ON METRIC ROBUSTNESS, GENERALIZATION AND OVER-PARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Metric learning has attracted significant attention since it can substantially improve the performance of subsequent tasks such as classification. Current theories of metric learning have mostly focused on the goodness of the learned metric, but how this metric affects subsequent classification performance remains an open problem. In this paper, we study the relationship between the generalization of metric learning and classification. Firstly, we introduce *metric robustness* of classification algorithms with respect to metrics, which quantifies the influence of metrics on classification performance. Secondly, we establish classification generalization guarantees through metric robustness for a range of commonly used metrics and classifiers. Thirdly, we prove that over-parameterization does not necessarily lead to overfitting for a class of distance metric learning methods, addressing an open problem that remains unresolved in the metric learning literature.

1 INTRODUCTION

The notion of metric (or similarity) plays a crucial role in many machine learning tasks such as classification, clustering, and ranking (Chen et al., 2009; Irani et al., 2016). For this reason, many researchers have explored empirically and theoretically to design metrics that are beneficial for the target task. Since manually tuning metrics can be challenging and ineffective for real-world tasks, *metric learning* has then been proposed to automatically learn a distance function using training examples. The learned metrics often lead to dramatic improvements in performance in subsequent tasks, and thus metric learning has attracted significant interest over the past two decades (Bellet et al., 2013; Kulis, 2013). Existing metric learning methods are mostly based on the observation that a pair of examples with the same label (or different labels) should be close to (or far apart from) each other under a good metric. Following this idea, the goal of metric learning is often to find the best metric function given a set of labeled data pairs. Among existing metric learning methods, Mahalanobis distance metric learning is commonly used, which aims to learn a metric parameterized by a positive semi-definite (PSD) matrix (Xing et al., 2002; Schultz & Joachims, 2003; Davis et al., 2007; Weinberger & Saul, 2009; McFee & Lanckriet, 2010). More flexible similarity functions, including formulations with non-PSD matrices, have also been considered (Qamar, 2010; Bellet et al., 2012). The learned metric is often used to improve the performance of algorithms such as k-nearest neighbors (k-NN) (Davis et al., 2007) or linear separators (Guo & Ying, 2014).

Although metric learning can empirically improve the performance of the resultant classifiers, a general theoretical understanding of why and how it affects subsequent classification remains incomplete. The main challenge lies in the fact that metric learning is inherently a two-stage method, that is, learning a metric first, then training the classifier with the help of the metric. Existing theoretical studies on metric learning generally focused on the goodness of *learning the metric*, which concerns the gap between the empirical error and the expected error of the learned metric (Jin et al., 2009; Cao et al., 2016; Bellet & Habrard, 2015), but it is still generally unclear whether and how the generalization of metric learning leads to the generalization of *the resultant classifiers*.

In this direction, Bellet et al. (2012); Guo & Ying (2014) studied regularized similarity learning approaches and established generalization guarantees for associated linear classifiers, using uniform stability (Bousquet & Elisseeff, 2002) and Rademacher complexity, respectively. Joint learning of similarities and landmark-based classifiers has also been analyzed in a semi-supervised setting (Nicolae et al., 2015). These results establish important connections for specific learning for-

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.

mulations. However, understanding how a learned metric affects the generalization of subsequent classifiers beyond these particular formulations remains a theoretical challenge.

In this paper, we propose the concept of *metric robustness* to characterize the connection between learned metrics and the generalization of resultant classifiers. Inspired by the principle of algorithmic robustness (Xu & Mannor, 2012), we study the classifier's loss variation over neighborhoods determined by the learned metric and derive classification generalization guarantees from this relation. We derive bounds on metric robustness for a range of commonly used metrics and classifiers, and connect metric-learning objectives to classification risk. We further apply our analysis to the question raised by Zhou (2021), showing that over-parameterization need not lead to overfitting in a class of distance metric learning methods.

In summary, the contributions of this paper are as follows:

- We introduce metric robustness as a joint property of a learned metric and its resultant classifier. It measures loss variation over learned-metric neighborhoods, providing a general condition for analyzing how metric learning supports subsequent classification.
- We develop the first unified framework that bounds the generalization gap and classification risk through metric robustness. It applies to a broad class of commonly used metrics and classifiers, connecting metric-learning objectives to downstream classification guarantees.
- We answer the open problem raised by Zhou (2021) for a class of LogDet metric learning methods, proving that empirical and expected classification errors can both vanish despite a parameter count quadratic in sample size.

The remainder of the paper is organized as follows. Section 2 introduces some preliminaries and notations. Section 3 presents our main result and its applications to concrete metrics and classifiers. In Section 4, we analyze the relation between the number of pairwise constraints and downstream classification generalization. Section 5 discusses the related work. We conclude in Section 6. All proofs, technical details, and further discussion of our results are provided in the appendices.

2 Preliminaries and Notations

We denote $\mathcal{X}$ as the feature space and assume that $\mathcal{X}$ is a compact subset of a normed vector space, with the ambient norm denoted by $\|\cdot\|_{\mathcal{X}}$. Let $\mathcal{Y}$ be a finite set of labels, and define $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We denote $z = (x, y) \in \mathcal{Z}$ to mean $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In this paper, we focus on learned distances of the form $d_f(x, x') = \|f(x) - f(x')\|_{\mathcal{X}'}$, where $f : \mathcal{X} \to \mathcal{X}'$ is a representation map into a normed embedding space $\mathcal{X}'$. For brevity, d_f denotes the induced pseudometric (distinct inputs may have zero distance when their representations coincide), and d_f^2 denotes the squared distance.

We denote $S = (Z_1, \ldots, Z_n)$ as the sequence of labeled training instances, where $Z_i = (X_i, Y_i)$ are independently and identically distributed (i.i.d.) according to a Borel probability distribution μ on $\mathcal{Z}$. A metric learning algorithm $\mathcal{A}_d$ learns a representation f_S, with induced distance d_{f_S}, from pairs of training instances in S. Subsequently, a classification algorithm $\mathcal{A}_c$ learns a classification rule or score function g_S from the transformed sample $((f_S(X_i), Y_i))_{i=1}^n$. All learning outputs and loss functions are assumed measurable.

For a pair $z_1, z_2 \in \mathcal{Z}$ and a representation f, we associate a nonnegative loss function $l_d(f, z_1, z_2)$. Generally speaking, metric learning methods define the loss function l_d to penalize sample pairs with the same label when their distance is large, encouraging them to be close to each other, and to penalize sample pairs with different labels when their distance is small, thereby pushing them further apart. For a labeled representation $(f(x), y)$ and a classifier g, we associate a nonnegative classification loss $l_c(g, (f(x), y))$. The generalization results below use bounded evaluation losses, with bounds stated in each result. Training objectives and evaluation losses are specified separately.

For integrable losses, we define the risks of the metric and its resultant classifier as follows:

$$\mathcal{L}_d(f) = \mathbb{E}_{Z,Z' \sim \mu}\, l_d(f, Z, Z'), \qquad \mathcal{L}_{c,f}(g) = \mathbb{E}_{(X,Y)\sim\mu}\, l_c\big(g, (f(X), Y)\big). \tag{1}$$

Here $Z = (X, Y)$ and Z' are independent test instances drawn from μ. The corresponding empirical risks are defined as follows:

$$\widehat{\mathcal{L}}_d(f) = \frac{1}{n^2} \sum_{i,j=1}^{n} l_d(f, Z_i, Z_j), \qquad \widehat{\mathcal{L}}_{c,f}(g) = \frac{1}{n} \sum_{i=1}^{n} l_c\big(g, (f(X_i), Y_i)\big). \tag{2}$$

Existing metric learning theory primarily studies the generalization gap of the metric loss, while we study the generalization gap $\left|\mathcal{L}_{c,f}(g) - \widehat{\mathcal{L}}_{c,f}(g)\right|$ and risk of the resultant classifier. We denote its misclassification probability by $R_{01}(g \circ f)$, with score ties resolved by a fixed rule. For example, in binary classification with $y \in \{-1, +1\}$, prediction is determined by the sign of the scalar score $g(f(x))$, using the fixed rule at zero. The clipped hinge evaluation loss $l_c(g, (f(x), y)) = \min\{1, [1 - y\, g(f(x))]_+\}$, where $[a]_+ = \max\{0, a\}$, takes values in $[0, 1]$ and upper-bounds the misclassification indicator. Consequently, $R_{01}(g \circ f) \leq \mathcal{L}_{c,f}(g)$; controlling both the empirical loss and its generalization gap gives a classification-risk guarantee.

3 Generalization in Metric Learning

In this section, we study how learned metrics affect the generalization of resultant classifiers. We introduce *metric robustness* to measure their local compatibility, establish a classification generalization bound, and bound the quantity for several learning methods. Throughout this section, $f = f_S$ and $g = g_S$ denote the learned outputs. All proofs for this section are provided in Appendix B.

3.1 Metric robustness

Algorithmic robustness relates local variation in prediction loss to generalization (Xu & Mannor, 2012). For classification after metric learning, this variation depends on the learned distance and the resultant classifier together: the distance determines which inputs lie in a neighborhood, and the classifier determines how their losses vary. Metric robustness measures this joint behavior at the local scales specified by a fixed partition of the sample space. A small value means that, on average over training points, the classification loss changes little throughout the corresponding learned-distance neighborhoods. The label is held fixed, so the response measures how the loss varies with the input.

Definition 3.1 (Metric robustness)**.** Fix a finite measurable partition $\Pi = \{C_1, \ldots, C_K\}$ of $\mathcal{Z}$ before observing S, with a single label in each cell, and write $C(z)$ for the cell containing z. For a learned pair (f, g) with evaluation loss in $[0, 1]$, let $a_i = \sup_{(x,Y_i)\in C(Z_i)} d_f(X_i, x)$ and $a = a(f, S; \Pi) = \max_i a_i$. Define the loss response at radius $t \in [0, \infty]$ by

$$\omega_i(t) = \sup_{\substack{x\in\mathcal{X} \\ d_f(X_i,x)\leq t}} \left|l_c(g, (f(x), Y_i)) - l_c(g, (f(X_i), Y_i))\right|. \tag{3}$$

Here the label is fixed at Y_i, and the ball at $t = \infty$ is $\mathcal{X}$. The metric robustness of (f, g) relative to S and Π is

$$\varepsilon_{\mathrm{mr}}(f, g, S; \Pi) = \frac{1}{n}\sum_{i=1}^{n} \omega_i(a_i). \tag{4}$$

The pair has ε-metric robustness relative to S, Π if this quantity is at most ε.

Remark 3.2 (Generality of Definition 3.1). Definition 3.1 applies to any measurable representation and classifier with evaluation loss in $[0, 1]$, and characterizes their local compatibility through the joint loss response. Section 3.3 gives bounds for Mahalanobis and kernel metrics with regularized classifiers, nearest neighbors, decision trees, and jointly trained neural representations. These calculations use score variation or decision margins to evaluate the effect of the learned metric on classification loss. The following proposition motivates this compatibility condition: empirical and population metric losses can both vanish while the classification generalization gap is maximal.

Proposition 3.3. *Let $X \sim \mathrm{Unif}[-1, 1]$, $Y = \mathrm{sign}(X)$, with $\mathrm{sign}(0) = 1$, and use the contrastive loss of Example 3.10 with $t = 4$. There is a measurable representation rule $f_S : \mathcal{X} \to \{-1, +1\}$ such that, for every $n \geq 1$, the linear score $g(v) = v$ satisfies, almost surely,*

$$\widehat{\mathcal{L}}_d(f_S) = \mathcal{L}_d(f_S) = 0, \qquad \widehat{R}_{01}(g \circ f_S) = 0, \qquad R_{01}(g \circ f_S) = 1.$$

When both classes are observed, every measurable classifier fitting the represented training labels has population error one.

The construction reverses the representations of the classes between training and testing while preserving distances within training pairs and within test pairs. This shows that controlling the empirical and population metric losses alone is insufficient to guarantee generalization of the resultant classifier. For each fixed partition as in Definition 3.1, its zero–one metric robustness is almost surely one for the displayed score and, when both classes are observed, for every interpolating classifier.

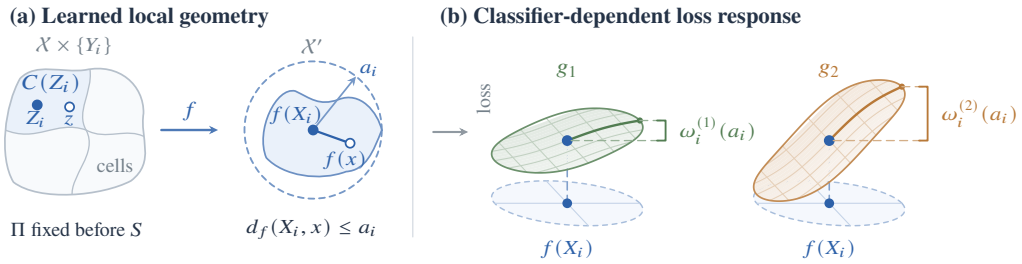


Figure 1: **Metric robustness.** (a) The cell $C(Z_i)$ determines the learned radius a_i. (b) Two classifiers have different loss responses over the same neighborhood at fixed label Y_i. Brackets indicate maximal changes from the loss at X_i, illustrating the joint dependence on the metric and classifier.

Remark 3.4 (Metric and partition). The learned distance determines the comparison neighborhoods, and the classifier determines their loss variation. For prediction-based losses, compensated rescaling $f_c = cf$, $g_c(v) = g(v/c)$, $c > 0$, leaves $\varepsilon_{\mathrm{mr}}$ unchanged. Distances and radii scale together, preserving the comparison neighborhoods in the input space and the classification loss at every input. The partition Π specifies the local scales and is fixed before observing S. Splitting cells by label ensures same-label comparisons within each cell, including when labels are noisy. Refining the partition can reduce the loss responses, at the cost of more cells in the generalization bound.

The balls in equation 3 can extend beyond the partition cell, as discussed further in Appendix B.3.

3.2 Generalization guarantees

We now establish a generalization guarantee for the resultant classifier in terms of metric robustness. The bound combines the local loss variation of the learned metric–classifier pair with a sampling term determined by the fixed partition.

Theorem 3.5 (Classification generalization). *Fix Π as in Definition 3.1 and $\delta \in (0,1)$, and set $\tau_n = \sqrt{(K \log 2 + \log(1/\delta))/(2n)}$. There exists a measurable event of probability at least $1-\delta$ over $S \sim \mu^n$ on which, simultaneously for every measurable learned pair with evaluation loss in $[0,1]$,*

$$\left| \mathcal{L}_{c,f}(g) - \widehat{\mathcal{L}}_{c,f}(g) \right| \le \varepsilon_{\mathrm{mr}}(f, g, S; \Pi) + \tau_n. \tag{5}$$

If the evaluation loss upper-bounds the misclassification indicator, then, on the same event,

$$R_{01}(g \circ f) \le \widehat{\mathcal{L}}_{c,f}(g) + \varepsilon_{\mathrm{mr}}(f, g, S; \Pi) + \tau_n. \tag{6}$$

Theorem 3.5 bounds the classification generalization gap by the average local response $\varepsilon_{\mathrm{mr}}$ and the sampling term τ_n. The first term measures classification-loss variation over neighborhoods determined by the learned distance; the second depends on the number of partition cells and independent training observations. For a fixed partition, sample size, and confidence level, smaller metric robustness gives a tighter bound. Combining this control with a small empirical classification loss yields the risk guarantee in equation 6.

For losses with a Lipschitz response to the learned distance, metric robustness admits a direct bound. If $a < \infty$ and, for some finite $L \ge 0$,

$$|l_c(g, (f(x), y)) - l_c(g, (f(x'), y))| \le L d_f(x, x') \quad (x, x' \in \mathcal{X},\ y \in \mathcal{Y}), \tag{7}$$

then $\omega_i(t) \le \min\{1, Lt\}$ for finite t, and hence

$$\varepsilon_{\mathrm{mr}} \le \frac{1}{n} \sum_i \min\{1, La_i\} \le \min\{1, La\}. \tag{8}$$

The constant L can depend on the learned pair. Equation 8 shows how the learned-distance radii and the classifier's loss sensitivity combine to control metric robustness. The examples below also obtain bounds from Hölder responses and local decision margins.

The same geometric radii yield a bound for pairwise metric losses, connecting the classification analysis to metric-learning robustness (Bellet & Habrard, 2015).

Corollary 3.6. *Suppose $a < \infty$ and $l_d(f, z, z') = \psi_{y,y'}(d_f(x, x')) \in [0, B_d]$, where each $\psi_{y,y'}$ is U-Lipschitz on the attained distances, for finite $B_d, U \geq 0$. On the event of Theorem 3.5,*

$$|\mathcal{L}_d(f) - \widehat{\mathcal{L}}_d(f)| \leq \min\{B_d, 2Ua\} + 2B_d\tau_n. \tag{9}$$

For losses U-Lipschitz in the squared distance and $d_f(x, x') \leq D < \infty$, replace $2Ua$ by $4UDa$.

Corollary 3.6 relates the generalization of the pairwise metric loss to the same local geometry. Its local term is determined by the learned-distance radii and the response of the pairwise loss, while the classification bound uses the resultant classifier's loss response. Passing to the product partition $\Pi \times \Pi$ gives the pairwise bound, whose sampling term still depends on the n independent training observations. Bounds on the learned geometry can therefore be used in both analyses, with the appropriate response calculation for each loss.

3.3 Examples

We next illustrate how metric robustness can be bounded for commonly used metrics and classifiers. Fix $r > 0$. A smallest radius-$r/2$ cover chosen before S gives a label-pure partition with input diameter at most r and $K \leq |\mathcal{Y}|\mathcal{N}(r/2, \mathcal{X}, \|\cdot\|_{\mathcal{X}})$, where $\mathcal{N}$ is the covering number. For each metric–classifier pair, we bound how far a partition cell extends in the learned distance and how much the classification loss varies over that distance. Combining these estimates gives metric robustness and classification guarantees through Theorem 3.5. Additional metric and classifier examples, together with margin refinements, are collected in Appendix B.5.

Example 3.7 (Mahalanobis metrics). Let $f(x) = A^{1/2}x$ on Euclidean inputs, with $A \succeq 0$. If $\widehat{J}_d(A) + \lambda_d\Omega(A) \leq E_d < \infty$, where $\widehat{J}_d \geq 0$, $\lambda_d > 0$, and Ω is the trace, Frobenius norm, or entrywise ℓ_1 norm, then $a_i \leq r\sqrt{\|A\|_{\mathrm{op}}} \leq r\sqrt{E_d/\lambda_d}$. Here the regularizer bounds the largest expansion of input distances; for a minimizer, E_d can be the objective at any feasible reference matrix. For binary labels, the affine score $g(v) = \langle w, v\rangle + b$ has clipped-hinge response $\omega_i(t) \leq \min\{1, \|w\|_2 t\}$ for finite t. An exact minimizer of $n^{-1}\sum_i[1 - Y_i g(f(X_i))]_+ + \lambda_c\|w\|_2^2/2$, with $\lambda_c > 0$, satisfies $\|w\|_2 \leq \sqrt{2/\lambda_c}$ by comparison with the zero score. The product of the radius bound and the score sensitivity therefore gives

$$|\mathcal{L}_{c,f}(g) - \widehat{\mathcal{L}}_{c,f}(g)| \leq \min\left\{1, r\sqrt{\frac{2E_d}{\lambda_d\lambda_c}}\right\} + \tau_n. \tag{10}$$

For Lasso minimizing the average squared loss plus $\lambda_c\|w\|_1$, comparison with the zero score gives $\|w\|_2 \leq \|w\|_1 \leq 1/\lambda_c$.

For LogDet regularization (Davis et al., 2007), write $D_{\mathrm{ld}}(A, A_0) = \mathrm{tr}(AA_0^{-1}) - \log\det(AA_0^{-1}) - d$, with $A, A_0 \succ 0$. If $D_{\mathrm{ld}}(A, A_0) \leq C$, then $a_i \leq r\sqrt{\|A_0\|_{\mathrm{op}} b(C)}$, where $b(C) \geq 1$ solves $b - \log b - 1 = C$. Combining this radius with the affine response gives the bound used in Section 4.

Example 3.8 (Kernel methods). For a fixed L_ϕ-Lipschitz Hilbert-space feature map ϕ and bounded positive operator A, the representation $f = A^{1/2}\phi$ gives $a_i \leq rL_\phi\sqrt{\|A\|_{\mathrm{op}}}$. A Gaussian-RKHS score g with kernel $k(v, v') = e^{-\|v-v'\|^2/(2\sigma^2)}$, $\sigma > 0$, has clipped-hinge response $\omega_i(t) \leq \min\{1, \|g\|_{\mathcal{H}_k} t/\sigma\}$ for finite t. The reproducing property converts feature-space displacement into score variation, so the learned operator and the classifier norm enter the bound together.

Example 3.9 (Nearest neighbors). Consider unweighted k-NN, $1 \leq k \leq N$, on labeled prototypes with fixed tie rules. Put $h = \lfloor k/2 \rfloor + 1$ and $c_i = g(f(X_i))$. Let s_i and o_i be the h-th nearest distance among prototypes labeled c_i and the $(k - h + 1)$-th nearest among the others, respectively; any missing rank is assigned distance ∞. Within the ball under consideration, each distance changes by at most a_i. Thus, for finite a_i, the condition $s_i + 2a_i < o_i$ is sufficient to preserve a strict majority for c_i. For zero–one loss, we therefore obtain $\varepsilon_{\mathrm{mr}} \leq n^{-1}\sum_i \mathbf{1}\{a_i = \infty \text{ or } s_i + 2a_i \geq o_i\}$. The prototypes may include training points; in this case, empirical evaluation includes self-neighbors.

To pass from a generalization-gap bound to a classification-risk guarantee, we also need to control the empirical classification loss. The following example obtains this control directly from a contrastive metric-learning objective.

Example 3.10 (Mean-distance classifiers). Let f have Hilbert-space image diameter at most $D < \infty$. For $t > 0$, take $l_d = d_f^2/t$ on same-class pairs and $l_d = [1 - d_f^2/t]_+$ otherwise. Assume at least two classes, each observed, and put $n_c = |\{j : Y_j = c\}|$. Define

$$q_c(x) = \frac{1}{n_c} \sum_{j:Y_j=c} d_f(x, X_j)^2, \qquad \widehat{Q} = \frac{1}{n} \sum_i \sum_{c \in \mathcal{Y}} \frac{1}{n_c} \sum_{j:Y_j=c} l_d(f, Z_i, Z_j). \tag{11}$$

Predict a class minimizing $q_c(x)$, with evaluation loss $\min\{1, [1 + (q_y(x) - \min_{c \neq y} q_c(x))/t]_+\}$. Averaging the pairwise losses bounds its empirical loss by $\widehat{Q}$. In each difference $q_y - q_c$, the squared norm of the query representation cancels, giving $\omega_i(a_i) \leq \min\{1, 2Da_i/t\}$. Thus

$$R_{01}(g \circ f) \leq \widehat{Q} + \varepsilon_{\mathrm{mr}} + \tau_n \leq \widehat{Q} + \frac{1}{n} \sum_i \min\{1, 2Da_i/t\} + \tau_n. \tag{12}$$

Here $\widehat{Q} \leq \widehat{\mathcal{L}}_d(f)/\min_c(n_c/n)$. For binary labels, the mean-distance score is affine in the representation, providing the comparison argument used in Section 4.2.

Example 3.11 (Neural representations). For Euclidean inputs and jointly trained parts $f_\theta(x) = \sum_{j=1}^q u_j[v_j^\top x + b_j]_+$, $g(v) = \langle w, v \rangle + b$, put $P(\theta) = \sum_j \|u_j\|_2 \|v_j\|_2$. An average hinge objective with penalty $\lambda(P(\theta)^2 + \|w\|_2^2)/2$, $\lambda > 0$, and value at most $E < \infty$ controls both stages: ReLU gives $a_i \leq rP(\theta)$, while the affine head has clipped-hinge response at most $\min\{1, \|w\|_2 t\}$. Since $2P(\theta)\|w\|_2 \leq P(\theta)^2 + \|w\|_2^2$,

$$\varepsilon_{\mathrm{mr}} \leq \min\{1, rP(\theta)\|w\|_2\} \leq \min\{1, rE/\lambda\}.$$

For a concrete example, take $\mathcal{X} = [-3, -2] \cup [-1, 1] \cup [2, 3]$, with positive outer intervals and a negative inner interval. The width-two representation $f_*(x) = [x]_+ + [-x]_+ = |x|$ and head $g_*(v) = 2v - 3$ have margin at least one and objective 4λ. The representation maps both positive intervals to $[2, 3]$ and the negative interval to $[0, 1]$, making the classes separable by the affine head. Adding zero-output neurons extends this comparator to every finite width $q \geq 2$. For returned objectives at most $(4 + c)\lambda$, with fixed $c \geq 0$ and $r = \lambda = n^{-1/3}$, both training and population errors are $O(n^{-1/3})$ with probability at least $1 - n^{-2}$, uniformly over these widths, including $q = n^2$. Partitions with $K(r) \leq 8/r + 6$ give a sampling term of order $n^{-1/3}$, matching the empirical-loss and metric-robustness bounds.

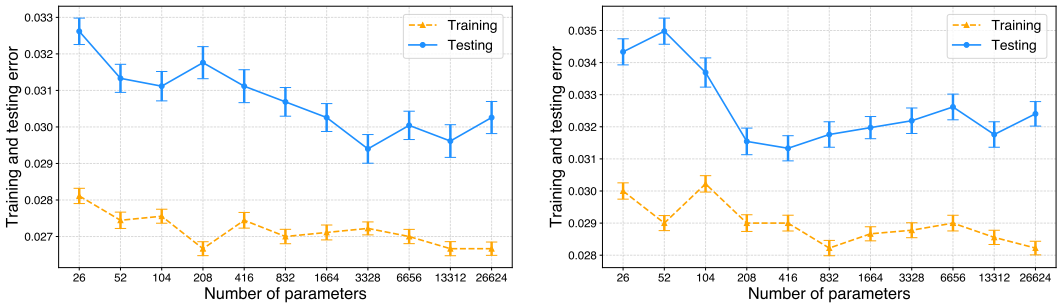
4 Over-parameterization without Overfitting

In this section, we apply the results of Section 3 to a question raised by Zhou (2021): why can metric learning generalize even when its parameter count exceeds the number of training examples? Metric learning aims to transform the original feature space into one better suited to a subsequent classifier. This resembles the role of representation layers in a neural network, which learn features for the final classification layer. Following Zhou's metric-learning example, we study LogDet metric learning and count the variables associated with pairwise constraints. We show how these constraints control classification loss and metric robustness, and establish conditions under which training and population classification errors both vanish as the parameter count grows quadratically. All proofs and further technical discussion in Section 4 are provided in Appendix C.

