combining the concentration and margin conditions permits efficient proper learning with desired guarantees.

Formally, let $\mathcal{X} \subseteq \mathbb{R}^d$ be the instance space, and $\mathcal{Y} = \{1, 2, \ldots, k\}$ be the label space. We consider the class of MLCs $\mathcal{H} := \{h_W : x \mapsto \arg\max_{y \in \mathcal{Y}} \langle W_y, x \rangle, W \in \mathbb{R}^{k \times d}\}$ and assume that the marginal distribution $\mathcal{D}_X$ is a mixture of bounded covariance distributions or logconcave distributions. Moreover, let $\mathcal{D}$ be the joint distribution over $\mathcal{X} \times \mathcal{Y}$ where the instances are drawn from $\mathcal{D}_X$ and labeled by the target classifier W^*. We assume that the uncorrupted $\mathcal{D}$ satisfies a margin condition. We consider the following learning problem in the presence of nasty noise.

Definition 1.1 (PAC learning of MLCs with nasty noise)**.** *Let $\epsilon, \delta \in (0, 1)$ be the target error rate and confidence parameter, respectively. A nasty adversary $EX_{nas}(\mathcal{D}_X, W^*, \eta)$ fixes $\mathcal{D}_X, W^*$ and $\eta \in (0, \frac{1}{k})$ throughout the learning process, takes a request from the learner, and first draws n instances independently from $\mathcal{D}_X$ and labels them by W^*, forming the original sample $\overline{S} = \{(x_i, \arg\max_{y \in \mathcal{Y}} W_y^* x_i)\}_{i=1}^n$. The adversary, fully aware of the learning algorithm, may then inspect the set $\overline{S}$ and replace at most an η fraction of it with any instance-label pairs from $\mathcal{X} \times \mathcal{Y}$. The resulting set S is returned to the learner. The goal of the learner is to output a multiclass linear classifier $\widehat{W} \in \mathbb{R}^{k \times d}$ such that with probability $1 - \delta$ (over the randomness of the samples and all internal random bits of the learning algorithm),* $\Pr_{x \sim \mathcal{D}_X}(\arg\max_{y \in \mathcal{Y}} \widehat{W}_y x \neq \arg\max_{y \in \mathcal{Y}} W_y^* x) \leq \epsilon$.

1.1 Main results

To ensure learning efficiency, we rely on the following structural assumptions regarding the margin and the marginal distribution.

Assumption 1.1 (Multiclass γ-margin condition)**.** *There exists a vector $w^* = (w_1^*, \ldots, w_k^*) \in \mathbb{R}^{kd}$ with $\|w^*\|_2 \leq 1$ and a margin $\gamma > 0$ such that for any clean sample (x, y) in the support of $\mathcal{D}$:*

$$\min_{y' \in \mathcal{Y} \setminus \{y\}} \left\langle w_y^* - w_{y'}^*, x \right\rangle \geq \gamma. \tag{1}$$

Assumption 1.2 (Mixtures of distributions)**.** *The marginal distribution $\mathcal{D}_X$ is a mixture of K distributions $\mathcal{D}_1, \ldots, \mathcal{D}_K$, i.e., $\mathcal{D}_X = \sum_{j=1}^K q_j \mathcal{D}_j$, where $\sum_{j=1}^K q_j = 1$ and $q_j \geq \alpha > 0$, $\forall j \in [K]$. Furthermore, for each component $j \in [K]$, $\mathcal{D}_j$ is a distribution over $\mathcal{X}$ with mean $\mu_j = \mathbb{E}_{x \sim \mathcal{D}_j}[x]$ and covariance matrix $\Sigma_j = \mathbb{E}_{x \sim \mathcal{D}_j}[(x - \mu_j)(x - \mu_j)^\top]$ satisfying $\Sigma_j \preceq \sigma^2 \mathbb{I}_d$, where $\sigma > 0$.*

Though our algorithm and analysis work for any mixture of bounded covariance distributions, the assumption of logconcavity still contributes to a more favorable margin condition.

Assumption 1.3 (Mixture of logconcaves)**.** *Under Assumption 1.2, we say that $\mathcal{D}_X$ is a mixture of logconcave distributions if, $\forall j \in [K]$, $\mathcal{D}_j$ is a logconcave distribution with mean μ_j and $\Sigma_j \preceq \sigma^2 \mathbb{I}_d$.*

We show that obtaining error guarantees for learning MLCs while tolerating a significant amount of nasty noise is possible under the margin condition and the mixture of bounded covariance distributions. More importantly, we show that such guarantees can be achieved with a sample complexity of $\tilde{O}(k^2 d)$.

Theorem 1.2 (Main theorem, mixture of bounded covariance)**.** *Let $C \geq 4$ be some universal constant. Suppose Assumption 1.2 holds with $\alpha \in [\frac{0.6}{k}, \frac{1}{k}]$, and Assumption 1.1 holds with some $\gamma > \max(4\tau, 4C\sigma)$, where $\tau = 2\sigma \cdot \sqrt{\frac{k}{\epsilon}}$. By taking at least $\Omega\big(k^2(d \log(d) + \log \frac{k}{\epsilon\delta}) + kd \log \frac{1}{\delta}\big)$ samples from EX_{nas} with $\eta \leq \frac{1}{2^{12} k^2}$, Algorithm 1 returns a $\hat{w}$ such that $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$ with probability at least $1 - \delta$ when $\hat{w}$ is minimizer of Eq.* (8).

The above theorem shows that as long as the clean samples are separable with an $\Omega\big(\sigma \cdot \sqrt{k/\epsilon}\big)$-margin, MLCs are robustly learnable even under the notorious nasty noise condition. We show that if the marginal distribution is additionally a mixture of logconcave distributions, similar guarantees can be achieved with a margin that grows logarithmically in k/ϵ.

Theorem 1.3 (Main theorem, mixture of logconcaves)**.** *Let $C \geq 4$ be some universal constant. Suppose Assumption 1.3 holds with $\alpha \in [\frac{0.6}{k}, \frac{1}{k}]$ and Assumption 1.1 holds with some $\gamma > \max(4\tau, 4C\sigma)$, where $\tau = 2\sigma(\log \frac{k}{\epsilon} + 1)$. By taking at least $\Omega\big(k^2(d \log(d) + \log \frac{k}{\epsilon\delta}) + kd \log \frac{1}{\delta}\big)$ samples from*

EX_{nas} with $\eta \leq \frac{1}{2^{12}k^2}$, Algorithm 1 returns a $\hat{w}$ such that $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$ with probability at least $1-\delta$ when $\hat{w}$ is minimizer of Eq. (8).

Remark 1.4 (Tail bounds)**.** *The margin required in both theorems depends on ϵ. We conjecture that this is unavoidable. Bounded covariance distributions can have heavier tails than logconcaves. Our two theorems have the same sufficient sample complexity under their respective margin conditions. The bounded covariance case requires a margin γ that grows faster as ϵ decreases.*

Remark 1.5 (Non-uniform weight mixtures)**.** *The mixture weights need not be equal. Under the main theorems' condition $\alpha \geq 0.6/k$, we have $1 = \sum_{j=1}^{K} q_j \geq K\alpha$. Hence $K \leq 1/\alpha \leq 5k/3 = O(k)$. The noise rate is constant for fixed k.*

Remark 1.6 (Implication for $k = 2$)**.** *For $k = 2$, our theorem also gives constant nasty-noise tolerance under the stated margin and mixture assumptions. This is because previous analysis essentially relies on the existence of a set of uncorrupted i.i.d. samples, which is possible under the malicious noise model (Shen (2025)). In this work, we incorporate a much more powerful outlier-removal scheme. Even when the clean samples are no longer i.i.d., we are still able to recover essential properties for robust learning.*

Remark 1.7. *The number K of mixture components in Assumptions 1.2 and 1.3 need not equal the number of classes k. Under the weight condition in the main theorems, however, $q_j \geq \alpha \geq 0.6/k$ and $1 = \sum_{j=1}^{K} q_j \geq K\alpha$. Hence $K \leq 1/\alpha \leq 5k/3 = O(k)$.*

1.2 Overview of techniques

Algorithmic design. Our main algorithm consists of a hard outlier pruning step followed by standard multiclass hinge loss minimization. The pruning step constructs a support graph, of which the edges between samples with the same observed label are assigned with nonnegative weights through a semidefinite program. The program enforces large total weight while controlling the weighted sum of the pairwise squared distances in every direction. Samples with small total weight are removed. We then perform multiclass hinge loss minimization on the remaining samples.

Support graph certificate. Some corrupted samples may survive the pruning. Our support graph constraints ensure that every surviving corrupted point assigns enough positive weight to clean points with the same label, which naturally defines a weighted average. The key insight is that the weighted average on clean samples would satisfy the target margin condition and can serves as a clean reference point. The matrix constraint further bounds the aggregate squared deviations of the surviving corrupted points from these reference points. This directly controls their contribution to the hinge loss subgradient. This is key to our analysis for showing that the gradient contribution from the surviving corrupted samples is well controlled.

Dense pancakes and multiclass gradient analysis. Our dense pancake analysis guarantees that, for all but an ϵ fraction of target points, the surviving sample contains sufficiently many uncorrupted points with the same label and close scores in every class direction. The multiclass extension is nontrivial because each point has $k-1$ competing labels, all of which must be controlled simultaneously. If $\hat{w}$ misclassifies such a point, the clean pancake arguments, together with the support graph bound on surviving corrupted points, force every valid subgradient to have a strictly negative inner product with w^*. This contradicts the zero-subgradient condition at an interior optimum. At a boundary optimum, we use a direction orthogonal to $\hat{w}$ for which the KKT condition requires zero inner product, while our gradient bound makes it strictly negative. The required pancake counts are established on the original independent sample and remain sufficient after corruption and the pruning step.

2 Related Work

Robust learning has been studied under several noise models. Random classification noise changes labels according to fixed probabilities that depend on the original class (Diakonikolas et al., 2025b). Under binary Massart noise, the probability of a label flip may depend on the instance but is bounded by a rate below $1/2$ (Chandrasekaran et al., 2024; Diakonikolas & Zarifis, 2024). Adversarial label noise allows arbitrary relabeling while leaving features unchanged (Talwar, 2020). Malicious noise may change both features and labels, but the positions chosen for corruption are independent of the observed examples (Kearns & Li, 1988; Awasthi et al., 2017). Under nasty noise, the adversary

inspects the original sample before choosing which pairs to replace. The unchanged points therefore need not form an independent sample (Bshouty et al., 1999; Blanc et al., 2026). For binary halfspaces, localization tolerates malicious noise of order ϵ under isotropic logconcavity (Awasthi et al., 2017). Under margin and logconcave mixture assumptions, Talwar (2020) analyzes the robustness of hinge loss using dense pancakes, Shen (2025) obtains constant malicious noise tolerance through soft removal and reweighted hinge loss, and Zeng & Shen (2025) extends it to attribute-efficient settings.

For multiclass linear classifiers, standard SVM formulations minimize multiclass hinge loss (Weston & Watkins, 1999; Crammer & Singer, 2001). Loss and prediction correction methods address corrupted labels (Patrini et al., 2017; Rooyen & Williamson, 2018; Zhang et al., 2021). Let OPT denote the minimum error of a multiclass linear classifier on the noisy labeled distribution. Under random classification noise, learning to error $\mathrm{OPT} + \epsilon$ without distributional assumptions has superpolynomial statistical query lower bounds, even with three classes and a known noise matrix (Diakonikolas et al., 2025b). Under Gaussian marginals, an improper learner achieves $O(\mathrm{OPT}) + \epsilon$ error with polynomial dependence on d and exponential dependence on a polynomial in k/ϵ (Diakonikolas et al., 2025a). More recent Gaussian results achieve $\widetilde{O}(k^{3/2}\sqrt{\mathrm{OPT}}) + \epsilon$ in $\mathrm{poly}(d, k, 1/\epsilon)$ time and $O(\mathrm{OPT}) + \epsilon$ for three classes (Diakonikolas et al., 2026). These guarantees measure error on the noisy labeled distribution under a Gaussian feature marginal.

Semidefinite programs have also been used in clustering and robust learning to control the effect of outliers through weights on sample points.Peng & Wei (2007) relax an assignment matrix for k-means to minimize the sum of squared distances within clusters, then round the solution to obtain a partition of the sample points. Kushagra et al. (2017) add a penalized noise cluster to a k-means SDP and recover separated clusters in the presence of outliers. For robust mean estimation, Steinhardt et al. (2017) limit the weight assigned to each sample when reconstructing a point and minimize the spectral reconstruction error. Their Euclidean optimization admits an SDP formulation and they also give an implementation using singular value decomposition in the Euclidean case.

Our program assigns symmetric weights to pairs with the same received label, retains high total weight, and bounds weighted squared differences in every direction. It uses these bounds to prune points before unweighted multiclass hinge loss minimization, without recovering mixture components or estimating their means. Our error analysis builds on dense pancake arguments for binary hinge loss learning Talwar (2020); Shen (2025). Under our margin and bounded covariance mixture assumption, we obtain error at most ϵ on the uncorrupted distribution while tolerating a constant fraction of nasty noise when the number of classes k is constant.

3 PRELIMINARIES

Hinge loss minimization for multiclass linear classifiers is a textbook algorithm. We use the multiclass hinge loss notation of Shalev-Shwartz & Ben-David (2014) and formally define it as follows. We use $[k] = \{1, \ldots, k\}$ and denote by $\langle u, v\rangle$ the inner product between vectors u, v. Recall the instance space $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space $\mathcal{Y} = [k]$ with $k \geq 2$. The number of mixture components is K, which need not equal k. We use $\|\cdot\|_2$ for the Euclidean norm. All logarithms are natural.

Feature mapping and predictors. Recall that $W \in \mathbb{R}^{kd}$ has class weights W_y as its rows, and $w \in \mathbb{R}^{kd}$ is its row-by-row vectorization. We use W_y and w_y interchangeably. The feature mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{kd}$ maps (x, y) to a vector whose y-th block is x and whose other blocks are zero. Thus,

$$h_w(x) = \arg\max_{y \in \mathcal{Y}} \big\langle w, \Psi(x, y)\big\rangle = \arg\max_{y \in \mathcal{Y}} \big\langle W_y, x\big\rangle = h_W(x). \tag{2}$$

We have $\|w\|_2^2 = \sum_{y \in \mathcal{Y}} \|w_y\|_2^2$. Prediction ties are resolved by the smallest label.

Generalized hinge loss. Denote by $\mathbb{1}(\cdot)$ the indicator function. The generalized hinge loss is $\ell(w; (x, y)) = \max_{y' \in \mathcal{Y}} \big(\mathbb{1}(y \neq y') - \big\langle w, \Psi(x, y) - \Psi(x, y')\big\rangle\big)$. To match the margin condition, we use the scaled loss

$$\ell_\gamma(w; (x, y)) := \max_{y' \in \mathcal{Y}} \left(\mathbb{1}(y \neq y') - \left\langle \frac{w}{\gamma}, \Psi(x, y) - \Psi(x, y') \right\rangle \right). \tag{3}$$

The loss over a sample S is

$$\ell_\gamma(w; S) = \sum_{(x,y)\in S} \ell_\gamma(w; (x, y)). \tag{4}$$

We omit the subscript γ when it is clear from the context. ∂_w denotes the subdifferential with respect to w. The dense pancake condition requires matching labels and closeness in every class score.

Definition 3.1 (Multiclass pancake)**.** *For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\|w\|_2 \le 1$, and $\tau > 0$, define the width-τ multiclass pancake centered at (x, y) by*

$$P_w^\tau(x, y) := \left\{ (x', y') \in \mathcal{X} \times \mathcal{Y} : y' = y \ \wedge \ \forall \bar{y} \in \mathcal{Y}, \left| \left\langle w_{\bar{y}}, x' - x \right\rangle \right| \le \tau \right\}. \tag{5}$$

For $0 < \rho \le 1$, the pancake is ρ-dense with respect to a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ if

$$\Pr_{(x',y')\sim\mathcal{D}} \big((x', y') \in P_w^\tau(x, y) \big) \ge \rho. \tag{6}$$

Let $\mathcal{D}_1, \mathcal{D}_2$ be distributions on $\mathcal{X} \times \mathcal{Y}$. For $0 < \beta < 1$, the pair $(\mathcal{D}_1, \mathcal{D}_2)$ satisfies the (τ, ρ, β)-dense pancake condition if, for every $w \in \mathbb{R}^{kd}$ with $\|w\|_2 \le 1$,

$$\Pr_{(x,y)\sim\mathcal{D}_2} \big(P_w^\tau(x, y) \text{ is } \rho\text{-dense w.r.t. } \mathcal{D}_1 \big) \ge 1 - \beta. \tag{7}$$

We alternatively say that $\mathcal{D}_1$ satisfies (τ, ρ, β)-dense pancake condition with respect to $\mathcal{D}_2$. We say (x, y) is the center of any pancake $P_w^\tau(x, y)$.

Additional notation. For a nonempty sample set S, μ_S denotes its empirical mean. We identify S with its uniform empirical distribution and write $(x, y) \sim S$ for a uniform draw. We write $(\mathcal{D}_X, w^*)$ for the distribution of features drawn from $\mathcal{D}_X$ and labeled by h_{w^*}. Denote by $\mathrm{err}_{\mathcal{D}_X}(w) = \Pr_{x\sim\mathcal{D}_X}(h_w(x) \ne h_{w^*}(x))$ the error function of a classifier induced by w. For symmetric matrices, $A \preceq B$ means $B - A$ is positive semidefinite. The symbols $\mathbb{I}_d$, tr, and $\lambda_{\max}$ denote the identity matrix, trace, and largest eigenvalue. $\theta(u, v)$ is the angle between nonzero vectors u, v.

4 Algorithm

Our algorithm has two main ingredients. The first one is a hard outlier pruning scheme (Algorithm 2) that constructs a support graph with bounded weighted sum of the pairwise squared distances in every direction. Specifically, it solves a semidefinite program that assigns proper weights to graph edges, such that the weighted average of the cross-sample squared distances is bounded by the covariance up to constant factors. As a result, any sample with an assigned weight less than $1/2$ will be pruned. This step is extremely helpful in restricting the influence from the dirty samples using second moment information. Following the pruning scheme is a standard multiclass hinge loss minimization (Eq. (8)), which provably returns a multiclass linear classifier $h_{\widehat{w}}$ that enjoys PAC guarantees.

Algorithm 1 Multiclass SVM

Require: Training set $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, parameter α, known scale $\sigma > 0$, margin $\gamma > 0$, fixed universal constant $C \ge 4$, and feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{dk}$.

Ensure: A multiclass linear classifier $h_{\widehat{w}}$.

1: ANS $\leftarrow$ Algorithm 2(S, α, σ, C).
2: **if** ANS $=$ failure **then**
3: **return** $h_0(x) \equiv 1$.
4: **end if**
5: Let $\{B_y\} \leftarrow$ ANS be the returned sets and $\widehat{S} \leftarrow \bigcup_{y\in\mathcal{Y}} B_y$.
6: Solve:

$$\widehat{w} \in \operatorname*{arg\,min}_{w\in\mathbb{R}^{dk}:\|w\|_2\le 1} \sum_{(x_i,y_i)\in\widehat{S}} \max_{y'\in\mathcal{Y}} \left\{ \mathbb{1}(y' \ne y_i) - \left\langle \frac{w}{\gamma}, \Psi(x_i, y_i) - \Psi(x_i, y') \right\rangle \right\}. \tag{8}$$

7: **return** $h_{\widehat{w}}(x) = \arg\max_{y\in\mathcal{Y}} \langle \widehat{w}, \Psi(x, y) \rangle$.

It is useful to think of the received sample S as a union of the set S_C of clean points and the set S_D of dirty points. Note that $|S_D| \le \eta n$ from Definition 1.1. For the filtered sample $\widehat{S} \subseteq S$ from Step 5, write $\widehat{S}_C = \widehat{S} \cap S_C$ and $\widehat{S}_D = \widehat{S} \cap S_D$.

Support graph certificate. One of the key technical contributions of this paper is a semidefinite program that gives a certificate that the influence from the dirty samples is restricted. Specifically, we will use the label information from the samples and construct a support graph. In this graph, every sample is a vertex and any edge connecting two vertices will be assigned a weight. Intuitively, samples from different classes should lie farther away from each other, such that assigning lower weights to the edges connecting them would lead to a smaller weighted sum of the squared distances. Since we concern about the pairwise distance between any pair of samples, it is useful to store all weights in a symmetric matrix.

Let $A = (a_{ii'})$ be a symmetric nonnegative pair weight matrix and define $a_i := \sum_{i'=1}^{n} a_{ii'}$. Find A such that

$$\begin{aligned} &0 \le a_i \le 1, \quad a_{ii'} = a_{i'i} \ge 0, \quad a_{ii'} = 0 \text{ if } y_i \ne y_{i'}, \quad a_{ii'} \le \tfrac{4a_i}{\alpha n}, \\ &\sum_i a_i \ge (1 - \alpha/16)n, \quad \tfrac{1}{2}\sum_{i,i'} a_{ii'}(x_i - x_{i'})(x_i - x_{i'})^\top \preceq \tfrac{C^2\sigma^2 n}{4} I_d. \end{aligned} \tag{9}$$

For $a_i > 0$, the bound $a_{ii'} \le 4a_i/(\alpha n)$ forces point i to assign positive weights to at least $\lceil \alpha n/4 \rceil$ sample points with the same label. If $\eta \le \alpha/64$, these points cannot all be corrupted. In particular, every surviving corrupted point assigns significant weights to clean points with the same label. In more details, a point with $a_i \ge 1/2$ may still be a dirty one. For such a point, the pair weight bound $a_{ii'} \le 4a_i/(\alpha n)$ and $|S_D| \le \eta n$ give $\sum_{i' \in S_C} a_{ii'} \ge a_i - |S_D| \frac{4a_i}{\alpha n} \ge \left(1 - \frac{4\eta}{\alpha}\right) a_i$. These clean points have the same label as point i. Under Assumption 1.1, their weighted average, as we will define as the reference point $\mu_{i,C}$, satisfies the target margin γ with respect to w^* as well. The matrix constraint bounds the squared deviations of the surviving corrupted points from these reference points. This is how the pruning program controls the influence from surviving dirty points in $\hat{S}$. It is worth noting that, we assign zero weights to every pair of points with different labels. This use of label information is central to our method and distinguishes it from unsupervised robustness techniques.

Algorithm 2 Support graph hard pruning

Require: Sample $S = \{(x_i, y_i)\}_{i=1}^n$, minimum component weight lower bound α, known scale $\sigma > 0$, and fixed universal constant $C \ge 4$.
Ensure: Disjoint same-label subsets $\{B_y\}$, or failure.
1: Set $T_y \leftarrow \{i \in [n] : y_i = y\}$ and $n_y \leftarrow |T_y|$ for each $y \in \mathcal{Y}$.
2: Solve Program (9).
3: **if** no feasible solution is obtained **then**
4: **return** "failure".
5: **end if**
6: Set $B_y \leftarrow \{(x_i, y_i) : i \in T_y,\ a_i \ge 1/2\}$ for each $y \in \mathcal{Y}$.
7: **return** $\{B_y\}$.

5 Analysis

We first derive a pointwise gradient condition for correct classification using the gradient analysis on hinge loss function. We then apply this condition to the surviving samples, analyze the pruning step, and establish the required sample size.

5.1 Pointwise Gradient Analysis

To analyze the performance of multiclass hinge loss minimization, we adapt the gradient argument developed for halfspace learning by Talwar (2020) and Shen (2025). Specifically, we establish a list of contradictory results which confirms the optimality of a $\hat{w}$ ensures its correctness. The challenge lies in that when $k > 2$, all sample points surviving the pruning step may have more competing labels to the true one. As a remedy, the margin condition in Assumption 1.1 ensures that for any clean point

(x, y), w_y^* has a score at least γ more than all of the remaining $k-1$ competing labels. Meanwhile, the multiclass pancake condition in Definition 3.1 is also made compatible with k different directions at the same time for a unit vector $w \in \mathbb{R}^{kd}$. In more details, the binary pancake condition in prior works only requires the points to be close along a single direction, namely $w_{+1} - w_{-1}$ when $\mathcal{Y} = \{-1, +1\}$. Our analysis generalizes the arguments in prior work by accounting for multiple directions.

Fix a point (x, y) misclassified by $\hat{w}$ and let set $S_P = \hat{S}_C \cap P_{\hat{w}}^{\tau}(x, y)$ include the clean sample points close to it. The following results compare the clean gradient contribution from S_P to the corrupted gradient contribution from $\hat{S}_D$ after applying Algorithm 2. If the size $|S_P|$ for (x, y) is sufficiently large, the resulting gradient bounds contradict the optimality of $\hat{w}$. Such an argument is applied both in the w^* direction for an interior optimum and in a direction perpendicular to $\hat{w}$ for a boundary optimum. Intuitively, when the minimization program runs, the negative gradient should always bring candidate closer to w^*, given that the clean samples are overwhelming. Hence, the component of any gradient (or subgradient) along w^* should always be negative. In the case when an optimum $\hat{w}$ is on the boundary, we choose a direction orthonogal to $\hat{w}$. The detailed proofs can be found in Section A.

Lemma 5.1 (Gradient bound along w^*)**.** *Let $\hat{w}$ be a minimizer of Eq.* (8) *in Algorithm 1. Suppose Assumption 1.1 holds with parameter γ, the pruning weights satisfy Program* (9)*, $C \geq 4$, and $\eta \leq \alpha/64$. Assume that point (x, y) is misclassified by $\hat{w}$. Consider pancake $P_{\hat{w}}^{\tau}(x, y)$ with $\tau < \gamma/4$ and $\frac{|S_P|}{4} > \frac{2C\sigma n\sqrt{\eta}}{\gamma}$, then every $g \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}}$ satisfies*

$$\langle g, w^* \rangle \leq -\frac{3}{4}|S_P|, \tag{10}$$

Lemma 5.2 (Gradient bound along w')**.** *Let $\hat{w}$ be a minimizer of Eq.* (8) *in Algorithm 1. Define*

$$w' := \frac{w^* - \kappa\hat{w}}{\sqrt{1-\kappa^2}}, \tag{11}$$

where $\kappa := \langle \hat{w}, w^* \rangle$, and assume that $\theta(\hat{w}, w^*) \in (0, \pi/2)$. Suppose Assumption 1.1 holds with parameter γ, the pruning weights satisfy Program (9)*, $C \geq 4$, and $\eta \leq \alpha/64$. Assume that sample (x, y) is misclassified by $\hat{w}$. If the pancake $P_{\hat{w}}^{\tau}(x, y)$ has width $\tau < \gamma/4$ and $\frac{|S_P|}{4} > \frac{2C\sigma n\sqrt{\eta}}{\gamma}$, then every $g \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}}$ satisfies*

$$\left\langle g, w' \right\rangle \leq -\frac{\sqrt{5}}{4}|S_P|. \tag{12}$$

It is easy to see from both Lemma 5.1 and 5.2 that as long as $|S_P| > 0$ and $\frac{|S_P|}{4} > \frac{2C\sigma n\sqrt{\eta}}{\gamma}$, both Eq. (10) and (12) to be strictly negative. If $\|\hat{w}\|_2 < 1$, the optimum has a zero subgradient, contradicting Eq. (10). If $\|\hat{w}\|_2 = 1$, optimality gives $g + \lambda\hat{w} = 0$ for some $\lambda \geq 0$. As Eq. (10) implies $\langle \hat{w}, w^* \rangle > 0$ and the margin and pancake conditions exclude positive collinearity, w' in Eq. (11) is well defined and is perpendicular to $\hat{w}$. Hence $\left\langle g, w' \right\rangle = 0$, contradicting Eq. (12). As a result, (x, y) must be correctly classified by $\hat{w}$. We summarize it in the following lemma.

Lemma 5.3. *Suppose Assumption 1.1 holds with parameter γ, the pruning weights satisfy Program* (9)*, $C \geq 4$, and $\eta \leq \alpha/64$. Let $\hat{w}$ be a minimizer of Eq.* (8)*. Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be any point, with $\tau < \gamma/4$. If $S_P \neq \emptyset$ and $|S_P| > \frac{8C\sigma n\sqrt{\eta}}{\gamma}$, then (x, y) is not misclassified by $\hat{w}$.*

5.2 Analysis of Algorithm 1

In this section, we show the learning guarantee for Algorithm 1 based on that Algorithm 2 does a proper pruning job. Let $S_j \subseteq S$ represent clean samples in S drawn from the j-th component of $\mathcal{D}_X$ and labeled by w^*. Then, $S_C = \cup_{j\in[K]} S_j$. Similarly, let $\hat{S}_j := S_j \cap \hat{S}$ denote the remaining clean samples after Step 5 of Algorithm 1. It is easy to see that $\hat{S}_C = \cup_{j\in[K]} \hat{S}_j$.

With a large enough sample size, the original uncorrupted sample set $\overline{S}$ has the pancake density required for the feasibility of Program (9). However, under the nasty noise condition, the adversary is allowed to inspect and replace up to an η fraction of the samples. The following proposition shows that the dense pancake structure is preserved after the pruning step.

Proposition 5.4 (Empirical dense pancake)**.** *Consider Algorithm 1. Suppose Assumptions 1.1 and 1.2 hold with $\alpha \in [0.6/k, 1/k]$, and $\eta \leq \alpha/64$. Let S be a set of samples from EX_{nas} with size*

$n \geq \Omega(k^2 d \cdot \log \frac{kd}{\delta})$. *With probability* $1-\delta$*, the Program* (9) *is feasible , and every feasible output* $\hat{S}$ *is* $(\tau, \frac{3\alpha}{20}, \epsilon)$*-dense with respect to* $(\mathcal{D}_X, w^*)$ *with appropriate* $\tau \in (0, \frac{\gamma}{4})$*. In addition,* $|\hat{S}| \geq \frac{|S|}{2}$ *and the* $\frac{3\alpha}{20}$*-density is contributed by* $\hat{S}_C$*.*

We first show that at least $3\alpha n/8$ original points are in each indicated pancake (see Proposition C.4). Then, the adversary replaces at most ηn of these points, and Program (9) removes another amount less than $\alpha n/8$ (see Lemma D.1). Thus, for every indicated (x, y),

$$\left|\hat{S}_C \cap \mathcal{P}_w^\tau(x,y)\right| \geq \frac{3\alpha n}{8} - \eta n - \frac{\alpha n}{8} \geq \frac{15\alpha n}{64} > \frac{3\alpha}{20}\left|\hat{S}\right|.$$

It remains to compare the clean pancake count with the corrupted contribution in Lemma 5.3. The following theorem shows that if $\hat{S}$ satisfies the dense pancake property (Proposition 5.4), then the error rate for the classifier $\hat{w}$ returned by Program (8) is at most ϵ.

Theorem 5.5 (Noise rate condition)**.** *Consider Algorithm 1 after feasible pruning, and let* $\hat{w}$ *be an exact minimizer of Eq.* (8)*. Suppose Assumption 1.1 holds with some* $\gamma > 0$ *and Assumption 1.2 holds with* σ*. Let* $k \geq 2$*,* $C \geq 4$*, and*$|S_D| \leq \eta n$*. Assume that* $\frac{\gamma}{4} > \max(\tau, C\sigma)$ *and* $\eta \leq \frac{1}{2^{12}k^2}$*. If set* $\hat{S}$ *is* $(\tau, \frac{3\alpha}{20}, \epsilon)$*-dense with respect to* $(\mathcal{D}_X, w^*)$ *with appropriate* $\tau \in (0, \frac{\gamma}{4})$ *for some* $\alpha \in [\frac{0.6}{k}, \frac{1}{k}]$*, the* $\frac{3\alpha}{20}$*-density is contributed by* $\hat{S}_C$*, and* $|\hat{S}| \geq \frac{|S|}{2}$*, then* $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$*.*

We include the computational complexity for Algorithm 1 in the following lemma. It gives polynomial-time bounds for pruning and SVM optimization. For more details, see Appendix B.

Lemma 5.6 (Runtime)**.** *Algorithm 1 runs in time polynomial in* n, d, k *and the input length.*

Suppose n satisfies Eq. (C.4), $\gamma > 4\max\{\tau, C\sigma\}$, and $\eta \leq 1/(2^{12}k^2)$. By Proposition 5.4 and Theorem 5.5, with probability at least $1-\delta$ over the original independent sample, the pruning program is feasible and every exact minimizer $\hat{w}$ of Eq. (8) on a surviving sample satisfies $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$.

5.3 Analysis of Algorithm 2

To prove feasibility of the program, we explicitly construct an assignment of pair weights. From each mixture component, we select a set of uncorrupted points that share the same label. Note that the label may differ across components. Within each selected set, we assign the same positive weight to every pair of points, and we assign zero weight to all other pairs. We choose these weights to satisfy the total weight constraint. The sample size bounds ensure that the selected sets contain enough pairs for each pair weight to remain within its prescribed bound. Finally, we verify the matrix constraint by bounding the weighted sum of outer products of the pairwise differences.

Moreover, the constructed weights assign zero weight to corrupted points. The sample bounds are first established on the original independent sample before the adversary chooses which points to replace. We show that removing points from the construction leaves enough uncorrupted points for the weight constraints, while the original matrix bound controls their pairwise squared differences. Algorithm 2 does not need to know which points were replaced. The proofs can be found in Appendix D.

Proposition 5.7 (Clean sample retention)**.** *Every feasible output of Algorithm 2 satisfies*

$$\left|S \setminus \hat{S}\right| < \frac{\alpha n}{8}, \qquad \left|\hat{S}\right| > (1-\alpha/8)n \geq \frac{n}{2}.$$

That is, an original pancake loses at most ηn points to corruption and fewer than $\alpha n/8$ to pruning.

Proposition 5.8 (Margin and squared deviations bounds on corrupted samples)**.** *Suppose Assumption 1.1 holds,* $\eta \leq \alpha/64$*, and the pruning weights are feasible. For every* $(x_i, y_i) \in \hat{S}_D$*, the clean same-label weighted average* $\mu_{i,C} := \frac{\sum_{i' \in S_C} a_{ii'} x_{i'}}{\sum_{i' \in S_C} a_{ii'}}$ *is well defined and satisfies* $\left\langle w^*_{y_i} - w^*_{\tilde{y}}, \mu_{i,C} \right\rangle \geq \gamma$ *for every* $\tilde{y} \neq y_i$*. Moreover,* $\sum_{i \in \hat{S}_D}(x_i - \mu_{i,C})(x_i - \mu_{i,C})^\top \preceq C^2\sigma^2 n I_d$*.*

The pruning analysis associates each surviving corrupted point with a weighted average $\mu_{i,C}$ of clean samples with the same label. The bounds on the pair weights ensure that these clean samples receive positive total weight, so the average is well defined and inherits their target margin. The matrix

constraint then bounds, in every direction, the sum of squared deviations of the surviving corrupted points from their corresponding clean averages. Together, these properties control the corrupted contribution to the gradient in Lemmas 5.1 and 5.2. In particular, grouping the gradient terms by label and using $\|w^*\|_2 \le 1$ avoids an additional factor of k, yielding a corrupted term strictly bounded by the size of S_P. This control of the corrupted contribution is a main contribution of the pruning step. The clean averages are used only in the analysis; the algorithm neither computes them nor identifies mixture components. The proofs appear in Appendix D.