4.1 Learning setting

Let $\mathcal{X} \subset \mathbb{R}^d$, $\|x\|_2 \leq R$, and $\mathcal{Y} = \{-1, +1\}$. For $n \geq 3$, write $n_c = |\{i : Y_i = c\}|$ and form the pair pools $\mathcal{T}_+ = \{(i, j) : i < j, Y_i = Y_j\}$ and $\mathcal{T}_- = \{(i, j) : i < j, Y_i \neq Y_j\}$, of sizes N_+ and N_-. Conditional on S, uniformly permute each pool and select its first m_σ pairs, denoted by $\mathcal{T}_{\sigma, m_\sigma}$, for $\sigma \in \{+, -\}$. The selections are nested, with total count $m = m_+ + m_- \leq \binom{n}{2}$. We consider both classes present and $1 \leq m_\sigma \leq N_\sigma$, using a fixed classifier otherwise.

Fix $A_0 \succ 0$ and $0 < u < l$. Write $f_A(x) = A^{1/2}x$, $q_A = d_{f_A}^2$, $\Delta = l - u$, and $v_e = X_i - X_j$ for $e = (i, j)$. Following Zhou (2021); Davis et al. (2007), consider the LogDet problem on the left; its optimizer $A_{\boldsymbol{m}}$ satisfies the dual relation on the right:



(a) SVM (b) Lasso

Figure 2: Training and testing errors for SVM and Lasso after metric learning. The horizontal axis counts pairwise constraints; bars show standard deviations over 20 random splits. Experimental details are provided in Appendix C.13.

$$\begin{aligned} &\min_{A \succ 0} \; D_{\mathrm{ld}}(A, A_0) \\ &\text{s.t. } q_A(e) \le u, \; e \in \mathcal{T}_{+,m_+}, \\ &\quad q_A(e) \ge l, \; e \in \mathcal{T}_{-,m_-}. \end{aligned} \tag{13}$$

$$A_{\boldsymbol{m}}^{-1} = A_0^{-1} + \sum_{e \in \mathcal{T}_{+,m_+}} \lambda_e v_e v_e^\top - \sum_{e \in \mathcal{T}_{-,m_-}} \nu_e v_e v_e^\top. \tag{14}$$

Here $\lambda_e, \nu_e \ge 0$, and $q_A(e) = v_e^\top A v_e$. The optimizer exists uniquely whenever the constraints are feasible; otherwise use the fixed classifier. There are m constraint-associated dual variables, the parameter count used by Zhou (2021); the matrix dimension is $p = d(d+1)/2$.

4.2 Pair constraints and classification risk

We first track how adding constraints affects the classification-risk bound. The analysis combines control of the learned geometry, approximation of the full pair pools, and a comparison classifier with small training loss. Fix $C \ge 0$ before observing S, and suppose the full training-pair pools admit a comparison matrix $A^\dagger \succ 0$ satisfying

$$D_{\mathrm{ld}}(A^\dagger, A_0) \le C, \qquad q_{A^\dagger} \le u \text{ on } \mathcal{T}_+, \qquad q_{A^\dagger} \ge l \text{ on } \mathcal{T}_-. \tag{15}$$

The same matrix is feasible for every selected subset, giving $D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C$. Put $\beta = \|A_0\|_{\mathrm{op}} b(C)$ and $f_{\boldsymbol{m}} = f_{A_{\boldsymbol{m}}}$. Fix a label-pure measurable partition Π_r before observing S, with $K(r)$ cells of input diameter at most r. LogDet then gives $A_{\boldsymbol{m}} \preceq \beta I$ and $a_i \le r\sqrt{\beta}$, uniformly over budgets. Thus the comparator bounds local expansion even as further pair constraints are imposed.

The selected pairs approximate the full training-pair pools. Let $L_+(A), L_-(A)$ be the full-pool means of $[q_A - u]_+, [l - q_A]_+$, and $\widehat{L}_{\sigma,k}(A)$ their means over the first k selected pairs. Fix $\delta \in (0,1)$, put $B = \max\{4R^2\beta, l\}$, and define

$$e_\sigma(k) = \frac{8R^2\sqrt{d}\,\beta}{n^2} + B\sqrt{\frac{p\log(1+2n^2) + \log(4n^2/\delta)}{2k}}, \quad 1 \le k < N_\sigma, \qquad e_\sigma(N_\sigma) = 0. \tag{16}$$

Conditional on S with both classes present, sampling without replacement (Hoeffding, 1963) gives $|L_\sigma(A) - \widehat{L}_{\sigma,k}(A)| \le e_\sigma(k)$ with probability at least $1 - \delta/2$, simultaneously for both pools, all prefixes, and all A with $D_{\mathrm{ld}}(A, A_0) \le C$. The error $e_\sigma(k)$ decreases with the number of selected pairs and becomes zero at the full pool. On this event, zero selected violations give $L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le V_{\boldsymbol{m}} := e_+(m_+) + e_-(m_-)$.

Fit the affine SVM of Example 3.7 to $f_{\boldsymbol{m}}$, with regularization $\lambda_c > 0$. Using the classwise mean squared distances q_c from Example 3.10, the score $(q_- - q_+)/\Delta$ is affine in the representation. Its hinge loss is controlled by pair violations, supplying a comparison score for the fitted SVM. Put $n_{\min} = \min(n_+, n_-)$ and $\alpha_S = \max\{1, 2N_+/(nn_{\min})\}$. Comparison with an affine mean-distance score and the zero score bounds the optimal hinge-plus-regularization objective by

$$H_{\boldsymbol{m}} = \min\left\{1, \frac{\alpha_S V_{\boldsymbol{m}}}{\Delta} + \frac{8\lambda_c R^2 \beta}{\Delta^2}\right\}, \qquad \tau_n = \sqrt{\frac{K(r)\log 2 + \log(2/\delta)}{2n}}. \tag{17}$$

Thus the empirical hinge loss is at most $H_{\boldsymbol{m}}$ and $\|w_{\boldsymbol{m}}\|_2 \le \sqrt{2H_{\boldsymbol{m}}/\lambda_c}$. Combining this norm bound with the radii above controls metric robustness and gives the following result.

Theorem 4.1 (Classification risk)**.** *For hard-constraint LogDet learning followed by the affine SVM above, there is an event of probability at least $1-\delta$ over S and the pair permutations on which, whenever both classes occur and equation 15 holds, every admissible budget satisfies*

$$R_{01}(g_{\boldsymbol{m}} \circ f_{\boldsymbol{m}}) \le \underbrace{H_{\boldsymbol{m}}}_{\textit{empirical loss}} + \underbrace{\min\left\{1, r\sqrt{\frac{2\beta H_{\boldsymbol{m}}}{\lambda_c}}\right\}}_{\textit{metric robustness}} + \underbrace{\tau_n}_{\textit{sampling}}. \tag{18}$$

The empirical zero–one error is at most $H_{\boldsymbol{m}}$; the clipped-hinge generalization gap is at most the last two terms. With all other settings fixed, these bounds are nonincreasing in each pair budget.

Theorem 4.1 quantifies the finite-sample effect of pair coverage. Adding pairs decreases $V_{\boldsymbol{m}}$, tightening the empirical-loss and classifier-norm bounds in $H_{\boldsymbol{m}}$. Together with the LogDet radius bound, this controls metric robustness. The classification-risk bound is therefore nonincreasing along the nested selections, while its sampling term depends on the n original observations. Its improvement comes from closer approximation of the full pair pools.

Introduce positive slacks ξ_e, with reference values $\xi_{0,e} = u$ for same-class pairs and $\xi_{0,e} = l$ for different-class pairs. The sum-slack formulation minimizes

$$\begin{aligned} &\min_{A\succ 0,\ \xi_e>0} D_{\mathrm{ld}}(A, A_0) + \gamma \sum_e \varphi(\xi_e/\xi_{0,e}), \\ &\text{s.t. } q_A(e) \le \xi_e \quad (e \in \mathcal{T}_{+,m_+}), \qquad q_A(e) \ge \xi_e \quad (e \in \mathcal{T}_{-,m_-}), \end{aligned} \tag{19}$$

where $\varphi(t) = t - \log t - 1$, and a fixed $\gamma > 0$ weights the unnormalized sum over selected pairs. Under equation 15, its optimizer $(A_{\boldsymbol{m}}, \xi_{\boldsymbol{m}})$ satisfies $D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C$ and $\sum_e \varphi(\xi_{\boldsymbol{m},e}/\xi_{0,e}) \le C/\gamma$, by comparison with $(A^\dagger, \xi_0)$. These bounds control the sum of squared threshold violations; Cauchy–Schwarz then bounds their average. Adding the pair-sampling errors yields $L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le V_{\boldsymbol{m}}$, where

$$V_{\boldsymbol{m}} = e_+(m_+) + e_-(m_-) + \sqrt{\frac{2C\max\{4uR^2\beta, l^2\}}{\gamma}}\left(m_+^{-1/2} + m_-^{-1/2}\right). \tag{20}$$

The conclusions of Theorem 4.1 hold on the same event with this $V_{\boldsymbol{m}}$ in equation 17. The additional slack term is nonincreasing in both pair budgets. Extensions to averaged hinge objectives and Gaussian SVMs are given in Appendices C.7–C.8.

4.3 Over-parameterization without overfitting

We now let n grow. Corollary 4.2 gives the main conclusion: training and population classification errors both vanish while the parameter count grows quadratically. Assume *population threshold realizability*: a fixed $A^\dagger \succ 0$ with $D_{\mathrm{ld}}(A^\dagger, A_0) \le C$ satisfies, for independent $Z = (X, Y), Z' = (X', Y') \sim \mu$, almost surely,

$$Y = Y' \implies q_{A^\dagger}(X, X') \le u, \qquad Y \ne Y' \implies q_{A^\dagger}(X, X') \ge l. \tag{21}$$

This positive within-class/between-class gap supplies the common sample comparator almost surely.

Corollary 4.2 (Vanishing errors)**.** *Consider hard-constraint or sum-slack ITML followed by the affine SVM. Suppose equation 21 holds, $\pi_0 = \min_c \mu(Y = c) > 0$, and a deterministic family of partitions Π_r as above satisfies $K(r) \le C_K r^{-s}$ for small r, with $s > 0$. Keep $d, R, A_0, u, l, C, C_K, s$ fixed, and also $\gamma > 0$ for sum-slack ITML. Take $m_{\sigma,n} = \lfloor \theta_\sigma N_\sigma \rfloor$, with fixed $\theta_\sigma \in (0, 1]$, and $r_n = \lambda_{c,n} = n^{-1/(s+2)}$. For all sufficiently large n, with probability at least $1 - 2n^{-2} - 2e^{-n\pi_0^2/2}$,*

$$m_n = \Theta(n^2), \qquad \max\{R_{01}(g_n \circ f_n), \widehat{R}_{01}(g_n \circ f_n)\} = O\left(n^{-1/(s+2)}\right), \tag{22}$$

where $(g_n, f_n) = (g_{\boldsymbol{m}_n}, f_{\boldsymbol{m}_n})$, with constants independent of n. Thus the zero–one generalization gap tends to zero in probability while $m_n/n \to \infty$ in probability.

Corollary 4.2 gives vanishing classification errors with $m_n/n \to \infty$. Fixed positive fractions of both pools yield $V_{\boldsymbol{m}_n} = O(\sqrt{\log n}/n)$ and $H_{\boldsymbol{m}_n} = O(\lambda_{c,n})$. The fitted classifier norm stays bounded, giving $\varepsilon_{\mathrm{mr}} = O(r_n)$. The partition complexity satisfies $\tau_n = O(r_n)$, so all three terms in equation 18 vanish at the stated rate. Here $V_{\boldsymbol{m}_n} = o(r_n)$, so the pair-approximation error does not slow the resulting convergence rate. Thus quadratically many constraints are compatible with the same regularization and local geometric control used to obtain the finite-sample risk guarantee.

Remark 4.3. The parameter count follows Zhou's dual coordinates in equation 14, with fixed matrix dimension. Under population threshold realizability, a bounded affine separator also gives the same consistency rate with fixed pair budgets or the reference metric. Appendix C.10 proves this comparison and combines it with Theorem 4.1: increasing pair coverage can improve the finite-sample bound's constants. The corollary establishes vanishing errors even with quadratically many constraints; actual risks can fluctuate as pairs are added. Example 3.11 gives a width-uniform neural guarantee through joint regularization of both the learned representation and the classification head.

5 Related Work

In this section, we discuss generalization analyses of metric learning and subsequent classification. Appendix D proves the special cases and comparisons described below, with all constants explicit.

Generalization gaps of metric learning. Theoretical studies have primarily analyzed the gap between the empirical and expected metric-learning losses. A central difficulty is the dependence between training pairs that share an observation. Bian & Tao (2011; 2012) studied empirical risk minimization and consistency in distance metric learning. Jin et al. (2009) applied the uniform stability framework of Bousquet & Elisseeff (2002) to regularized Mahalanobis learning, while Cao et al. (2016) used Rademacher complexity and tools for pairwise averages to analyze several matrix-norm regularizers. Bellet & Habrard (2015) adapted algorithmic robustness (Xu & Mannor, 2012) to metric learning, relating pairwise loss variation to generalization through a partition of the sample space. Our product-partition analysis recovers their finite-sample robustness bounds. Metric robustness further measures the classification-loss response to learned geometry, connecting the analysis of the learned metric with that of its resultant classifier through a common sample-space partition.

Generalization of resultant classifiers. The goodness framework of Balcan et al. (2008) gives classification guarantees through similarities to landmarks. Bellet et al. (2012) learned non-PSD similarities for linear classification using uniform stability; Guo & Ying (2014) developed classification guarantees under general matrix-norm regularizers using Rademacher analysis. Nicolae et al. (2015) analyzed a jointly learned similarity and landmark-based linear classifier in a semi-supervised setting, using algorithmic robustness. For bounded Lipschitz landmark similarities, our theorem recovers their robustness and generalization bounds as a special case. Averaging the local responses can sharpen the uniform local bound. For the bilinear similarities studied by Bellet et al. (2012); Guo & Ying (2014), we also recover the empirical loss comparison and derive a geometric classification-risk bound. The joint response further accommodates the nonlinear classifiers and neural representations in Section 3.3, and supports our analysis of additional pair constraints in LogDet metric learning, where empirical fitting and local loss responses determine the risk bound.

6 Conclusion

We introduce metric robustness to characterize the relation between learned metrics and resultant classifiers, and establish bounds on classification generalization and risk. The analysis covers common metric-learning and classification methods, including jointly trained neural representations. The examples relate these bounds to metric regularization, classifier norms, decision margins, and empirical fitting, making the condition explicit for the methods considered. For LogDet metric learning, the classification-risk bound is nonincreasing with added pairs. Under threshold realizability and the stated regularization, training and expected errors vanish with quadratically many constraints. The fixed-budget comparison identifies when pair coverage improves the bound's constants. Extending the analysis to broader neural architectures requires understanding how joint training controls the learned geometry and classification-loss response. Our regularized neural example offers a starting point for studying these quantities along optimization trajectories and relating their evolution to classification generalization in jointly trained networks of varying widths and depths.

AI USE STATEMENT

In this work, we used generative AI tools to obtain reviewer-style feedback on the manuscript. Based on this feedback, we corrected typographical errors and improved grammar, wording, and readability. We did not use generative AI tools for any other purposes in this work. We reviewed all AI-assisted revisions to ensure that they preserved the intended meaning and accurately reflected our scientific content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies theoretical guarantees for classification after metric learning. The empirical illustration reuses the Breast Cancer Wisconsin (Original) benchmark to examine classification errors as the number of pairwise constraints increases. These experiments serve as methodological illustrations. Any clinical application would require separate validation on representative data and assessment of errors across relevant patient populations. The theoretical guarantees are stated with their assumptions, and the discussion distinguishes the proved risk bounds from observed finite-sample behavior.

REPRODUCIBILITY STATEMENT

The notation, learning procedures, and assumptions are specified in Sections 2–4. Appendices B and C provide detailed proofs of the theoretical results and calculations for the examples, including optimization comparisons and probability bounds. Appendix D gives the reductions to prior results and identifies the corresponding assumptions and constants. For Figure 2, Appendix C.13 records the dataset and missing-value handling, training and test sample sizes, repeated random splits, nested pair sampling, and computing environment. It also explains the correspondence between the experimental setting and the theoretical analysis, and identifies the information unavailable in the original experimental record.

REFERENCES

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 287–298, 2008.

Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. Similarity learning for provably accurate sparse linear classification. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1871–1878, 2012.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709, 2013.

Wei Bian and Dacheng Tao. Learning a distance metric by empirical loss minimization. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 1186–1191, 2011.

Wei Bian and Dacheng Tao. Constrained empirical risk minimization framework for distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1194–1205, 2012.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(27):747–776, 2009.

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216, 2007.

Zheng-Chu Guo and Yiming Ying. Guaranteed classification via regularized similarity learning. *Neural Computation*, 26(3):497–522, 2014.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jasmine Irani, Nitin Pise, and Madhura Phatak. Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications*, 134(7):9–14, 2016.

Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems*, volume 22, pp. 862–870, 2009.

Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

Brian McFee and Gert Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 775–782, 2010.

Maria-Irina Nicolae, Marc Sebban, Amaury Habrard, Eric Gaussier, and Massih-Reza Amini. Algorithmic robustness for semi-supervised (ϵ, γ, τ)-good metric learning. In *Proceedings of the 22nd International Conference on Neural Information Processing, Part I*, volume 9489 of *Lecture Notes in Computer Science*, pp. 253–263. Springer, 2015.

Ali Mustafa Qamar. *Generalized cosine and similarity metrics: A supervised learning approach based on nearest neighbors*. PhD thesis, Université de Grenoble, 2010.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems*, volume 16, pp. 41–48, 2003.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009.

William Wolberg. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 1990. Dataset.

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15, 2002.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.

Zhi-Hua Zhou. Why over-parameterization of deep neural networks does not overfit? *Science China Information Sciences*, 64(1):116101, 2021.

A NOTATION

Symbols are grouped by use; repeated symbols have the meanings specified in their sections.

Symbol	Meaning
Data, losses, and robustness	
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{X}'$	Spaces
$\mu;\ Z = (X, Y), Z'$	Law; observations
$S, Z_i, n;\ \mathcal{A}_d, \mathcal{A}_c$	Sample; algorithms
$f_S, g_S;\ d_f, d_f^2$	Outputs; distances
$l_d, l_c;\ \mathcal{L}_d, \widehat{\mathcal{L}}_d$	Losses; metric risks
$\mathcal{L}_{c,f}, \widehat{\mathcal{L}}_{c,f}$	Classification risks
$R_{01}, \widehat{R}_{01}, \widehat{R}_h$	Error/hinge risks
$[a]_+, \chi(a);\ \delta$	Loss maps; confidence
$\Pi, \Pi_r;\ C_j, C(z)$	Partitions; cells
$K, K(r);\ r, \mathcal{N}$	Counts; scale/cover
$a_i, a;\ B_f(X_i, t)$	Radii; metric ball
$\omega_i(t), \varepsilon_{\mathrm{mr}}$	Response; robustness
$\tau_n, L;\ B, B_d, U$	Sampling; loss control
$p_j, \widehat{p}_j, n_j;\ T$	Masses/counts; TV
$h, c_i;\ \bar{h}_j, \widehat{h}_j$	Loss/variation; means
$\bar{a}_i, \bar{\omega}_i;\ D$	Bounds; diameter
Metrics and classifiers	
$A, A_0;\ d, p$	Matrices; dimensions
$\|\cdot\|_{\mathrm{op}}, \|\cdot\|_F, \|\cdot\|_1$	Matrix/vector norms
$\Omega, \widehat{J}_d, E_d$	Penalty/loss/budget
λ_d, λ_c	Regularization
$D_{\mathrm{ld}}, C, \mathcal{A}_C$	LogDet; budget/set
$b_-(C), b(C), \beta, \underline{\beta}$	Spectral roots/bounds
$R, f_A, q_A;\ \phi, L_\phi$	Geometry; feature map
$k, \mathcal{H}_k, \sigma;\ H, \alpha$	Kernel; Hölder data
$w, b;\ m_i, b_i(t)$	Head; score bounds
γ_i, b_i	Boundary/tree bounds
$N, k, h;\ c_i, s_i, o_i, u_i$	Counts; voting margins
$n_c, q_c(x);\ t, \widehat{Q}$	Class statistics; loss
$\bar{f}_c, v_c, a_c$	Class moments
$f_\theta, q;\ u_j, v_j, b_j$	Network parameters
$P(\theta), J, \lambda, E;\ B, \eta, c$	Objective/norm bounds
$s, \zeta_j, \alpha_j;\ L_s, \gamma$	Landmark quantities
$\mathbb{E}, \mathbb{P}, I, \mathbf{1}\{\cdot\}, \log$	Basic notation
$O(\cdot), \Theta(\cdot), o(\cdot), \inf \varnothing$	Orders; $\inf \varnothing = \infty$

Symbol	Meaning
Pair constraints and risk bounds	
$\mathcal{T}_\pm, N_\pm, \mathcal{T}_{\sigma,k}$	Pools, sizes, prefixes
$m_+, m_-, m, \boldsymbol{m}$	Pair budgets
$A_{\boldsymbol{m}}, f_{\boldsymbol{m}}, v_e$	Matrix/map; difference
$u, l, \Delta;\ \lambda_e, \nu_e$	Thresholds; duals
$A^\dagger;\ \psi_+, \psi_-$	Comparator; losses
$L_\sigma, \widehat{L}_{\sigma,k}, e_\sigma(k)$	Pair means; errors
$B, Q;\ \xi_e, \xi_{0,e}, \gamma$	Bounds; slacks/weight
$\varphi(t), h(t)$	$t - \log t - 1$
$V_{\boldsymbol{m}}, H_{\boldsymbol{m}}, \widetilde{H}, \overline{H}_{\boldsymbol{m}}, G$	Violation/SVM bounds; slack factor
$n_{\min}, \alpha_S;\ w_0, b_0, s_0$	Class size/scaling; comparator
$\pi_c, \pi_0;\ C_K, s$	Class masses; partition growth
$\theta_+, \theta_-;\ r_n, \lambda_{c,n}, \kappa$	Fractions; rates
m_n, f_n, g_n, δ_n	Budgets, maps, confidence
$q_c^*, s_*;\ \vartheta$	Affine comparator
$\bar{X}_c, q_c^\dagger, s^\dagger$	Comparator statistics
$\vartheta_S, w_A^\dagger, b_S^\dagger;\ a, \chi, v$	Comparison constants
Extensions and proof notation	
$\widehat{F}_{\boldsymbol{m}}, F_S, O_S$	Objectives; optimum
Q_0, C_{λ_d}	Reference loss/budget
$V_{\boldsymbol{m}}^{\mathrm{avg}}, H_{\boldsymbol{m}}^{\mathrm{avg}}, \widetilde{\psi}_\sigma$	Averaged-loss bounds
$\kappa(q), \eta, g_0$	Gaussian quantities
$H_{\boldsymbol{m}}^k, B_{\boldsymbol{m}}^k$	Kernel-SVM bounds
$\mathbf{K}, \mathbf{K}^{+1/2}$	Gram; inverse square root
$\mathcal{V}, J_{\mathbf{K}};\ a, \alpha$	Kernel span/objective; weights
η_d, η_c, ϵ	Optimization errors; bias penalty
$T_S;\ \Lambda, \rho, M, Z_n$	Input set; net terms
$B_e, c_e, I_*, \mathcal{K}$	Normals, offsets; active set/cone
$E_{\mathrm{pair}}, E_{\mathrm{part}}$	Probability events
$\mu_c, \bar{v}_c, a_c^*$	Class law/moments
$p_{\min,S};\ \langle\cdot,\cdot\rangle_F$	Class mass; Frobenius product
W_c, W_{diff}	Class violation sums
p_-, p_1, p_2, x_q, η	Counterexample values
$a, D, C_{\mathrm{svm}};\ v_S(A)$	Constants; violations
$\widehat{Q}(s), \widehat{Q}(A), E_c$	Similarity loss; fit bound
$p_j, b_j, \alpha_j^0, c_{\mathrm{op}}$	Landmark weights/labels; norm constant

B PROOFS FOR SECTION 3

Throughout this appendix, population expectations are taken over fresh observations with the learned representation and classifier held fixed. A loss, a radius, or a norm can depend on the training sample. The generalization bounds use a single event involving the counts of a fixed partition; the comparisons on that event are deterministic. We first establish the properties of metric robustness and the generalization theorem, and then give the calculations for each example.

B.1 PROPERTIES OF METRIC ROBUSTNESS

Let $\Pi = \{C_1, \dots, C_K\}$ be a finite measurable label-pure partition of $\mathcal{Z}$, fixed independently of the training sample. For a fixed learned pair, write

$$h(x, y) = l_c(g, (f(x), y)), \qquad B_f(X_i, t) = \{x \in \mathcal{X} : d_f(X_i, x) \le t\}.$$

As in Definition 3.1, $B_f(X_i, \infty) = \mathcal{X}$. Every cell containing Z_i contains (X_i, Y_i), so the set defining a_i is nonempty. Its supremum is a well-defined element of $[0, \infty]$. The loss response is also well defined: its supremum is over a set containing X_i, and all of its terms lie in $[0, 1]$.

The local comparison. For every $z = (x, Y_i) \in C(Z_i)$, the definition of the supremum gives

$$d_f(X_i, x) \le a_i, \qquad |h(x, Y_i) - h(X_i, Y_i)| \le \omega_i(a_i).$$

The first inequality holds even when the supremum is not attained. When $a_i = \infty$, the second follows from the stipulated whole-space ball. Thus every same-cell comparison required by the generalization argument is included in the corresponding metric response. The response is nondecreasing in its radius, because the balls are nested, and $0 \le \varepsilon_{\mathrm{mr}} \le 1$.

The distance can be a pseudometric. If $d_f(x, x') = 0$, then $f(x) = f(x')$, so $h(x, y) = h(x', y)$ for every fixed label. Consequently, $\omega_i(0) = 0$. No injectivity of the representation is required for the preceding comparison.

Dependence on both outputs. The representation determines the balls and their radii; the classifier determines the loss values compared on those balls. Both enter the definition. For instance, let $\mathcal{X} = [0, 1]$, $f(x) = x$, and consider a training anchor $X_i = 0$ with label $+1$. Use clipped hinge loss and a cell whose input part is $[0, 1]$, so $a_i = 1$. The constant score $g_1(v) = 1$ has $\omega_i(t) = 0$ at every radius. The score $g_2(v) = 1 - v$ has loss x on this interval and therefore $\omega_i(t) = \min\{t, 1\}$. The same learned geometry can therefore have zero or maximal response, depending on the classifier. Definition 3.1 records this distinction directly.

Compensated rescaling. Suppose the evaluation loss has the prediction-based form as following $l_c(g, (v, y)) = \ell(g(v), y)$. For $c > 0$, put $f_c = cf$ and $g_c(v) = g(v/c)$. Then

$$g_c(f_c(x)) = g(f(x)), \qquad d_{f_c}(x, x') = c\, d_f(x, x'), \qquad a_{i,c} = c\, a_i.$$

The last identity follows by taking the supremum of the scaled distances; it includes $c \cdot \infty = \infty$. For every finite t, the ball for f_c at radius t is the ball for f at radius t/c. Pointwise losses are unchanged, so

$$\omega_{i,c}(t) = \omega_i(t/c), \qquad \omega_{i,c}(a_{i,c}) = \omega_i(a_i).$$

The second equality also holds at infinite radius, since both balls are $\mathcal{X}$. Averaging proves the rescaling assertion in Remark 3.4.