Program (9) has at most $n(n+1)/2$ pair weights, $O(n^2)$ scalar constraints, and one $d \times d$ matrix constraint. Its computation is analyzed in Appendix D; Lemma 5.6 covers the runtime of Algorithm 1.

5.4 Sample Complexity

We bound the sample size needed to obtain dense pancakes in the original independent sample $\overline{S}$. We consider logconcave and bounded-covariance components separately, then their mixtures. Proposition C.4 combines the original sample counts with the pruning bounds to obtain the guarantee for the surviving sample. The proofs are in Appendix C.

Theorem 5.9 (Sample complexity for a single log-concave distribution)**.** *Suppose that Assumption 1.1 is satisfied with parameter γ. If distribution $\mathcal{D}_j$ is logconcave with mean μ_j and covariance $\Sigma_j \preceq \sigma^2 \mathbb{I}_d$, let $k \ge 2$, $0 < \beta < 1/k$, $0 < \delta < 1$, and $\tau = 2\sigma(\log(1/\beta) + 1) \le \gamma/2$. Then a set $\overline{S}_j$ of at least*

$$\frac{128}{1-k\beta}\left[kd\log\left(1+\frac{256\sqrt{d}}{(1-k\beta)^{3/2}}\right)+\log\frac{1}{\delta}\right] \tag{13}$$

independent samples from $(\mathcal{D}_j, w^*)$ satisfies the $(\tau, (1-k\beta)/2, k\beta)$-dense pancake condition with respect to $(\mathcal{D}_j, w^*)$ with probability at least $1-\delta$, simultaneously for all $\|w\|_2 \le 1$.

Theorem 5.10 (Sample complexity for a single bounded covariance distribution)**.** *Suppose that Assumption 1.1 is satisfied with parameter γ. If $\mathcal{D}_j$ is a distribution with mean μ_j and covariance $\Sigma_j \preceq \sigma^2 \mathbb{I}_d$, let $k \ge 2$, $0 < \beta < 1/k$, $0 < \delta < 1$, and $\tau = 2\sigma/\sqrt{\beta} \le \gamma/2$. Then a set $\overline{S}_j$ of at least the size in Eq. (13), drawn independently from $(\mathcal{D}_j, w^*)$, satisfies the $(\tau, (1-k\beta)/2, k\beta)$-dense pancake condition with respect to $(\mathcal{D}_j, w^*)$ with probability at least $1-\delta$.*

The bounded covariance width grows as $1/\sqrt{\beta}$, whereas the logconcave width grows logarithmically in $1/\beta$. Their sample bounds are the same, but bounded covariance requires a larger margin.

The two preceding component theorems give density $(1-k\beta)/2$ and allow a $k\beta$ fraction of centers to fail. When applying Theorem 5.11 below, use these values for its density and exception parameters, and use $\delta/(2K)$ as the failure allowance for each component. The resulting bound concerns the original independent mixture sample $\overline{S}$.

Theorem 5.11 (Sample complexity for mixtures)**.** *Suppose that Assumption 1.2 is satisfied with $\mathcal{D}_X = \sum_{j=1}^K q_j \mathcal{D}_j$, $\sum_j q_j = 1$, and $q_j \ge \alpha$. Fix $\tau > 0$, $0 < \rho \le 1$, and $0 < \beta, \delta < 1$. For each $j \in [K]$, let $n_j \ge 1$ be an integer such that an independent sample of any size $m \ge n_j$ from $(\mathcal{D}_j, w^*)$ satisfies the (τ, ρ, β)-dense pancake condition with respect to $(\mathcal{D}_j, w^*)$ with probability at least $1-\delta/(2K)$, uniformly over $\|w\|_2 \le 1$. Then an independent mixture sample $\overline{S}$ of size $n \ge \max\left\{2\max_{j\in[K]} \frac{n_j}{q_j}, \frac{8}{\alpha}\log\frac{2K}{\delta}\right\}$ satisfies the $(\tau, \alpha\rho/2, \beta)$-dense pancake condition with respect to $(\mathcal{D}_X, w^*)$ with probability at least $1-\delta$, uniformly over $\|w\|_2 \le 1$.*

6 Conclusion and Limitations

We prove a PAC guarantee for multiclass linear classifiers under a margin condition and a mixture of bounded covariance distributions. The algorithm tolerates $\eta \le 1/(2^{12}k^2)$ nasty noise, a constant rate for fixed k. The guarantee also applies to malicious, adversarial label, or random classification noise corruption whenever at most ηn sample points are changed.

The proof combines clean pancake counts with bounds on the directional squared deviations of retained replacement points from same-label unchanged point averages. The sufficient noise rate decreases as $1/k^2$, and the required margin depends on the target accuracy. Improving these conditions remains open; our analysis does not establish a lower bound for other second-order methods.

AI USE STATEMENT

Generative AI assisted with developing the feasibility program, drafting proofs for Algorithm 2, and polishing writing.

REPRODUCIBILITY STATEMENT

The assumptions, algorithms, and theoretical guarantees required to reproduce the theoretical results of this work are stated in the main paper. Complete proofs of the technical results are provided in the appendix.

REFERENCES

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63:50:1–50:27, 2017.

Guy Blanc, Yizhi Huang, Tal Malkin, and Rocco A. Servedio. Is nasty noise actually harder than malicious noise? In *Proceedings of the 37th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2026.

Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22:35–52, 1998.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. In Osamu Watanabe and Takashi Yokomori (eds.), *Proceedings of the 10th International Conference on Algorithmic Learning Theory*, Lecture Notes in Computer Science, pp. 206–218. Springer, 1999.

Gautam Chandrasekaran, Vasilis Kontonis, Konstantinos Stavropoulos, and Kevin Tian. Learning noisy halfspaces with a margin: Massart is no harder than random. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 2024.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Ilias Diakonikolas and Nikos Zarifis. A near-optimal algorithm for learning margin halfspaces with Massart noise. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 2024.

Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Nikos Zarifis. Robust learning of multi-index models via iterative subspace approximation. In *66th Annual Symposium on Foundations of Computer Science*, 2025a.

Ilias Diakonikolas, Mingchen Ma, Lisheng Ren, and Christos Tzamos. Statistical query hardness of multiclass linear classification with random classification noise. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Ilias Diakonikolas, Giannis Iakovidis, and Mingchen Ma. Polynomial-time robust multiclass linear classification under Gaussian marginals. *arXiv preprint arXiv:2605.21428*, 2026.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pp. 267–280, 1988.

Shrinu Kushagra, Nicole McNabb, Yaoliang Yu, and Shai Ben-David. Provably noise-robust, regularised k-means clustering. *CoRR*, abs/1711.11247, 2017.

Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd, and Hervé Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1–3):193–228, 1998. International Linear Algebra Society (ILAS) Symposium on Fast Algorithms for Control, Signals and Image Processing.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 2007.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

Brendan Van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

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

Jie Shen. Efficient PAC learning of halfspaces with constant malicious noise rate. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*. PMLR, 2025.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *CoRR*, abs/1703.04940, 2017.

Kunal Talwar. On the error resistance of hinge-loss minimization. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jason Weston and Chris Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the 7th European Symposium on Artificial Neural Networks*, 1999.

Shiwei Zeng and Jie Shen. Attribute-efficient PAC learning of sparse halfspaces with constant malicious noise rate. *CoRR*, abs/2505.21430, 2025.

Mingyuan Zhang, Jane H. Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

A Pointwise Gradient Analysis

Here i, i' denote sample indices, and j is a label index in the label groupings below. Membership such as $i \in \hat{S}_D$ means that the corresponding sample (x_i, y_i) belongs to $\hat{S}_D$. Recall that the feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{kd}$ has the y-th block equal to x and all others zero, i.e. $\langle w, \Psi(x, y)\rangle = \langle w_y, x\rangle$. The multiclass hinge loss is

$$\ell(w;\, x_i, y_i) = \max_{y' \in \mathcal{Y}} \Big(\mathbb{1}(y' \neq y_i) - \tfrac{1}{\gamma} \langle w, \Psi(x_i, y_i) - \Psi(x_i, y') \rangle \Big). \tag{14}$$

At the optimum $\hat{w}$, a subgradient of ℓ w.r.t. w is $\frac{1}{\gamma}\big(\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)\big)$, where we define a pseudo-label for the sample with index $i \in \hat{S}$

$$\hat{y}_i \in \arg\max_{y_i' \in \mathcal{Y}} \Big(\mathbb{1}(y_i' \neq y_i) - \tfrac{1}{\gamma} \langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, y_i') \rangle \Big), \tag{15}$$

which specifies the displayed subgradient. Therefore, choosing one active pseudo-label for each summand gives

$$g = \frac{1}{\gamma} \sum_{i \in \hat{S}} \big(\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \big) \in \partial_w \ell(w; \hat{S}) \Big|_{w=\hat{w}}. \tag{16}$$

The pseudo-labels may be chosen arbitrarily among the active maximizers. Prediction, as opposed to the auxiliary pseudo-label choice, uses the fixed tie rule of Algorithm 1.

In our proof, for g of the form Eq. (16), we use the inner product

$$\langle g, w \rangle = \frac{1}{\gamma} \sum_{i \in \hat{S}} \langle w, \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \rangle = \frac{1}{\gamma} \sum_{i \in \hat{S}} \langle w_{\hat{y}_i} - w_{y_i}, x_i \rangle, \tag{17}$$

which will be used interchangeably. Note that if $\hat{y}_i = y_i$ then $\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) = 0$, so such a sample contributes nothing to g; in the proofs below we may therefore restrict attention to indices with $\hat{y}_i \neq y_i$. We use $S_C, S_D, \hat{S}_C, \hat{S}_D$ from Section 3 and Section 4. Thus S_C contains the unchanged sample points and S_D contains the replacement points, with $\left|\hat{S}_D\right| \leq \eta n$. Recall that for any set $\hat{S}$, we can view it as a union of a subset of the clean samples and that of the dirty samples, i.e. $\hat{S} = \hat{S}_C \cup \hat{S}_D$. To perform pointwise gradient analysis, for any (x, y), we further partition the sample set $\hat{S}$ into three subsets $\hat{S} = S_P \cup S_O \cup \hat{S}_D$. That is, given any point (x, y) and its pancake $P^{\tau}_{\hat{w}}(x, y)$ as given in Definition 3.1, define

$$S_P := \hat{S}_C \cap P^{\tau}_{\hat{w}}(x, y), \quad S_O := \hat{S}_C \setminus P^{\tau}_{\hat{w}}(x, y), \quad \hat{S}_D = \hat{S} \cap S_D. \tag{18}$$

For $\|\hat{w}\|_2 \leq 1$, $P^{\tau}_{\hat{w}}(x, y)$ uses the same inequalities as in Definition 3.1. The point (x, y) need not belong to S or $\hat{S}$; no margin inequality at (x, y) is assumed. Based on this partition, we can also split g in Eq. (16) into three parts, i.e. $g = g_1 + g_2 + g_3$, where

$$g_1 = \frac{1}{\gamma} \sum_{i \in S_P} \big(\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \big),$$

$$g_2 = \frac{1}{\gamma} \sum_{i \in S_O} \big(\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \big),$$

$$g_3 = \frac{1}{\gamma} \sum_{i \in \hat{S}_D} \big(\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \big).$$

Following Talwar (2020) and Shen (2025), we will prove the robustness of the minimization program and a minimizer $\hat{w}$ using contradictions. That is, assume an instance-label pair (x, y) is misclassified by $\hat{w}$; if (x, y) has a large number of clean samples lying close to it, we are able to show that $\hat{w}$ must not be the optimum. We reproduce proofs from prior work wherever the multiclass setting requires substantive modification, and cite them otherwise. We remark that the structure of the proof might be similar to prior works, but many details for the multiclass learning setting (including the new pruning subroutine) cannot be extended naively from the binary setting.

Lemma A.1 (Restatement of gradient bound along w^*)**.** *Let $\hat{w}$ be a minimizer of Eq.* (8) *in Algorithm 1. Suppose Assumption 1.1 holds with parameter γ, the pruning weights satisfy Program* (9)*, $C \geq 4$, and $\eta \leq \alpha/64$. Assume that a point (x, y) is misclassified by $\hat{w}$. Consider pancake $P^{\tau}_{\hat{w}}(x, y)$ with $\tau < \gamma/4$, and define subsets as in Eq.* (18)*, then, for every valid subgradient g,*

$$\langle g, w^* \rangle \leq -|S_P| + \frac{2C\sigma n\sqrt{\eta}}{\gamma}. \tag{19}$$

Proof. For g of the form Eq. (16), we bound $\langle g_1, w^* \rangle$, $\langle g_2, w^* \rangle$, and $\langle g_3, w^* \rangle$ individually, and then combine the bounds.

Part 1: $\langle g_1, w^* \rangle \leq -|S_P|$. The term g_1 sums the subgradients of clean samples in the pancake, i.e. S_P:

$$\langle g_1, w^* \rangle = \frac{1}{\gamma} \sum_{i \in S_P} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle. \tag{20}$$

For each $i \in S_P$, we have that $y_i = y$. Let $y' := h_{\hat{w}}(x) \in \arg\max_{y'' \in \mathcal{Y}} \langle \hat{w}_{y''}, x \rangle$; since (x, y) is misclassified by $\hat{w}$, we have $y' \neq y$. In addition, since $(x_i, y_i) \in P^{\tau}_{\hat{w}}(x, y)$, Definition 3.1 gives that, for any $\bar{y} \in \mathcal{Y}$, $\left|\langle \hat{w}, \Psi(x_i, \bar{y}) - \Psi(x, \bar{y}) \rangle\right| \leq \tau$; we can then show that $\hat{y}_i \neq y_i$ for all $i \in S_P$. Specifically, it suffices to exhibit the existence of some y'_i such that

$$\mathbb{1}(y'_i \neq y_i) - \tfrac{1}{\gamma} \langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, y'_i) \rangle > 0, \tag{21}$$

since the maximum in Eq. (15) is then attained at a label other than y_i (as the choice $y'_i = y_i$ contributes exactly 0), and thus $y_i \neq \hat{y}_i$. Taking $y'_i = y'$, Eq. (21) holds for y' if

$$\langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, y') \rangle < \gamma.$$

We then show that the above inequality holds because

$$\begin{aligned} &\langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, y') \rangle \\ &= \langle \hat{w}, \Psi(x_i, y_i) - \Psi(x, y) + \Psi(x, y) \rangle - \langle \hat{w}, \Psi(x_i, y') - \Psi(x, y') + \Psi(x, y') \rangle \\ &\leq \langle \hat{w}, \Psi(x_i, y_i) - \Psi(x, y) \rangle - \langle \hat{w}, \Psi(x_i, y') - \Psi(x, y') \rangle \\ &\leq 2\tau < \gamma, \end{aligned} \tag{22}$$

where the second transition follows because $\langle \hat{w}, \Psi(x, y) - \Psi(x, y') \rangle \leq 0$ (by the choice of y'), and the third transition applies the pancake bound in the two directions $\bar{y} = y$ and $\bar{y} = y'$. Hence, we know that $\hat{y}_i$ and y_i are different. By the γ-margin condition on the clean samples in S_C, $\left\langle w^*_{y_i} - w^*_{\hat{y}_i}, x_i \right\rangle \geq \gamma$ for any $i \in S_P$, so each summand $\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle \leq -\gamma$. Summing over S_P gives $\langle g_1, w^* \rangle \leq -|S_P|$.

Part 2: $\langle g_2, w^* \rangle \leq 0$. The term g_2 sums the subgradients of the off-pancake clean samples:

$$\langle g_2, w^* \rangle = \frac{1}{\gamma} \sum_{i \in S_O} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle. \tag{23}$$

For any $i \in S_O \subseteq \hat{S}_C$ we distinguish two cases. If $\hat{y}_i = y_i$, the summand is exactly 0. If $\hat{y}_i \neq y_i$, Eq. (1) gives $\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle \leq -\gamma < 0$. In either case the summand is ≤ 0. Therefore $\langle g_2, w^* \rangle \leq 0$.

Part 3: $\langle g_3, w^* \rangle \leq \frac{2C\sigma n\sqrt{\eta}}{\gamma}$. In this part we use the following notation. For each label index $y' \in [k]$, let

$$S_{D,y'} := \{(x_i, y_i) \in \hat{S}_D : \hat{y}_i = y'\}, \tag{24}$$

so that $\{S_{D,y'}\}_{y' \in [k]}$ partitions $\hat{S}_D$. Recall also that w^* is a vector of norm at most one whose blocks satisfy $\sum_{y' \in [k]} \left\| w^*_{y'} \right\|_2^2 = \|w^*\|_2^2 \leq 1$.

By Proposition 5.8 in the analysis of Algorithm 2, the clean same-label reference $\mu_{i,\mathrm{C}}$ is defined by Eq. (72) and satisfies Eq. (73). Since $\eta \le \alpha/64$, Eq. (74) implies, for every unit vector $v \in \mathbb{R}^d$,

$$\sum_{i\in\hat{S}_D} \big((x_i - \mu_{i,\mathrm{C}}) \cdot v\big)^2 \le C^2\sigma^2 n. \tag{25}$$

The term g_3 sums the subgradients of the surviving dirty samples:

$$\langle g_3, w^*\rangle = \frac{1}{\gamma} \sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle. \tag{26}$$

For each $i \in \hat{S}_D$, write $x_i = \mu_{i,\mathrm{C}} + (x_i - \mu_{i,\mathrm{C}})$, where $\mu_{i,\mathrm{C}}$ is the reference point defined in Eq. (72). By Eq. (73), $\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, \mu_{i,\mathrm{C}} \right\rangle \le 0$. When $\hat{y}_i = y_i$, both sides are zero. So we will drop the negative term. Thus, we are only concerned about contribution from dirty samples whose pseudo-label $\hat{y}_i$ differs from their received label y_i. This gives

$$\begin{aligned}
&\sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle \\
&= \sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, \mu_{i,\mathrm{C}} \right\rangle + \sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i - \mu_{i,\mathrm{C}} \right\rangle && (27)\\
&\le \underbrace{\sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i}, x_i - \mu_{i,\mathrm{C}} \right\rangle}_{\text{(I)}} + \underbrace{\sum_{i\in\hat{S}_D} \left\langle -w^*_{y_i}, x_i - \mu_{i,\mathrm{C}} \right\rangle}_{\text{(II)}}. && (28)
\end{aligned}$$

We now bound (I) and (II) separately; each is at most $C\sigma n\sqrt{\eta}$. Note that we may *not* extract $w^*_{\hat{y}_i} - w^*_{y_i}$ from the sum, as it depends on i. Instead, we group (I) by $\hat{y}_i$ and (II) by y_i.

Bounding (I). Regrouping directly by $\hat{y}_i$ using Eq. (24),

$$\text{(I)} = \sum_{i\in\hat{S}_D} \left\langle w^*_{\hat{y}_i}, x_i - \mu_{i,\mathrm{C}} \right\rangle = \sum_{y'=1}^{k} \Big\langle w^*_{y'}, \sum_{i\in S_{D,y'}} (x_i - \mu_{i,\mathrm{C}}) \Big\rangle. \tag{29}$$

Applying the Cauchy–Schwarz inequality across the k blocks of w^* and using $\|w^*\|_2 \le 1$,

$$\begin{aligned}
\text{(I)} &\le \left(\sum_{y'=1}^{k} \left\| w^*_{y'} \right\|_2^2 \right)^{1/2} \left(\sum_{y'=1}^{k} \Big\| \sum_{i\in S_{D,y'}} (x_i - \mu_{i,\mathrm{C}}) \Big\|_2^2 \right)^{1/2} \\
&\le \left(\sum_{y'=1}^{k} \Big\| \sum_{i\in S_{D,y'}} (x_i - \mu_{i,\mathrm{C}}) \Big\|_2^2 \right)^{1/2}.
\end{aligned} \tag{30}$$

Fix y'. If the vector sum below is zero, the desired bound is immediate. Otherwise choose v to be that vector divided by its norm. Cauchy–Schwarz and the aggregate directional bound Eq. (25) give

$$\begin{aligned}
\Big\| \sum_{i\in S_{D,y'}} (x_i - \mu_{i,\mathrm{C}}) \Big\|_2 &= \sum_{i\in S_{D,y'}} (x_i - \mu_{i,\mathrm{C}}) \cdot v \\
&\le \sqrt{|S_{D,y'}|} \left(\sum_{i\in S_{D,y'}} ((x_i - \mu_{i,\mathrm{C}}) \cdot v)^2 \right)^{1/2} \\
&\le \sqrt{|S_{D,y'}|} \left(\sum_{i\in \hat{S}_D} ((x_i - \mu_{i,\mathrm{C}}) \cdot v)^2 \right)^{1/2} \\
&\le C\sigma\sqrt{n|S_{D,y'}|}.
\end{aligned} \tag{31}$$

The intermediate inequality uses nonnegative squares and $S_{D,y'} \subseteq \hat{S}_D$. Substituting Eq. (31) into Eq. (30) and using that the $S_{D,y'}$ partition $\hat{S}_D$,

$$\text{(I)} \le \left(\sum_{y'=1}^{k} C^2\sigma^2 n |S_{D,y'}| \right)^{1/2} = C\sigma\sqrt{n|\hat{S}_D|} \le C\sigma n\sqrt{\eta}, \tag{32}$$

where the last step uses $\left|\hat{S}_D\right| \le \eta n$.

Bounding (II). Each B_j returned by Algorithm 2 has observed label j. For a nonzero vector sum over $\hat{S}_D \cap B_j$, the argument in Eq. (31), with $S_{D,y'}$ replaced by $\hat{S}_D \cap B_j$, gives norm at most $C\sigma\sqrt{n|\hat{S}_D \cap B_j|}$; zero sums satisfy the same bound. Therefore,

$$\begin{aligned}
\text{(II)} &= \sum_j \Big\langle -w_j^*, \sum_{i \in \hat{S}_D \cap B_j} (x_i - \mu_{i,\mathrm{C}}) \Big\rangle \\
&\le \left(\sum_j \left\| w_j^* \right\|_2^2 \right)^{1/2} \left(\sum_j \Big\| \sum_{i \in \hat{S}_D \cap B_j} (x_i - \mu_{i,\mathrm{C}}) \Big\|_2^2 \right)^{1/2} \\
&\le \left(\sum_j C^2\sigma^2 n |\hat{S}_D \cap B_j| \right)^{1/2} = C\sigma\sqrt{n|\hat{S}_D|} \le C\sigma n\sqrt{\eta}.
\end{aligned}$$

Here, the disjoint subsets $\hat{S}_D \cap B_j$ partition the surviving dirty samples.

As a result, we have that

$$\langle g_3, w^* \rangle \le \frac{2C\sigma n\sqrt{\eta}}{\gamma}. \tag{33}$$

Combining. Adding the three parts:

$$\langle g, w^* \rangle = \langle g_1, w^* \rangle + \langle g_2, w^* \rangle + \langle g_3, w^* \rangle \le -|S_P| + \frac{2C\sigma n\sqrt{\eta}}{\gamma}, \tag{34}$$

which yields Eq. (19). Since the loss is a finite sum of maxima of affine functions, every subgradient is a convex combination of the vectors obtained by choosing one maximizing pseudo-label $\hat{y}_i$ per sample in Eq. (16). The bound holds for each choice and is linear in g, so it holds for every subgradient.

□

Lemma A.2 (Restatement of gradient bound on orthogonal w' direction)**.** *Let $\hat{w}$ be a minimizer of Eq.* (8)*, with $\|\hat{w}\|_2 = 1$. Define*

$$w' := \frac{w^* - \kappa\hat{w}}{\sqrt{1-\kappa^2}}, \tag{35}$$

where $\kappa := \langle \hat{w}, w^* \rangle$, and assume that $\theta(\hat{w}, w^*) \in (0, \pi/2)$. Suppose Assumption 1.1 holds with parameter γ, the pruning weights satisfy Program (9)*, $C \ge 4$, and $\eta \le \alpha/64$. Assume that a point (x, y) is misclassified by $\hat{w}$. Consider pancake $P^{\tau}_{\hat{w}}(x, y)$ with $\tau < \gamma/4$ and define subsets as in Eq.* (18)*, if $\frac{|S_P|}{4} > \frac{2C\sigma n\sqrt{\eta}}{\gamma}$, then, for every valid subgradient g,*

$$\langle g, w' \rangle \le -\frac{\sqrt{5}}{4}|S_P|. \tag{36}$$

Proof. For g of the form Eq. (16), we again bound each of $\langle g_1, w' \rangle$, $\langle g_2, w' \rangle$, $\langle g_3, w' \rangle$ in turn. Now we consider the case used in the orthogonal-direction argument, $\theta(\hat{w}, w^*) \in (0, \pi/2)$. Since $\|\hat{w}\|_2 = 1$ and $0 < \|w^*\|_2 \le 1$, we have

$$\kappa = \langle \hat{w}, w^* \rangle = \|w^*\|_2 \cos\theta(\hat{w}, w^*) \in (0, 1).$$

Moreover,

$$\langle w', \hat{w} \rangle = 0, \qquad \|w'\|_2^2 = \frac{\|w^*\|_2^2 - \kappa^2}{1 - \kappa^2} \leq 1.$$

Part 1: $\langle g_1, w' \rangle \leq -\frac{1}{2}|S_P|$**.** The term g_1 sums over the pancake clean samples S_P: $\langle g_1, w' \rangle = \frac{1}{\gamma} \sum_{i \in S_P} \left\langle w'_{\hat{y}_i} - w'_{y_i}, x_i \right\rangle$. For each $i \in S_P$, expanding $w' = (w^* - \kappa \hat{w})/\sqrt{1-\kappa^2}$ gives

$$\left\langle w'_{y_i} - w'_{\hat{y}_i}, x_i \right\rangle = \frac{\left\langle w^*_{y_i} - w^*_{\hat{y}_i}, x_i \right\rangle - \kappa \left\langle \hat{w}_{y_i} - \hat{w}_{\hat{y}_i}, x_i \right\rangle}{\sqrt{1-\kappa^2}}. \tag{37}$$

By Eq. (1) and the analysis in Lemma A.1, $y_i \neq \hat{y}_i$ and $\left\langle w^*_{y_i} - w^*_{\hat{y}_i}, x_i \right\rangle \geq \gamma$. Since $\hat{y}_i$ is the active maximizer of the hinge loss $\ell(\hat{w}; x_i, y_i)$, according to Eq. (22), we have that

$$\langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, \hat{y}_i) \rangle \leq \langle \hat{w}, \Psi(x_i, y_i) - \Psi(x_i, y') \rangle \leq 2\tau \leq \frac{\gamma}{2},$$

where $y' = h_{\hat{w}}(x)$, as in the proof of Lemma A.1. Therefore, $\left\langle w'_{y_i} - w'_{\hat{y}_i}, x_i \right\rangle \geq \frac{\gamma - \kappa\gamma/2}{\sqrt{1-\kappa^2}} = \frac{(1-\kappa/2)\gamma}{\sqrt{1-\kappa^2}}$. Equivalently, $\left\langle w'_{\hat{y}_i} - w'_{y_i}, x_i \right\rangle \leq -\frac{(1-\kappa/2)\gamma}{\sqrt{1-\kappa^2}}$. Summing over S_P gives

$$\langle g_1, w' \rangle \leq -\frac{1 - \kappa/2}{\sqrt{1-\kappa^2}}|S_P| \leq -\frac{1}{2}|S_P|, \tag{38}$$

since optimizing over $0 < \kappa < 1$ gives

$$\frac{1 - \kappa/2}{\sqrt{1-\kappa^2}} \geq \frac{\sqrt{3}}{2} \geq \frac{1}{2}. \tag{39}$$

Part 2: $\langle g_2, w' \rangle \leq 0$**.** The term g_2 sums the subgradients of the off-pancake clean samples:

$$\langle g_2, w' \rangle = \frac{1}{\gamma} \sum_{i \in S_O} \left\langle w'_{\hat{y}_i} - w'_{y_i}, x_i \right\rangle. \tag{40}$$

For any $i \in S_O \subseteq \hat{S}_C$, we consider two cases. If $\hat{y}_i = y_i$, then the active hinge label agrees with the true label, so this sample contributes 0 to g_2. If $\hat{y}_i \neq y_i$, then since $\hat{y}_i$ is the hinge-active label, its hinge value is at least the value attained at $y'_i = y_i$, which is 0, we have $1 - \frac{1}{\gamma} \langle \hat{w}_{y_i} - \hat{w}_{\hat{y}_i}, x_i \rangle \geq 0$, and hence $\langle \hat{w}_{y_i} - \hat{w}_{\hat{y}_i}, x_i \rangle \leq \gamma$. Using $w' = (w^* - \kappa\hat{w})/\sqrt{1-\kappa^2}$, we get

$$\left\langle w'_{y_i} - w'_{\hat{y}_i}, x_i \right\rangle = \frac{\left\langle w^*_{y_i} - w^*_{\hat{y}_i}, x_i \right\rangle - \kappa \left\langle \hat{w}_{y_i} - \hat{w}_{\hat{y}_i}, x_i \right\rangle}{\sqrt{1-\kappa^2}} \geq \frac{\gamma - \kappa\gamma}{\sqrt{1-\kappa^2}} > 0. \tag{41}$$

Therefore $\left\langle w'_{\hat{y}_i} - w'_{y_i}, x_i \right\rangle < 0$. Thus, if $\hat{y}_i = y_i$, the summand is 0, and if $\hat{y}_i \neq y_i$, the summand is negative. Hence each summand is ≤ 0, giving $\langle g_2, w' \rangle \leq 0$.

Part 3: $\langle g_3, w' \rangle \leq \frac{1}{\sqrt{1-\kappa^2}} \cdot \frac{2C\sigma n\sqrt{\eta}}{\gamma}$**.** The term g_3 sums over the surviving dirty samples:

$$\langle g_3, w' \rangle = \frac{1}{\gamma} \sum_{i \in \hat{S}_D} \left\langle w'_{\hat{y}_i} - w'_{y_i}, x_i \right\rangle. \tag{42}$$

Using $w' = \frac{w^* - \kappa\hat{w}}{\sqrt{1-\kappa^2}}$, we have

$$\langle g_3, w' \rangle = \frac{1}{\gamma\sqrt{1-\kappa^2}} \sum_{i \in \hat{S}_D} \left[\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, x_i \right\rangle - \kappa \left\langle \hat{w}_{\hat{y}_i} - \hat{w}_{y_i}, x_i \right\rangle \right]. \tag{43}$$

For each $i \in \hat{S}_D$, write $x_i = \mu_{i,\mathrm{C}} + (x_i - \mu_{i,\mathrm{C}})$, where $\mu_{i,\mathrm{C}}$ is the clean same-label average in Eq. (72), as used in Part 3 of Lemma A.1. If $\hat{y}_i = y_i$, both the target-mean and the $\hat{w}$ terms in Eq. (43) are zero. If $\hat{y}_i \neq y_i$, Eq. (73) gives $\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, \mu_{i,\mathrm{C}} \right\rangle \leq -\gamma$; since $\hat{y}_i$ maximizes Eq. (15), its hinge value is at least zero and $\left\langle \hat{w}_{\hat{y}_i} - \hat{w}_{y_i}, x_i \right\rangle \geq -\gamma$. Thus, in either case, since $0 < \kappa < 1$,

$$\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, \mu_{i,\mathrm{C}} \right\rangle - \kappa \left\langle \hat{w}_{\hat{y}_i} - \hat{w}_{y_i}, x_i \right\rangle \leq 0. \tag{44}$$

The remaining two sums are exactly (I) and (II) in Part 3 of Lemma A.1; the bounds established there give (I) + (II) $\leq 2C\sigma n\sqrt{\eta}$. Substituting into Eq. (43) gives

$$\begin{aligned} \left\langle g_3, w' \right\rangle &= \frac{1}{\gamma\sqrt{1-\kappa^2}} \sum_{i \in \hat{S}_D} \left[\left\langle w^*_{\hat{y}_i} - w^*_{y_i}, \mu_{i,\mathrm{C}} \right\rangle - \kappa \left\langle \hat{w}_{\hat{y}_i} - \hat{w}_{y_i}, x_i \right\rangle \right] \\ &\quad + \frac{(\mathrm{I}) + (\mathrm{II})}{\gamma\sqrt{1-\kappa^2}} \\ &\leq \frac{2C\sigma n\sqrt{\eta}}{\gamma\sqrt{1-\kappa^2}}. \end{aligned} \tag{45}$$

The last step uses Eq. (44) and the two bounds from Part 3 of Lemma A.1.

Combining. Adding the three parts, $\left\langle g, w' \right\rangle = \left\langle g_1, w' \right\rangle + \left\langle g_2, w' \right\rangle + \left\langle g_3, w' \right\rangle$, and using Eq. (38), $\left\langle g_2, w' \right\rangle \leq 0$, and Eq. (45), we obtain

$$\left\langle g, w' \right\rangle \leq -\frac{1 - \kappa/2}{\sqrt{1-\kappa^2}} |S_P| + \frac{1}{\sqrt{1-\kappa^2}} \cdot \frac{2C\sigma n\sqrt{\eta}}{\gamma}. \tag{46}$$

Since $\frac{|S_P|}{4} > \frac{2C\sigma n\sqrt{\eta}}{\gamma}$, then

$$\begin{aligned} \left\langle g, w' \right\rangle &\leq -\frac{1 - \kappa/2}{\sqrt{1-\kappa^2}} |S_P| + \frac{|S_P|}{4\sqrt{1-\kappa^2}} \\ &= -\frac{3/4 - \kappa/2}{\sqrt{1-\kappa^2}} |S_P|. \end{aligned} \tag{47}$$

Since $\min_{\kappa \in (0,1)} \frac{\frac{3}{4} - \frac{\kappa}{2}}{\sqrt{1-\kappa^2}} = \frac{\sqrt{5}}{4}$ (attained at $\kappa = 2/3$), we get

$$\left\langle g, w' \right\rangle \leq -\frac{\sqrt{5}}{4} |S_P|. \tag{48}$$

By the convex-hull argument in Lemma A.1, this bound for every choice of $\hat{y}_i$ holds for every subgradient. □

The following deterministic result gives us the guarantee that, whenever the required conditions of Eq. (1), the program in Eq. (9), $\eta \leq \alpha/64$ and Definition 3.1 are satisfied, a point (x, y) will not be misclassified by the classifier $\hat{w}$. It follows the same structure as that of Theorem 11 of Shen (2025), but we have extended it to the robust multiclass setting.