Refinement and labels. Let Π' refine Π, meaning that every cell of Π' is contained in a cell of Π. For the same sample and learned pair, $C'(Z_i) \subseteq C(Z_i)$ gives $a'_i \le a_i$. The functions ω_i are unchanged, since they are defined by the learned distance and losses, and their monotonicity gives

$$\varepsilon_{\mathrm{mr}}(f, g, S; \Pi') \le \varepsilon_{\mathrm{mr}}(f, g, S; \Pi).$$

A refinement can have more cells, increasing the counting term in the generalization bound. The comparison holds for the same learned pair evaluated with the two partitions.

Crossing an input partition with the finite label set produces a label-pure partition of the entire space $\mathcal{X} \times \mathcal{Y}$. This construction accommodates noisy labels: the points (x, y) and (x, y') with $y \ne y'$ belong to different cells. Each local comparison fixes the observed label of its anchor, as required in the definition.

B.2 Proof of Proposition 3.3

Proof. Let $T_S = \{X_1, \ldots, X_n\}$, and define

$$f_S(x) = \begin{cases} \operatorname{sign}(x), & x \in T_S, \\ -\operatorname{sign}(x), & x \notin T_S, \end{cases} \qquad \operatorname{sign}(0) = 1.$$

This is jointly Borel measurable in (S, x). Indeed, $\{(S, x) : x \in T_S\}$ is a finite union of the closed equality sets $\{x = X_i\}$, and the sign function is Borel measurable. The contrastive loss with scale four is

$$l_d(f, z, z') = \mathbf{1}\{y = y'\}\frac{|f(x) - f(x')|^2}{4} + \mathbf{1}\{y \neq y'\}\left[1 - \frac{|f(x) - f(x')|^2}{4}\right]_+.$$

Every training representation equals its training label. Hence the squared distance is zero for an equal-label training pair and four for an unequal-label training pair. Every summand in $\widehat{\mathcal{L}}_d(f_S)$ is therefore zero, including the diagonal summands.

Fix the training sample. Its input set T_S is finite, whereas the input distribution is nonatomic. A fresh input lies outside T_S with probability one, and consequently $f_S(X) = -Y$ almost surely. Two independent test points again have squared represented distance zero when their labels agree and four when they differ. Their pairwise loss is zero almost surely, proving $\mathcal{L}_d(f_S) = 0$. This expectation is evaluated at the fixed map learned from S.

The score $g(v) = v$ is never zero on the attained representations. At every training point its sign is Y_i, and at almost every test point its sign is $-Y$. Thus

$$\widehat{R}_{01}(g \circ f_S) = 0, \qquad R_{01}(g \circ f_S) = 1.$$

If both classes occur in S, fitting the represented training labels requires predicting $+1$ at representation $+1$ and -1 at representation -1. Every fresh point almost surely receives the opposite represented label, so every classifier satisfying these two interpolation conditions has population error one. This argument is pointwise in the classifier and therefore applies simultaneously to all such measurable classifiers. The probability that both classes occur is $1 - 2^{1-n}$. □

Metric robustness of the construction. Fix any partition allowed in Definition 3.1. Almost surely, no training observation belongs to a cell of zero μ-measure: the union of the finitely many such cells has measure zero. Consequently, every occupied cell has positive measure. Removing the finite set of inputs T_S from an occupied cell leaves the same positive measure, and $Y = \operatorname{sign}(X)$ holds almost surely. For each i, there is therefore an $x \notin T_S$ satisfying

$$(x, Y_i) \in C(Z_i), \qquad \operatorname{sign}(x) = Y_i.$$

Its representation is $-Y_i$, whereas $f_S(X_i) = Y_i$. All represented distances are at most two, so this witness proves $a_i = 2$. For the displayed score, the zero–one loss changes from zero at the anchor to one at this point. Hence $\omega_i(a_i) = 1$ for every i, and $\varepsilon_{\mathrm{mr}} = 1$. If both classes are observed, every interpolating classifier has the same two predictions and the same conclusion. The almost-sure statement is for each fixed partition, in agreement with the quantifiers following Proposition 3.3.

B.3 Proof of Theorem 3.5

We prove a finite-partition statement for a general bounded loss. It also supplies the concentration event used by the pairwise result and every example.

Let $h : \mathcal{Z} \to [0, B]$ be measurable, with $0 < B < \infty$. The function can be selected after observing the sample. Suppose nonnegative numbers c_i satisfy

$$|h(z) - h(Z_i)| \leq c_i \qquad \text{for every } z \in C(Z_i).$$

There is an event of probability at least $1 - \delta$, depending only on the fixed partition counts, on which simultaneously for all such h and (c_i),

$$\left|\mathbb{E}_\mu h - \frac{1}{n}\sum_{i=1}^{n} h(Z_i)\right| \leq \frac{1}{n}\sum_{i=1}^{n} \min\{B, c_i\} + B\tau_n, \qquad \tau_n = \sqrt{\frac{K \log 2 + \log(1/\delta)}{2n}}. \tag{23}$$

If $B = 0$, the loss is identically zero and this assertion holds deterministically.

Proof of the finite-partition statement. Write

$$p_j = \mu(C_j), \qquad n_j = \sum_{i=1}^{n} \mathbf{1}\{Z_i \in C_j\}, \qquad \widehat{p}_j = \frac{n_j}{n}, \qquad T = \frac{1}{2}\sum_{j=1}^{K} |p_j - \widehat{p}_j|.$$

Since $\sum_j (p_j - \widehat{p}_j) = 0$, the sum of the positive differences equals the negative of the sum of the negative differences. Both are T. It follows that

$$T = \max_{J \subseteq \{1,\dots,K\}} \sum_{j \in J} (p_j - \widehat{p}_j).$$

For a fixed subset J, the variables $U_i = \mathbf{1}\{Z_i \in \bigcup_{j\in J} C_j\}$ are independent Bernoulli variables of mean $p_J = \sum_{j\in J} p_j$. The one-sided bounded-variable inequality gives

$$\Pr\left\{ p_J - \frac{1}{n}\sum_i U_i > t \right\} \le e^{-2nt^2} \qquad (t > 0).$$

For completeness, this estimate follows directly from the exponential moment. If U is Bernoulli with mean p, the log moment generating function of $p - U$ has value and first derivative zero at zero, and its second derivative is a Bernoulli variance under exponential tilting, at most $1/4$. Integrating twice gives $\log \mathbb{E}e^{s(p-U)} \le s^2/8$ for $s \ge 0$. Independence and Markov's inequality then give $\exp(-nst + ns^2/8)$, minimized at $s = 4t$. The degenerate cases $p = 0, 1$ satisfy the same estimate.

A union bound over the 2^K subsets yields

$$\Pr\{T > \tau_n\} \le 2^K e^{-2n\tau_n^2} = \delta.$$

With probability one, all cells of population mass zero have empirical mass zero. Let E be the intersection of this probability-one event with $\{T \le \tau_n\}$. It is measurable and has probability at least $1 - \delta$.

Fix a sample in E, and fix any admissible h. Define its population mean within cell j by

$$\bar{h}_j = \begin{cases} p_j^{-1} \int_{C_j} h(z)\, d\mu(z), & p_j > 0, \\ 0, & p_j = 0, \end{cases}$$

and its empirical mean within an occupied cell by $\widehat{h}_j = n_j^{-1} \sum_{i: Z_i \in C_j} h(Z_i)$; set $\widehat{h}_j = 0$ when $n_j = 0$. Every occupied cell has $p_j > 0$ on E. Integrating the local inequality over that cell gives

$$|\bar{h}_j - h(Z_i)| \le \min\{B, c_i\} \qquad \text{whenever } Z_i \in C_j.$$

We now decompose the difference of risks exactly:

$$\begin{aligned} \mathbb{E}_\mu h - \frac{1}{n}\sum_i h(Z_i) &= \sum_j p_j \bar{h}_j - \sum_j \widehat{p}_j \widehat{h}_j \\ &= \sum_j (p_j - \widehat{p}_j)\bar{h}_j + \sum_j \widehat{p}_j (\bar{h}_j - \widehat{h}_j). \end{aligned}$$

The first sum has absolute value at most BT, because $0 \le \bar{h}_j \le B$, and the positive and negative parts of the coefficient vector each have total mass T. For the second sum,

$$\left| \sum_j \widehat{p}_j (\bar{h}_j - \widehat{h}_j) \right| \le \frac{1}{n}\sum_{i=1}^{n} |\bar{h}_{C(Z_i)} - h(Z_i)| \le \frac{1}{n}\sum_{i=1}^{n} \min\{B, c_i\}.$$

Here $\bar{h}_{C(Z_i)}$ denotes the mean in the cell containing Z_i. Empty empirical cells have zero weight in this sum; their population contributions are already included in the first sum. Combining the two bounds with $T \le \tau_n$ proves equation 23. Every step after fixing $S \in E$ is deterministic, so the same event works for every admissible loss and local bound. □

Proof of Theorem 3.5. Set $h(x, y) = l_c(g, (f(x), y))$ and $B = 1$. For $z = (x, y) \in C(Z_i)$, label purity gives $y = Y_i$, and the local comparison established above gives

$$|h(z) - h(Z_i)| \leq \omega_i(a_i).$$

Take $c_i = \omega_i(a_i)$ in equation 23. Since each response is at most one, the average local term is exactly $\varepsilon_{\mathrm{mr}}$. This proves equation 5 on E, simultaneously for every measurable learned pair.

If the evaluation loss upper-bounds the misclassification indicator pointwise, integration gives $R_{01}(g \circ f) \leq \mathcal{L}_{c,f}(g)$. The one-sided consequence of equation 5 is then equation 6. For clipped hinge, a wrong sign implies a nonpositive signed margin, so its value is one. At a score tie, its value is also one, whatever the fixed prediction rule. This verifies the stated pointwise domination, including ties.

The event E depends only on S and Π. Once it holds, the result applies to every output of both learning stages and every realized value of algorithmic randomness. In particular, use of the same observations in both stages requires no further independence assumption. □

Neighborhoods and partition cells. The inclusion used in the proof is

$$\{x : (x, Y_i) \in C(Z_i)\} \subseteq B_f(X_i, a_i).$$

It can be strict. A representation can place inputs from distinct partition cells within the same learned-distance ball. The loss is evaluated throughout this ball at the fixed label Y_i, including at inputs whose actual labels might differ. This produces an upper bound on the same-cell variation needed in the theorem. The whole-ball response permits the norm and decision-margin calculations below, because those calculations depend on represented distance alone. Replacing the ball by the cell would usually give a smaller response but a different quantity. Infinite radii are permitted in the theorem because the evaluation loss is bounded.

Measurability and probability statements. Measurability of the learned losses suffices for their population integrals. Supremums over uncountable sets need not be Borel measurable as functions of the training sample merely because the functions being maximized are measurable. The theorem is therefore stated using the measurable event E, and its inequality holds pointwise on every sample in that event. Probability is assigned to E, while exact responses enter the deterministic inequality on that event.

When the bound is reported as a measurable statistic, one may use measurable majorants. If $\bar{a}_i \geq a_i$, $\bar{\omega}_i(t) \geq \omega_i(t)$, and $\bar{\omega}_i$ is nondecreasing with measurable evaluation $\bar{\omega}_i(\bar{a}_i)$, then

$$\varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i \bar{\omega}_i(\bar{a}_i).$$

The norm, fitted-margin, finite-tree, and prototype expressions below provide such majorants whenever the fitted parameters are measurable.

For a loss in $[0, B]$, use the unnormalized response with values in $[0, B]$; the same proof gives the average response plus $B\tau_n$. If a partition is selected from a finite family fixed before observing the sample, apply the theorem to every partition with confidence levels whose sum is at most δ. Their events then hold simultaneously with probability at least $1 - \delta$. For a single partition, all statements above use that partition fixed in advance.

The Lipschitz specialization. Under equation 7, every point in a ball of finite radius t satisfies

$$|h(x, Y_i) - h(X_i, Y_i)| \leq Lt.$$

The loss range also bounds this difference by one. Taking the supremum gives $\omega_i(t) \leq \min\{1, Lt\}$. Since $a < \infty$, all radii can be substituted into this estimate, and

$$\varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i \min\{1, La_i\} \leq \min\{1, La\}.$$

This proves equation 8, including $L = 0$. The argument fixes the learned pair, so L can depend on that pair.

B.4 Proof of Corollary 3.6

Proof. Fix a sample in the event E from the preceding proof. First suppose $B_d > 0$; if $B_d = 0$, the metric loss is identically zero. If $z = (x, y) \in C(Z_i)$ and $z' = (x', y') \in C(Z_j)$, their labels satisfy $y = Y_i$, $y' = Y_j$. The reverse triangle inequality for the learned pseudometric gives

$$\begin{aligned} |d_f(X_i, X_j) - d_f(x, x')| &\le d_f(X_i, x) + d_f(X_j, x') \\ &\le a_i + a_j \le 2a. \end{aligned}$$

The first inequality follows by applying the triangle inequality in both directions. Since the loss function for these two labels is U-Lipschitz on attained distances,

$$|l_d(f, Z_i, Z_j) - l_d(f, z, z')| \le \min\{B_d, U(a_i + a_j)\} \le \min\{B_d, 2Ua\} =: c.$$

On the product partition $\Pi \times \Pi$, the population masses are $p_j p_k$, and the masses of the empirical product measure are $\widehat{p}_j \widehat{p}_k$. Their total variation satisfies

$$\begin{aligned} \frac{1}{2} \sum_{j,k} |p_j p_k - \widehat{p}_j \widehat{p}_k| &\le \frac{1}{2} \sum_{j,k} \big(|p_j - \widehat{p}_j| p_k + \widehat{p}_j |p_k - \widehat{p}_k| \big) \\ &= 2T. \end{aligned}$$

The deterministic decomposition in the proof of equation 23 now applies to the loss $(z, z') \mapsto l_d(f, z, z')$. To see the empirical weights explicitly, a product cell $C_j \times C_k$ contains $n_j n_k$ ordered index pairs. Averaging their losses and weighting by $n_j n_k / n^2$ gives exactly the empirical risk in equation 2. In each occupied product cell, integration of the preceding local inequality bounds the difference between its population and empirical conditional means by c. Cells with zero population mass cannot be occupied on E.

The resulting local contribution is at most c, and the difference of product masses contributes at most $2B_d T$. Since $T \le \tau_n$,

$$|\mathcal{L}_d(f) - \widehat{\mathcal{L}}_d(f)| \le \min\{B_d, 2Ua\} + 2B_d \tau_n,$$

which proves equation 9. The concentration step used only the n independent observations. No independence among the n^2 empirical pairs has been invoked.

If the loss is U-Lipschitz as a function of squared distance and all attained distances are at most D, let $s = d_f(X_i, X_j)$ and $t = d_f(x, x')$. Then

$$|s^2 - t^2| = (s + t)|s - t| \le 2D(a_i + a_j) \le 4Da.$$

Replacing c by $\min\{B_d, 4UDa\}$ in the same decomposition proves the squared-distance assertion. The case $D = 0$ is included. □

B.5 Verification of the examples

B.5.1 Partitions from covers

Fix $r > 0$. Compactness of $\mathcal{X}$ implies total boundedness, so finitely many radius-$r/2$ balls cover $\mathcal{X}$. Take a cover of minimum cardinality $N = \mathcal{N}(r/2, \mathcal{X}, \|\cdot\|_{\mathcal{X}})$, with a fixed ordering of its balls $B_1, \ldots, B_N$. Define

$$V_1 = \mathcal{X} \cap B_1, \qquad V_j = (\mathcal{X} \cap B_j) \setminus \bigcup_{k<j} B_k \quad (j \ge 2),$$

and discard empty sets. These cells are Borel measurable, disjoint, and cover $\mathcal{X}$. For two inputs in one cell, their distances to its ball center are at most $r/2$, so their mutual distance is at most r. Taking $V_j \times \{y\}$ for every label gives a label-pure partition with

$$K \le |\mathcal{Y}| \mathcal{N}(r/2, \mathcal{X}, \|\cdot\|_{\mathcal{X}}).$$

The space, cover, ordering, and radius are fixed before sampling. The partition can depend on a deterministic sample size through a preselected radius r_n.

When $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \le R\}$, the explicit estimate

$$\mathcal{N}(r/2, \mathcal{X}, \|\cdot\|_2) \le (1 + 4R/r)^d$$

follows from a packing argument. A maximal set of centers in $\mathcal{X}$ with pairwise distances strictly greater than $r/2$ is finite and covers $\mathcal{X}$ by radius-$r/2$ balls; otherwise another center could be added. The open radius-$r/4$ Euclidean balls around its centers are disjoint and contained in the radius-$R + r/4$ ball around the origin. Comparison of their volumes gives the claimed cardinality bound. Existence of a finite maximal set follows, for example, from the same volume bound applied during successive selection.

B.5.2 Mahalanobis metrics and affine classifiers

Proof of Example 3.7. For a positive semidefinite matrix A, its symmetric square root satisfies

$$d_f(x, x')^2 = \|A^{1/2}(x - x')\|_2^2 = (x - x')^\top A(x - x') \le \|A\|_{\mathrm{op}} \|x - x'\|_2^2.$$

Thus the fixed covering partition yields $a_i \le r\sqrt{\|A\|_{\mathrm{op}}}$. Let the eigenvalues of A be $\lambda_1, \dots, \lambda_d \ge 0$. Then

$$\|A\|_{\mathrm{op}} = \max_j \lambda_j \le \sum_j \lambda_j = \mathrm{tr}(A), \qquad \|A\|_{\mathrm{op}} \le \|A\|_F \le \sum_{j,k} |A_{jk}|.$$

The last inequality is the elementary comparison between the Euclidean and ℓ_1 norms of the vector of matrix entries. For each of the three penalties in the example, nonnegativity of $\widehat{J}_d$ gives

$$\|A\|_{\mathrm{op}} \le \Omega(A) \le E_d/\lambda_d.$$

Consequently $a_i \le r\sqrt{E_d/\lambda_d}$. A feasible reference matrix bounds the objective of a minimizer by direct comparison. If an approximate output has objective at most the reference objective plus η_d, their sum can be used as E_d.

For binary labels define $\chi(u) = \min\{1, [1-u]_+\}$. Positive-part and truncation are both 1-Lipschitz, so χ is 1-Lipschitz. For an affine score $g(v) = \langle w, v\rangle + b$,

$$\begin{aligned} |\chi(yg(f(x))) - \chi(yg(f(x')))| &\le |g(f(x)) - g(f(x'))| \\ &= |\langle w, f(x) - f(x')\rangle| \\ &\le \|w\|_2 d_f(x, x'). \end{aligned}$$

The intercept cancels. Together with boundedness of the loss, this proves $\omega_i(t) \le \min\{1, \|w\|_2 t\}$ for finite t.

The zero score $(w, b) = (0, 0)$ has average hinge loss one and no penalty. Therefore every exact SVM minimizer satisfies

$$\frac{1}{n}\sum_i [1 - Y_i g(f(X_i))]_+ + \frac{\lambda_c}{2}\|w\|_2^2 \le 1, \qquad \|w\|_2 \le \sqrt{2/\lambda_c}.$$

Substituting the radius and response bounds into the definition gives

$$\varepsilon_{\mathrm{mr}} \le \frac{1}{n}\sum_i \min\{1, \|w\|_2 a_i\} \le \min\left\{1, r\sqrt{\frac{2E_d}{\lambda_d \lambda_c}}\right\}.$$

Theorem 3.5 proves equation 10. If the SVM objective is at most its minimum plus η_c, the same argument gives $\|w\|_2 \le \sqrt{2(1 + \eta_c)/\lambda_c}$. □

Squared Frobenius regularization. For $\Omega(A) = \|A\|_F^2$, the objective comparison gives $\|A\|_F \le (E_d/\lambda_d)^{1/2}$. Hence

$$a_i \le r\sqrt{\|A\|_{\mathrm{op}}} \le r(E_d/\lambda_d)^{1/4}.$$

This is the radius to combine with the same classifier responses when the metric penalty is squared.

Squared-loss Lasso. For binary labels, consider the explicitly normalized objective

$$\frac{1}{n}\sum_i \big(Y_i - \langle w, f(X_i)\rangle - b\big)^2 + \lambda_c \|w\|_1.$$

At the zero score its value is $n^{-1}\sum_i Y_i^2 = 1$. Any minimizer therefore has $\|w\|_1 \le 1/\lambda_c$, and $\|w\|_2 \le \|w\|_1$. Using clipped hinge for evaluation, the affine response just proved yields

$$\varepsilon_{\mathrm{mr}} \le \min\left\{1, \frac{r}{\lambda_c}\sqrt{\frac{E_d}{\lambda_d}}\right\}.$$

This is a bounded-loss generalization calculation for a score trained by squared loss. In particular, no boundedness of the squared training loss over the whole input space is needed. A different multiplicative normalization of the squared training loss changes the zero-score comparison and should be carried through this norm bound.

Existence of the fitted affine heads. For the finite-dimensional Mahalanobis representation, let $M = \max_i \|f(X_i)\|_2 < \infty$. The SVM objective is continuous. Its sublevel set at the zero-score value one satisfies $\|w\|_2 \le \sqrt{2/\lambda_c}$. If both labels occur, choose one observation of each label. Nonnegativity of all hinge terms implies that each individual hinge is at most n. The positive observation therefore gives

$$b \ge 1 - n - \langle w, f(X_i)\rangle \ge 1 - n - M\sqrt{2/\lambda_c},$$

whereas the negative observation gives

$$b \le n - 1 - \langle w, f(X_j)\rangle \le n - 1 + M\sqrt{2/\lambda_c}.$$

Thus $|b| \le n - 1 + M\sqrt{2/\lambda_c}$ on this sublevel set. It is nonempty, closed, and bounded in a finite-dimensional space, so it is compact; continuity gives an attained global minimum there. If only one label y_0 occurs, $w = 0$ and $b = y_0$ have zero hinge and zero penalty and attain the minimum zero.

For the squared-loss Lasso, the sublevel set at one has $\|w\|_1 \le 1/\lambda_c$, and each squared residual is at most n. At any training observation,

$$|b| \le |Y_i| + |\langle w, f(X_i)\rangle| + |Y_i - \langle w, f(X_i)\rangle - b| \le 1 + M/\lambda_c + \sqrt{n}.$$

This sublevel set is likewise nonempty and compact, and the continuous objective attains its global minimum on it. These existence arguments justify the exact minimizers used in Example 3.7; the preceding norm bounds hold for every such minimizer.

Fitted-margin refinement. Write $m_i = Y_i g(f(X_i))$. Suppose score variation on $B_f(X_i, t)$, with finite t, is at most a finite $b_i(t) \ge 0$. Every signed margin on this ball lies in $[m_i - b_i(t), m_i + b_i(t)]$. Since χ is nonincreasing, the possible losses lie in $[\chi(m_i + b_i(t)), \chi(m_i - b_i(t))]$. It follows that

$$\begin{aligned}\omega_i(t) &\le \max\{\chi(m_i - b_i(t)) - \chi(m_i),\ \chi(m_i) - \chi(m_i + b_i(t))\} \\ &\le \min\{1, b_i(t)\}.\end{aligned} \tag{24}$$

The second inequality follows from the range and Lipschitz constant of χ. If $m_i \ge 1 + b_i(a_i)$, then every margin in the relevant ball is at least one, and its loss response is exactly zero. For an affine head, take $b_i(t) = \|w\|_2 t$. The kernel calculation below supplies another valid choice. Thus the average response can be strictly smaller than the bound obtained by applying the same worst-case sensitivity at every anchor.

B.5.3 LogDet geometry

Let $A, A_0 \succ 0$ and $B = A_0^{-1/2} A A_0^{-1/2}$. Similarity and the cyclic property of trace give

$$D_{\mathrm{ld}}(A, A_0) = \mathrm{tr}(B) - \log\det B - d = \sum_{j=1}^{d} h(b_j), \qquad h(b) = b - \log b - 1,$$

where $b_1, \dots, b_d > 0$ are the eigenvalues of B. The derivative $h'(b) = 1 - 1/b$ is negative on $(0, 1)$, zero at one, and positive on $(1, \infty)$; also $h(1) = 0$. Hence all summands are nonnegative. If their

sum is at most C, then $C \geq 0$ and every summand is at most C. The restriction of h to $[1, \infty)$ is continuous, strictly increasing after one, and tends to infinity. Thus there is a unique $b(C) \geq 1$ with $h(b(C)) = C$, including $b(0) = 1$.

An eigenvalue exceeding $b(C)$ would have $h(b_j) > C$, which is impossible. Consequently

$$B \preceq b(C)I, \qquad A \preceq b(C)A_0, \qquad \|A\|_{\mathrm{op}} \leq b(C)\|A_0\|_{\mathrm{op}}.$$

The Mahalanobis calculation now gives

$$a_i \leq r\sqrt{b(C)\|A_0\|_{\mathrm{op}}}.$$

Together with the affine response, this proves the LogDet bound stated after Example 3.7. Section 4 establishes objective comparisons that make C uniform over the selected pair budgets.

B.5.4 Kernel representations and classifiers

Proof of Example 3.8. Let ϕ map into a Hilbert space $\mathcal{H}$, and let A be a bounded self-adjoint positive operator on $\mathcal{H}$. Its positive square root satisfies $\|A^{1/2}v\|_{\mathcal{H}} \leq \sqrt{\|A\|_{\mathrm{op}}}\|v\|_{\mathcal{H}}$. Therefore

$$d_f(x, x') = \|A^{1/2}(\phi(x) - \phi(x'))\|_{\mathcal{H}} \leq \sqrt{\|A\|_{\mathrm{op}}}L_\phi\|x - x'\|_{\mathcal{X}}.$$

Taking a same-cell supremum proves $a_i \leq rL_\phi\sqrt{\|A\|_{\mathrm{op}}}$.

For the Gaussian kernel on the representation space, let $k_v = k(v, \cdot)$. The reproducing property and the Hilbert-space Cauchy–Schwarz inequality give

$$|g(v) - g(v')| = |\langle g, k_v - k_{v'}\rangle_{\mathcal{H}_k}| \leq \|g\|_{\mathcal{H}_k}\|k_v - k_{v'}\|_{\mathcal{H}_k}.$$

The kernel-section norm is determined exactly by the kernel:

$$\begin{aligned}\|k_v - k_{v'}\|^2_{\mathcal{H}_k} &= k(v, v) + k(v', v') - 2k(v, v') \\ &= 2\left(1 - e^{-\|v - v'\|^2_{\mathcal{H}}/(2\sigma^2)}\right).\end{aligned}$$

Since $1 - e^{-u} \leq u$ for $u \geq 0$,

$$|g(v) - g(v')| \leq \|g\|_{\mathcal{H}_k}\sqrt{2\left(1 - e^{-\|v - v'\|^2_{\mathcal{H}}/(2\sigma^2)}\right)} \leq \frac{\|g\|_{\mathcal{H}_k}}{\sigma}\|v - v'\|_{\mathcal{H}}.$$

Clipped hinge is 1-Lipschitz in the score, so taking the supremum over a radius-t learned-distance ball gives both

$$\omega_i(t) \leq \min\left\{1, \|g\|_{\mathcal{H}_k}\sqrt{2(1 - e^{-t^2/(2\sigma^2)})}\right\} \leq \min\{1, \|g\|_{\mathcal{H}_k}t/\sigma\}.$$

This proves the response asserted in the example and the sharper response promised there. For a score $g(v) + b$ with an unpenalized intercept, the same difference calculation holds because the intercept cancels. □

Regularization and Gaussian input features. If A is positive and trace class, its nonnegative eigenvalues give $\|A\|_{\mathrm{op}} \leq \mathrm{tr}(A)$. A nonnegative metric objective with trace penalty $\lambda_d\,\mathrm{tr}(A)$ and value at most E_d therefore gives $a_i \leq rL_\phi\sqrt{E_d/\lambda_d}$. Similarly, a kernel SVM objective with penalty $\lambda_c\|g\|^2_{\mathcal{H}_k}/2$ and value at most one gives $\|g\|_{\mathcal{H}_k} \leq \sqrt{2/\lambda_c}$. Combining these two comparisons yields

$$\varepsilon_{\mathrm{mr}} \leq \min\left\{1, \frac{rL_\phi}{\sigma}\sqrt{\frac{2E_d}{\lambda_d\lambda_c}}\right\}.$$

For the canonical feature map of an input Gaussian kernel of bandwidth $\sigma_0 > 0$, the same kernel-section computation, with $v = x$ and $v' = x'$, gives

$$\|\phi(x) - \phi(x')\|^2 = 2(1 - e^{-\|x - x'\|^2_2/(2\sigma_0^2)}) \leq \|x - x'\|^2_2/\sigma_0^2.$$

Thus $L_\phi \leq 1/\sigma_0$. These calculations also hold for a finite-dimensional feature map with the stated Lipschitz constant.

B.5.5 NEAREST-NEIGHBOR CLASSIFICATION

Proof of Example 3.9. Fix the learned list of $N \geq k$ prototypes after training. Write their representations and labels as $(v_j, b_j)_{j=1}^{N}$; repetitions are allowed. Break distance ties by a fixed order of prototype indices and vote ties by a fixed order of classes. All distances below are measured in the norm of the embedding space. For every prototype index j and $x \in B_f(X_i, t)$, with $t < \infty$, the reverse triangle inequality gives

$$\big|\|f(x) - v_j\| - \|f(X_i) - v_j\|\big| \leq \|f(x) - f(X_i)\| \leq t. \tag{25}$$

The prototypes can therefore be represented training observations or points learned directly in the embedding space.

Put $c_i = g(f(X_i))$, $h = \lfloor k/2 \rfloor + 1$, and $q = k - h + 1$. The integer h is a strict majority of k, and $h + q = k + 1$. Order the distances to prototypes labeled c_i and to prototypes with another label separately, using the same fixed index order for ties. Let s_i be the h-th distance in the first list and o_i the q-th distance in the second. A missing order statistic is defined to be ∞. This convention causes no indeterminate subtraction, since the condition in the example compares $s_i + 2t$ directly with o_i.

Suppose that $t < \infty$ and $s_i + 2t < o_i$. In particular, $s_i < \infty$, so at least h prototypes carry label c_i. For every $x \in B_f(X_i, t)$, at least h of these prototypes have distance at most $s_i + t$, by equation 25. If there are at least q oppositely labeled prototypes, their q-th nearest distance at x is at least $o_i - t$. To see this, a q-th distance smaller than $o_i - t$ would give q prototype distances smaller than o_i at X_i, contradicting its definition. Hence

$$s_i + t < o_i - t.$$

There are at most $q - 1$ opposing prototypes at distance no greater than the h-th nearest prototype labeled c_i. If there are fewer than q opposing prototypes in the whole list, the same conclusion holds directly, and $o_i = \infty$.

In either case, the h-th nearest prototype labeled c_i has rank at most $h + q - 1 = k$ in the combined ordered list. The first k neighbors consequently contain at least h votes for c_i. This is a strict majority, so $g(f(x)) = c_i$ throughout the ball, independently of the vote tie rule. With the evaluation label fixed at Y_i, the zero–one loss is constant on that ball, whether or not the anchor prediction is correct. Therefore

$$\omega_i(t) = 0 \quad \text{if } t < \infty \text{ and } s_i + 2t < o_i.$$

In every other case, the loss response is at most one. In particular, the bound remains valid for an infinite radius. Thus, with

$$u_i = \mathbf{1}\{a_i = \infty \text{ or } s_i + 2a_i \geq o_i\}, \qquad \varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i u_i,$$

Theorem 3.5 gives, on its probability event,

$$|R_{01}(g \circ f) - \widehat{R}_{01}(g \circ f)| \leq \frac{1}{n} \sum_i u_i + \tau_n. \tag{26}$$

Here the condition $a_i = \infty$ is checked first, so no arithmetic involving $2a_i$ is needed in that case. All finite-list order statistics are measurable when the fitted prototype parameters are measurable. Since the failure indicator is nondecreasing in the radius, measurable upper bounds for a_i can be substituted. If prototypes are selected from S, the empirical risk evaluates the final rule on that same sample, including self-neighbors when present. □

The one-neighbor case. For $k = 1$, one has $h = q = 1$. The quantities s_i and o_i are the nearest distances to a prototype with the anchor's predicted label and with any other label, respectively. Thus $s_i + 2a_i < o_i$ certifies label stability on the whole ball. In particular, when all training observations are prototypes and their representations separate the observed labels, the self-neighbor gives $s_i = 0$. It then suffices that every differently labeled prototype have distance greater than $2a_i$. A nearest neighbor can change identity within the ball while the predicted label remains unchanged.

A nonvacuous example with all training points as prototypes. Let $\mathcal{X} = [-2, -1] \cup [1, 2]$, let X be uniformly distributed on this union, and set $Y = -1$ on the first interval and $Y = +1$ on the second. Use the identity representation and all n training observations as prototypes. Fix any positive integer k and let $h = \lfloor k/2 \rfloor + 1$. Partition each interval into four disjoint subintervals of length $1/4$, assigning shared endpoints to one adjoining subinterval. Taking products of these eight subintervals with the two labels gives a fixed label-pure partition of $\mathcal{Z}$ with $K = 16$. Every radius satisfies $a_i \le 1/4$.

Suppose that each class occurs at least h times. At any input, at least h prototypes of its true class are at distance at most one, whereas all prototypes of the other class are at distance at least two. The k-NN rule therefore predicts the true class everywhere on $\mathcal{X}$. In particular, $c_i = Y_i$, and

$$s_i \le 1, \qquad o_i \ge 2, \qquad s_i + 2a_i \le \frac{3}{2} < 2 \le o_i.$$

If an opposing order statistic is missing, the last inequality uses $o_i = \infty$ and is still valid. Hence all u_i vanish, $\varepsilon_{\mathrm{mr}} = 0$, and both empirical and population classification errors are zero. When $n \ge 4h$, the binomial class counts and Hoeffding's inequality give

$$\Pr\{\min(n_{-1}, n_{+1}) < h\} \le 2e^{-n/8}.$$

Thus the zero-response property holds with probability tending to one for every fixed k. On its intersection with the event of Theorem 3.5, the generalization-gap bound reduces to

$$\tau_n = \sqrt{\frac{16 \log 2 + \log(1/\delta)}{2n}}.$$

That intersection has probability at least $1 - \delta - 2e^{-n/8}$. The exact zero population error established above uses only the class-count event, rather than the concentration bound.

Comparison with a neighbor-index gap. A sufficient condition for preserving the entire set of neighbor indices is $d_{(k+1)}(X_i) - d_{(k)}(X_i) > 2a_i$, when $k < n$. This can be much stronger than preserving the predicted label. Indeed, let n_j denote the number of training observations in partition cell C_j, and take every training point as a prototype. If $n_j \ge k + 1$, each anchor in C_j has at least $k + 1$ prototypes within its defining radius a_i, including itself. Consequently,

$$d_{(k+1)}(X_i) - d_{(k)}(X_i) \le a_i \le 2a_i.$$

At most kK training observations belong to cells containing at most k observations. The resulting index-gap upper bound therefore satisfies

$$\frac{1}{n} \sum_i \mathbf{1}\{d_{(k+1)}(X_i) - d_{(k)}(X_i) \le 2a_i\} \ge \max\{0, 1 - kK/n\}.$$

For fixed k, K, this quantity tends to one, including in the preceding separated-class example. The label-based calculation instead gives zero local responses on an event of probability tending to one. It controls the classification decision directly and allows neighbors with the same label to exchange their order.

B.5.6 Hölder responses and decision boundaries

Hölder scores. Suppose $a < \infty$ and, on the represented inputs,

$$|g(v) - g(v')| \le H\|v - v'\|^\alpha, \qquad 0 \le H < \infty, \quad 0 < \alpha < 1.$$

Clipped hinge is 1-Lipschitz in the score, so for finite t,

$$\omega_i(t) \le \min\{1, Ht^\alpha\}, \qquad \varepsilon_{\mathrm{mr}} \le \frac{1}{n} \sum_i \min\{1, Ha_i^\alpha\}.$$

To give a concrete response of this type, take $\mathcal{X} = [0, 1]$, $f(x) = x$, and $g(v) = \sqrt{[v]_+} - 1/2$. For $u \ge v \ge 0$,

$$(\sqrt{u} - \sqrt{v})^2 \le (\sqrt{u} - \sqrt{v})(\sqrt{u} + \sqrt{v}) = u - v.$$

By symmetry this proves $|\sqrt{u} - \sqrt{v}| \le \sqrt{|u - v|}$; composing with the 1-Lipschitz positive-part map proves that g is $1/2$-Hölder on $\mathbb{R}$ with constant one. For label -1, its clipped hinge loss on $[0, 1]$ is $\min\{1, 1/2 + \sqrt{v}\}$. At $0 < v \le 1/4$, the change from its value at zero is $\sqrt{v}$. The ratio of this change to v equals $1/\sqrt{v}$ and is unbounded as $v \downarrow 0$. Its Hölder response gives the finite bound $\varepsilon_{\mathrm{mr}} \le \min\{1, \sqrt{r}\}$ for the covering partition.

Distance to a prediction change. For any fixed prediction rule $h = g \circ f : \mathcal{X} \to \mathcal{Y}$, define

$$\gamma_i = \inf\{d_f(X_i, x) : h(x) \neq h(X_i)\}, \qquad \inf \varnothing = \infty.$$

If $t < \gamma_i$, no point in the closed radius-t ball can have a different prediction. Its zero–one loss at the fixed label Y_i is then constant, whether that prediction is correct or incorrect. Hence

$$\omega_i(a_i) \leq \mathbf{1}\{\gamma_i \leq a_i\}.$$

The weak inequality on the right is needed because the infimum need not be attained; at equality we use the bound one. The convention $\infty \leq \infty$ also makes this a valid, possibly conservative bound at an infinite radius. A measurable lower bound on γ_i, combined with a measurable upper bound on a_i, gives a measurable upper bound on the response.

Decision trees. Consider a finite axis-aligned tree on a Euclidean representation. Follow the path of $v_i = f(X_i)$ from root to leaf. For a split on coordinate j at threshold u, its distance in that coordinate from the anchor is $|v_{ij} - u|$. Let b_i be the minimum of these distances along the path, taking the minimum of an empty set to be ∞. If $\|v - v_i\|_2 < b_i$, then at every split on the path,

$$|v_j - v_{ij}| \leq \|v - v_i\|_2 < |v_{ij} - u|.$$

Thus v_j stays on the same strict side of that threshold as v_{ij}. Induction from the root shows that v follows the entire original path and reaches the same leaf. At a threshold anchor, $b_i = 0$ gives the bound one for every fixed threshold-tie convention. Since no prediction change is possible at represented distance less than b_i, we have $b_i \leq \gamma_i$, and therefore

$$\omega_i(a_i) \leq \mathbf{1}\{b_i \leq a_i\}, \qquad \varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i \mathbf{1}\{b_i \leq a_i\}.$$

Adjacent leaves may have the same prediction, so this path calculation can be conservative. For a learned finite tree with measurable split parameters and tie rules, the path and its finite minimum are measurable. Using $\bar{a}_i \geq a_i$ gives the measurable bound $n^{-1} \sum_i \mathbf{1}\{b_i \leq \bar{a}_i\}$.

Binary score margins for zero–one evaluation. If a score varies by at most $b_i(t) < \infty$ on $B_f(X_i, t)$, the strict condition $|g(f(X_i))| > b_i(t)$ keeps its sign unchanged throughout the ball. Thus

$$\omega_i(t) \leq \mathbf{1}\{|g(f(X_i))| \leq b_i(t)\}$$

for zero–one evaluation with fixed score ties. For the affine head of a neural representation, $b_i(t) = \|w\|_2 t$. The radius estimate proved below then gives

$$\varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i \mathbf{1}\{|g(f_\theta(X_i))| \leq rP(\theta)\|w\|_2\}.$$

This last expression uses absolute prediction margins, whereas the clipped-hinge refinement in equation 24 uses signed margins relative to the true label.

B.6 The objective-to-risk calculation in Example 3.10

All statements in this subsection assume $|\mathcal{Y}| \geq 2$ and $n_c > 0$ for every class. They hold on the fixed partition-count event of Theorem 3.5 whenever this class-coverage condition is met. A learning rule defined on every sample can use any fixed measurable fallback when a class is absent. Class coverage appears as an antecedent on the original probability event.

Proof of Example 3.10. Fix the learned representation and put

$$q_c(x) = \frac{1}{n_c} \sum_{j : Y_j = c} \|f(x) - f(X_j)\|^2, \quad \ell(x, y) = \min\left\{1, \left[1 + \frac{q_y(x) - \min_{c \neq y} q_c(x)}{t}\right]_+\right\}.$$

The minimum is over a finite nonempty set. The quantities are measurable, and $\ell \in [0, 1]$. If the mean-distance rule makes an error, its selected wrong class c satisfies $q_c(x) \leq q_y(x)$, including a misclassified tie. It follows that $\ell(x, y) = 1$, proving pointwise domination of the zero–one loss.

Empirical loss. Let $a = q_y(x)/t \geq 0$ and $b_c = q_c(x)/t$. For any real b and $a \geq 0$, $[1 + a - b]_+ \leq a + [1 - b]_+$: when $1 - b \geq 0$ the two sides are equal, and otherwise the right-hand side is a, which bounds the positive part on the left. Since positive part commutes with a finite maximum,

$$\begin{aligned}\ell(x, y) &\leq \max_{c \neq y} [1 + a - b_c]_+ \\ &\leq a + \max_{c \neq y} [1 - b_c]_+ \\ &\leq a + \sum_{c \neq y} [1 - b_c]_+.\end{aligned}$$

The last inequality uses nonnegativity. The hinge is convex, so for each wrong class,

$$\left[1 - \frac{q_c(x)}{t}\right]_+ \leq \frac{1}{n_c} \sum_{j:Y_j=c} \left[1 - \frac{\|f(x) - f(X_j)\|^2}{t}\right]_+.$$

For the true class, $q_y(x)/t$ is exactly the classwise average same-class contrastive loss. Combining these observations gives

$$\ell(x, y) \leq \sum_{c \in \mathcal{Y}} \frac{1}{n_c} \sum_{j:Y_j=c} l_d(f, (x, y), Z_j).$$

Taking $(x, y) = Z_i$ and averaging over i proves $\widehat{\mathcal{L}}_{c,f}(g) \leq \widehat{Q}$. If $p_{\min,S} = \min_c(n_c/n) > 0$, then $1/n_c \leq 1/(np_{\min,S})$. Nonnegativity of the pair loss yields

$$\widehat{Q} \leq \frac{1}{n^2 p_{\min,S}} \sum_{i,j} l_d(f, Z_i, Z_j) = \frac{\widehat{\mathcal{L}}_d(f)}{p_{\min,S}}.$$

Loss response. Let

$$\bar{f}_c = \frac{1}{n_c} \sum_{j:Y_j=c} f(X_j), \qquad v_c = \frac{1}{n_c} \sum_{j:Y_j=c} \|f(X_j)\|^2.$$

Expansion of the Hilbert-space square gives

$$q_c(x) = \|f(x)\|^2 - 2\langle f(x), \bar{f}_c \rangle + v_c.$$

In $q_y(x) - q_c(x)$, the query term $\|f(x)\|^2$ cancels. The diameter assumption also gives

$$\begin{aligned}\|\bar{f}_y - \bar{f}_c\| &= \left\| \frac{1}{n_y n_c} \sum_{i:Y_i=y} \sum_{j:Y_j=c} (f(X_i) - f(X_j)) \right\| \\ &\leq \frac{1}{n_y n_c} \sum_{i:Y_i=y} \sum_{j:Y_j=c} \|f(X_i) - f(X_j)\| \leq D.\end{aligned}$$

Consequently, for all x, x',

$$|(q_y - q_c)(x) - (q_y - q_c)(x')| \leq 2D\|f(x) - f(x')\| = 2D\, d_f(x, x').$$

A finite maximum preserves a common Lipschitz constant. Indeed, $\max_c u_c - \max_c v_c \leq \max_c(u_c - v_c) \leq \max_c |u_c - v_c|$, and the reverse inequality follows by interchanging the two lists. Since $q_y - \min_{c \neq y} q_c = \max_{c \neq y}(q_y - q_c)$, and the map $s \mapsto \min\{1, [1 + s]_+\}$ is 1-Lipschitz, we obtain

$$|\ell(x, y) - \ell(x', y)| \leq \frac{2D}{t} d_f(x, x').$$

The response and robustness bounds are therefore

$$\omega_i(a_i) \leq \min\{1, 2Da_i/t\}, \qquad \varepsilon_{\mathrm{mr}} \leq \frac{1}{n} \sum_i \min\{1, 2Da_i/t\}.$$

All radii are finite because the image diameter is at most D. The empirical comparison and Theorem 3.5 now give equation 12. □

The role of the mean-distance score. The expansion above also gives

$$q_c(x) = \|f(x) - \bar{f}_c\|^2 + v_c - \|\bar{f}_c\|^2.$$

Thus the empirical comparison applies to the complete mean-distance scores, including their within-class variance terms. The factor $2D/t$ uses cancellation in the Hilbert-space square. For a general pseudometric with diameter D, bounding each squared-distance variation separately gives the constant $4D/t$ for their difference.

An affine SVM trained after the metric. For binary labels, define the comparison score

$$s_0(x) = \frac{q_-(x) - q_+(x)}{t} = \langle w_0, f(x)\rangle + b_0, \quad w_0 = \frac{2(\bar{f}_+ - \bar{f}_-)}{t}, \quad b_0 = \frac{v_- - v_+}{t}.$$

Then $\|w_0\| \le 2D/t$. For either label, its hinge loss equals $[1 + (q_y - q_{-y})/t]_+$. The empirical comparison in the preceding proof holds for this untruncated hinge as well, since there is only one wrong class. Thus the average hinge loss of s_0 is at most $\widehat{Q}$.

An affine SVM optimum also exists when the representation takes values in an infinite-dimensional Hilbert space. Projecting its coefficient vector onto the finite-dimensional span of the represented training points preserves all training scores and cannot increase the penalty. In that span, the norm and intercept bounds used for Example 3.7 give a compact sublevel set, since both classes are present. Continuity then gives a minimum.

Let (w, b) minimize average hinge plus $\lambda_c\|w\|^2/2$ over affine scores on the fixed representation. Comparison with both s_0 and the zero score gives

$$J_c(w, b) \le H_0 := \min\left\{1, \widehat{Q} + \frac{2\lambda_c D^2}{t^2}\right\}.$$

The empirical clipped hinge is at most H_0, and $\|w\| \le \sqrt{2H_0/\lambda_c}$. Hence the theorem in fact gives

$$R_{01}(g \circ f) \le H_0 + \frac{1}{n}\sum_i \min\{1, \sqrt{2H_0/\lambda_c}\, a_i\} + \tau_n.$$

Using $H_0 \le 1$ in the response and $H_0 \le \widehat{Q} + 2\lambda_c D^2/t^2$ in the empirical term yields the stated extension:

$$R_{01}(g \circ f) \le \widehat{Q} + \frac{2\lambda_c D^2}{t^2} + \frac{1}{n}\sum_i \min\{1, \sqrt{2/\lambda_c}\, a_i\} + \tau_n. \tag{27}$$

For an output within objective error $\eta_c \ge 0$ of the minimum, replace H_0 by $H_0 + \eta_c$ in the empirical and norm bounds. The error is added after the minimum defining H_0.

Threshold-loss extension. Fix $0 \le u < t$, put $\Delta = t - u > 0$, and use

$$l_d(f, z, z') = \mathbf{1}\{y = y'\}\frac{[d_f(x, x')^2 - u]_+}{\Delta} + \mathbf{1}\{y \ne y'\}\frac{[t - d_f(x, x')^2]_+}{\Delta}.$$

Keep the same classwise definition of $\widehat{Q}$, and use margin scale Δ in the mean-distance classification loss. For any wrong class c,

$$\begin{aligned}\left[1 + \frac{q_y - q_c}{\Delta}\right]_+ &= \frac{[(q_y - u) + (t - q_c)]_+}{\Delta} \\ &\le \frac{[q_y - u]_+ + [t - q_c]_+}{\Delta}.\end{aligned}$$

The inequality is subadditivity of the positive part. Taking a maximum over wrong classes, then bounding it by a sum of nonnegative terms, gives one true-class term and a sum of wrong-class terms. Convexity of the positive part bounds each by the corresponding average of pairwise violations. This proves the same empirical comparison $\widehat{\mathcal{L}}_{c,f}(g) \le \widehat{Q}$.

The Hilbert-space cancellation is unchanged, so the response constant is $2D/\Delta$. In the binary case, $(q_- - q_+)/\Delta$ is an affine comparator with norm at most $2D/\Delta$, and its untruncated hinge is also bounded by $\widehat{Q}$. The risk and SVM comparisons therefore hold with Δ in place of t. Throughout this extension the classwise normalization is retained; it is the normalization used to obtain the downstream empirical-loss bound.

B.7 A JOINTLY TRAINED NEURAL-NETWORK EXAMPLE

We analyze a finite-width representation with Euclidean input and output spaces,

$$f_\theta(x) = \sum_{j=1}^{q} u_j[v_j^\top x + b_j]_+, \qquad g(v) = \langle w, v\rangle + b, \qquad P(\theta) = \sum_{j=1}^{q} \|u_j\|_2\|v_j\|_2.$$

Here u_j is an output vector, v_j is an input vector, and b_j, b are scalar biases. All parameters are finite. The loss for metric robustness is clipped hinge, and the training loss is the ordinary hinge in the joint objective

$$J(\theta, w, b) = \frac{1}{n}\sum_{i=1}^{n}[1 - Y_i(\langle w, f_\theta(X_i)\rangle + b)]_+ + \frac{\lambda}{2}\big(P(\theta)^2 + \|w\|_2^2\big), \qquad \lambda > 0. \tag{28}$$

Proof of the metric-robustness bound in Example 3.11. For scalars s, t, $|[s]_+ - [t]_+| \le |s - t|$. Applying this fact neuron by neuron, followed by the triangle and Cauchy–Schwarz inequalities, gives

$$\begin{aligned}\|f_\theta(x) - f_\theta(x')\|_2 &\le \sum_{j=1}^{q} \|u_j\|_2\big|[v_j^\top x + b_j]_+ - [v_j^\top x' + b_j]_+\big| \\ &\le \sum_{j=1}^{q} \|u_j\|_2|v_j^\top(x - x')| \\ &\le P(\theta)\|x - x'\|_2.\end{aligned}$$

Thus $a_i \le rP(\theta)$. The affine-head calculation gives $\omega_i(t) \le \min\{1, \|w\|_2 t\}$ for finite t. If a returned output has $J \le E < \infty$, nonnegativity of the hinge and $(P - \|w\|_2)^2 \ge 0$ imply

$$P(\theta)\|w\|_2 \le \frac{P(\theta)^2 + \|w\|_2^2}{2} \le \frac{E}{\lambda}.$$

Consequently

$$\varepsilon_{\mathrm{mr}} \le \frac{1}{n}\sum_i \min\{1, \|w\|_2 a_i\} \le \min\{1, rP(\theta)\|w\|_2\} \le \min\{1, rE/\lambda\}.$$

The empirical clipped hinge and empirical zero–one error are both at most the empirical training hinge, which is at most E. On the event of Theorem 3.5, this proves

$$R_{01}(g \circ f_\theta) \le E + \min\{1, rE/\lambda\} + \tau_n.$$

The two learned stages are controlled by the same joint objective in this argument. □

A finite comparator and consistency. Suppose a fixed finite-width parameter choice (θ_*, w_*, b_*) satisfies population margin at least one almost surely and

$$B = \frac{P(\theta_*)^2 + \|w_*\|_2^2}{2} < \infty.$$

With probability one its training hinge is zero on every observation in a finite sample. Its objective is then λB. Adding neurons whose output weights are zero leaves this comparator and its penalty unchanged, so it is available at every finite width at least its original width. For any measurable returned output satisfying $J \le \lambda B + \eta$, $\eta \ge 0$, the preceding proof gives

$$R_{01}(g \circ f_\theta) \le \lambda B + \eta + \min\{1, r(B + \eta/\lambda)\} + \tau_n. \tag{29}$$

The empirical zero–one error is at most $\lambda B + \eta$.

For deterministic sequences $\lambda_n \downarrow 0$, $r_n \downarrow 0$, and $\eta_n = O(\lambda_n)$, assume the fixed partitions satisfy $K(r_n)/n \to 0$. Choose $\delta_n \downarrow 0$ with $\log(1/\delta_n)/n \to 0$. Each term on the right of equation 29 then tends to zero, on an event whose probability tends to one. The same conclusion holds simultaneously for all finite widths and outputs satisfying the common objective comparison, because the event in the generalization theorem depends only on the partition counts. The comparator proves feasibility of the stated objective level; an optimization algorithm is covered when its returned objective meets that comparison.

A nonlinear example with growing width. Let $\mathcal{X} = [-3,-2] \cup [-1,1] \cup [2,3]$, and let any input distribution be supported there. Assign label $+1$ on the two outer intervals and -1 on the inner interval. Consider the scalar representation and score

$$f_*(x) = [x]_+ + [-x]_+ = |x|, \qquad g_*(v) = 2v - 3.$$

For x in an outer interval, $|x| \in [2,3]$, so the score is at least one and the label is positive. For $x \in [-1,1]$, $|x| \in [0,1]$, so the score is at most minus one and the label is negative. The signed margin is therefore at least one at every point of $\mathcal{X}$.