Lemma A.3 (Restatement of Lemma 5.3)**.** *Suppose the margin inequalities in Eq.* (1) *hold on S_C. Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be any point. Assume the constraints of the program in Eq.* (9) *and $\eta \leq \alpha/64$, let $\hat{w}$ be a optimum of Eq.* (8)*, and take $\tau < \gamma/4$ in Eq.* (18)*. If $S_P \neq \emptyset$ and*

$$|S_P| > \frac{8C\sigma n\sqrt{\eta}}{\gamma}, \tag{49}$$

then (x, y) is not misclassified by $\hat{w}$.

Proof. Recall that $\hat{w}$ is a minimizer of the program in Eq. (8). Assume for contradiction that (x, y) is misclassified by $\hat{w}$. By Lemma A.1 and condition in Eq. (49), $\langle g, w^* \rangle < 0$ for all $g \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}}$. We analyze the conditions in two cases. The objective is finite and convex, and the unit ball has the strictly feasible point $w = 0$. Thus the subgradient optimality and normal-cone conditions used

below apply. In the boundary case, the normal cone consists of $\lambda\hat{w}$ with $\lambda \geq 0$; this yields the stated stationarity equation.

Case 1 ($\|\hat{w}\|_2 < 1$). $\hat{w}$ is in the interior of the constraint set. For the optimization problem, first-order optimality gives $0 \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}}$, so there exists g with $\langle g, w^* \rangle = 0$, contradicting $\langle g, w^* \rangle < 0$. Thus (x, y) cannot be misclassified.

Case 2 ($\|\hat{w}\|_2 = 1$). $\hat{w}$ is on the boundary of the constraint set. By the Karush–Kuhn–Tucker (KKT) conditions, there exist a subgradient $g \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}}$ and a multiplier $\lambda \geq 0$ such that $g + \lambda\hat{w} = 0$. Taking the inner product with w^* gives $\lambda \langle \hat{w}, w^* \rangle = -\langle g, w^* \rangle$. Since $\langle g, w^* \rangle < 0$, we obtain $\lambda \langle \hat{w}, w^* \rangle > 0$. Because $\lambda \geq 0$, this gives

$$\langle \hat{w}, w^* \rangle > 0.$$

Moreover, $\hat{w}$ and w^* cannot be positively collinear. If $w^* = t\hat{w}$ for some $t > 0$, then $t = \|w^*\|_2 \leq 1$, since $\|\hat{w}\|_2 = 1$. Condition (49) implies $S_P \neq \emptyset$. Choose $(x_i, y_i) \in S_P$ and let $y' = h_{\hat{w}}(x)$ be the predicted label at the misclassified center (x, y). Then $y' \neq y = y_i$. The margin at the clean point and the pancake calculation in Eq. (22) give

$$\gamma \leq \left\langle w^*_{y_i} - w^*_{y'}, x_i \right\rangle = t \left\langle \hat{w}_{y_i} - \hat{w}_{y'}, x_i \right\rangle \leq 2t\tau \leq 2\tau < \gamma,$$

a contradiction. Thus $\hat{w}$ and w^* are not collinear.

The angle satisfies $\theta(\hat{w}, w^*) \in (0, \pi/2)$ and w' is well defined. Now define w' as in Eq. (35). By the orthogonal relation between $\hat{w}$ and w', $\langle w', \hat{w} \rangle = 0$. Applying the KKT condition again, $\langle g, w' \rangle = -\lambda \langle \hat{w}, w' \rangle = 0$. However, Lemma A.2 with condition in Eq. (49) implies that

$$\langle g, w' \rangle < 0 \quad \text{for any } g \in \partial_w \ell(w; \hat{S})|_{w=\hat{w}},$$

which contradicts the previous equality. Thus (x, y) cannot be misclassified. Both cases give a contradiction, so (x, y) is correctly classified by $\hat{w}$. □

B Analysis of Algorithm 1

For each $j \in [K]$, let $\overline{S}_j$ be the original draws from component $\mathcal{D}_j$. Let $S_j \subseteq S$ be the clean samples in S drawn from the j-th component of $\mathcal{D}_X$ and labeled by w^*. That is, $\cup_{j\in[K]} S_j = S_C$. Similarly, denote by $\hat{S}_j := S_j \cap \hat{S}$ the remaining clean samples after Step 5 of Algorithm 1. In addition, let $\hat{S}_C = \cup_{j\in[K]} \hat{S}_j$. Write $n = |S| = \left|\overline{S}\right|$, where $\overline{S}$ is the original independent sample from $(\mathcal{D}_X, w^*)$, and S is obtained by at most ηn corruption. Thus S_C contains unchanged samples and $|S_D| \leq \eta n$.

Proposition B.1 (Restatement of empirical dense pancake, Proposition 5.4)**.** *Consider Algorithm 1. Suppose Assumptions 1.1 and 1.2 hold with $\alpha \in [0.6/k, 1/k]$, and $\eta \leq \alpha/64$. Let S be a set of samples from EX_{nas} with size $n \geq \Omega(k^2 d \cdot \log \frac{kd}{\delta})$. With probability $1 - \delta$, the Program (9) is feasible , and every feasible output $\hat{S}$ is $(\tau, \frac{3\alpha}{20}, \epsilon)$-dense with respect to $(\mathcal{D}_X, w^*)$ with appropriate $\tau \in (0, \frac{\gamma}{4})$. In addition, $|\hat{S}| \geq \frac{|S|}{2}$ and the $\frac{3\alpha}{20}$-density is contributed by $\hat{S}_C$.*

Proof. Due to the sample-size requirement and the covering argument in Proposition C.4, with failure probability at most $\delta/4$, the following holds simultaneously for all $\|w\|_2 \leq 1$ and $j \in [K]$: for centers satisfying Eq. (56),

$$\left|\overline{S}_j \cap \mathcal{P}^{\tau}_w(x, y)\right| \geq \frac{3nq_j}{8} \geq \frac{3\alpha n}{8}.$$

For each component, these centers have probability at least $1 - 2\epsilon/k \geq 1 - \epsilon$. This concentration is on the original sample $\overline{S}$, not the subset S_C chosen by the adversary. The same sample-size requirement implies both sample-size bounds of Proposition D.2. Its feasibility failure probability is at most $\delta/2$.

Note that we adjust parameter τ based on the distribution assumption: if $\mathcal{D}_X$ is a mixture of logconcaves, $\tau = 2\sigma \cdot (\log \frac{k}{\epsilon} + 1)$; if $\mathcal{D}_X$ is a mixture of bounded covariance distributions, then $\tau = 2\sigma \cdot \sqrt{\frac{k}{\epsilon}}$.

Fix such an original sample, any permitted corruption, and any feasible pruning output. For each center counted above, at most ηn original samples are corrupted, and Lemma D.1 permits fewer than $\alpha n/8$ additional deletions. Therefore,

$$\sum_{(x_i,y_i)\in\hat{S}_C} \mathbb{1}\left\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\right\} \geq \frac{3\alpha n}{8} - \eta n - \frac{\alpha n}{8} \geq \frac{15\alpha n}{64}.$$

Since $\left|\hat{S}\right| \leq n$ and $15/64 > 3/20$, the same whole-sample density event has probability at least $1-\epsilon$ under every component. Averaging with the mixture probabilities gives

$$\begin{aligned}
&\Pr_{(x,y)\sim(\mathcal{D}_X,w^*)}\left(\frac{1}{\left|\hat{S}\right|}\sum_{(x_i,y_i)\in\hat{S}_C}\mathbb{1}\left\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\right\}\geq\frac{3\alpha}{20}\right)\\
&=\sum_{j=1}^{K} q_j \Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(\frac{1}{\left|\hat{S}\right|}\sum_{(x_i,y_i)\in\hat{S}_C}\mathbb{1}\left\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\right\}\geq\frac{3\alpha}{20}\right)\\
&\geq\sum_{j=1}^{K} q_j(1-\epsilon)=1-\epsilon.
\end{aligned}$$

As a result, we have that $\hat{S}$ is $(\tau, \frac{3\alpha}{20}, \epsilon)$-dense with respect to $(\mathcal{D}_X, w^*)$. Note that since $\cup_{j\in[K]}\hat{S}_j = \hat{S}_C$, the density lower bound $\frac{3\alpha}{20}$ is all contributed by $\hat{S}_C$.

In addition, Lemma D.1 gives

$$|\hat{S}| > (1-\alpha/8)n \geq n/2 = |S|/2.$$

The proof is complete. □

Theorem B.2 (Noise rate condition)**.** *Consider Algorithm 1. After feasible pruning, let $\hat{w}$ be a minimizer of Eq.* (8)*. Suppose Assumption 1.1 holds with some $\gamma > 0$ and Assumption 1.2 holds with σ. Let $k \geq 2$, $C \geq 4$, and $|S_D| \leq \eta n$. Assume that $\frac{\gamma}{4} > \max(\tau, C\sigma)$ and $\eta \leq \frac{1}{2^{12}k^2}$. If set $\hat{S}$ is $(\tau, \frac{3\alpha}{20}, \epsilon)$-dense with respect to $(\mathcal{D}_X, w^*)$ with appropriate $\tau \in (0, \frac{\gamma}{4})$ for some $\alpha \in [\frac{0.6}{k}, \frac{1}{k}]$, the $\frac{3\alpha}{20}$-density is contributed by $\hat{S}_C$, and $|\hat{S}| \geq \frac{|S|}{2}$, then $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$.*

Proof. If $\hat{S}$ is $(\tau, \frac{3\alpha}{20}, \epsilon)$-dense with respect to $(\mathcal{D}_X, w^*)$ with $\tau < \frac{\gamma}{4}$ and the $\frac{3\alpha}{20}$-density is contributed by $\hat{S}_C$, we know that there exists at least $1-\epsilon$ probability mass of (x, y) over the distribution $(\mathcal{D}_X, w^*)$ such that

$$|S_P| = \left|P_{\hat{w}}^\tau(x,y)\cap\hat{S}_C\right| \geq \frac{3\alpha}{20}\cdot\left|\hat{S}\right| \geq \frac{3\alpha}{40}\cdot n.$$

From the condition established in Lemma 5.3, it suffices to verify that $\frac{3\alpha}{40}\cdot n > \frac{8C\sigma n\sqrt{\eta}}{\gamma}$. By assumption, $\frac{\gamma}{4} > C\sigma$. The sufficient condition is $\frac{3\alpha}{40} > 2\sqrt{\eta}$. Consider that $\alpha \in [\frac{0.6}{k}, \frac{1}{k}]$, then it suffices to choose $\eta \leq \frac{1}{2^{12}k^2}$. In fact,

$$2\sqrt{\eta} \leq \frac{1}{32k} < \frac{9}{200k} \leq \frac{3\alpha}{40}.$$

Moreover, $\eta \leq 1/(4096k^2) \leq \alpha/64$, as required by Lemma 5.3. Lemma 5.3 therefore classifies all the indicated centers correctly, and $\mathrm{err}_{\mathcal{D}_X}(\hat{w}) \leq \epsilon$. □

Lemma B.3 (Runtime of Algorithm 1)**.** *Algorithm 1 runs in polynomial time. More precisely, its runtime is polynomial in n, d, k and the input length.*

Proof. First, Section D.3 gives the polynomial-time pruning step under its input conditions. If pruning reports failure, Algorithm 1 returns h_0. Otherwise, forming $\hat{S}$ from the returned sets takes $O(n+k)$ time.

Next, each hinge loss in Eq. (8) is the maximum of k affine functions of w. Introducing one variable for each surviving sample's loss gives $kd+\left|\hat{S}\right|$ variables, $O(k\left|\hat{S}\right|)$ linear inequalities, and the unit-ball constraint. The loss variables may be restricted to $[0, n+2]$ without changing the minimum, since $w=0$ has total loss $\left|\hat{S}\right| \le n$. Setting $w=0$ and all loss variables to 2 satisfies the constraints strictly. Thus, Eq. (8) is a convex optimization problem. Standard convex optimization methods optimize it in polynomial time (Lobo et al., 1998; Grötschel et al., 1981). Combining this with the pruning runtime proves the claim. □

C Sample Complexity

In this section, we provide a tighter analysis for the sample complexity required for the dense pancake condition for a mixture of distributions. We remark that the existing statistical analysis in Talwar (2020) assumes that the instances have bounded L_2 norm, i.e. $\|x\|_2 \le 1$. Under our bounded covariance Assumption 1.2, such a uniform norm bound does not necessarily hold. We therefore impose a distance bound only on the original sample points used in the covering argument below.

The following theorem gives a sample complexity for a mixture of distributions, provided that the sample complexity of obtaining the dense pancake condition for each mixture component is already available. Unlike the corresponding mixture result in Talwar (2020), our statement also accounts for the random number of original sample points drawn from each mixture component.

Theorem C.1 (Restatement of Theorem 5.11)**.** *Suppose that Assumption 1.2 is satisfied. Fix $\tau \ge 0$, $0 < \rho \le 1$, and $0 < \beta, \delta < 1$, and let $n_j \ge 1$ be integers. If $\forall j \in [K]$, an independent sample $\overline{S_j}$ of any size $\left|\overline{S_j}\right| \ge n_j$ from $(\mathcal{D}_j, w^*)$ satisfies (τ, ρ, β)-dense pancake condition with respect to $(\mathcal{D}_j, w^*)$ with probability at least $1-\delta/(2K)$, simultaneously for all $\|w\|_2 \le 1$, then an independent sample $\overline{S}$ from $(\mathcal{D}_X, w^*)$ of size n satisfying*

$$n \ge \max\left\{2\max_{j\in[K]}\frac{n_j}{q_j}, \frac{8}{\alpha}\log\frac{2K}{\delta}\right\} \tag{50}$$

satisfies $(\tau, \frac{\alpha\rho}{2}, \beta)$-dense pancake condition with respect to $(\mathcal{D}_X, w^*)$ with probability at least $1-\delta$, simultaneously for all $\|w\|_2 \le 1$.

Proof. Each original draw is generated by component j with probability q_j, independently across draws. The Chernoff bound (Shen, 2025, Lemma 26) and a union bound give

$$\Pr\left(\exists j \in [K] : \left|\overline{S}_j\right| < \frac{nq_j}{2}\right) \le \sum_{j=1}^{K} e^{-nq_j/8} \le Ke^{-\alpha n/8} \le \frac{\delta}{2}.$$

By (50), $nq_j/2 \ge n_j$ for every j. Given $\left|\overline{S}_j\right| = m$, these m points are independent samples from $(\mathcal{D}_j, w^*)$. The theorem's hypothesis therefore bounds the probability that $\left|\overline{S}_j\right| \ge n_j$ but $\overline{S}_j$ fails the component dense pancake condition by $\delta/(2K)$. A union bound gives total failure probability at most $\delta/2 + K\delta/(2K) = \delta$. Fix a sample for which, for every j, $\left|\overline{S}_j\right| \ge nq_j/2$ and the component dense pancake condition holds. Fix any $\|w\|_2 \le 1$.

Since for all j, $\overline{S_j}$ satisfies (τ, ρ, β)-dense pancake condition with respect to $(\mathcal{D}_j, w^*)$, by Definition 3.1 we have

$$\Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(\frac{1}{\left|\overline{S_j}\right|}\sum_{(x',y')\in\overline{S_j}} \mathbb{1}\left(y' = y \wedge \forall \bar{y} \in \mathcal{Y}, \left|\langle w_{\bar{y}}, x'-x\rangle\right| \le \tau\right) \ge \rho\right) \ge 1-\beta.$$

Now consider the mixture distribution. Since $\left|\overline{S}_j\right| \geq \alpha n/2$, a pancake containing at least $\rho\left|\overline{S}_j\right|$ points of $\overline{S}_j$ contains at least $\alpha\rho n/2$ such points. Therefore,

$$q_j \Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(\frac{1}{\left|\overline{S}\right|}\sum_{(x',y')\in\overline{S_j}}\mathbb{1}\left(y'=y\wedge\forall\bar{y}\in\mathcal{Y},\left|\left\langle w_{\bar{y}},x'-x\right\rangle\right|\leq\tau\right)\geq\frac{\alpha\rho}{2}\right)\geq q_j(1-\beta).$$

Summing these inequalities over j, using $\overline{S}_j\subseteq\overline{S}$ and $\mathcal{D}_X=\sum_j q_j\mathcal{D}_j$, gives

$$\Pr_{(x,y)\sim(\mathcal{D}_X,w^*)}\left(\frac{1}{\left|\overline{S}\right|}\sum_{(x',y')\in\overline{S}}\mathbb{1}\left(y'=y\wedge\forall\bar{y}\in\mathcal{Y},\left|\left\langle w_{\bar{y}},x'-x\right\rangle\right|\leq\tau\right)\geq\frac{\alpha\rho}{2}\right)\geq 1-\beta,$$

which means that $\overline{S}$ satisfies the $(\tau,\frac{\alpha\rho}{2},\beta)$-dense pancake condition with respect to $(\mathcal{D}_X,w^*)$. Since the component conditions hold simultaneously for all $\|w\|_2\leq 1$, so does this conclusion.