The representation uses $u_1 = u_2 = 1$, $v_1 = 1$, $v_2 = -1$, and zero hidden biases. Hence $P(\theta_*) = 2$, while $|w_*| = 2$. It follows that $B = 4$, and its objective is exactly 4λ for every sample. The representation identifies the two positive input intervals in $[2,3]$, while sending the negative interval to $[0,1]$; the affine head separates these two represented sets.

We next construct a deterministic partition of the entire labeled space. Divide each outer interval into $\lceil 1/r \rceil$ subintervals and the inner interval into $\lceil 2/r \rceil$ subintervals. Use half-open pieces except for the final endpoint, so the pieces form a Borel partition. Each piece has diameter at most r. Cross every input piece with both labels, including the label of probability zero on that interval. This gives

$$K(r) = 2\big(2\lceil 1/r \rceil + \lceil 2/r \rceil\big) \le 8/r + 6.$$

Crossing with both labels ensures that this is a partition of $\mathcal{X} \times \{-1,+1\}$, as required by the general theorem.

Fix $c \ge 0$, and take $n \ge 2$, $r_n = \lambda_n = n^{-1/3}$, and $\delta_n = n^{-2}$. Every returned output satisfying $J \le (4+c)\lambda_n$ has

$$\widehat{R}_{01} \le (4+c)n^{-1/3}, \qquad \varepsilon_{\mathrm{mr}} \le \min\{1, (4+c)n^{-1/3}\}.$$

On the common partition-count event, whose probability is at least $1 - n^{-2}$,

$$\begin{aligned} R_{01} &\le 2(4+c)n^{-1/3} + \sqrt{\frac{(8n^{1/3}+6)\log 2 + 2\log n}{2n}} \\ &= O(n^{-1/3}). \end{aligned}$$

To verify the final order, the squared sampling term is a constant multiple of $n^{-2/3}$ plus $(3\log 2 + \log n)/n$; the latter is $O(n^{-2/3})$, since $\log n = O(n^{1/3})$. The resulting constants depend on c and are uniform in the network width and the distribution within the three intervals.

Both errors are nonnegative, so $|R_{01} - \widehat{R}_{01}| \le \max\{R_{01}, \widehat{R}_{01}\}$, proving the same rate for their gap. The width-two comparator extends to every finite $q \ge 2$ by zero-output neurons. The count event is the same for all of these widths. In particular, one may take $q_n = n^2$: the displayed scalar-input, scalar-output parameterization has q_n values each of u_j, v_j, b_j, together with w, b, for a total of $3q_n + 2$ scalar parameters. The rate applies to every output at that width satisfying the stated joint-objective bound.

A two-interval example. For comparison, take $\mathcal{X} = [-2,-1] \cup [1,2]$, with the left interval labeled -1, the right labeled $+1$, and any distribution on these intervals. The width-two representation

$$f_*(x) = \frac{[x+1]_+ - [x-1]_+}{2}, \qquad g_*(v) = 2v - 1$$

is zero on the left interval and one on the right. At the endpoints -1 and 1 these identities also hold. Its signed classification margin is exactly one throughout the support. In this parameterization, the two output weights are $1/2, -1/2$, both input weights equal one, and the hidden biases are $1, -1$. Thus $P(\theta_*) = 1$, $|w_*| = 2$, and $B = 5/2$.

Dividing both input intervals into $\lceil 1/r \rceil$ pieces and crossing with both labels gives $K(r) = 4\lceil 1/r \rceil \le 4/r + 4$. Taking $r_n = \lambda_n = n^{-1/3}$, $\eta_n = O(\lambda_n)$, and $\delta_n = n^{-2}$ in equation 29 gives the same $O(n^{-1/3})$ training and population error bounds under its objective comparison. Zero-output neurons again allow arbitrary larger finite widths.

C Proofs and Extensions for Section 4

C.1 LogDet geometry and the constrained optimization problem

We work in the Euclidean space of real symmetric $d \times d$ matrices, equipped with the Frobenius inner product $\langle U, V \rangle_F = \operatorname{tr}(UV)$. Throughout this appendix, $A_0 \succ 0$, $0 < u < l$, and the deterministic constant $C \geq 0$ have the meanings specified in Section 4. Define

$$h(t) = t - \log t - 1 \quad (t > 0), \qquad \mathcal{A}_C = \{A \succ 0 : D_{\mathrm{ld}}(A, A_0) \leq C\}.$$

We first establish the geometric properties of this set and then prove existence, uniqueness, and the dual relation for the hard-constraint problem.

Spectral bounds. The derivatives $h'(t) = 1 - t^{-1}$ and $h''(t) = t^{-2}$ show that h is nonnegative, decreases on $(0, 1]$, and increases on $[1, \infty)$, with its unique minimum $h(1) = 0$. It tends to infinity as $t \downarrow 0$ and as $t \to \infty$. For $C > 0$, let $b_-(C) \in (0, 1)$ and $b(C) > 1$ be the two solutions of $h(t) = C$; set $b_-(0) = b(0) = 1$.

For $A \succ 0$, put $M = A_0^{-1/2} A A_0^{-1/2}$. Cyclicity of the trace and multiplicativity of the determinant give

$$D_{\mathrm{ld}}(A, A_0) = \operatorname{tr} M - \log\det M - d = \sum_{j=1}^{d} h(\lambda_j(M)).$$

All summands are nonnegative. Thus $A \in \mathcal{A}_C$ implies $h(\lambda_j(M)) \leq C$ for each j, and hence

$$b_-(C) I \preceq M \preceq b(C) I.$$

Congruence by $A_0^{1/2}$ yields

$$b_-(C) A_0 \preceq A \preceq b(C) A_0 \preceq \beta I, \qquad \beta = \|A_0\|_{\mathrm{op}} b(C).$$

In particular, every matrix in $\mathcal{A}_C$ stays a positive distance from singularity. The set is bounded, and any convergent sequence in it has a positive-definite limit; continuity of D_{ld} then puts that limit in $\mathcal{A}_C$. Therefore $\mathcal{A}_C$ is compact. When $C = 0$, the preceding calculation gives $\mathcal{A}_0 = \{A_0\}$.

For $A \in \mathcal{A}_C$, the spectral bound implies

$$\|A\|_F \leq \sqrt{d}\,\beta, \qquad q_A(x, x') = (x - x')^\top A (x - x') \leq \beta \|x - x'\|_2^2 \leq 4R^2 \beta.$$

If (x, Y_i) belongs to the partition cell of Z_i, the input-diameter condition on Π_r gives $\|x - X_i\|_2 \leq r$. Consequently,

$$a_i = \sup_{(x, Y_i) \in C(Z_i)} d_{f_A}(X_i, x) \leq r\sqrt{\beta}, \qquad a(f_A, S; \Pi_r) \leq r\sqrt{\beta}.$$

The same bounds hold for every selected-pair solution whenever its LogDet value is at most C.

Existence and uniqueness. Consider one fixed selection of pair constraints and suppose that it has a feasible matrix $A^{\mathrm{f}} \succ 0$. Set $C_{\mathrm{f}} = D_{\mathrm{ld}}(A^{\mathrm{f}}, A_0)$. Every feasible point with a smaller objective lies in the compact set $\mathcal{A}_{C_{\mathrm{f}}}$. The selected inequalities are closed affine conditions on A, so their intersection with this compact set is nonempty and compact. The continuous objective attains its minimum on that intersection, and its minimum there is also the minimum over the entire feasible positive-definite domain.

The feasible domain is convex. Moreover, for any nonzero symmetric matrix H, the second directional derivative of the objective at $A \succ 0$ is

$$\frac{d^2}{dt^2} D_{\mathrm{ld}}(A + tH, A_0)\bigg|_{t=0} = \operatorname{tr}(A^{-1} H A^{-1} H) = \|A^{-1/2} H A^{-1/2}\|_F^2 > 0.$$

Thus the objective is strictly convex and the minimizing matrix is unique. In particular, the common comparator in equation 15 is feasible for every prefix selection and has objective at most C; all the resulting optima belong to the same set $\mathcal{A}_C$.

Dual relation. Let A_* be the optimum for the fixed selection. Write all selected constraints as $\langle B_e, A\rangle_F \le c_e$, where

$$(B_e, c_e) = \begin{cases} (v_e v_e^\top, u), & e \in \mathcal{T}_{+,m_+}, \\ (-v_e v_e^\top, -l), & e \in \mathcal{T}_{-,m_-}. \end{cases}$$

Let I_* be the set of active constraints at A_*, and let $\mathcal{K} = \{\sum_{e\in I_*} t_e B_e : t_e \ge 0\}$ be the cone generated by their normal matrices. A cone generated by finitely many vectors is a closed convex cone. We claim that

$$-\nabla D_{\mathrm{ld}}(A_*, A_0) \in \mathcal{K}.$$

If this inclusion failed, separation from a closed convex cone in a finite-dimensional Euclidean space would give a symmetric matrix H such that

$$\langle B_e, H\rangle_F \le 0 \quad (e \in I_*), \qquad \langle \nabla D_{\mathrm{ld}}(A_*, A_0), H\rangle_F < 0.$$

The positive-definite cone is open, so $A_* + tH \succ 0$ for all sufficiently small $t > 0$. Each active affine inequality remains satisfied by the first displayed condition. Every inactive inequality has strictly positive slack, and there are finitely many such inequalities; reducing t if necessary preserves all of them. Finally, differentiability and the strictly negative directional derivative imply $D_{\mathrm{ld}}(A_* + tH, A_0) < D_{\mathrm{ld}}(A_*, A_0)$ for sufficiently small positive t. This contradicts optimality and proves the claim.

Choose nonnegative coefficients in this conic representation and assign coefficient zero to inactive constraints. Since $\nabla D_{\mathrm{ld}}(A_*, A_0) = A_0^{-1} - A_*^{-1}$, stationarity gives exactly equation 14:

$$A_*^{-1} = A_0^{-1} + \sum_{e\in\mathcal{T}_{+,m_+}} \lambda_e v_e v_e^\top - \sum_{e\in\mathcal{T}_{-,m_-}} \nu_e v_e v_e^\top, \qquad \lambda_e, \nu_e \ge 0.$$

The coefficients also obey complementary slackness,

$$\lambda_e\{q_{A_*}(e) - u\} = 0, \qquad \nu_e\{l - q_{A_*}(e)\} = 0.$$

Conversely, feasibility, these complementary-slackness equations, and stationarity are sufficient for optimality. If A is any feasible matrix and t_e denotes the multiplier of its selected inequality, convexity gives

$$D_{\mathrm{ld}}(A, A_0) - D_{\mathrm{ld}}(A_*, A_0) \ge -\sum_e t_e \langle B_e, A - A_*\rangle_F \ge 0.$$

For an active constraint the inner product is nonpositive, and an inactive constraint has multiplier zero. Thus the displayed dual relation holds for every feasible selected problem, including cases with active redundant constraints. There is one dual coordinate for each selected inequality, giving the count $m = m_+ + m_-$; some coordinates can be zero, and their values need not be unique. The primal optimum remains unique in its $p = d(d+1)/2$-dimensional matrix space.

C.2 Uniform approximation of the full training-pair pools

This subsection proves the simultaneous sampling statement used in Section 4.2. Fix $n \ge 3$, $\delta \in (0, 1)$, and a realized sample containing both labels. Then

$$N_- = n_+ n_- > 0, \qquad N_+ = \binom{n_+}{2} + \binom{n_-}{2} > 0.$$

The second inequality uses $n \ge 3$. All pair means below are therefore well defined, including when one class has a single observation.

For $\sigma \in \{+, -\}$, define

$$\psi_+(A; e) = [q_A(e) - u]_+, \qquad \psi_-(A; e) = [l - q_A(e)]_+,$$

and

$$L_\sigma(A) = \frac{1}{N_\sigma} \sum_{e\in\mathcal{T}_\sigma} \psi_\sigma(A; e), \qquad \widehat{L}_{\sigma,k}(A) = \frac{1}{k} \sum_{e\in\mathcal{T}_{\sigma,k}} \psi_\sigma(A; e).$$

Conditional on this sample, only the uniformly random permutations of the two finite pools are random. We prove that their joint law assigns probability at least $1 - \delta/2$ to the event

$$|\widehat{L}_{\sigma,k}(A) - L_\sigma(A)| \le e_\sigma(k) \quad \text{for every } A \in \mathcal{A}_C,\ \sigma \in \{+, -\},\ 1 \le k \le N_\sigma, \tag{30}$$

where e_σ is defined in equation 16. The event is uniform over the matrix and over all prefixes of both permutations.

A finite-population concentration inequality. We give the bounded-variable calculation to specify the constants. Let $0 < B < \infty$, let $a_1, \dots, a_N \in [0, B]$ have mean $\bar{a}$, and let $U_1, \dots, U_k$ be a uniformly ordered sample of k distinct population members. For every real t,

$$\mathbb{E}\exp\left(t\sum_{j=1}^{k}(U_j - \bar{a})\right) \le \left\{\frac{1}{N}\sum_{j=1}^{N} e^{t(a_j-\bar{a})}\right\}^k.$$

Indeed, put $z_j = e^{t(a_j-\bar{a})} > 0$. The left side is the normalized elementary symmetric polynomial

$$\binom{N}{k}^{-1} \sum_{1\le j_1<\cdots<j_k\le N} z_{j_1}\cdots z_{j_k},$$

which is at most $(N^{-1}\sum_j z_j)^k$. One direct way to see the latter inequality is to fix the sum of the z_j's and replace any two coordinates by their average. In the elementary symmetric polynomial, terms involving neither coordinate are unchanged, terms involving exactly one depend only on their sum, and terms involving both have their product multiplied by a nonnegative coefficient. Such a replacement cannot decrease the polynomial. Repeatedly average a largest and a smallest coordinate. If their difference is D, the sum of squared deviations from the common mean decreases by $D^2/2$; that sum is at most ND^2 before the replacement. It therefore converges geometrically to zero, so the coordinate vector converges to the constant vector at its mean. Continuity proves the claimed inequality. For $k = 1$, it is an equality. This is the moment-generating-function comparison underlying the without-replacement bound of Hoeffding (1963, Theorem 4).

For completeness, if W is a centered variable whose range lies in an interval of length B, write $F(t) = \log \mathbb{E}e^{tW}$. Boundedness permits differentiation, and $F''(t)$ is the variance of W under the exponentially tilted probability law. The variance of any variable in an interval of length B is at most $B^2/4$: subtracting the interval midpoint bounds its second moment by $B^2/4$, and variance is no larger than that second moment. Since $F(0) = F'(0) = 0$, integration gives $F(t) \le t^2B^2/8$, for either sign of t. Applied to a uniform draw from the finite population and combined with the preceding comparison, this yields

$$\mathbb{E}\exp\left(t\sum_{j=1}^{k}(U_j - \bar{a})\right) \le e^{kt^2B^2/8}.$$

For $\eta > 0$, Markov's inequality with $t > 0$ therefore bounds the upper tail by $\exp(-tk\eta + kt^2B^2/8)$. Taking $t = 4\eta/B^2$, and applying the same argument to the lower tail, proves

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{j=1}^{k} U_j - \bar{a}\right| > \eta\right\} \le 2\exp(-2k\eta^2/B^2).$$

At $k = N$, the sampled mean equals $\bar{a}$ exactly.

A finite matrix cover. Set $\Lambda = \sqrt{d}\,\beta$ and $\rho = \Lambda/n^2$. The compact set $\mathcal{A}_C$ lies in a Frobenius ball of radius Λ in the p-dimensional space of symmetric matrices. Choose a maximal finite subset with mutual distances strictly larger than ρ. Maximality makes it an internal ρ-net. The open radius-$\rho/2$ balls centered at these net points are disjoint and lie inside a ball of radius $\Lambda + \rho/2$. Comparing Euclidean volumes shows that the number M of net points satisfies

$$M \le \left(1 + \frac{2\Lambda}{\rho}\right)^p = (1 + 2n^2)^p.$$

This also justifies that a maximal such subset is finite. The set $\mathcal{A}_C$ and this net are determined by A_0, C, n, independently of the training sample.

For any pair $e = (i, j)$, $\|v_e\|_2 \le 2R$, so

$$|q_A(e) - q_{A'}(e)| = |\langle A - A', v_e v_e^\top\rangle_F| \le \|A - A'\|_F \|v_e v_e^\top\|_F \le 4R^2\|A - A'\|_F.$$

The positive-part map is 1-Lipschitz. Thus the same estimate holds for $|\psi_\sigma(A; e) - \psi_\sigma(A'; e)|$, and for the differences of either full-pool or prefix means. Moreover, Appendix C.1 implies, for every $A \in \mathcal{A}_C$,

$$0 \le \psi_+(A; e) \le 4R^2\beta, \qquad 0 \le \psi_-(A; e) \le l.$$

Both populations are therefore bounded by $B = \max\{4R^2\beta, l\} > 0$.

Uniformity over matrices and pair budgets. Put

$$Z_n = p\log(1+2n^2) + \log(4n^2/\delta).$$

At one net point and one proper prefix of length $k < N_\sigma$, the finite-population bound with $\eta = B\sqrt{Z_n/(2k)}$ has failure probability at most $2e^{-Z_n}$, conditional on S. Across the two pools there are at most $N_+ + N_- = \binom{n}{2} \le n^2$ prefixes. A union bound over both pools, their prefixes, and the M net points has failure probability at most

$$2Mn^2 e^{-Z_n} \le \delta/2.$$

No independence among the events in this union bound is required.

On the complementary event, take any $A \in \mathcal{A}_C$ and a net point A' with $\|A - A'\|_F \le \rho$. For every proper prefix,

$$\begin{aligned}
|\widehat{L}_{\sigma,k}(A) - L_\sigma(A)| &\le |\widehat{L}_{\sigma,k}(A') - L_\sigma(A')| + |\widehat{L}_{\sigma,k}(A) - \widehat{L}_{\sigma,k}(A')| + |L_\sigma(A) - L_\sigma(A')| \\
&\le B\sqrt{\frac{Z_n}{2k}} + 8R^2\rho \\
&= B\sqrt{\frac{p\log(1+2n^2) + \log(4n^2/\delta)}{2k}} + \frac{8R^2\sqrt{d}\,\beta}{n^2} = e_\sigma(k).
\end{aligned}$$

For $k = N_\sigma$, equality of the full and sampled pools gives zero error, which justifies the separate endpoint definition $e_\sigma(N_\sigma) = 0$. If $N_\sigma = 1$, this endpoint is the only admissible prefix. This proves equation 30, including the cases $C = 0$ and singleton pools.

Each conclusion is pointwise for every $A \in \mathcal{A}_C$ once the finite net event holds. It therefore applies to matrices fitted from the same selected pairs and to pair budgets chosen after examining the sample and its permutations. Uniform marginal permutations for the two pools suffice; their mutual independence is unnecessary for this argument. Pairs can share training observations, since the conditional argument samples from a fixed finite population.

The finite net event is measurable. Define it to hold automatically when a class is absent. Its conditional failure probability is at most $\delta/2$ for every sample containing both classes, so integration over S gives an unconditional failure probability at most $\delta/2$. In subsequent arguments, we intersect this event with the fixed-partition event at failure level $\delta/2$. This construction proves a statement of joint probability at least $1 - \delta$ over S and the pair permutations, with the feasibility assumption imposed as an implication on that event.

C.3 General classification rules

Fix the deterministic, label-pure partition Π_r used in Section 4. When both classes occur and the common comparator condition equation 15 holds, the preceding geometric argument gives, for every admissible budget,

$$A_{\boldsymbol{m}} \preceq \beta I, \qquad a(f_{A_{\boldsymbol{m}}}, S; \Pi_r) \le r\sqrt{\beta}. \tag{31}$$

Let $g_{\boldsymbol{m}}$ be any measurable resultant classifier, and use an evaluation loss taking values in $[0,1]$ that upper-bounds its misclassification indicator. Write $\omega_{\boldsymbol{m},i}$ for the anchored response in equation 3. By its definition as a supremum over nested balls, this response is nondecreasing in the radius. Hence

$$\varepsilon_{\mathrm{mr}}(f_{\boldsymbol{m}}, g_{\boldsymbol{m}}, S; \Pi_r) = \frac{1}{n}\sum_i \omega_{\boldsymbol{m},i}(a_i) \le \frac{1}{n}\sum_i \omega_{\boldsymbol{m},i}(r\sqrt{\beta}).$$

Apply Theorem 3.5 with failure probability $\delta/2$. The resulting event depends only on the cell counts of S and holds simultaneously for all measurable learned outputs. Thus, on an event of probability at least $1 - \delta/2$, whenever the comparator and class-presence conditions hold,

$$R_{01}(g_{\boldsymbol{m}} \circ f_{\boldsymbol{m}}) \le \widehat{\mathcal{L}}_{c,f_{\boldsymbol{m}}}(g_{\boldsymbol{m}}) + \frac{1}{n}\sum_i \omega_{\boldsymbol{m},i}(r\sqrt{\beta}) + \tau_n, \qquad \tau_n = \sqrt{\frac{K(r)\log 2 + \log(2/\delta)}{2n}}. \tag{32}$$

The same average-response term plus τ_n bounds the absolute generalization gap of the evaluation loss. These inequalities are simultaneous over budgets and classifiers, including choices depending on the pair permutations or additional algorithmic randomness. For the affine SVM, the next two subsections bound the empirical loss and the response by explicit functions of the pair budgets.

C.4 From pair violations to an affine comparison classifier

Fix a sample with both labels observed and a matrix $A \in \mathcal{A}_C$. Write $v = f_A(x) = A^{1/2}x$, and let

$$q_c(x) = \frac{1}{n_c} \sum_{j:Y_j=c} q_A(x, X_j), \qquad s_0(x) = \frac{q_-(x) - q_+(x)}{\Delta}, \qquad \Delta = l - u > 0.$$

The class averages include the anchor when its label is c. Let

$$\widehat{R}_h(s) = \frac{1}{n} \sum_{i=1}^{n} [1 - Y_i s(X_i)]_+$$

be the empirical hinge loss. We prove both statements in

$$\widehat{R}_h(s_0) \le \frac{\alpha_S}{\Delta} \{L_+(A) + L_-(A)\}, \qquad \|w_0\|_2 \le \frac{4R\sqrt{\beta}}{\Delta}, \tag{33}$$

where $s_0(x) = \langle w_0, f_A(x) \rangle + b_0$ and $\alpha_S = \max\{1, 2N_+/(nn_{\min})\}$.

Hinge loss and pool normalizations. For $Y_i = c$, the sign in the score definition gives $Y_i s_0(X_i) = \{q_{-c}(X_i) - q_c(X_i)\}/\Delta$. Therefore

$$\begin{aligned}
[1 - Y_i s_0(X_i)]_+ &= \frac{[\Delta + q_c(X_i) - q_{-c}(X_i)]_+}{\Delta} \\
&= \frac{[(q_c(X_i) - u) + (l - q_{-c}(X_i))]_+}{\Delta} \\
&\le \frac{[q_c(X_i) - u]_+ + [l - q_{-c}(X_i)]_+}{\Delta}.
\end{aligned}$$

Convexity of the positive-part map bounds the two numerator terms by

$$[q_c(X_i) - u]_+ \le \frac{1}{n_c} \sum_{j:Y_j=c} [q_A(X_i, X_j) - u]_+,$$

and

$$[l - q_{-c}(X_i)]_+ \le \frac{1}{n_{-c}} \sum_{j:Y_j=-c} [l - q_A(X_i, X_j)]_+.$$

Let W_c be the sum of same-class violations over unordered pairs in class c, and let W_{diff} be the sum of different-class violations over $\mathcal{T}_-$. Averaging the first bound over i, each off-diagonal pair in class c is counted twice, while its diagonal terms vanish because $q_A(X_i, X_i) = 0 < u$. Thus the total within-class contribution is

$$\frac{2}{n} \sum_{c \in \{-1,+1\}} \frac{W_c}{n_c} \le \frac{2(W_+ + W_-)}{nn_{\min}} = \frac{2N_+}{nn_{\min}} L_+(A).$$

Here the subscripts on W_c refer to class labels, whereas L_+ denotes the pool of all same-class pairs. For the different-class contribution, each unordered pair is used once with each endpoint as anchor, giving

$$\frac{W_{\text{diff}}}{n} \left(\frac{1}{n_+} + \frac{1}{n_-} \right) = \frac{W_{\text{diff}}}{n_+ n_-} = L_-(A),$$

where we used $n_+ + n_- = n$ and $N_- = n_+ n_-$. Multiplying by $1/\Delta$ and using $\alpha_S \ge 1$ proves the hinge bound in equation 33. This calculation remains valid when a class contains only one training observation, since then its within-class sum is zero.

Affine form and norm. Define the represented class means and second moments by

$$\bar{f}_c = \frac{1}{n_c} \sum_{j:Y_j=c} f_A(X_j), \qquad a_c = \frac{1}{n_c} \sum_{j:Y_j=c} \|f_A(X_j)\|_2^2.$$

Expanding the square in the distance average gives

$$q_c(x) = \|f_A(x)\|_2^2 - 2\langle f_A(x), \bar{f}_c\rangle + a_c.$$

The first term is identical for the two classes and cancels in their difference. Hence

$$w_0 = \frac{2(\bar{f}_+ - \bar{f}_-)}{\Delta}, \qquad b_0 = \frac{a_- - a_+}{\Delta}.$$

Every represented training point has norm at most $R\sqrt{\beta}$, so each class mean has that norm bound and

$$\|w_0\|_2 \le \frac{2(\|\bar{f}_+\|_2 + \|\bar{f}_-\|_2)}{\Delta} \le \frac{4R\sqrt{\beta}}{\Delta}.$$

This completes the proof of equation 33. In particular, the comparison classifier uses the same representation as the fitted SVM, and its regularization cost is at most $8\lambda_c R^2\beta/\Delta^2$.

C.5 Proof of Theorem 4.1

The common event. Let E_{pair} be the measurable net event constructed in Appendix C.2, and let E_{part} be the cell-count event of Theorem 3.5 for Π_r at failure level $\delta/2$. We proved $\mathbb{P}(E_{\mathrm{pair}}^c) \le \delta/2$, after integrating over S; the generalization theorem gives $\mathbb{P}(E_{\mathrm{part}}^c) \le \delta/2$. Therefore

$$\mathbb{P}(E_{\mathrm{pair}} \cap E_{\mathrm{part}}) \ge 1 - \delta.$$

Fix a sample and its pair permutations in this intersection, and suppose both classes occur and equation 15 holds. All subsequent statements are deterministic on this event and hold for every admissible pair budget.

Control of full-pool violations. The common comparator and exact hard optimization imply $A_{\boldsymbol{m}} \in \mathcal{A}_C$. Each selected inequality is satisfied, so $\widehat{L}_{+,m_+}(A_{\boldsymbol{m}}) = \widehat{L}_{-,m_-}(A_{\boldsymbol{m}}) = 0$. Applying equation 30 to this data-dependent matrix gives

$$L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le e_+(m_+) + e_-(m_-) = V_{\boldsymbol{m}}.$$

The uniform event is what allows the fitted matrix to depend on the same selected pairs. For sum-slack ITML, Appendix C.6 proves this inequality with the value of $V_{\boldsymbol{m}}$ in equation 20; the rest of the proof is unchanged.