□

We show the sample complexity for the dense pancake condition for a single log-concave distribution (a single mixture component in Assumption 1.3), and then give the corresponding bounded covariance proof.

Theorem C.2 (Restatement sample complexity for a single log-concave distribution Theorem 5.9)**.** *Suppose that Assumption 1.1 is satisfied with parameter γ. Fix $0<\beta<1/k$ and $0<\delta'<1$, and take $\tau=2\sigma(\log(1/\beta)+1)\leq\gamma/2$. If distribution $\mathcal{D}_j$ is logconcave with mean μ_j and covariance $\Sigma_j\preceq\sigma^2\mathbb{I}_d$, then a set $\overline{S_j}$ of at least*

$$\frac{128}{1-k\beta}\left[kd\log\left(1+\frac{256\sqrt{d}}{(1-k\beta)^{3/2}}\right)+\log\frac{1}{\delta'}\right] \tag{51}$$

independent samples from $(\mathcal{D}_j,w^*)$ satisfies $(\tau,\frac{1-k\beta}{2},k\beta)$-dense pancake condition with respect to $(\mathcal{D}_j,w^*)$ with probability at least $1-\delta'$, simultaneously for all $\|w\|_2\leq 1$.

Proof. We first show that $(\mathcal{D}_j,w^*)$ satisfies $(2\sigma(\log\frac{1}{\beta}+1),1-k\beta,k\beta)$-dense pancake condition.

It is well known that for any vector $\bar{w}\in\mathbb{R}^d$, $\|\bar{w}\|_2\leq 1$, any random vector from a logconcave distribution $\mathcal{D}_j$ satisfies that

$$\Pr_{x\sim\mathcal{D}_j}\left(\left|\bar{w}\cdot(x-\mu_j)\right|\leq\sigma\cdot\left(\log\frac{1}{\beta}+1\right)\right)\geq 1-\beta. \tag{52}$$

This result immediately implies that $\forall w\in\mathbb{R}^{kd}$, $\|w\|_2\leq 1$

$$\Pr_{x\sim\mathcal{D}_j}\left(\forall\bar{y}\in\mathcal{Y},\left|w_{\bar{y}}\cdot(x-\mu_j)\right|\leq\sigma\cdot\left(\log\frac{1}{\beta}+1\right)\right)\geq 1-k\beta.$$

The scalar tail bound follows by applying Lovász & Vempala (2007, Theorem 5.1 and Lemma 5.17) to the centered projection, whose second moment is at most σ^2. The preceding class-wise display uses a union bound over the k class coordinates; no independence between those coordinates is needed.

We will also need the stronger bound for their joint Euclidean norm. For $\|w\|_2\leq 1$,

$$\mathbb{E}_{x\sim\mathcal{D}_j}\sum_{\bar{y}\in\mathcal{Y}}\left\langle w_{\bar{y}},x-\mu_j\right\rangle^2=\sum_{\bar{y}\in\mathcal{Y}}w_{\bar{y}}^\top\Sigma_j w_{\bar{y}}\leq\sigma^2\sum_{\bar{y}\in\mathcal{Y}}\left\|w_{\bar{y}}\right\|_2^2\leq\sigma^2. \tag{53}$$

The same calculation holds for w^*. Each score vector is a fixed linear image of $x - \mu_j$ and is logconcave on its linear span. The same norm-tail lemma therefore gives

$$\begin{aligned}\Pr_{x\sim\mathcal{D}_j}\left(\sum_{\bar{y}\in\mathcal{Y}}\left\langle w_{\bar{y}}, x-\mu_j\right\rangle^2 > \tau^2/4\right) \le \beta,\\ \Pr_{x\sim\mathcal{D}_j}\left(\sum_{\bar{y}\in\mathcal{Y}}\left\langle w^*_{\bar{y}}, x-\mu_j\right\rangle^2 > \tau^2/4\right) \le \beta.\end{aligned} \tag{54}$$

The multiplier is $1+\log(1/\beta) > 1$. A zero second moment is handled directly by Markov's inequality at the positive threshold. Thus the probability-β assertion below follows from the joint norm bound.

That means, except for a β probability mass,

$$\left|w^*_{\bar{y}}\cdot(x-\mu_j)\right| \le \tau/2, \forall \bar{y}\in\mathcal{Y} \tag{55}$$

In the following, we show that for any w^* that satisfies Assumption 1.1 with γ, a large fraction of $\mathcal{D}_j$ has the same label as its mean μ_j under the classification of w^*. Note that an x may have a different label than $y_j = \arg\max_{y\in\mathcal{Y}} w^*_y\cdot\mu_j$ under w^* only when $\exists y'\in\mathcal{Y}\setminus\{y_j\}$,

$$\begin{aligned} w^*_{y'}\cdot x &\ge w^*_{y_j}\cdot x,\\ w^*_{y'}\cdot\mu_j + w^*_{y'}\cdot(x-\mu_j) &\ge w^*_{y_j}\cdot\mu_j + w^*_{y_j}\cdot(x-\mu_j),\\ (w^*_{y_j}-w^*_{y'})\cdot\mu_j + (w^*_{y_j}-w^*_{y'})\cdot(x-\mu_j) &\le 0\end{aligned}$$

A logconcave component has convex support. Under the margin Assumption 1.1, this support lies in the union of the finitely many disjoint closed regions in which one label has margin at least γ. A connected set cannot meet two of these regions while remaining in their union. Hence the component support lies in one such region. Integrating its linear margin inequalities shows that $\left\langle w^*_{y_j} - w^*_{y'}, \mu_j\right\rangle \ge \gamma$ for every $y'\neq y_j$. This argument is only for the original margin assumption and logconcave component; it is not inferred from covariance alone. Convexity of logconcave support is described in Lovász & Vempala (2007, Section 2.1).

Since $(w^*_{y_j}-w^*_{y'})\cdot\mu_j \ge \gamma$ for any $y'\in\mathcal{Y}\setminus\{y_j\}$, it can only happen when $(w^*_{y_j}-w^*_{y'})\cdot(x-\mu_j) \le -\gamma$. However, due to Eq. (55), we know that for any $y_j, y'\in\mathcal{Y}$

$$\left|(w^*_{y_j}-w^*_{y'})\cdot(x-\mu_j)\right| \le \left|w^*_{y_j}\cdot(x-\mu_j)\right| + \left|w^*_{y'}\cdot(x-\mu_j)\right| \le \tau.$$

Hence, $\forall y'\in\mathcal{Y}\setminus\{y_j\}$, $(w^*_{y_j}-w^*_{y'})\cdot(x-\mu_j) \ge -\tau > -\gamma$. That is, $w^*_{y'}\cdot x \ge w^*_{y_j}\cdot x$ cannot happen and $\arg\max_{y\in\mathcal{Y}} w^*_y\cdot x = y_j$. Under the margin Assumption 1.1, the support argument above already gives the same label throughout this component's support. Thus intersecting with the preceding class-wise projection event for w does not consume a second probability allowance. More formally,

$$\Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(y=y_j\wedge\forall\bar{y}\in\mathcal{Y}, \left|\left\langle w_{\bar{y}}, x-\mu_j\right\rangle\right| \le \tau/2\right) \ge 1-k\beta.$$

For later use under the sample only margin premise, we record a label comparison that does not use a margin at μ_j. If (x_i, y_i) satisfies the margin inequalities and $\sum_{\bar{y}}\left\langle w^*_{\bar{y}}, x_i-\mu_j\right\rangle^2 \le \tau^2/4$, then any point (x, y) labeled by w^* and satisfying the same projection inequality obeys, for every $\tilde{y}\neq y_i$,

$$\begin{aligned}\left\langle w^*_{y_i}-w^*_{\tilde{y}}, x\right\rangle &\ge \left\langle w^*_{y_i}-w^*_{\tilde{y}}, x_i\right\rangle - \sqrt{2}\left[\left(\sum_{\bar{y}}\left\langle w^*_{\bar{y}}, x_i-\mu_j\right\rangle^2\right)^{1/2} + \left(\sum_{\bar{y}}\left\langle w^*_{\bar{y}}, x-\mu_j\right\rangle^2\right)^{1/2}\right]\\ &\ge \gamma - \sqrt{2}\tau > 0.\end{aligned}$$

Thus $y = y_i$. This comparison uses the margin at x_i, not at the evaluated point x or the mean.

For fixed w, let $S^{\text{good}}(w)$ consist of points satisfying

$$\sum_{\bar{y}\in\mathcal{Y}} \langle w_{\bar{y}}, x - \mu_j\rangle^2 \le \tau^2/4, \qquad \sum_{\bar{y}\in\mathcal{Y}} \left\langle w^*_{\bar{y}}, x - \mu_j\right\rangle^2 \le \tau^2/4. \tag{56}$$

The two joint norm bounds give

$$(\mathcal{D}_j, w^*)\big(S^{\text{good}}(w)\big) \ge 1 - 2\beta \ge 1 - k\beta. \tag{57}$$

Under the original margin assumption, these points have label y_j by the mean comparison above. This proves the following original population conclusion. Under a sample only margin premise, the same label conclusion instead follows once a point x_i as above is supplied by the nonempty counted sample set.

Moreover,

$$\Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(\Pr_{(x',y')\sim(\mathcal{D}_j,w^*)}\left(y' = y \wedge \forall \bar{y}\in\mathcal{Y}, \left|\langle w_{\bar{y}}, x' - x\rangle\right| \le \tau\right) \ge 1 - k\beta\right) \ge 1 - k\beta. \tag{58}$$

For the last display, two points in $S^{\text{good}}(w)$ have equal labels and, for every $\bar{y}$,

$$\left|\langle w_{\bar{y}}, x' - x\rangle\right| \le \left|\langle w_{\bar{y}}, x' - \mu_j\rangle\right| + \left|\langle w_{\bar{y}}, x - \mu_j\rangle\right| \le \tau.$$

Thus a pancake around any point of this set contains the entire set. This is a population statement for each fixed w; it is not yet the empirical sampling event.

We then show the sample complexity for obtaining an empirical set that satisfies a dense pancake condition for $(\mathcal{D}_j, w^*)$. For ease of presentation, let $\rho = 1 - k\beta$ and $\beta' = k\beta$, then for any $w \in \mathbb{R}^{kd}$, $\|w\|_2 \le 1$,

$$\Pr_{(x,y)\sim(\mathcal{D}_j,w^*)}\left(\mathcal{P}^\tau_w(x,y) \text{ is } \rho\text{-dense w.r.t.}(\mathcal{D}_j, w^*)\right) \ge 1 - \beta',$$

due to Definition 3.1.

From now on, we use $\mathcal{D}$ instead of $(\mathcal{D}_j, w^*)$ to further ease the notation. The following part follows the idea of Talwar (2020), but we include it for clarity and completeness. Now, fix a w. For any $(x, y) \in S^{\text{good}}(w)$, apply Chernoff bound (Lemma E.2),

where $n = \left|\overline{S}_j\right|$ in this component calculation. The set $S^{\text{good}}(w)$ is defined by (56); every one of its points has a population pancake of probability at least ρ under the original margin assumption.

$$\Pr_{\overline{S}_j\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left((x_i, y_i) \in \mathcal{P}^\tau_w(x,y)\right) < \rho/2\right) \le \exp\left(-\frac{\rho n}{8}\right).$$

Then,

$$\mathbb{E}_{\overline{S}_j\sim\mathcal{D}^n}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left\{(x,y) \in S^{\text{good}}(w) \ \wedge \ \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{(x_i, y_i) \in \mathcal{P}^\tau_w(x,y)\} < \rho/2\right\}\right]\right] \le e^{-\rho n/8}.$$

Equivalently, exchanging the order of expectation over the independent sample and evaluated point gives

$$\mathbb{E}_{\overline{S}_j\sim\mathcal{D}^n,(x,y)\sim\mathcal{D}}\left[\mathbb{1}\left\{(x,y) \in S^{\text{good}}(w) \ \wedge \ \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{(x_i, y_i) \in \mathcal{P}^\tau_w(x,y)\} < \rho/2\right\}\right] \le e^{-\rho n/8},$$

which gives

$$\mathbb{E}_{\overline{S}_j\sim\mathcal{D}^n}\left[\Pr_{(x,y)\sim\mathcal{D}}\left((x,y)\in S^{\text{good}}(w)\ \wedge\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\}<\rho/2\right)\right]\le e^{-\rho n/8}.$$

By taking Markov's inequality, we have

$$\Pr_{\overline{S}_j\sim\mathcal{D}^n}\left[\Pr_{(x,y)\sim\mathcal{D}}\left((x,y)\in S^{\text{good}}(w)\ \wedge\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\}<\rho/2\right)>\beta'\right]\le\frac{1}{\beta'}e^{-\rho n/8}.$$

Together with the points in the complement set of $S^{\text{good}}(w)$, which contributes at most another β' probability mass, we have that

$$\Pr_{\overline{S}_j\sim\mathcal{D}^n}\left[\Pr_{(x,y)\sim\mathcal{D}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{(x_i,y_i)\in\mathcal{P}_w^\tau(x,y)\}<\rho/2\right)>2\beta'\right]\le\frac{1}{\beta'}e^{-\rho n/8}.$$

To obtain the stated $k\beta$ exception bound simultaneously for all $\|w\|_2\le 1$. It remains to show that the empirical conclusion holds simultaneously for every $w\in\mathbb{R}^{kd}$, $\|w\|_2\le 1$. The following covering calculation is valid for any $0<\rho\le 1-2\beta$. The component theorem uses $\rho=1-k\beta$; the Proposition C.4 will use $\rho=1/2$ with $\beta=\epsilon/k$. No use of β' is made in this covering calculation.

For the covering net, set

$$b=8\sigma\sqrt{d/\rho},\qquad \tau'=\frac{\sigma\rho}{16},\qquad \frac{\tau'}{b}=\frac{\rho^{3/2}}{128\sqrt{d}}. \tag{59}$$

The covariance bound and Markov's inequality give

$$\Pr_{x\sim\mathcal{D}_j}\left(\|x-\mu_j\|_2>b/2\right)\le\frac{d\sigma^2}{b^2/4}=\frac{\rho}{16}.$$

This bound is used only for the sample points counted below.

It suffices to construct a $\frac{\tau'}{b}$-net over the unit ball. Let $\mathcal{N}$ be such a net in $\mathbb{R}^{kd}$ with its points inside the ball. Then

$$|\mathcal{N}|\le\left(1+\frac{2b}{\tau'}\right)^{kd}=\left(1+\frac{256\sqrt{d}}{\rho^{3/2}}\right)^{kd}. \tag{60}$$

To see this, choose a maximal set whose pairwise distances exceed τ'/b. Balls of radius $\tau'/(2b)$ around its points are disjoint and lie inside the ball of radius $1+\tau'/(2b)$. Comparing their kd-dimensional volumes gives (60); maximality gives the covering property.

For each $w'\in\mathcal{N}$, let $\overline{S}_j(w')$ contain the original component sample points satisfying

$$\begin{aligned}&\sum_{\bar y\in\mathcal{Y}}\left\langle w^*_{\bar y},x-\mu_j\right\rangle^2\le\tau^2/4,\qquad \|x-\mu_j\|_2\le b/2,\\ &\left[\sum_{\bar y\in\mathcal{Y}}\left\langle w'_{\bar y},x-\mu_j\right\rangle^2\right]^{1/2}\le(\tau-\tau')/2.\end{aligned} \tag{61}$$

The probability that the last inequality fails is at most

$$e^{\,1-(1+\log(1/\beta)-\rho/32)}=\beta e^{\rho/32}\le\beta+\rho/16.$$

Indeed, the multiplier is greater than one, and $e^s\le(1-s)^{-1}\le 1+2s$ for $0\le s\le 1/2$, applied with $s=\rho/32$, proves the last bound. The remaining two conditions fail with probabilities at most β and $\rho/16$. Consequently, a component draw satisfies (61) with probability at least

$$1-\beta-\frac{\rho}{16}-\left(\beta+\frac{\rho}{16}\right)\ge\frac{7\rho}{8}. \tag{62}$$

For each fixed w', the count uses independent original draws. The lower-tail Chernoff bound (Shen, 2025, Lemma 26), with relative decrease $1-(3/4)/(7/8)=1/7$, gives

$$\Pr\left(\left|\overline{S}_j(w')\right| < 3n\rho/4\right) \le \exp\left[-\frac{(1/7)^2}{2}\frac{7n\rho}{8}\right] = e^{-n\rho/112} \le e^{-n\rho/128}.$$

A union bound over (60) therefore gives

$$\Pr\left(\exists w' \in \mathcal{N} : \left|\overline{S}_j(w')\right| < 3n\rho/4\right) \le \exp\left[kd\log\left(1+\frac{256\sqrt{d}}{\rho^{3/2}}\right) - \frac{n\rho}{128}\right]. \tag{63}$$

The right side is at most δ' whenever

$$n \ge \frac{128}{\rho}\left[kd\log\left(1+\frac{256\sqrt{d}}{\rho^{3/2}}\right) + \log\frac{1}{\delta'}\right]. \tag{64}$$

On this sample event, suppose (1) holds on $\overline{S}_j$. Fix any $\|w\|_2 \le 1$ and any $(x,y) \in S^{\mathrm{good}}(w)$. Choose $w' \in \mathcal{N}$ with $\|w-w'\|_2 \le \tau'/b$. For every counted point $(x',y') \in \overline{S}_j(w')$ and every $\bar{y} \in \mathcal{Y}$,

$$\begin{aligned}\left|\langle w_{\bar{y}}, x'-\mu_j\rangle\right| &\le \left|\left\langle w'_{\bar{y}}, x'-\mu_j\right\rangle\right| + \left\|w_{\bar{y}}-w'_{\bar{y}}\right\|_2 \left\|x'-\mu_j\right\|_2 \\ &\le \frac{\tau-\tau'}{2} + \frac{\tau'}{b}\frac{b}{2} = \frac{\tau}{2}.\end{aligned}$$

The count $\left|\overline{S}_j(w')\right| \ge 3n\rho/4 > 0$ supplies a sample point (x_i,y_i) satisfying the margin and the w^* inequality in (56). The preceding label comparison applies to both the center and every counted point, so $y'=y$. The center's own centered inner products are at most $\tau/2$. Their difference is therefore at most τ, and

$$\overline{S}_j(w') \subseteq \overline{S}_j \cap \mathcal{P}_w^{\tau}(x,y), \qquad \left|\overline{S}_j \cap \mathcal{P}_w^{\tau}(x,y)\right| \ge 3n\rho/4 \ge n\rho/2. \tag{65}$$

This event is simultaneous in w and in the indicated centers. By (57), those centers have probability at least $1-k\beta$. Substituting $\rho = 1-k\beta$ into (64) proves (51) and the stated dense-pancake condition at width τ.

The count event just constructed uses only the original independent draws and the fixed projection and distance tests. It does not condition those draws on satisfying the margin. On that event, the sample margins supply the point used in the label comparison, proving the more general sample level implication. Under Assumption 1.1 as printed, the original sample margins hold by that assumption, so the original unconditional sample confidence statement follows as well.

□

The proof for the sample complexity of dense pancake condition for a single distribution with bounded covariance is similar to the above, but with $\tau = 2\sigma\cdot\frac{1}{\sqrt{\beta}}$. We give the changed probability estimates below. The same b, τ', and net yield the same sufficient sample bound. To see that $\tau = 2\sigma\cdot\frac{1}{\sqrt{\beta}}$, note that for $\|\bar{w}\|_2 \le 1$ and $t>0$,

$$\Pr_{x\sim\mathcal{D}_j}\left(\left|(x-\mu_j)\cdot\bar{w}\right| \ge \sqrt{t}\right) \le \frac{\sigma^2}{t}.$$

Let the right hand side be β and we obtain that $t=\sigma^2/\beta$ and $\sqrt{t}=\sigma/\sqrt{\beta}$.

Theorem C.3 (Restatement of sample complexity for a single bounded covariance distribution Theorem 5.10)**.** *Suppose that Assumption 1.1 is satisfied with parameter γ. If $\mathcal{D}_j$ is a distribution with mean μ_j and covariance $\Sigma_j \preceq \sigma^2\mathbb{I}_d$, let $k\ge 2$, $0<\beta<1/k$, $0<\delta<1$, and $\tau = 2\sigma/\sqrt{\beta} \le \gamma/2$. Then a set $\overline{S}_j$ of at least the size in Eq.* (13)*, drawn independently from $(\mathcal{D}_j, w^*)$, satisfies the $(\tau,(1-k\beta)/2,k\beta)$-dense pancake condition with respect to $(\mathcal{D}_j,w^*)$ with probability at least $1-\delta$.*

Proof. For this component theorem take $\rho = 1 - k\beta$. As above, the following covering estimates also hold for every $0 < \rho \le 1 - 2\beta$. The second moment calculation (53) requires only the covariance and coefficient-norm bounds. Markov's inequality gives

$$\Pr\left(\sum_{\bar{y}} \langle w_{\bar{y}}, x - \mu_j \rangle^2 > \tau^2/4\right) \le 4\sigma^2/\tau^2 = \beta,$$

$$\Pr\left(\sum_{\bar{y}} \left\langle w^*_{\bar{y}}, x - \mu_j \right\rangle^2 > \tau^2/4\right) \le 4\sigma^2/\tau^2 = \beta.$$

Thus (57) remains valid. We do not claim a margin at μ_j for a bounded covariance component: its support need not be convex. Instead use the comparison with a margin satisfying point established in the preceding proof. Under the original margin assumption such a point satisfying the target projection bound exists because the latter set has probability at least $1 - \beta > 0$. For the sample-only form the nonempty counted sample set supplies it.

Use the same b, τ', and net from (59)–(60). The distance test still fails with probability at most $\rho/16$. The contracted score test fails with probability at most

$$\frac{\sigma^2}{((\tau - \tau')/2)^2} = \frac{\beta}{(1 - \rho\sqrt{\beta}/32)^2} \le \beta + \frac{\rho\beta^{3/2}}{8} \le \beta + \frac{\rho}{16}.$$

Indeed, $(1-u)^{-2} \le 1 + 4u$ for $0 \le u \le 1/4$, because $(1+4u)(1-u)^2 - 1 = u(2 - 7u + 4u^2) \ge 0$; apply this with $u = \rho\sqrt{\beta}/32$. Also $\beta < 1/k \le 1/2$ implies $\beta^{3/2} < 1/2$.

Combining the three test failures gives (62), unchanged. For each fixed net vector the original-draw indicators are independent, so the same Chernoff bound and union bound give (63)–(64). If the indicated sample margins hold, a counted point supplies the label comparison. The same perturbation estimate gives (65). The center probability is at least $1 - k\beta$, and the empirical density is at least $(1 - k\beta)/2$. □

The following proposition applies the covering calculation to the original sample and uses the feasibility and retention results in Section D to establish the density contributed by clean points in $\widehat{S}$.

Proposition C.4 (Sample size for surviving dense pancakes)**.** *Suppose Assumptions 1.1 and 1.2 hold, with $k \ge 2$, $K, d, n \ge 1$, $\sigma > 0$, $C \ge 4$, $0 < \delta < 1$, $\alpha \in [0.6/k, 1/k]$, and $0 < \epsilon < 1/2$. Let $\overline{S}$ contain n independent original draws, as in Section D. Set*

$$\tau = \begin{cases} 2\sigma\sqrt{k/\epsilon}, & \textit{bounded covariance components}, \\ 2\sigma\big(1 + \log(k/\epsilon)\big), & \textit{logconcave components}, \end{cases} \qquad \gamma > 4\tau. \tag{66}$$

Suppose

$$\eta \le \frac{\alpha}{64}. \tag{67}$$

It suffices to take $n \ge \Omega\Big(k^2 d \log(d) + kd \log \frac{Kd}{\delta}\Big)$. With probability at least $1 - \delta$ over $\overline{S}$, after every permitted corruption, the pruning program is feasible, and every feasible output $\widehat{S}$ satisfies

$$\left|\widehat{S}\right| > (1 - \alpha/8)n \ge n/2, \qquad \widehat{S} \textit{ is } (\tau, 3\alpha/20, \epsilon)\textit{-dense} \tag{68}$$

with respect to $(\mathcal{D}_X, w^*)$. The density is supplied by $\widehat{S}_C$, simultaneously for all $\|w\|_2 \le 1$.

Proof. For the component probability calculations, use $\beta = \epsilon/k$. The points in (56) have component probability at least

$$1 - \frac{2\epsilon}{k} \ge 1 - \epsilon > \frac{1}{2}.$$

Thus the covering calculation applies with $\rho = 1/2$ to every component. Averaging over the normalized component probabilities, for every w, at least $1 - \epsilon$ of the mixture's labeled draws satisfy (56) for their generating component.

For a fixed component j and net vector w', each original mixture draw is generated by j and satisfies (61) with probability at least $7q_j/16$, by (62) at $\rho = 1/2$. Applying Chernoff to these n independent indicators gives

$$\Pr\left(\left|\overline{S}_j(w')\right| < 3nq_j/8\right) \le e^{-nq_j/224} \le e^{-n\alpha/256}.$$

The same net works for all components, since b and τ' in (59) do not depend on j. A union bound over the K components and the net therefore bounds the count-failure probability by

$$K\big(1 + 512\sqrt{2d}\big)^{kd} e^{-n\alpha/256} \le \delta/4,$$

using the first term of (C.4). On this event, every $\overline{S}_j(w')$ contains at least $3nq_j/8 \ge 3\alpha n/8$ original sample points satisfying (61).

In the bounded covariance case, $\epsilon < 1/2 < \alpha k$ and Eq. (66) give

$$\gamma > 4\tau = 8\sigma\sqrt{k/\epsilon} > 8\sigma/\sqrt{\alpha}.$$

In the logconcave case, Proposition D.2 requires no additional numerical lower bound on γ.

The first term of Eq. (C.4) implies

$$n \ge \frac{128}{\alpha} \log \frac{4(K+1)}{\delta}.$$

For R chosen in Proposition D.2, its last term also implies

$$n \ge \frac{R}{2\sigma^2} \log \frac{4d}{\delta}.$$

Indeed, $R/(2\sigma^2) = 64d/\alpha$ in the bounded covariance case, while in the logconcave case

$$\frac{R}{2\sigma^2} = \frac{d}{2}\big(1 + \log(128/\alpha)\big)^2 \le \frac{384d}{\alpha}.$$

Proposition D.2 therefore gives feasibility with probability at least $1-\delta/2$. Every permitted corruption satisfies

$$\frac{|S_D|}{n} \le \eta \le \frac{\alpha}{64}.$$

For each such corruption for which Assumption 1.1 holds on S_C, Proposition D.2 therefore proves feasibility.

Fix both successful sample events and a permitted corruption for which (1) holds on S_C. Fix a component j and a center satisfying (56) for that component, and choose a net vector w' within τ'/b of w. There is an unchanged point $(x_i, y_i) \in \overline{S}_j(w') \cap S_C$, because

$$\left|\overline{S}_j(w') \cap S_C\right| \ge \frac{3nq_j}{8} - \eta n \ge \frac{23\alpha n}{64} > 0.$$

This point satisfies the margin condition and the w^* inequality in (56). The label comparison in the component proof shows that the center and every point of $\overline{S}_j(w')$ have label y_i. The net comparison used for (65) bounds their score differences by τ. Thus the center's pancake contains $\overline{S}_j(w')$. At most ηn corruptions prevent these original points from belonging to S_C. Proposition D.1 bounds the additional deletions by $\alpha n/8$. Hence

$$\left|\widehat{S}_C \cap \mathcal{P}_w^\tau(x,y)\right| \ge \left|\overline{S}_j(w')\right| - \eta n - \frac{\alpha n}{8} \ge \frac{3\alpha n}{8} - \frac{\alpha n}{64} - \frac{\alpha n}{8} = \frac{15\alpha n}{64}. \tag{69}$$

This counts only unchanged surviving points and uses $|S_D| \le \eta n$. Since $\left|\widehat{S}\right| \le n$ and $15/64 > 3/20$, the indicated centers have the claimed density in $\widehat{S}$. Their mixture probability is at least $1 - \epsilon$. The net event is uniform in w, and the retention inequalities hold for every permitted corruption satisfying the margin condition on S_C and every feasible output. The two failure probabilities total at most $\delta/4 + \delta/2 = 3\delta/4 \le \delta$. □

These statements establish density and feasibility under Assumption 1.1. The classification conclusion additionally requires the separate pointwise loss bound.

D Analysis of Algorithm 2

We prove that the pruning program has a feasible solution under Assumption 1.1 when the number of corruptions is sufficiently small. We also prove that every feasible solution removes fewer than $\alpha n/8$ sample points.

The pruning program. Algorithm 2 finds weights satisfying Eq. (9). Pairs with $i = i'$ are allowed and contribute zero to the matrix sum. The algorithm retains exactly the points with $a_i \geq 1/2$:

$$\widehat{S} = \{(x_i, y_i) \in S : a_i \geq 1/2\}.$$

points are grouped by their received labels. The coefficients w^* and the parameter γ are not inputs to this program.

D.1 Feasibility and retention

Proposition D.1 (Restatement of Proposition 5.7)**.** *Every feasible output satisfies*

$$\left|S \setminus \widehat{S}\right| < \alpha n/8. \tag{70}$$

Consequently,

$$\left|S_C \setminus \widehat{S}_C\right| < \alpha n/8. \tag{71}$$

In particular $\left|\widehat{S}\right| > (1 - \alpha/8)n \geq n/2$.

Proof. The total-weight constraint gives $\sum_i (1 - a_i) \leq \alpha n/16$. Every deleted point has $1 - a_i > 1/2$. If any point is deleted, its deletion count is therefore less than $2(\alpha n/16) = \alpha n/8$; if none is deleted the same strict inequality holds since $\alpha n > 0$. The set $S_C \setminus \widehat{S}_C$ is a subset of all deleted points. Subtracting the total deletion count from n gives the size bound. □

Proposition D.2 (Program feasibility)**.** *Suppose Assumptions 1.1 and 1.2 hold. Let $\eta \leq \alpha/64$ and $C \geq 4$, and take either*

$$R = 128 d\sigma^2/\alpha, \qquad \text{if } \gamma > 8\sigma/\sqrt{\alpha};$$
$$R = d\sigma^2(1 + \log(128/\alpha))^2, \quad \text{if further Assumption 1.3 holds}.$$

With probability at least $1 - (K+1)e^{-\alpha n/128} - de^{-2n\sigma^2/R}$ over n independent clean draws $\overline{S}$, Program (9) *is feasible under an η-fraction of nasty corruptions. Then, a sample size*

$$n \geq \max\left\{\frac{128}{\alpha}\log\frac{4(K+1)}{\delta}, \quad \frac{R}{2\sigma^2}\log\frac{4d}{\delta}\right\}$$

suffices to guarantee that with probability at least $1 - \delta/2$, Program (9) *is feasible under an η-fraction of nasty corruptions.*

Proof. Recalling the feature-map definition in Section 3 and Assumption 1.1, $\left\langle w^*, \Psi(x, y)\right\rangle = \left\langle w_y^*, x\right\rangle$ and $\sum_{y\in\mathcal{Y}} \left\|w_y^*\right\|_2^2 \leq 1$. We first choose a label y_j for each component so that a clean draw from that component is excluded by the program for either $y \neq y_j$ or $\left\|x - \mu_j\right\|_2^2 > R$ with probability at most $3\alpha/128$.

Bounded-covariance case. Fix component j and choose its most probable clean label y_j. Let (X, Y) and (X', Y') be independent clean draws from this component; conditioning on j below refers to this sampling law. The independence of two draws gives

$$\begin{aligned}
\Pr(Y \neq y_j \mid j) &= 1 - \Pr(Y = y_j \mid j) \\
&= 1 - \Pr(Y = y_j \mid j) \sum_{y\in\mathcal{Y}} \Pr(Y' = y \mid j) \\
&\leq 1 - \sum_{y\in\mathcal{Y}} \Pr(Y = y,\ Y' = y \mid j) \\
&\leq 1 - \sum_{y\in\mathcal{Y}} \Pr(Y = y \mid j)^2 = \Pr(Y \neq Y' \mid j),
\end{aligned}$$

where the second transition applies the law of total probability; the third transition uses the fact that y_j is the most probable label, i.e. $\Pr(Y = y_j \mid j) = \max_{y \in \mathcal{Y}} \Pr(Y = y \mid j)$. If $Y \neq Y'$, applying the margin condition at both draws gives

$$\langle w_Y^* - w_{Y'}^*, X \rangle \geq \gamma,$$
$$\langle w_{Y'}^* - w_Y^*, X' \rangle \geq \gamma.$$

Combining these inequalities gives

$$\langle w_Y^* - w_{Y'}^*, X - X' \rangle \geq 2\gamma > 0.$$

We may therefore square both sides. Using $(u - v)^2 \leq 2u^2 + 2v^2$ and the fact that $Y \neq Y'$, we obtain

$$\begin{aligned} 4\gamma^2 &\leq \langle w_Y^* - w_{Y'}^*, X - X' \rangle^2 \\ &\leq 2 \langle w_Y^*, X - X' \rangle^2 + 2 \langle w_{Y'}^*, X - X' \rangle^2 \\ &\leq 2 \sum_{y \in \mathcal{Y}} \left\langle w_y^*, X - X' \right\rangle^2. \end{aligned}$$

Consequently, for all outcomes,

$$2\gamma^2 \, \mathbf{1}_{\{Y \neq Y'\}} \leq \sum_{y \in \mathcal{Y}} \left\langle w_y^*, X - X' \right\rangle^2.$$

Since X, X' are independent with mean μ_j and covariance Σ_j, $\mathbb{E}[(X - X')(X - X')^\top \mid j] = 2\Sigma_j$. Taking expectations and applying the covariance bound gives

$$\begin{aligned} 2\gamma^2 \Pr(Y \neq Y' \mid j) &\leq \sum_{y \in \mathcal{Y}} \mathbb{E}\left[\left\langle w_y^*, X - X' \right\rangle^2 \,\middle|\, j \right] \\ &= 2 \sum_{y \in \mathcal{Y}} (w_y^*)^\top \Sigma_j w_y^* \\ &\leq 2\sigma^2 \sum_{y \in \mathcal{Y}} \left\| w_y^* \right\|_2^2 \leq 2\sigma^2. \end{aligned}$$

Dividing by $2\gamma^2$ on both sides yields

$$\Pr(Y \neq Y' \mid j) \leq \frac{\sigma^2}{\gamma^2} < \frac{\alpha}{64}.$$

Separately, $\mathbb{E}[\|X - \mu_j\|_2^2 \mid j] = \operatorname{tr}(\Sigma_j) \leq d\sigma^2$, so Markov's inequality gives

$$\Pr(\|X - \mu_j\|_2^2 > R \mid j) \leq \frac{d\sigma^2}{R} = \frac{\alpha}{128},$$

by choosing a threshold $R = 128 d\sigma^2/\alpha$. A union bound over the events of distinct labels and large distance gives an exclusion probability of $3\alpha/128$.