The downstream optimum. For the representation $f_{\boldsymbol{m}} = f_{A_{\boldsymbol{m}}}$, the affine SVM minimizes

$$J_{\boldsymbol{m}}(w, b) = \frac{1}{n}\sum_i [1 - Y_i\{\langle w, f_{\boldsymbol{m}}(X_i)\rangle + b\}]_+ + \frac{\lambda_c}{2}\|w\|_2^2.$$

We first verify that an exact minimum exists. The zero score has objective 1. On the sublevel set $J_{\boldsymbol{m}} \le 1$, nonnegativity gives $\|w\|_2 \le \sqrt{2/\lambda_c}$, and each individual hinge term is at most n. Choosing one positive observation i and one negative observation j gives

$$1 - n - \langle w, f_{\boldsymbol{m}}(X_i)\rangle \le b \le n - 1 - \langle w, f_{\boldsymbol{m}}(X_j)\rangle.$$

Since all represented norms are at most $R\sqrt{\beta}$, it follows that

$$|b| \le n - 1 + R\sqrt{2\beta/\lambda_c}.$$

This sublevel set is closed and bounded in a finite-dimensional space, hence compact. It is nonempty, so continuity yields an attained global minimum. Convexity of hinge loss and strict convexity of the penalty in w imply that all minimizers have the same w. Their intercepts form a nonempty compact interval, so there is a unique intercept of minimum absolute value. Appendix C.12 justifies a measurable selection; all inequalities below hold for every minimizing intercept.

Let $(w_{\boldsymbol{m}}, b_{\boldsymbol{m}})$ be a minimizing pair. Comparison with zero and with the mean-distance score from Appendix C.4 gives

$$\begin{aligned} J_{\boldsymbol{m}}(w_{\boldsymbol{m}}, b_{\boldsymbol{m}}) &\le \min\{1, J_{\boldsymbol{m}}(w_0, b_0)\} \\ &\le \min\left\{1, \frac{\alpha_S\{L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}})\}}{\Delta} + \frac{8\lambda_c R^2\beta}{\Delta^2}\right\} \\ &\le \min\left\{1, \frac{\alpha_S V_{\boldsymbol{m}}}{\Delta} + \frac{8\lambda_c R^2\beta}{\Delta^2}\right\} = H_{\boldsymbol{m}}. \end{aligned}$$

Writing $s_{\boldsymbol{m}} = g_{\boldsymbol{m}} \circ f_{\boldsymbol{m}}$, nonnegativity of the two objective terms gives separately

$$\widehat{R}_h(s_{\boldsymbol{m}}) \le H_{\boldsymbol{m}}, \qquad \|w_{\boldsymbol{m}}\|_2 \le \sqrt{2H_{\boldsymbol{m}}/\lambda_c}.$$

Metric robustness and classification risk. Use the clipped hinge for evaluation,

$$l_c(g,(v,y)) = \min\{1, [1 - yg(v)]_+\}.$$

For a fixed label y, clipping and taking a positive part are both 1-Lipschitz operations. Cauchy–Schwarz therefore gives

$$\begin{aligned} |l_c(g_{\boldsymbol{m}},(f_{\boldsymbol{m}}(x),y)) - l_c(g_{\boldsymbol{m}},(f_{\boldsymbol{m}}(x'),y))| &\le |\langle w_{\boldsymbol{m}}, f_{\boldsymbol{m}}(x) - f_{\boldsymbol{m}}(x')\rangle| \\ &\le \|w_{\boldsymbol{m}}\|_2 d_{f_{\boldsymbol{m}}}(x,x'). \end{aligned}$$

The left side is also at most one. Thus each anchored response satisfies $\omega_{\boldsymbol{m},i}(t) \le \min\{1, \|w_{\boldsymbol{m}}\|_2 t\}$. The geometric bound equation 31 now implies

$$\begin{aligned} \varepsilon_{\mathrm{mr}}(f_{\boldsymbol{m}}, g_{\boldsymbol{m}}, S; \Pi_r) &\le \frac{1}{n}\sum_i \min\{1, \|w_{\boldsymbol{m}}\|_2 a_i\} \\ &\le \min\left\{1, r\sqrt{\frac{2\beta H_{\boldsymbol{m}}}{\lambda_c}}\right\}. \end{aligned}$$

On E_{part}, Theorem 3.5 therefore gives

$$|\mathcal{L}_{c,f_{\boldsymbol{m}}}(g_{\boldsymbol{m}}) - \widehat{\mathcal{L}}_{c,f_{\boldsymbol{m}}}(g_{\boldsymbol{m}})| \le \min\left\{1, r\sqrt{\frac{2\beta H_{\boldsymbol{m}}}{\lambda_c}}\right\} + \tau_n.$$

The clipped empirical loss is at most the empirical hinge loss. The pointwise clipped loss upper-bounds the corresponding misclassification indicator: if a scalar score is misclassified, its signed margin is nonpositive, and the clipped hinge equals one; the same statement holds at a zero score under either fixed tie rule. Consequently,

$$\widehat{R}_{01}(s_{\boldsymbol{m}}) \le H_{\boldsymbol{m}}, \qquad R_{01}(s_{\boldsymbol{m}}) \le H_{\boldsymbol{m}} + \min\left\{1, r\sqrt{\frac{2\beta H_{\boldsymbol{m}}}{\lambda_c}}\right\} + \tau_n.$$

This proves equation 18 and every asserted empirical-error and clipped-hinge-gap bound. If $B_{\boldsymbol{m}}$ denotes the right side of equation 18, both zero–one errors lie in $[0, \min\{1, B_{\boldsymbol{m}}\}]$. In particular,

$$|R_{01}(s_{\boldsymbol{m}}) - \widehat{R}_{01}(s_{\boldsymbol{m}})| \le \min\{1, B_{\boldsymbol{m}}\}.$$

Dependence on the pair budgets. For each pool, $e_\sigma(k)$ is a constant plus a nonnegative multiple of $k^{-1/2}$ for proper prefixes, followed by its drop to zero at the full pool. It is therefore nonincreasing on its entire admissible range. Holding the sample, partition, confidence level, geometric budget, and regularization fixed makes $\alpha_S, \beta, \lambda_c, \tau_n$ fixed. Thus $V_{\boldsymbol{m}}$ and $H_{\boldsymbol{m}}$ are nonincreasing in either budget, and the square-root and clipping operations preserve this order. This proves the monotonicity of all the stated upper bounds. The assertion concerns the explicit bounds; the fitted solutions and their realized risks can change nonmonotonically as the constraints are enlarged.

The probability statement just proved can equivalently be written as

$$\mathbb{P}\{\text{both classes and equation 15 hold, but some admissible budget violates the theorem}\} \le \delta.$$

It follows from an unconditional event and an implication on that event. This formulation permits sample-dependent feasibility without replacing the sampling law by a law conditional on feasibility.

The mean-distance classifier itself. The same event also gives a direct risk bound for the comparison classifier. Apply equation 32 with the clipped hinge and the score $s_{0,\boldsymbol{m}}$ constructed from $A_{\boldsymbol{m}}$. The bridge bounds its empirical hinge by $\alpha_S V_{\boldsymbol{m}}/\Delta$, and its coefficient norm by $4R\sqrt{\beta}/\Delta$. Its response at radius $r\sqrt{\beta}$ is therefore at most $\min\{1, 4R\beta r/\Delta\}$. Hence

$$R_{01}(s_{0,\boldsymbol{m}}) \le \frac{\alpha_S V_{\boldsymbol{m}}}{\Delta} + \min\left\{1, \frac{4R\beta r}{\Delta}\right\} + \tau_n. \tag{34}$$

This conclusion uses the explicit comparison score and requires no further classifier optimization.

C.6 STANDARD SUM-SLACK ITML

We next prove the extension to the LogDet slack objective used in ITML (Davis et al., 2007). For each selected pair introduce a positive slack ξ_e, and put $\xi_{0,e} = u$ for same-class pairs and $\xi_{0,e} = l$ for different-class pairs. For fixed $\gamma > 0$, consider

$$\begin{aligned} &\min_{A \succ 0,\ \xi_e > 0} D_{\mathrm{ld}}(A, A_0) + \gamma \sum_e \varphi(\xi_e/\xi_{0,e}), \\ &\text{s.t. } q_A(e) \le \xi_e \quad (e \in \mathcal{T}_{+,m_+}), \qquad q_A(e) \ge \xi_e \quad (e \in \mathcal{T}_{-,m_-}), \end{aligned} \tag{35}$$

where $\varphi(t) = t - \log t - 1$, and the sum is over the selected pairs. The penalty is an unnormalized sum, with the same γ for all pair budgets.

Existence, uniqueness, and the common budget. Under equation 15, $(A^\dagger, \xi_0)$ is feasible for every selected problem and has objective at most C. On the sublevel set with objective at most C, nonnegativity gives

$$A \in \mathcal{A}_C, \qquad \sum_e \varphi(\xi_e/\xi_{0,e}) \le C/\gamma.$$

In particular, each ratio $\xi_e/\xi_{0,e}$ belongs to $[b_-(C/\gamma), b(C/\gamma)]$. All matrix eigenvalues and all slacks thus stay bounded away from zero, and all variables are bounded above. Intersecting these bounds with the closed affine constraints gives a nonempty compact set on which the objective is continuous. The minimum is attained. The feasible domain is convex, the LogDet term is strictly convex in A, and each slack penalty is strictly convex in its scalar variable. Their sum is strictly convex in (A, ξ), so this minimizing pair is unique. Write it as $(A_{\boldsymbol{m}}, \xi_{\boldsymbol{m}})$. We have established

$$D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C, \qquad \sum_e \varphi(\xi_{\boldsymbol{m},e}/\xi_{0,e}) \le C/\gamma.$$

For $C = 0$, all nonnegative objective terms must vanish, giving $A_{\boldsymbol{m}} = A_0$ and $\xi_{\boldsymbol{m}} = \xi_0$; the following estimates remain valid with zero slack contribution.

The common hard comparator also guarantees $X_i \neq X_j$ for each selected opposite-label pair, since their comparator squared distance is at least $l > 0$. Consequently $q_A(e) > 0$ for such a pair and every $A \succ 0$. More generally, distinct endpoints for every selected opposite-label pair suffice for feasibility of the sum-slack problem: choose any $A \succ 0$, then use the slacks $\max\{u, q_A(e)\}$ for same-class pairs and $\min\{l, q_A(e)\}$ for different-class pairs.

Eliminating the slacks. For fixed A, $\varphi(\xi/u)$ is minimized over the item $\xi \ge q_A(e)$ by $\xi = \max\{u, q_A(e)\}$. Similarly, $\varphi(\xi/l)$ is minimized over $0 < \xi \le q_A(e)$ by $\xi = \min\{l, q_A(e)\}$. Thus the equivalent matrix penalties are

$$\phi_+(q) = \varphi(\max\{1, q/u\}), \qquad \phi_-(q) = \varphi(\min\{1, q/l\}), \quad \phi_-(0) = +\infty.$$

On each active interval the second derivative is $1/q^2 > 0$; at $q = u$ or $q = l$, the first derivative matches the zero derivative of the adjacent constant piece. These functions are convex and lower semicontinuous on their domains. The resulting matrix objective is strictly convex after adding LogDet, consistently with the joint uniqueness proof above. These formulas also specify the exact optimal slacks for later calculations.

Slack penalties control squared threshold violations. Set $Q = 4R^2\beta$ and $G^2 = 2\max\{uQ, l^2\}$. We prove, for every selected pair at the minimizing output,

$$\psi_\sigma(A_{\boldsymbol{m}}; e)^2 \le G^2 \varphi(\xi_{\boldsymbol{m},e}/\xi_{0,e}). \tag{36}$$

For a same-class pair with $q = q_{A_{\boldsymbol{m}}}(e) > u$, feasibility gives $\xi \ge q > u$, and the geometric bound gives $q \le Q$. For $t \ge 1$,

$$\varphi(t) \ge \frac{(t-1)^2}{2t}.$$

To verify this, the difference of the two sides is zero at one and has derivative $(t-1)^2/(2t^2) \ge 0$. Since φ is increasing on $[1, \infty)$, it follows that

$$(q-u)^2 \le 2uq\,\varphi(q/u) \le 2uQ\,\varphi(\xi/u) \le G^2\varphi(\xi/u).$$

For a different-class pair with $q < l$, feasibility gives $0 < \xi \le q < l$. For $0 < t \le 1$, Taylor's formula with integral remainder and $\varphi''(t) = t^{-2} \ge 1$ gives

$$\varphi(t) = \int_t^1 \frac{s-t}{s^2}\, ds \ge \int_t^1 (s-t)\, ds = \frac{(1-t)^2}{2}.$$

The function φ decreases on $(0, 1]$, and therefore

$$(l-q)^2 \le 2l^2\varphi(q/l) \le 2l^2\varphi(\xi/l) \le G^2\varphi(\xi/l).$$

If a selected pair satisfies its original threshold, its hinge violation is zero and equation 36 holds by nonnegativity. This covers every selected pair.

Average violations and classification. For either selected pool, Cauchy–Schwarz and the common sum budget imply

$$\begin{aligned}\widehat{L}_{\sigma,m_\sigma}(A_{\boldsymbol{m}}) &= \frac{1}{m_\sigma} \sum_{e\in\mathcal{T}_{\sigma,m_\sigma}} \psi_\sigma(A_{\boldsymbol{m}}; e) \\ &\le \left\{ \frac{1}{m_\sigma} \sum_{e\in\mathcal{T}_{\sigma,m_\sigma}} \psi_\sigma(A_{\boldsymbol{m}}; e)^2 \right\}^{1/2} \\ &\le G \left\{ \frac{1}{m_\sigma} \sum_{e\in\mathcal{T}_{\sigma,m_\sigma}} \varphi(\xi_{\boldsymbol{m},e}/\xi_{0,e}) \right\}^{1/2} \le G\sqrt{\frac{C}{\gamma m_\sigma}}.\end{aligned}$$

The matrix still lies in $\mathcal{A}_C$, so the same uniform finite-pool event applies. Adding its two approximation errors yields

$$L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le V_{\boldsymbol{m}},$$

where

$$V_{\boldsymbol{m}} = e_+(m_+) + e_-(m_-) + \sqrt{\frac{2C\max\{4uR^2\beta, l^2\}}{\gamma}} \left(m_+^{-1/2} + m_-^{-1/2}\right). \tag{37}$$

This is precisely the full-pool inequality required by the proof of Theorem 4.1. The affine comparison, SVM objective bound, and metric-robustness calculation therefore establish all its conclusions with equation 20, on the same event of probability at least $1-\delta$. Since C, γ, β are fixed across budgets, the additional term is nonincreasing in both m_+ and m_-. At a full pool the approximation error is zero, while the displayed slack term continues to control the selected violations. This completes the sum-slack extension.

C.7 An averaged hinge objective

This subsection treats an objective that averages the violations within each selected pool. Its normalization differs from the sum of slack penalties in equation 19, so we give its objective comparison separately. Let $\lambda_d > 0$, and let $A_{\boldsymbol{m}}$ minimize

$$\widehat{F}_{\boldsymbol{m}}(A) = \widehat{L}_{+,m_+}(A) + \widehat{L}_{-,m_-}(A) + \lambda_d D_{\mathrm{ld}}(A, A_0), \qquad A \succ 0. \tag{38}$$

The losses are $\psi_+(A; i, j) = [q_A(X_i, X_j) - u]_+$ and $\psi_-(A; i, j) = [l - q_A(X_i, X_j)]_+$, with the same thresholds as in Section 4. Throughout each finite-sample comparison, λ_d is fixed across the pair budgets. Both classes and both positive pair budgets are understood to be present; the fixed fallback completes the algorithm on the remaining samples.

Existence and a deterministic geometric bound. Set

$$Q_0 = 4R^2\|A_0\|_{\mathrm{op}}, \qquad C_{\lambda_d} = \frac{Q_0 + l}{\lambda_d}, \qquad F_S(A) = L_+(A) + L_-(A) + \lambda_d D_{\mathrm{ld}}(A, A_0).$$

For every pair, $0 \le q_{A_0}(X_i, X_j) \le Q_0$. Consequently, each same-class violation at A_0 is at most Q_0, and each opposite-class violation is at most l. The regularizer vanishes at A_0, giving

$$\widehat{F}_{\boldsymbol{m}}(A_0) \le Q_0 + l, \qquad F_S(A_0) \le Q_0 + l.$$

Both objectives are continuous and strictly convex on the positive-definite cone: the hinge terms are convex functions of the affine pairwise squared distances, and $\lambda_d D_{\mathrm{ld}}$ is strictly convex. Every sublevel set below the reference objective is contained in the compact LogDet set $\mathcal{A}_{C_{\lambda_d}}$ from Appendix C.1. Minimization over that set attains the global minimum, since points outside it have a larger regularization term than the reference objective. Thus the selected-pool and full-pool problems each have a unique optimizer. Write A_* for the latter and $O_S = F_S(A_*)$. In particular,

$$D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C_{\lambda_d}, \qquad D_{\mathrm{ld}}(A_*, A_0) \le C_{\lambda_d}.$$

This argument also covers coincident inputs with different labels: the hinge objective remains finite at the reference matrix.

Full-pool comparison and downstream risk. Compute β, B, and e_σ using C_{λ_d} in place of C. The budget is deterministic, so the uniform pair-sampling event of Appendix C.2 applies. On this event, for every $A \in \mathcal{A}_{C_{\lambda_d}}$ and every admissible budget,

$$|F_S(A) - \widehat{F}_{\boldsymbol{m}}(A)| \le e_+(m_+) + e_-(m_-) =: e_{\boldsymbol{m}}.$$

The regularizer cancels in this difference. Applying the inequality at the two optimizers gives the complete comparison

$$\begin{aligned} L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) &\le F_S(A_{\boldsymbol{m}}) \\ &\le \widehat{F}_{\boldsymbol{m}}(A_{\boldsymbol{m}}) + e_{\boldsymbol{m}} \\ &\le \widehat{F}_{\boldsymbol{m}}(A_*) + e_{\boldsymbol{m}} \\ &\le F_S(A_*) + 2e_{\boldsymbol{m}}. \end{aligned}$$

We have therefore proved, simultaneously over the selected pair budgets,

$$D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C_{\lambda_d}, \qquad L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le V_{\boldsymbol{m}}^{\mathrm{avg}} := O_S + 2\{e_+(m_+) + e_-(m_-)\}. \tag{39}$$

The full-pool value O_S is common to all budgets. It records the best tradeoff between the two full-pool violations and the regularizer for this sample.

Fit the affine SVM with penalty $\lambda_c \|w\|_2^2/2$, and put

$$H_{\boldsymbol{m}}^{\mathrm{avg}} = \min\left\{1, \frac{\alpha_S V_{\boldsymbol{m}}^{\mathrm{avg}}}{\Delta} + \frac{8\lambda_c R^2 \beta}{\Delta^2}\right\}.$$

The mean-distance comparison in Appendix C.4 gives empirical hinge loss at most $\alpha_S V_{\boldsymbol{m}}^{\mathrm{avg}}/\Delta$ and comparison-head norm at most $4R\sqrt{\beta}/\Delta$. Comparing the fitted SVM with this head and with the zero score bounds its objective by $H_{\boldsymbol{m}}^{\mathrm{avg}}$. Hence its empirical hinge loss is at most this quantity and its norm is at most $\sqrt{2H_{\boldsymbol{m}}^{\mathrm{avg}}/\lambda_c}$. The radius bound $a_i \le r\sqrt{\beta}$ then gives the clipped-hinge response bound

$$\varepsilon_{\mathrm{mr}} \le \min\left\{1, r\sqrt{2\beta H_{\boldsymbol{m}}^{\mathrm{avg}}/\lambda_c}\right\}.$$

Intersecting the pair event, whose failure probability is at most $\delta/2$, with the fixed-partition event at confidence $\delta/2$, proves

$$R_{01}(g_{\boldsymbol{m}} \circ f_{\boldsymbol{m}}) \le H_{\boldsymbol{m}}^{\mathrm{avg}} + \min\left\{1, r\sqrt{2\beta H_{\boldsymbol{m}}^{\mathrm{avg}}/\lambda_c}\right\} + \tau_n, \qquad \tau_n = \sqrt{\frac{K(r)\log 2 + \log(2/\delta)}{2n}}. \tag{40}$$

The joint event has probability at least $1 - \delta$, and the inequalities hold there whenever both classes occur, simultaneously over budgets. The empirical zero–one error is at most $H_{\boldsymbol{m}}^{\mathrm{avg}}$. The last two terms bound the clipped-hinge generalization gap. The risk bound may be truncated at one. Its dependence on the pair budgets is nonincreasing because O_S and the geometric constants are fixed across those budgets.

A common zero-violation comparator. Suppose $A^\dagger$ satisfies equation 15 with a fixed budget C. Comparison with this matrix yields

$$\widehat{F}_{\boldsymbol{m}}(A_{\boldsymbol{m}}) \le \widehat{F}_{\boldsymbol{m}}(A^\dagger) = \lambda_d D_{\mathrm{ld}}(A^\dagger, A_0) \le \lambda_d C.$$

All objective terms are nonnegative. Dividing the regularization bound by λ_d, and separately bounding the sum of the two hinge means, gives

$$D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C, \qquad \widehat{L}_{+,m_+}(A_{\boldsymbol{m}}) + \widehat{L}_{-,m_-}(A_{\boldsymbol{m}}) \le \lambda_d C.$$

Compute the sampling errors using this fixed geometric budget. Adding the two full-pool deviations proves the sharper bounds

$$D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) \le C, \qquad L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le \lambda_d C + e_+(m_+) + e_-(m_-). \tag{41}$$

Thus equation 40 also holds with

$$V_{\boldsymbol{m}}^{\mathrm{avg}} = \lambda_d C + e_+(m_+) + e_-(m_-), \qquad \beta = \|A_0\|_{\mathrm{op}} b(C).$$

The fixed budget matters when λ_d tends to zero: the automatic reference budget C_{λ_d} alone would grow, whereas the zero-violation comparator controls geometry uniformly.

Under the population assumptions of Corollary 4.2, take

$$\lambda_{d,n} = \lambda_{c,n} = r_n = n^{-1/(s+2)}, \qquad m_{\sigma,n} = \lfloor \theta_\sigma N_\sigma \rfloor, \qquad \delta = 2n^{-2}.$$

Appendix C.9 proves that the population comparator satisfies all sample thresholds almost surely and that, on an event of probability at least $1 - 2n^{-2} - 2e^{-n\pi_0^2/2}$, the class counts give $\alpha_S = O(1)$, $m_{\sigma,n} = \Theta(n^2)$, and $e_+(m_{+,n}) + e_-(m_{-,n}) = O(\sqrt{\log n}/n)$. Put $\kappa = 1/(s+2) < 1/2$. Equation equation 41 then gives

$$V_{\boldsymbol{m}_n}^{\mathrm{avg}} \le C n^{-\kappa} + O(\sqrt{\log n}/n) = O(n^{-\kappa}).$$

It follows that $H_{\boldsymbol{m}_n}^{\mathrm{avg}} = O(\lambda_{c,n})$. The fitted classifier norm is bounded, its metric robustness is $O(r_n)$, and the counting term is $O(n^{-\kappa})$, as verified in the corollary proof. Substitution into equation 40 proves the same classification-error rate and success probability as in that corollary. The selected-pair violation term vanishes because $\lambda_{d,n} C \to 0$.

Unit-offset hinge losses. Consider instead

$$\widetilde{\psi}_+(A; i, j) = [1 + q_A(X_i, X_j) - u]_+, \qquad \widetilde{\psi}_-(A; i, j) = [1 + l - q_A(X_i, X_j)]_+.$$

If $u > 1$, these are exactly the unshifted losses with thresholds $u' = u - 1$, $l' = l + 1$, and gap $\Delta' = l - u + 2$. The preceding results then apply with these effective thresholds. In particular, a zero-violation comparison matrix must satisfy $q_{A^\dagger} \le u - 1$ within class and $q_{A^\dagger} \ge l + 1$ across classes.

A full-pool comparison is also available for every $u > 0$. Denote the corresponding pool means by $\widetilde{L}_\sigma$ and $\widetilde{\widehat{L}}_{\sigma,m_\sigma}$, and minimize the two selected means plus $\lambda_d D_{\mathrm{ld}}$. At A_0, their sum is at most $Q_0 + l + 2$, so both selected and full-pool minimizers belong to

$$\mathcal{A}_{\widetilde{C}_{\lambda_d}}, \qquad \widetilde{C}_{\lambda_d} = (Q_0 + l + 2)/\lambda_d.$$

Compute β from this budget, put $Q = 4R^2\beta$, and use the loss bound $\widetilde{B} = \max\{Q + 1, l + 1\}$. Each offset hinge is $4R^2$-Lipschitz in A under the Frobenius norm, since its affine argument changes by at most $4R^2\|A - A'\|_F$. The finite-pool proof therefore supplies errors $\widetilde{e}_\sigma(k)$ with the formula in equation 16, replacing B by $\widetilde{B}$; at a full pool the error is zero.

Let

$$\widetilde{F}_S(A) = \widetilde{L}_+(A) + \widetilde{L}_-(A) + \lambda_d D_{\mathrm{ld}}(A, A_0), \qquad \widetilde{O}_S = \min_{A \succ 0} \widetilde{F}_S(A).$$

Its minimum exists uniquely by the same strict-convexity and compact-sublevel argument. On the uniform event, its selected counterpart $\widetilde{\widehat{F}}_{\boldsymbol{m}}$ differs from it by at most $\widetilde{e}_{\boldsymbol{m}} = \widetilde{e}_+(m_+) + \widetilde{e}_-(m_-)$. With $\widetilde{A}_*$ its full-pool minimizer, exact optimization gives

$$\begin{aligned} \widetilde{L}_+(A_{\boldsymbol{m}}) + \widetilde{L}_-(A_{\boldsymbol{m}}) &\le \widetilde{F}_S(A_{\boldsymbol{m}}) \\ &\le \widetilde{\widehat{F}}_{\boldsymbol{m}}(A_{\boldsymbol{m}}) + \widetilde{e}_{\boldsymbol{m}} \\ &\le \widetilde{\widehat{F}}_{\boldsymbol{m}}(\widetilde{A}_*) + \widetilde{e}_{\boldsymbol{m}} \le \widetilde{O}_S + 2\widetilde{e}_{\boldsymbol{m}}. \end{aligned}$$

The positive-part function is nondecreasing, so $\psi_\sigma \le \widetilde{\psi}_\sigma$ for both pools. Consequently the right-hand side also bounds $L_+(A_{\boldsymbol{m}})+L_-(A_{\boldsymbol{m}})$. Substituting this value for $V_{\boldsymbol{m}}$ in equation 17 and equation 18, with the recomputed geometry and the original gap Δ, proves the downstream bound for the offset objective. This argument includes $u \le 1$, where a positive effective same-class threshold is unavailable.

C.8 A DOWNSTREAM GAUSSIAN-KERNEL SVM

Fix $\sigma > 0$, and use the Gaussian kernel

$$k(v, v') = \exp\{-\|v - v'\|_2^2/(2\sigma^2)\}$$

on the learned representation $f_A(x) = A^{1/2}x$. Its reproducing kernel Hilbert space is denoted by $\mathcal{H}_k$. For the scalar squared-distance argument, write

$$\kappa(q) = e^{-q/(2\sigma^2)}, \qquad \eta = \kappa(u) - \kappa(l) > 0.$$

The positivity follows from $u < l$ and strict decrease of the exponential. At each pair budget, fit

$$(g_{\boldsymbol{m}}, b_{\boldsymbol{m}}) \in \operatorname*{argmin}_{g\in\mathcal{H}_k,\, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^{n}[1 - Y_i\{g(f_A(X_i)) + b\}]_+ + \frac{\lambda_c}{2}\|g\|_{\mathcal{H}_k}^2 \right\}, \qquad A = A_{\boldsymbol{m}}. \tag{42}$$

The intercept is unpenalized and $\lambda_c > 0$. We use a fixed measurable optimal intercept, justified below, and write $s_{\boldsymbol{m}}(x) = g_{\boldsymbol{m}}(f_A(x)) + b_{\boldsymbol{m}}$. Let $\widehat{R}_h$ denote empirical hinge loss.

Assume the metric step supplies $A_{\boldsymbol{m}} \in \mathcal{A}_C$ and $L_+(A_{\boldsymbol{m}}) + L_-(A_{\boldsymbol{m}}) \le V_{\boldsymbol{m}}$ on the simultaneous pair-sampling event. These statements follow from the hard and sum-slack analyses whenever the common comparator exists. Define

$$H_{\boldsymbol{m}}^k = \min\left\{1, \frac{\alpha_S V_{\boldsymbol{m}}}{2\sigma^2\eta} + \frac{2\lambda_c}{\eta^2}\right\}, \tag{43}$$

$$B_{\boldsymbol{m}}^k = H_{\boldsymbol{m}}^k + \min\left\{1, \frac{r}{\sigma}\sqrt{\frac{2\beta H_{\boldsymbol{m}}^k}{\lambda_c}}\right\} + \tau_n, \qquad \tau_n = \sqrt{\frac{K(r)\log 2 + \log(2/\delta)}{2n}}. \tag{44}$$

On the intersection with the fixed-partition event, simultaneously over the admissible budgets,

$$\widehat{R}_{01}(s_{\boldsymbol{m}}) \le \widehat{R}_h(s_{\boldsymbol{m}}) \le H_{\boldsymbol{m}}^k, \qquad R_{01}(s_{\boldsymbol{m}}) \le \min\{1, B_{\boldsymbol{m}}^k\},$$
$$|R_{01}(s_{\boldsymbol{m}}) - \widehat{R}_{01}(s_{\boldsymbol{m}})| \le \min\{1, B_{\boldsymbol{m}}^k\}.$$

The joint event has probability at least $1 - \delta$, with the feasibility implication of Theorem 4.1. When $V_{\boldsymbol{m}}$ is nonincreasing in each budget, these upper bounds have the same property with all other settings fixed.

Kernel means give a comparison classifier. Put $v_i = f_A(X_i)$, and define

$$g_0(v) = \frac{1}{\eta}\left\{\frac{1}{n_+}\sum_{j:Y_j=+1} k(v, v_j) - \frac{1}{n_-}\sum_{j:Y_j=-1} k(v, v_j)\right\}.$$

This is a finite linear combination of kernel sections and therefore belongs to $\mathcal{H}_k$. Each section has norm one. Applying the triangle inequality first within each average and then to their difference gives $\|g_0\|_{\mathcal{H}_k} \le 2/\eta$.

For $Y_i = c$, define the two kernel averages

$$M_{i,c} = \frac{1}{n_c}\sum_{j:Y_j=c} \kappa(q_A(X_i, X_j)), \qquad M_{i,-c} = \frac{1}{n_{-c}}\sum_{j:Y_j=-c} \kappa(q_A(X_i, X_j)).$$

Then $Y_i g_0(v_i) = (M_{i,c} - M_{i,-c})/\eta$, and

$$\eta - M_{i,c} + M_{i,-c} = \{\kappa(u) - M_{i,c}\} + \{M_{i,-c} - \kappa(l)\}.$$

Using $[a+b]_+ \le [a]_+ + [b]_+$ and convexity of the positive-part function for each average yields

$$[1 - Y_i g_0(v_i)]_+ \le \frac{1}{\eta}\Bigg\{ \frac{1}{n_c} \sum_{j:Y_j=c} [\kappa(u) - \kappa(q_A(X_i, X_j))]_+ + \frac{1}{n_{-c}} \sum_{j:Y_j=-c} [\kappa(q_A(X_i, X_j)) - \kappa(l)]_+ \Bigg\}.$$

For $q \ge 0$, the derivative satisfies $|\kappa'(q)| \le 1/(2\sigma^2)$. If $q \le u$, the first positive part is zero; if $q > u$, the mean value theorem bounds it by $(q-u)/(2\sigma^2)$. The opposite-class expression is treated with the cases $q \ge l$ and $q < l$. Thus

$$[\kappa(u) - \kappa(q)]_+ \le \frac{[q-u]_+}{2\sigma^2}, \qquad [\kappa(q) - \kappa(l)]_+ \le \frac{[l-q]_+}{2\sigma^2}.$$

Combining the inequalities and averaging over i gives the same classwise normalization as in Appendix C.4. More explicitly, the same-class contribution is

$$\frac{2}{n} \sum_{c\in\{-1,+1\}} \frac{1}{n_c} \sum_{\substack{i<j \\ Y_i=Y_j=c}} \psi_+(A; i, j) \le \frac{2N_+}{n n_{\min}} L_+(A).$$

Its diagonal terms vanish because $q_A(X_i, X_i) = 0 < u$. Every cross-class pair receives coefficient $n^{-1}(n_+^{-1} + n_-^{-1}) = (n_+ n_-)^{-1} = N_-^{-1}$, so its contribution equals $L_-(A)$. We conclude that

$$\widehat{R}_h(g_0 \circ f_A) \le \frac{\alpha_S}{2\sigma^2\eta}\{L_+(A) + L_-(A)\}.$$

The comparison uses intercept zero and has regularization cost at most $2\lambda_c/\eta^2$. The zero score has objective one. Exact minimization in equation 42 therefore bounds its objective by $H_{\boldsymbol{m}}^k$, and nonnegativity of its two terms gives

$$\widehat{R}_h(s_{\boldsymbol{m}}) \le H_{\boldsymbol{m}}^k, \qquad \|g_{\boldsymbol{m}}\|_{\mathcal{H}_k} \le \sqrt{2H_{\boldsymbol{m}}^k/\lambda_c}.$$

Existence of an exact solution. Let $\mathcal{V} = \operatorname{span}\{k(\cdot, v_i) : 1 \le i \le n\}$. Decompose $g = g_{\mathcal{V}} + g_\perp$. By the reproducing property, $g_\perp(v_i) = \langle g_\perp, k(\cdot, v_i)\rangle = 0$, so projection onto $\mathcal{V}$ preserves every training score. Pythagoras gives $\|g\|^2 = \|g_{\mathcal{V}}\|^2 + \|g_\perp\|^2$. An optimum may therefore be sought in the finite-dimensional space $\mathcal{V}$.

Consider its objective sublevel set at one; it is nonempty because $(g, b) = (0, 0)$ belongs to it. In this set, $\|g\| \le \sqrt{2/\lambda_c}$, and every individual hinge is at most n. For a positive example, the latter implies $b \ge 1 - n - g(v_i)$; for a negative example it implies $b \le n - 1 - g(v_j)$. Since $|g(v_i)| \le \|g\|$, both classes together give

$$|b| \le n - 1 + \sqrt{2/\lambda_c}.$$

The sublevel set is closed and bounded in $\mathcal{V} \times \mathbb{R}$, hence compact. The continuous objective attains its minimum there and consequently its global minimum. If two optima had different RKHS components, their midpoint would strictly reduce the norm penalty while convexity would not increase the hinge term, a contradiction. The RKHS component is therefore unique. Its optimal intercepts form a nonempty closed bounded convex subset of $\mathbb{R}$, hence a compact interval. That interval has a unique point of minimum absolute value. Appendix C.12 proves that this choice is measurable, including when the Gram matrix is singular.

Metric robustness and population risk. For any two representations,

$$\|k(\cdot, v) - k(\cdot, v')\|_{\mathcal{H}_k}^2 = k(v, v) + k(v', v') - 2k(v, v') = 2\left(1 - e^{-\|v-v'\|_2^2/(2\sigma^2)}\right) \le \frac{\|v - v'\|_2^2}{\sigma^2}.$$

The last inequality uses $1 - e^{-t} \le t$ for $t \ge 0$. Cauchy–Schwarz and the reproducing property imply

$$|g(v) - g(v')| \le \frac{\|g\|_{\mathcal{H}_k}}{\sigma}\|v - v'\|_2.$$

The constant intercept cancels in this difference. For a fixed label, the clipped hinge is one-Lipschitz as a function of its score and takes values in $[0, 1]$. Its response therefore satisfies

$$\omega_i(t) \le \min\{1, \|g_{\boldsymbol{m}}\|_{\mathcal{H}_k} t/\sigma\}, \qquad \varepsilon_{\mathrm{mr}} \le \min\left\{1, \frac{r}{\sigma}\sqrt{\frac{2\beta H^k_{\boldsymbol{m}}}{\lambda_c}}\right\},$$

where the second inequality uses $a_i \le r\sqrt{\beta}$ for every anchor. Theorem 3.5, applied with failure probability $\delta/2$, bounds the clipped-hinge generalization gap by this response plus τ_n. The empirical clipped hinge is at most the empirical hinge, and the population clipped hinge dominates zero–one risk. These observations give $R_{01} \le B^k_{\boldsymbol{m}}$. Separately, $\widehat{R}_{01} \le H^k_{\boldsymbol{m}} \le B^k_{\boldsymbol{m}}$. For any two nonnegative numbers bounded by $B^k_{\boldsymbol{m}}$, their absolute difference is also bounded by $B^k_{\boldsymbol{m}}$. This proves the stated zero–one gap inequality, including its truncation at one. The argument distinguishes this bound from the smaller response-plus-counting bound for the clipped-hinge gap.

Growing pair budgets. Under the assumptions of Corollary 4.2, keep σ fixed and use the stated fixed positive pool fractions. On the event proved below, $\eta > 0$ is fixed, $\alpha_S = O(1)$, and $V_{\boldsymbol{m}_n} = O(\sqrt{\log n}/n)$. With $\kappa = 1/(s+2)$ and $r_n = \lambda_{c,n} = n^{-\kappa}$,

$$H^k_{\boldsymbol{m}_n} \le \frac{\alpha_S V_{\boldsymbol{m}_n}}{2\sigma^2\eta} + \frac{2n^{-\kappa}}{\eta^2} = O(n^{-\kappa}), \qquad H^k_{\boldsymbol{m}_n}/\lambda_{c,n} = O(1).$$

The robustness term is consequently $O(r_n)$, and the partition count gives $\tau_n = O(n^{-\kappa})$. Thus $B^k_{\boldsymbol{m}_n} = O(n^{-\kappa})$, proving vanishing empirical error, population error, and zero–one gap while $m_n = \Theta(n^2)$. The success probability is the same $1 - 2n^{-2} - 2e^{-n\pi_0^2/2}$ as in the corollary. The constants depend on the fixed bandwidth through η and σ.

C.9 Consistency with quadratically many pair constraints

We prove Corollary 4.2, then describe the population condition and the role of the pair count.

Proof of Corollary 4.2. Write $\pi_c = \Pr(Y = c)$, $\pi_0 = \min_c \pi_c > 0$, and $\kappa = 1/(s+2)$. Since $s > 0$, we have $0 < \kappa < 1/2$. All constants below depend only on the fixed problem parameters and the fixed sampling fractions.

First, for each $i < j$, (Z_i, Z_j) has law $\mu \otimes \mu$. The population assumption therefore holds for this pair with probability one. Taking a finite union over the $\binom{n}{2}$ exceptional events proves that $A^\dagger$ satisfies every full-pool threshold almost surely. This establishes equation 15 simultaneously for all selected subsets. Independence between the different training pairs is not needed for this finite union.

Second, each n_c is a sum of independent Bernoulli variables with mean $n\pi_c$. The one-sided Hoeffding inequality gives

$$\Pr\{n_c/n < \pi_0/2\} \le \exp\{-2n(\pi_c - \pi_0/2)^2\} \le e^{-n\pi_0^2/2}.$$

Let $E^{\mathrm{count}}_n = \{n_{\min} \ge n\pi_0/2\}$. A union bound over the two labels gives

$$\Pr(E^{\mathrm{count}}_n) \ge 1 - 2e^{-n\pi_0^2/2}.$$

On this event and for sufficiently large n, both class counts are at least two. Thus

$$N_- = n_+ n_- \ge \pi_0^2 n^2/4,$$
$$N_+ = \frac{1}{2}\sum_c n_c(n_c - 1) \ge \frac{1}{4}\sum_c n_c^2 \ge \pi_0^2 n^2/8,$$
$$N_+ + N_- = \binom{n}{2}, \qquad \alpha_S = \max\{1, 2N_+/(n n_{\min})\} \le 2/\pi_0.$$

In the last inequality, $2N_+ \le n(n-1)$ and $n_{\min} \ge n\pi_0/2$ were used. Fix $\theta_+, \theta_- \in (0, 1]$. Once $\theta_\sigma N_\sigma \ge 2$,

$$\frac{1}{2}\theta_\sigma N_\sigma \le \lfloor \theta_\sigma N_\sigma \rfloor \le \theta_\sigma N_\sigma.$$

The displayed lower bounds on the pools therefore give deterministic positive constants bounding $m_{\sigma,n}/n^2$ above and below. In particular, both budgets are admissible, and $m_n = \Theta(n^2)$ on E_n^{count}.

Third, apply Theorem 4.1 with $\delta = 2n^{-2}$. The associated geometric constants β, B and $p = d(d+1)/2$ are fixed. For a budget that is not a full pool, the numerator of its square-root term in equation 16 is

$$p\log(1+2n^2) + \log(2n^4) = O(\log n).$$

Since $m_{\sigma,n} = \Theta(n^2)$, its sampling error is $O(\sqrt{\log n}/n)$; the additive covering error is $O(n^{-2})$. For a full pool the error is exactly zero. Therefore

$$e_+(m_{+,n}) + e_-(m_{-,n}) = O(\sqrt{\log n}/n).$$

For hard constraints this is $V_{\boldsymbol{m}_n}$. For sum-slack ITML, the additional term in equation 20 is

$$G\sqrt{C/\gamma}\{m_{+,n}^{-1/2} + m_{-,n}^{-1/2}\} = O(n^{-1}), \qquad G^2 = 2\max\{4uR^2\beta, l^2\}.$$

Here $\gamma > 0$ and all other quantities in the coefficient are fixed. For either metric learner,

$$V_{\boldsymbol{m}_n} = O(\sqrt{\log n}/n) = o(n^{-\kappa}),$$

because $n^{\kappa-1}\sqrt{\log n} \to 0$.

With $r_n = \lambda_{c,n} = n^{-\kappa}$, the bound on α_S gives

$$H_{\boldsymbol{m}_n} \le \frac{\alpha_S V_{\boldsymbol{m}_n}}{\Delta} + \frac{8R^2\beta}{\Delta^2} n^{-\kappa} = O(n^{-\kappa}).$$

Consequently $H_{\boldsymbol{m}_n}/\lambda_{c,n} = O(1)$, the fitted SVM norm is bounded, and the robustness contribution obeys

$$\min\left\{1, r_n\sqrt{2\beta H_{\boldsymbol{m}_n}/\lambda_{c,n}}\right\} = O(n^{-\kappa}).$$

For all sufficiently large n, the partition assumption applies at r_n. Using $\log(2/\delta) = 2\log n$,

$$\tau_n^2 \le \frac{C_K \log 2\, n^{\kappa s} + 2\log n}{2n} = O(n^{-2\kappa}).$$

Indeed, $\kappa s - 1 = -2\kappa$, while $\log n/n = o(n^{-2\kappa})$ because $2\kappa < 1$. Taking square roots gives $\tau_n = O(n^{-\kappa})$.

Finally, intersect the theorem event with E_n^{count} and the probability-one comparator event. The probability of the intersection is at least

$$1 - 2n^{-2} - 2e^{-n\pi_0^2/2}.$$

All admissible pair budgets are covered by the theorem event, so it also covers the sample-dependent budgets $\lfloor \theta_\sigma N_\sigma \rfloor$. On this intersection, equation 18 gives $R_{01}(g_n \circ f_n) = O(n^{-\kappa})$, and the empirical zero–one error is at most the empirical hinge loss, hence at most $H_{\boldsymbol{m}_n} = O(n^{-\kappa})$. Enlarging a deterministic constant if necessary gives the asserted bound for the maximum of the two errors. Their absolute difference is bounded by this maximum.

The exceptional probability tends to zero. Thus both errors and their gap converge to zero in probability. The deterministic lower bound $m_n \ge cn^2$ on the same events, for some $c > 0$, implies that for every fixed $M < \infty$, $\Pr(m_n/n > M) \to 1$. This proves the asserted divergence in probability of the parameter-to-sample ratio. □

A continuous realizable distribution. Let $\mathcal{X} = [-2,-1] \cup [1,2]$. Draw X uniformly from each interval with probability $1/2$, and assign labels -1 on the left and $+1$ on the right. Take $A^\dagger = 1$, $A_0 = 1/4$, $u = 1$, and $l = 4$. Within either interval, $|x - x'| \le 1$, so $q_{A^\dagger}(x, x') \le 1$. Across intervals, $|x - x'| \ge 2$, so $q_{A^\dagger}(x, x') \ge 4$. The population thresholds therefore hold everywhere on the labeled support.

The one-dimensional LogDet value is

$$D_{\mathrm{ld}}(1, 1/4) = 4 - \log 4 - 1 = 3 - \log 4 =: C.$$

The upper solution of $t - \log t - 1 = C$ is $b(C) = 4$; consequently $\beta = \|A_0\|_{\mathrm{op}} b(C) = 1$. To exhibit the required fixed partitions of the entire labeled space, divide each interval into $\lceil 1/r \rceil$ disjoint Borel subintervals of length at most r, assigning shared endpoints consistently. For each input subinterval, take its product with each of the two labels. This gives

$$K(r) = 4\lceil 1/r \rceil \le 8/r \qquad (0 < r \le 1).$$

It includes the zero-probability label combinations, so it is a partition of $\mathcal{X} \times \mathcal{Y}$. The corollary applies with $s = 1$ and yields $O(n^{-1/3})$.

At the reference matrix, an opposite-class squared distance is $(X - X')^2/4$. It is strictly less than four almost surely, since equality requires the two extreme endpoints, an event of probability zero. Thus the reference matrix violates the opposite-class threshold almost surely conditional on opposite labels, while the displayed comparator satisfies it. The learned representation can retain continuous within-class variation; the population condition does not require each class to collapse to a point.

Affine margin separation implied by the thresholds. Let μ_c be the conditional input distribution given $Y = c$, which exists because $\pi_c > 0$. Define $v(x) = (A^\dagger)^{1/2} x$ and

$$q_c^*(x) = \int \|v(x) - v(x')\|_2^2 \, \mu_c(dx'), \qquad s_*(x) = \frac{q_-^*(x) - q_+^*(x)}{\Delta}.$$

Bounded inputs make all these integrals finite. The population assumption, restricted to the two same-class product measures and to each cross-class product measure, holds almost everywhere under those measures. Fubini's theorem implies that for μ_c-almost every x, its within-class integrand is at most u for μ_c-almost every x', and its cross-class integrand is at least l for μ_{-c}-almost every x'. Integrating therefore gives

$$q_c^*(x) \le u, \qquad q_{-c}^*(x) \ge l \quad \text{for } \mu_c\text{-almost every } x.$$

For $c = +1$, the numerator of s_* is at least Δ; for $c = -1$, it is at most $-\Delta$. Thus $Y s_*(X) \ge 1$ almost surely.

To identify the classifier, let

$$\bar{v}_c = \int v(x') \, \mu_c(dx'), \qquad a_c^* = \int \|v(x')\|_2^2 \, \mu_c(dx').$$

Expanding the squared norm gives

$$q_c^*(x) = \|v(x)\|_2^2 - 2\langle v(x), \bar{v}_c \rangle + a_c^*.$$

Subtracting the two expressions cancels $\|v(x)\|_2^2$, hence

$$s_*(x) = \frac{2\langle v(x), \bar{v}_+ - \bar{v}_- \rangle + a_-^* - a_+^*}{\Delta}.$$

This is an affine score of margin at least one. In particular, population threshold realizability is a separability assumption of sufficient strength to permit vanishing classification error.

Every learned $A \succ 0$ is invertible. Therefore the affine scores $w^\top A^{1/2} x + b$, with unrestricted w, b, form the same class as $\vartheta^\top x + b$, under the one-to-one change of variables $\vartheta = A^{1/2} w$. The regularizer becomes

$$\|w\|_2^2 = \vartheta^\top A^{-1} \vartheta.$$

The metric changes the regularization geometry and the fitted solution within that class. Corollary 4.2 controls the fitted classifiers through objective comparison and metric robustness. The lower spectral bound also permits a uniformly bounded affine comparison score at every pair budget, as proved in Appendix C.10.

Pair budgets and independent coordinates. The quantity m_n counts pair constraints and their associated dual coordinates; $p = d(d+1)/2$ counts independent entries of the symmetric matrix. The corollary keeps p and the other geometric constants fixed. Its conclusion shows that quadratically many pair coordinates are compatible with vanishing errors under the stated conditions.

The pair-dependent proof itself also gives the stated rate for linear budgets. Suppose each of $m_{+,n}$ and $m_{-,n}$ lies between fixed positive multiples of n, on the class-count event. The sampling errors are then $O(\sqrt{\log n/n})$, and the sum-slack contribution is $O(n^{-1/2})$. For $\kappa < 1/2$,

$$n^{\kappa}\sqrt{\log n/n} = n^{\kappa-1/2}\sqrt{\log n} \longrightarrow 0.$$

Thus $V_{\boldsymbol{m}_n} = o(n^{-\kappa})$, and the rest of the proof gives the same covering-based rate. At the other extreme, taking every pair sets the finite-pool sampling errors to zero, while the counting term from the n observations remains. For sum-slack ITML, the fixed γ multiplies an unnormalized sum; its bounded total slack cost yields the $m_\sigma^{-1/2}$ mean-violation bounds. Rescaling that penalty as the budget grows would require changing this part of the analysis.

C.10 Comparison with fixed pair budgets

The population assumption in Corollary 4.2 permits a useful comparison with fixed pair budgets. We first give a finite-sample bound that holds uniformly over the entire LogDet sublevel set, then compare its constants with the bounds obtained from selected pairs. Throughout this subsection, the partition is fixed independently of the sample and is label-pure, as in Definition 3.1.

Define

$$\underline{\beta} = b_-(C)\lambda_{\min}(A_0) > 0, \qquad \widetilde{H} = \min\left\{1, \frac{8\lambda_c R^2\beta^2}{\underline{\beta}\Delta^2}\right\}. \tag{45}$$

Appendix C.1 gives $\underline{\beta} I \preceq A \preceq \beta I$ for every $A \in \mathcal{A}_C$. All constants in equation 45 are independent of the selected pairs.

Proposition C.1 (Fixed-budget comparison)**.** *Fix $r, \lambda_c > 0$ and $\delta \in (0,1)$, and use Π_r and τ_n from Section 4.2. There is an event over S of probability at least $1 - \delta/2$ on which the following implication holds. If both labels occur and equation 15 holds, then, for every $A \in \mathcal{A}_C$, any exact affine SVM fitted to $f_A(x) = A^{1/2}x$ satisfies*

$$\widehat{R}_{01}(g_A \circ f_A) \le \widetilde{H}, \tag{46}$$

$$R_{01}(g_A \circ f_A) \le \widetilde{H} + \min\left\{1, r\sqrt{\frac{2\beta\widetilde{H}}{\lambda_c}}\right\} + \tau_n. \tag{47}$$

The last two terms of equation 47 also bound the clipped-hinge generalization gap. On the probability-$1-\delta$ event of Theorem 4.1, its hard-constraint and sum-slack bounds remain valid after replacing $H_{\boldsymbol{m}}$ by

$$\overline{H}_{\boldsymbol{m}} = \min\{H_{\boldsymbol{m}}, \widetilde{H}\}. \tag{48}$$

Proof. Fix a sample containing both classes and satisfying equation 15. Choose a comparison matrix $A^\dagger$ from that condition. For each class, define

$$\bar{X}_c = \frac{1}{n_c}\sum_{j:Y_j=c} X_j, \qquad q_c^\dagger(x) = \frac{1}{n_c}\sum_{j:Y_j=c}(x - X_j)^\top A^\dagger(x - X_j), \qquad s^\dagger(x) = \frac{q_-^\dagger(x) - q_+^\dagger(x)}{\Delta}.$$

For every anchor with label c, the same-class thresholds, including the zero diagonal distance, give $q_c^\dagger(X_i) \le u$. The different-class thresholds give $q_{-c}^\dagger(X_i) \ge l$. Consequently $Y_i s^\dagger(X_i) \ge 1$ for all i, so this score has zero empirical hinge loss.

Expanding the quadratic terms cancels the common $x^\top A^\dagger x$ term and gives

$$s^\dagger(x) = \vartheta_S^\top x + b_S^\dagger, \qquad \vartheta_S = \frac{2A^\dagger(\bar{X}_+ - \bar{X}_-)}{\Delta},$$

$$b_S^\dagger = \frac{1}{\Delta}\left(\frac{1}{n_-}\sum_{j:Y_j=-1} X_j^\top A^\dagger X_j - \frac{1}{n_+}\sum_{j:Y_j=+1} X_j^\top A^\dagger X_j\right).$$

Since $\|\bar{X}_c\|_2 \le R$ and $A^\dagger \preceq \beta I$,

$$\|\vartheta_S\|_2 \le \frac{4R\beta}{\Delta}.$$

Now let A be any element of $\mathcal{A}_C$. The same input score is represented by $w_A^\dagger = A^{-1/2}\vartheta_S$ and intercept $b_S^\dagger$, since $\langle A^{-1/2}\vartheta_S, A^{1/2}x\rangle = \langle \vartheta_S, x\rangle$. The lower spectral bound gives

$$\|w_A^\dagger\|_2^2 = \vartheta_S^\top A^{-1}\vartheta_S \le \frac{16R^2\beta^2}{\underline{\beta}\Delta^2}.$$

Comparing the fitted SVM with this zero-hinge score and with the zero score therefore yields

$$J_A(w_A, b_A) \le \min\left\{1, \frac{\lambda_c}{2}\|w_A^\dagger\|_2^2\right\} \le \widetilde{H}, \qquad \widehat{R}_h(g_A \circ f_A) \le \widetilde{H}, \qquad \|w_A\|_2 \le \sqrt{2\widetilde{H}/\lambda_c}.$$

These inequalities are deterministic and simultaneous over $A \in \mathcal{A}_C$. They use only the existence of the comparison matrix; a measurable selection of that matrix is unnecessary. Existence of the fitted SVM minimizers follows from the compact-sublevel argument in Appendix C.5.