Logconcave case. If Assumption 1.3 also holds, the support of each $\mathcal{D}_j$ is convex and hence connected. The hard margin gives a unique maximizing label at every support point. Continuity of the linear scores makes this label locally constant, hence constant on the support; denote it by y_j. Thus no points are excluded because of their labels, and no additional quantitative lower bound on γ beyond positivity is needed. The logconcave norm-tail bound (Lemma E.4), together with $\mathbb{E}\|X - \mu_j\|_2^2 \leq d\sigma^2$, gives an exclusion probability

$$\Pr(\|X - \mu_j\|_2^2 > R \mid j) \leq \exp\left(1 - \sqrt{\frac{R}{d\sigma^2}}\right) = \frac{\alpha}{128},$$

by choosing a threshold $R = d\sigma^2(1 + \log(128/\alpha))^2$.

Feasible weights construction. Let $\overline{S}_j$ be the original points generated by component j, and let $\overline{S}'_j$ retain those with label y_j and $\left\|x_i - \mu_j\right\|_2^2 \le R$. Applying Chernoff bound (Lemma E.2) shows that

$$n - \sum_j \left|\overline{S}'_j\right| \le \frac{3\alpha n}{64}, \text{ and } \left|\overline{S}_j\right| \ge \frac{\alpha n}{2} \text{ for every } j,$$

except with probability at most $e^{-\alpha n/128} + Ke^{-\alpha n/8} \le (K+1)e^{-\alpha n/128}$. This combines the two scalar failure probabilities into the first term of the stated bound.

Moreover, let matrix

$$M = \sum_j \left(\sum_{i \in \overline{S}'_j} (x_i - \mu_j)(x_i - \mu_j)^\top + \sum_{i \in \overline{S}_j \setminus \overline{S}'_j} \mathbf{0}_{d\times d} \right)$$

be a sum of n independent positive semidefinite matrices bounded by $R\mathbb{I}_d$, with $\mathbb{E}M \preceq n\sigma^2\mathbb{I}_d$. Note that replacing samples in $\overline{S}_j \setminus \overline{S}'_j$ by zero does not void the independence and only decreases the expectation as any $(x_i - \mu_j)(x_i - \mu_j)^\top$ is positive semi-definite. Matrix Chernoff (Lemma E.3) yields

$$\Pr(M \not\preceq 4n\sigma^2\mathbb{I}_d) \le d \exp\left(-\frac{(4\log 4 - 3)n\sigma^2}{R}\right) \le de^{-2n\sigma^2/R}.$$

On these events, fix any nasty noise corruption and put $S'_j = \overline{S}'_j \cap S_C$. At most ηn original points were removed, so

$$\sum_j \left|S'_j\right| \ge (1 - \alpha/16)n, \qquad \left|S'_j\right| \ge \frac{\alpha n}{2} - \frac{3\alpha n}{64} - \eta n \ge \frac{7\alpha n}{16} > \frac{\alpha n}{4}.$$

Set $a_{ii'} = 1/\left|S'_j\right|$ when $i, i' \in S'_j$ for some j, and $a_{ii'} = 0$ otherwise. Its row sums a_i are 1 on $\bigcup_j S'_j$ and 0 elsewhere, including every injected point. In particular, for $i, i' \in S'_j$,

$$a_{ii'} = \frac{1}{\left|S'_j\right|} \le \frac{4}{\alpha n} = \frac{4a_i}{\alpha n},$$

and the pair-weight bound also holds when $a_{ii'} = 0$. The other scalar and same-label constraints follow directly from the construction. Writing $\mu_{S'_j}$ for the empirical mean of S'_j, the pairwise-variance identity gives

$$\begin{aligned}\frac{1}{2}\sum_{i,i'} a_{ii'}(x_i - x_{i'})(x_i - x_{i'})^\top &= \sum_j \sum_{i \in S'_j} (x_i - \mu_{S'_j})(x_i - \mu_{S'_j})^\top \\ &\preceq M \preceq 4n\sigma^2\mathbb{I}_d \preceq \frac{C^2\sigma^2 n}{4}\mathbb{I}_d.\end{aligned}$$

This proves feasibility. All concentration events concern only $\overline{S}$, so the conclusion holds simultaneously for every permitted nasty noise corruption.

As a result, it suffices to choose

$$(K+1)e^{-\alpha n/128} \le \frac{\delta}{4}, \text{ and } de^{-2n\sigma^2/R} \le \frac{\delta}{4},$$

which yields the sample bound as stated, i.e.

$$n \ge \max\Big(\frac{128}{\alpha}\log\frac{4(K+1)}{\delta}, \frac{R}{2\sigma^2}\log\frac{4d}{\delta}\Big).$$

Specifically, if consider $\alpha = \Theta(\frac{1}{K})$, then for the bounded covariance condition, it suffices to choose $n \ge \Omega\Big(K\log\frac{K}{\delta} + Kd\cdot\log\frac{d}{\delta}\Big)$; and for the logconcave condition, it suffices to choose $n \ge \Omega\Big(K\log\frac{K}{\delta} + d\log^2 K\cdot\log\frac{d}{\delta}\Big)$. □

The feasibility guarantee requires Assumption 1.1 in addition to Assumption 1.2.

D.2 BOUNDS FOR THE SURVIVING CORRUPTED SAMPLE POINTS

Proposition D.3 (Restatement of Proposition 5.8)**.** *Suppose Assumption 1.1 holds. Assume that* $|S_D| \le \eta n$*, and* $4\eta/\alpha < 1$*. For any feasible solution and surviving corrupted sample* $(x_i, y_i) \in \widehat{S}_D$*, the weighted average of its clean paired sample points is well defined:*

$$\mu_{i,\mathrm{C}} = \frac{\sum_{i' \in S_C} a_{ii'} x_{i'}}{\sum_{i' \in S_C} a_{ii'}}. \tag{72}$$

It satisfies

$$\left\langle w^*_{y_i} - w^*_{\tilde{y}}, \mu_{i,\mathrm{C}} \right\rangle \ge \gamma \quad \textit{for every } \tilde{y} \in \mathcal{Y} \setminus \{y_i\}, \tag{73}$$

and the surviving corrupted points satisfy

$$\sum_{i \in \widehat{S}_D} (x_i - \mu_{i,\mathrm{C}})(x_i - \mu_{i,\mathrm{C}})^\top \preceq \frac{C^2\sigma^2 n}{2(1 - 4\eta/\alpha)} \mathbb{I}_d. \tag{74}$$

In particular, the right side is at most $C^2\sigma^2 n \mathbb{I}_d$ *if* $4\eta/\alpha \le 1/2$*. These are deterministic consequences of the program and the clean-point margins; logconcavity is not needed.*

Proof. **Step 1: the average exists and preserves the label margin.** The total weight from the sample indexed by i to corrupted samples is at most their number times the largest allowed pair weight:

$$\begin{aligned} \sum_{i' \in S_D} a_{ii'} &\le |S_D| \frac{4a_i}{\alpha n} \le \frac{4\eta}{\alpha} a_i, \\ \sum_{i' \in S_C} a_{ii'} &= a_i - \sum_{i' \in S_D} a_{ii'} \ge \left(1 - \frac{4\eta}{\alpha}\right) a_i \ge \frac{1}{2}\left(1 - \frac{4\eta}{\alpha}\right) > 0. \end{aligned} \tag{75}$$

The last step uses the retention threshold $a_i \ge 1/2$. Every clean sample indexed by i' with positive $a_{ii'}$ has label y_i. For each $\tilde{y} \ne y_i$, multiply the corresponding clean-point margin inequality by $a_{ii'}$, sum over S_C, and divide by the positive denominator:

$$\left\langle w^*_{y_i} - w^*_{\tilde{y}}, \mu_{i,\mathrm{C}} \right\rangle = \frac{\sum_{i' \in S_C} a_{ii'} \left\langle w^*_{y_i} - w^*_{\tilde{y}}, x_{i'} \right\rangle}{\sum_{i' \in S_C} a_{ii'}} \ge \gamma.$$

We remark that the algorithm does not compute $\mu_{i,\mathrm{C}}$, which need not belong to S or $\widehat{S}$. In addition, the contributing points to $\mu_{i,\mathrm{C}}$ need not survive pruning.

Step 2: the pairwise bound controls deviations from the average. By the definition of $\mu_{i,\mathrm{C}}$, $\sum_{i' \in S_C} a_{ii'}(x_{i'} - \mu_{i,\mathrm{C}}) = 0$. Expand $x_i - x_{i'} = (x_i - \mu_{i,\mathrm{C}}) - (x_{i'} - \mu_{i,\mathrm{C}})$. The two cross terms vanish after summation, leaving the identity

$$\begin{aligned} &\sum_{i' \in S_C} a_{ii'}(x_i - x_{i'})(x_i - x_{i'})^\top \\ &\quad = \left(\sum_{i' \in S_C} a_{ii'} \right) (x_i - \mu_{i,\mathrm{C}})(x_i - \mu_{i,\mathrm{C}})^\top + \sum_{i' \in S_C} a_{ii'}(x_{i'} - \mu_{i,\mathrm{C}})(x_{i'} - \mu_{i,\mathrm{C}})^\top. \end{aligned}$$

The final sum is positive semidefinite: its quadratic form in any direction v is $\sum_{i' \in S_C} a_{ii'} \left\langle v, x_{i'} - \mu_{i,\mathrm{C}} \right\rangle^2 \ge 0$. Omitting that sum gives a lower bound on the left side. Divide by the clean pair weight, whose reciprocal is at most $2/(1 - 4\eta/\alpha)$ by Eq. (75), and sum over

surviving corrupted points. Then

$$\begin{aligned}
&\sum_{i\in\widehat{S}_D}(x_i-\mu_{i,\mathrm{C}})(x_i-\mu_{i,\mathrm{C}})^\top \\
&\quad\preceq \frac{2}{1-4\eta/\alpha}\sum_{\substack{i\in\widehat{S}_D\\ i'\in S_C}} a_{ii'}(x_i-x_{i'})(x_i-x_{i'})^\top \\
&\quad\preceq \frac{2}{1-4\eta/\alpha}\cdot\frac{1}{2}\sum_{i,i'=1}^{n} a_{ii'}(x_i-x_{i'})(x_i-x_{i'})^\top \\
&\quad\preceq \frac{C^2\sigma^2 n}{2(1-4\eta/\alpha)}\mathbb{I}_d.
\end{aligned}$$

For the middle inequality, each clean–corrupted pair in the restricted sum occurs in both orders in the full sum. Symmetry makes their contributions equal, and the factor $1/2$ counts the pair exactly once. Every other contribution is positive semidefinite. The final inequality is the program's matrix constraint. □

D.3 Computational complexity

In this section, we show that Algorithm 2 is computationally efficient. It is easy to see that all steps except for Step 2 are in polynomial time. It remains to show the computational efficiency of Step 2.

Lemma D.4. *Program* (9) *is a semidefinite program and can be solved in polynomial time.*

Proof. Fix the sample and parameters, with $\alpha > 0$. Eliminate entries with $y_i \neq y_{i'}$ and use the remaining upper-triangular entries $a_{ii'}$, $i \leq i'$, as independent variables, imposing symmetry by definition. There are at most $n(n+1)/2$ variables and $O(n^2)$ scalar constraints. Each row sum a_i is linear in these variables, so all scalar constraints are linear.

By symmetry, the matrix constraint is equivalent to

$$F(A) := \frac{C^2\sigma^2 n}{4} I_d - \sum_{\substack{i<i'\\ y_i=y_{i'}}} a_{ii'}(x_i-x_{i'})(x_i-x_{i'})^\top \succeq 0. \tag{76}$$

The coefficient matrices are fixed, so F is affine. Thus minimizing the constant objective 0 subject to these constraints is a semidefinite program.

For separation, check the scalar constraints directly. Any violated constraint supplies a separating hyperplane. Otherwise, test whether $F(A) \succeq 0$. If this fails, rational symmetric elimination produces a rational vector v with $v^\top F(A) v < 0$. The linear inequality

$$\sum_{\substack{i<i'\\ y_i=y_{i'}}} a_{ii'}\langle v, x_i - x_{i'}\rangle^2 \leq \frac{C^2\sigma^2 n}{4}\|v\|_2^2 \tag{77}$$

holds for every feasible matrix and is violated at the candidate. These operations have polynomial bit complexity in the rational input and candidate encoding lengths, yielding a polynomial-time separation oracle.

Finally, nonnegativity and the row-sum bounds imply $0 \leq a_{ii'} \leq 1$. Hence the feasible region has a known enclosing ball of radius $O(n)$. The ellipsoid running-time bound follows from the remark below. □

Remark D.5 (Numerical precision)**.** *For rational input of encoding length L and $0 < \varepsilon < 1$, whenever the original system is feasible, the ellipsoid method returns a point satisfying each scalar inequality within additive ε and $F(A) \succeq -\varepsilon I_d$ in $\mathrm{poly}(n, d, L, \log(1/\varepsilon))$ bit operations. Indeed, restrict the relaxed search to $-1 \leq a_{ii'} \leq 2$ for the independent entries. Since all coefficient norms have polylogarithmic size, the relaxed system contains, around every originally feasible point, a ball of radius at least $\varepsilon\, 2^{-\mathrm{poly}(n,d,L)}$, while its enclosing radius is $O(n)$. The ellipsoid method depends logarithmically on the ratio of these radii. Symmetry and the prescribed zero entries remain exact by construction; the guarantee hence concerns approximate feasibility.*

E USEFUL LEMMAS

Definition E.1 (Multivector feature map)**.** *Recall $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = [k]$ from Section 3. The multivector (class-sensitive) feature map $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{kd}$ is*

$$\Psi(x, y) := (0, \ldots, 0, \underbrace{x}_{y\text{-th block}}, 0, \ldots, 0). \tag{78}$$

For any $w \in \mathbb{R}^{kd}$, write $w = (w_1, \ldots, w_k)$ with $w_y \in \mathbb{R}^d$. Then

$$\langle w, \Psi(x, y) \rangle \;=\; \langle w_y, x \rangle . \tag{79}$$

Moreover, $\|\Psi(x, y)\|_2 = \|x\|_2$ for any y and $\|\Psi(x, y') - \Psi(x, y)\|_2 = \sqrt{2}\,\|x\|_2$ for $y' \neq y$.

In the concentration bounds below, α and η are local parameters of their respective lemmas.

Lemma E.2 (Chernoff bounds)**.** *Fix $\eta \in [0, 1]$. Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables taking values in $\{0, 1\}$ and let*

$$Z \;=\; \sum_{i=1}^{n} Z_i. \tag{80}$$

If $\Pr(Z_i = 1) \leq \eta$ for every $i \in [n]$, then for any $\alpha \in [0, 1]$,

$$\Pr\big(Z \geq (1 + \alpha)\eta n\big) \;\leq\; \exp\left(-\frac{\alpha^2 \eta n}{3}\right). \tag{81}$$

If $\Pr(Z_i = 1) \geq \eta$ for every $i \in [n]$, then for any $\alpha \in [0, 1]$,

$$\Pr\big(Z \leq (1 - \alpha)\eta n\big) \;\leq\; \exp\left(-\frac{\alpha^2 \eta n}{2}\right). \tag{82}$$

Lemma E.3 (Matrix Chernoff inequality)**.** *Let $M_1, M_2, \ldots, M_n$ be independent random self-adjoint matrices of dimension d. Fix $\Lambda > 0$. Assume $M_i \succeq 0$ and $\lambda_{\max}(M_i) \leq \Lambda$ almost surely for every $i \in [n]$. Let*

$$\mu_{\max} \;=\; \lambda_{\max}\left(\sum_{i=1}^{n} \mathbb{E}[M_i]\right). \tag{83}$$

Then for all $\alpha \geq 0$, with probability at least

$$1 - d\left(\frac{e^{\alpha}}{(1 + \alpha)^{1+\alpha}}\right)^{\mu_{\max}/\Lambda}, \tag{84}$$

we have

$$\lambda_{\max}\left(\sum_{i=1}^{n} M_i\right) \;\leq\; (1 + \alpha)\mu_{\max}. \tag{85}$$

This is the upper-tail bound of Tropp (2012, Corollary 5.2).

Lemma E.4 (Properties of isotropic log-concave distributions)**.** *Let D be an isotropic log-concave distribution over $\mathbb{R}^d$. Here isotropic means $\mathbb{E}_{x \sim D}[x] = 0$ and $\mathbb{E}_{x \sim D}[xx^\top] = \mathbb{I}_d$. The projection and tail bounds follow from Lovász & Vempala (2007, Theorem 5.1 and Lemmas 5.2 and 5.17). Then:*

(1) Orthogonal projections of D onto any subspace are isotropic log-concave when expressed in orthonormal coordinates on that subspace.

(2) For any unit vector $u \in \mathbb{R}^d$ and any $\alpha > 0$,

$$\Pr_{x \sim D}\big(|\langle u, x \rangle| \geq \alpha\big) \;\leq\; e^{-\alpha+1}. \tag{86}$$

(3) For any $\alpha \geq 0$,

$$\Pr_{x \sim D}\Big(\|x\|_2 \geq \alpha\sqrt{d}\Big) \;\leq\; e^{-\alpha+1}. \tag{87}$$

For $0 \leq \alpha \leq 1$, the displayed tail bounds are immediate since $e^{1-\alpha} \geq 1$.