For the clipped hinge, Cauchy–Schwarz and the 1-Lipschitz positive part and clipping maps give $\omega_i(t) \le \min\{1, \|w_A\|_2 t\}$. The cell radii satisfy $a_i \le r\sqrt{\beta}$, whence

$$\varepsilon_{\mathrm{mr}}(f_A, g_A, S; \Pi_r) \le \min\left\{1, r\sqrt{2\beta\widetilde{H}/\lambda_c}\right\}.$$

The cell-count event at failure level $\delta/2$ has probability at least $1-\delta/2$. It is uniform over the learned functions, so Theorem 3.5 applies to every one of these outputs on the same event. The empirical clipped hinge is at most the empirical hinge, and the clipped hinge bounds the zero–one loss. This proves equation 46–equation 47 and the gap assertion. The probability statement is an unconditional event with a feasibility implication; it does not condition the sample law on feasibility.

On the event of Theorem 4.1, the same fitted objective is bounded by both $H_{\boldsymbol{m}}$ and $\widetilde{H}$. Its empirical hinge and squared head norm are therefore controlled by their minimum. Substitution into the preceding metric-robustness calculation proves equation 48 for hard constraints and for the stated sum-slack formulation. No additional probability event is needed. □

Consistency with fixed or zero pair budgets. Assume the population threshold condition and the fixed-parameter and covering assumptions of Corollary 4.2. Put $\kappa = 1/(s+2)$ and $r_n = \lambda_{c,n} = n^{-\kappa}$. The population comparison matrix satisfies all full-sample thresholds almost surely, as proved in Appendix C.9. Every exact hard or sum-slack solution belongs to $\mathcal{A}_C$ for every admissible selection, regardless of its size. The reference matrix A_0 also belongs to this set, since its divergence is zero. Thus Proposition C.1 applies to any of these outputs and, in particular, to $A_n = A_0$, which uses no training pairs.

Use the proposition with $\delta = 2n^{-2}$. Its count event then has probability at least $1 - n^{-2}$. Both labels occur with probability at least $1 - 2(1-\pi_0)^n$, because $\Pr(n_c = 0) = (1-\pi_c)^n \le (1-\pi_0)^n$. On the intersection,

$$\widetilde{H} = O(n^{-\kappa}), \qquad \min\{1, r_n\sqrt{2\beta\widetilde{H}/\lambda_{c,n}}\} = O(n^{-\kappa}), \qquad \tau_n = O(n^{-\kappa}).$$

The last assertion follows from

$$\tau_n^2 \le \frac{C_K \log 2\, n^{\kappa s} + 2\log n}{2n} = O(n^{-2\kappa}),$$

using $\kappa s - 1 = -2\kappa$ and $2\kappa < 1$. Consequently, with probability at least $1 - n^{-2} - 2(1-\pi_0)^n$,

$$\max\{R_{01}(g_n \circ f_{A_n}), \widehat{R}_{01}(g_n \circ f_{A_n})\} = O(n^{-1/(s+2)}).$$

This statement includes fixed budgets $m_+ = m_- = 1$ for $n \ge 3$ on the event that both labels occur, as well as the reference matrix without metric training. The constants do not depend on the selected budgets. The proof also covers arbitrary selections for this comparison bound; the pair-dependent improvement in Theorem 4.1 continues to require its specified random-prefix sampling scheme.

What additional pairs improve. The comparison identifies two sources of control. The full-sample comparator and the LogDet lower spectral bound give a bounded affine separator in every admissible representation. This suffices for the consistency rate. The selected-pair calculation instead bounds the empirical loss of the mean-distance classifier formed in the learned representation itself. Its regularization contribution is $8\lambda_c R^2\beta/\Delta^2$, whereas the transported comparator uses $8\lambda_c R^2\beta^2/(\underline{\beta}\Delta^2)$.

To make this numerical comparison explicit, write $a = 8\lambda_c R^2\beta/\Delta^2$, $\chi = \beta/\underline{\beta} \geq 1$, and $v = \alpha_S V_{\boldsymbol{m}}/\Delta$. Then

$$H_{\boldsymbol{m}} = \min\{1, a+v\}, \qquad \widetilde{H} = \min\{1, a\chi\}, \qquad \overline{H}_{\boldsymbol{m}} = \min\{1, a+v, a\chi\}.$$

For fixed remaining quantities, increasing the pair budgets decreases v and therefore makes the combined bound nonincreasing. The pair-dependent objective bound is strictly smaller than the comparison bound whenever

$$a + v < \min\{1, a\chi\}.$$

For example, under hard constraints with both complete pools, $v = 0$, so strict improvement holds if $\chi > 1$ and $a < 1$. The untruncated risk expression $h + \min\{1, r\sqrt{2\beta h/\lambda_c}\} + \tau_n$ is strictly increasing in $h \geq 0$, because of its first term. The same strict improvement therefore holds for that expression. If both final expressions are truncated at one, strict improvement of the objective bound need not produce a nontrivial risk guarantee. These are comparisons between valid upper bounds; they assert no monotonicity of the fitted classifier's actual risk.

The quadratic-budget corollary establishes compatibility between a growing number of constraint-associated coordinates and vanishing classification errors. The finite-sample result further quantifies when increased pair coverage improves the explicit guarantee relative to the comparison available at every budget.

C.11 A FINITE-SAMPLE INCREASE IN CLASSIFICATION RISK

The upper bounds in Theorem 4.1 are nonincreasing in the pair budgets, while the fitted classifier's actual risk can increase after a constraint is added. We give a finite-sample construction satisfying the common-comparator condition, with both total pair counts larger than the sample size. The construction has positive probability under the nested uniform sampling protocol.

Distribution, sample, and selected pairs. Take four possible input points,

$$p_- = (-1, 0), \quad p_1 = (1, 1), \quad p_2 = (3, 1), \quad x_q = (1, 0),$$

with labels $-1, +1, +1, +1$, respectively. Give each of the first three points probability η, and give x_q probability $1 - 3\eta$, where $0 < \eta < 1/3$. Let $n = 5$. The training multiset $\{p_-, p_1, p_1, p_1, p_2\}$ has probability

$$\frac{5!}{1!\,3!\,1!}\eta^5 = 20\eta^5.$$

There are six same-class index pairs: three between distinct copies of p_1, and three between a copy of p_1 and p_2. There are four opposite-class pairs, all incident to p_-. Use all four opposite-class pairs. Under a uniform permutation of the same-class pool, the probability that its first three elements are precisely the three zero-distance pairs is $3!3!/6! = 1/20$. On that event compare the budgets $\boldsymbol{m}_1 = (3, 4)$ and $\boldsymbol{m}_2 = (4, 4)$. The additional same-class pair joins p_1 to p_2. The total counts are seven and eight, both greater than five, and the joint event has probability $\eta^5 > 0$.

The two hard-constraint metric solutions. Use $A_0 = I$, $u = 0.4$, and $l = 1$. At the first budget, I has zero same-class distances on all selected same-class pairs and opposite-class distances five or seventeen. It is feasible and is the unique unconstrained minimizer of LogDet, so $A_1 = I$.

At the second budget, the additional constraint is $4A_{11} \leq 0.4$. Write a positive-definite matrix as

$$A = \begin{pmatrix} a & t \\ t & c \end{pmatrix}, \qquad a > 0,\ c > 0,\ ac - t^2 > 0.$$

Its LogDet objective is $a + c - \log(ac - t^2) - 2$. For fixed a, c, this is minimized uniquely at $t = 0$. The relaxed problem retaining only $a \leq 0.1$ therefore minimizes $h(a) + h(c)$, where $h(z) = z - \log z - 1$.

The second term has unique minimum at $c = 1$. Since $h'(a) = 1 - 1/a < 0$ for $0 < a \le 0.1$, the first is minimized uniquely at $a = 0.1$. Thus the relaxed optimizer is $\mathrm{diag}(0.1, 1)$.

At this matrix, $q_A(p_-, p_1) = 4(0.1) + 1 = 1.4$ and $q_A(p_-, p_2) = 16(0.1) + 1 = 2.6$, both above one. It is feasible for the full second problem and hence is its unique optimum:

$$A_2 = \mathrm{diag}(0.1, 1).$$

It also satisfies every full-pool same-class constraint, including the unselected copies of the (p_1, p_2) pair. Consequently it is a common comparator with

$$C = D_{\mathrm{ld}}(A_2, I) = 0.1 - \log 0.1 - 1 = \log 10 - 0.9.$$

Exact downstream SVM solutions. Fit the average-hinge affine SVM with $\lambda_c = 0.1$. For $A = \mathrm{diag}(a, 1)$, put $D = 4a + 1$. We claim that whenever $\lambda_c < D/10$, its unique score has

$$w = \frac{(4\sqrt{a}, 2)}{D}, \qquad b = -\frac{1}{D}. \tag{49}$$

The represented points are $(-\sqrt{a}, 0)$, $(\sqrt{a}, 1)$, and $(3\sqrt{a}, 1)$. Direct substitution gives signed margins one at the negative point and every copy of p_1, and margin $(12a + 1)/D > 1$ at p_2. All hinge losses are therefore zero.

To verify global optimality, divide the SVM objective by λ_c and introduce nonnegative slacks ζ_i. The resulting convex quadratic program is

$$\min_{w,b,\zeta \ge 0} \frac{1}{2}\|w\|_2^2 + C_{\mathrm{svm}} \sum_i \zeta_i, \qquad Y_i(w^\top v_i + b) \ge 1 - \zeta_i, \quad C_{\mathrm{svm}} = \frac{1}{5\lambda_c}.$$

Assign margin-constraint multipliers $2/D$ to the negative point and one copy of p_1, and zero to the other observations. The inequality $\lambda_c < D/10$ is exactly $2/D < C_{\mathrm{svm}}$. Thus the box constraints $0 \le \alpha_i \le C_{\mathrm{svm}}$ hold. These multipliers satisfy $\sum_i \alpha_i Y_i = 0$ and

$$\sum_i \alpha_i Y_i v_i = \frac{2}{D}(\sqrt{a}, 0) + \frac{2}{D}(\sqrt{a}, 1) = \frac{(4\sqrt{a}, 2)}{D} = w.$$

Set all slacks to zero and assign the slack-nonnegativity multipliers $C_{\mathrm{svm}} - \alpha_i$. The stationarity equations in w, b, ζ now hold. The two positive margin multipliers are attached to unit margins, so margin complementary slackness holds; slack complementary slackness holds because every slack is zero. All primal and dual constraints are satisfied. These equations are sufficient for optimality of the convex program, proving the claim.

Strict convexity of the quadratic norm makes w unique across optima. To see that the intercept is unique as well, use the displayed optimal dual solution: its two positive multipliers lie strictly below C_{svm}. Complementary slackness with any primal optimum forces zero slacks and unit margins at both corresponding observations, which determine b. Both $a = 1$ and $a = 0.1$ satisfy the strict condition because $D/10$ is respectively 0.5 and 0.14.

At the query point, the score is

$$s_A(x_q) = \frac{4a - 1}{4a + 1}.$$

It is $3/5$ for A_1 and $-3/7$ for A_2. The three observed support points remain correctly classified, while the positive query point changes from correct to incorrect. Thus the empirical zero–one and hinge losses stay zero, but the population error increases from zero to $1 - 3\eta$.

The population threshold condition fails in this example. The same-class pair (p_1, p_2) requires $4A_{11} \le 0.4$, whereas the opposite-class pair (p_-, x_q) requires $4A_{11} \ge 1$. Both pairs have positive product probability. This identifies precisely the distinction between the finite-sample comparator, which exists here, and the stronger population assumption used in the consistency corollary.

The sum-slack formulation. The same construction gives an increase for equation 19 with $\gamma = 1$. At the first budget, $A = I$ and all reference slacks have objective zero, so I is again the unique

solution. At the second budget, the zero-distance same-class penalties vanish. Drop the nonnegative opposite-class penalties temporarily. The remaining objective is

$$D_{\mathrm{ld}}(A, I) + h(\max\{1, 10A_{11}\}).$$

For fixed $a = A_{11}$, the same determinant argument uniquely minimizes the other entries at $t = 0, c = 1$. The scalar objective is therefore

$$h(a) + h(\max\{1, 10a\}).$$

For $0 < a \leq 0.1$, its derivative is $1 - 1/a < 0$. For $a > 0.1$, its derivative is

$$(1 - 1/a) + 10(1 - 1/(10a)) = 11 - 2/a.$$

It is negative up to $a = 2/11$, zero there, and positive afterwards. The unique relaxed minimizer is consequently $A = \mathrm{diag}(2/11, 1)$. Its two opposite-class squared distances are $19/11$ and $43/11$, both larger than one; all omitted penalties are zero. It is therefore the unique full sum-slack solution. Now $D = 19/11$, so $\lambda_c = 0.1 < D/10$, and equation 49 applies. The query score is $-3/19$. The population error again increases from zero to $1 - 3\eta$, with both empirical losses still zero.

C.12 Measurable outputs and optimization accuracy

The probability statements concern measurable outputs. We describe explicit selections for the convex algorithms and then give the changes needed for approximate solutions. The prediction rule uses a fixed tie convention throughout.

Hard LogDet optimization. Fix the labels and a selected index pattern. Let $v_S(A)$ be the sum of its selected hinge violations. It is jointly continuous in the input coordinates and A, is convex in A, and is nonnegative. For an integer $h \geq 1$, the set $\mathcal{A}_h$ is nonempty, compact, and convex. For each integer $j \geq 1$, let

$$A_{j,h}(S) = \operatorname*{argmin}_{A \in \mathcal{A}_h} \{D_{\mathrm{ld}}(A, A_0) + j v_S(A)\}.$$

Strict convexity and compactness give a unique minimizer. To prove continuity in the input coordinates, take $S_k \to S$. Compactness gives a convergent subsequence of $A_{j,h}(S_k)$. Its defining optimality inequality holds against every fixed $A \in \mathcal{A}_h$; passing to the limit by joint continuity shows that the subsequential limit minimizes the limiting objective. Uniqueness identifies it with $A_{j,h}(S)$. Every convergent subsequence has this limit, so the entire sequence converges. This proves continuity.

The function $d_h(S) = \min_{A \in \mathcal{A}_h} v_S(A)$ is also continuous. For example, joint continuity is uniform on $\mathcal{A}_h$ when $S_k \to S$, and therefore

$$|d_h(S_k) - d_h(S)| \leq \sup_{A \in \mathcal{A}_h} |v_{S_k}(A) - v_S(A)| \longrightarrow 0.$$

Thus $E_h = \{S : d_h(S) = 0\}$ is measurable. It is exactly the event that a feasible matrix exists in $\mathcal{A}_h$. On E_h, let A_* be the unique hard optimizer. Its objective is at most h, so $A_* \in \mathcal{A}_h$. Comparison gives

$$D_{\mathrm{ld}}(A_{j,h}, A_0) + j v_S(A_{j,h}) \leq D_{\mathrm{ld}}(A_*, A_0) \leq h.$$

Hence $v_S(A_{j,h}) \leq h/j$. Every limit point as $j \to \infty$ is feasible by continuity and has LogDet value at most the hard optimum. Uniqueness forces every such limit point to equal A_*, proving $A_{j,h} \to A_*$. It follows that the optimizer restricted to E_h is measurable as a pointwise limit.

The feasible-data event is $E = \bigcup_{h \geq 1} E_h$: any feasible positive-definite matrix has finite LogDet value and belongs to some integer sublevel. On E, choose the least h for which $S \in E_h$, and use the preceding limit. This is a measurable construction on countably many measurable pieces, and it returns the same unique global hard optimum. On E^c, use the fixed fallback. For each fixed n, there are finitely many label and index patterns, so combining them preserves measurability.

Sum-slack and averaged-hinge optimization. For sum-slack ITML, its feasible-data domain consists exactly of samples for which every selected opposite-label pair has distinct inputs. Necessity follows because a positive slack cannot be bounded above by zero. For sufficiency, take any

positive-definite A; its distance on a nonzero pair difference is positive, so slacks $\max\{u, q_A\}$ and $\min\{l, q_A\}$ satisfy all constraints and are positive.

On a fixed label/index pattern this domain is open. The slack-eliminated objective in Appendix C.6 is jointly continuous there and strictly convex in A. Around any sample in the domain, every selected opposite-class distance at A_0 stays bounded away from zero. The reference objective is consequently locally bounded. Optimizers for samples in that neighborhood all belong to one compact LogDet sublevel, by comparison with A_0. For a converging sequence of samples, a convergent subsequence of optimizers can therefore be taken in this common set. Passing their optimality inequalities to the limit against any fixed positive-definite matrix proves that the limit is the unique limiting optimizer. The optimizer is continuous on the feasible-data domain and is measurable after assigning the fallback on its complement. The optimal slacks are continuous functions of that matrix and the data on this domain.

The averaged-hinge objective is jointly continuous for all input configurations and has a finite reference objective everywhere. On the bounded input domain, its reference bound is uniform for fixed λ_d. The same compactness and uniqueness argument proves continuity of its optimizer for each fixed label/index pattern. Offset hinge losses have these properties as well.

Affine SVM selection. For a fixed two-class label pattern, let the represented inputs be $v_1, \dots, v_n$. Add ϵb^2, where $\epsilon > 0$, to the average-hinge SVM objective. Its regularizer is strictly convex in (w, b), so it has a unique minimizer (w_ϵ, b_ϵ). Comparison with the zero score bounds its objective by one and gives $\|w_\epsilon\| \le \sqrt{2/\lambda_c}$. Every hinge is at most n, and observations from both labels yield

$$|b_\epsilon| \le n - 1 + \sqrt{2/\lambda_c} \max_i \|v_i\|.$$

The same bound applies to any unperturbed optimum. It is locally uniform in the represented inputs and independent of ϵ. The joint-continuity argument just given therefore proves continuity of (w_ϵ, b_ϵ) for fixed ϵ.

The unperturbed objective has a unique minimizing w_*, because of its strictly convex quadratic penalty. Its minimizing intercepts form a compact interval. Let b_* be the unique point of this interval with smallest absolute value. If J_* is the unperturbed minimum, optimality of the perturbed solution gives

$$J_* + \epsilon b_\epsilon^2 \le J(w_\epsilon, b_\epsilon) + \epsilon b_\epsilon^2 \le J_* + \epsilon b_*^2.$$

Thus $|b_\epsilon| \le |b_*|$ and $J(w_\epsilon, b_\epsilon) \to J_*$ as $\epsilon \downarrow 0$. The uniform compact bound makes every subsequential limit an unperturbed optimum with intercept of absolute value at most $|b_*|$. By its defining uniqueness, that limit must be (w_*, b_*). Consequently $(w_{1/j}, b_{1/j}) \to (w_*, b_*)$, which gives a measurable selection as a pointwise limit of continuous maps. Composition with a measurable metric output preserves measurability. Samples missing a class receive the stipulated fixed fallback.

Gaussian SVM selection with singular Gram matrices. Let $\mathbf{K} = (k(v_i, v_j))_{i,j=1}^n$. It is positive semidefinite with diagonal entries one. Projection onto the training-kernel span, already proved above, reduces the problem to

$$J_{\mathbf{K}}(a, b) = \frac{1}{n} \sum_i [1 - Y_i\{(\mathbf{K}^{1/2}a)_i + b\}]_+ + \frac{\lambda_c}{2}\|a\|_2^2, \qquad a \in \mathbb{R}^n,\ b \in \mathbb{R}.$$

Here a is a coefficient variable for this finite-dimensional construction. If it has a component in $\ker \mathbf{K}$, removing that component preserves all scores and strictly decreases the penalty unless the component is zero. A minimizer therefore lies in $\operatorname{range} \mathbf{K}$. Define $\alpha = \mathbf{K}^{+1/2}a$, where the superscript denotes the pseudoinverse square root. The RKHS element $g(v) = \sum_j \alpha_j k(v_j, v)$ then has training scores $\mathbf{K}\alpha = \mathbf{K}^{1/2}a$ and norm

$$\|g\|_{\mathcal{H}_k}^2 = \alpha^\top \mathbf{K}\alpha = \|a\|_2^2.$$

Conversely, for any coefficient vector α, taking $a = \mathbf{K}^{1/2}\alpha$ gives these scores and this norm. This proves the finite-dimensional equivalence even when $\mathbf{K}$ has deficient rank.

For $\epsilon > 0$, the objective $J_{\mathbf{K}}(a, b) + \epsilon b^2$ has a unique minimizer. The square-root map is continuous on the positive-semidefinite cone. Moreover,

$$|(\mathbf{K}^{1/2}a)_i| \le \|\mathbf{K}^{1/2}e_i\|_2 \|a\|_2 = \sqrt{\mathbf{K}_{ii}}\|a\|_2 = \|a\|_2.$$

Comparison with zero therefore gives $\|a\|_2 \le \sqrt{2/\lambda_c}$ and, when both classes occur, $|b| \le n - 1 + \sqrt{2/\lambda_c}$. Compactness and uniqueness prove continuity of the perturbed minimizer in $\mathbf{K}$ on this fixed-diagonal domain. The same minimum-absolute-intercept argument proves that, as $\epsilon \downarrow 0$, these minimizers converge to the unperturbed minimizer with the chosen intercept. That pair is consequently measurable in $\mathbf{K}$.

Finally, spectral decomposition verifies the pointwise identity

$$\mathbf{K}^{+1/2} = \lim_{j\to\infty} (\mathbf{K} + j^{-1}I)^{-1}\mathbf{K}^{1/2}.$$

For a positive eigenvalue t, the corresponding scalar expression $\sqrt{t}/(t + j^{-1})$ tends to $1/\sqrt{t}$; at $t = 0$ it is identically zero. Each finite-j matrix function is continuous, so the limit is measurable. Thus α, the intercept, and the evaluated score are measurable even across changes in Gram-matrix rank. This selection concerns the unique RKHS function; its kernel-expansion coefficients need not themselves be unique before a convention is imposed.

Random pair selection. For fixed n, the possible label patterns, admissible index sets, and within-pool permutations are finite. They can be represented by a finite auxiliary probability space conditional on each label pattern. The preceding measurable maps can therefore be applied on each pattern and permutation, giving a measurable algorithm on the joint sample-and-permutation space. The simultaneous events used in the proofs include every prefix, so selecting a budget measurably from this same output family does not require another probability bound.

Approximate metric and classifier solutions. The proof uses objective and constraint bounds, so it also applies to measurable approximate outputs satisfying explicit numerical error bounds. These upper bounds are deterministic, or are bounded by deterministic quantities specified in advance; this ensures that the LogDet class in the pair-sampling event remains fixed.

For a hard-problem output $\widetilde{A}_{\boldsymbol{m}} \succ 0$, suppose

$$D_{\mathrm{ld}}(\widetilde{A}_{\boldsymbol{m}}, A_0) \le D_{\mathrm{ld}}(A_{\boldsymbol{m}}, A_0) + \eta_{d,\boldsymbol{m}}, \qquad 0 \le \eta_{d,\boldsymbol{m}} \le \bar{\eta}_d.$$

On the common-comparator event its LogDet value is at most $C + \bar{\eta}_d$. Recompute β and the errors e_σ using that enlarged deterministic budget. The pair event then gives the full-pool bound

$$L_+(\widetilde{A}_{\boldsymbol{m}}) + L_-(\widetilde{A}_{\boldsymbol{m}}) \le \widehat{L}_{+,m_+}(\widetilde{A}_{\boldsymbol{m}}) + \widehat{L}_{-,m_-}(\widetilde{A}_{\boldsymbol{m}}) + e_+(m_+) + e_-(m_-).$$

The selected means are the measured constraint residuals; they vanish when the approximate output is exactly feasible. This formulation permits residual violations provided they are included in the downstream comparison.

For sum-slack ITML, suppose the returned matrix and slacks satisfy its positive-slack constraints and have objective at most the exact optimum plus $\eta_{d,\boldsymbol{m}} \le \bar{\eta}_d$. Comparison with the common hard-feasible matrix gives

$$D_{\mathrm{ld}}(\widetilde{A}_{\boldsymbol{m}}, A_0) \le C + \bar{\eta}_d, \qquad \sum_e \varphi(\widetilde{\xi}_e/\xi_{0,e}) \le (C + \bar{\eta}_d)/\gamma.$$

The entire slack-violation proof therefore holds after replacing C by $C + \bar{\eta}_d$, including the induced values of β, Q, G. The factor $1/\gamma$ continues to apply to the unnormalized sum.

For the averaged-hinge objective, an output within $\eta_{d,\boldsymbol{m}} \le \bar{\eta}_d$ of its minimum has the automatic budget $(Q_0 + l + \bar{\eta}_d)/\lambda_d$. On the event for this class, the earlier comparison chain changes only in its optimization step:

$$F_S(\widetilde{A}_{\boldsymbol{m}}) \le \widehat{F}_{\boldsymbol{m}}(\widetilde{A}_{\boldsymbol{m}}) + e_{\boldsymbol{m}} \le \widehat{F}_{\boldsymbol{m}}(A_*) + \eta_{d,\boldsymbol{m}} + e_{\boldsymbol{m}} \le O_S + \eta_{d,\boldsymbol{m}} + 2e_{\boldsymbol{m}}.$$

Under the common zero-violation comparator, the sharper inequalities are

$$D_{\mathrm{ld}}(\widetilde{A}_{\boldsymbol{m}}, A_0) \le C + \eta_{d,\boldsymbol{m}}/\lambda_d, \qquad \widehat{L}_{+,m_+}(\widetilde{A}_{\boldsymbol{m}}) + \widehat{L}_{-,m_-}(\widetilde{A}_{\boldsymbol{m}}) \le \lambda_d C + \eta_{d,\boldsymbol{m}}.$$

If the downstream affine SVM is within objective error $\eta_{c,\boldsymbol{m}} \ge 0$ of its minimum, the comparison objective bound becomes $H_{\boldsymbol{m}} + \eta_{c,\boldsymbol{m}}$. Hence

$$\widehat{R}_h \le H_{\boldsymbol{m}} + \eta_{c,\boldsymbol{m}}, \qquad \|w\|_2 \le \sqrt{2(H_{\boldsymbol{m}} + \eta_{c,\boldsymbol{m}})/\lambda_c}.$$

Substitute this enlarged bound into the empirical-loss and metric-robustness terms of equation 18. The optimization error is added after the minimum with one, since an approximate solution need not have objective at most the zero-score objective. For a Gaussian SVM, the identical reasoning uses $H^k_{\boldsymbol{m}} + \eta_{c,\boldsymbol{m}}$ and its RKHS norm.

Accuracy sufficient for the asymptotic rate. Let $\kappa = 1/(s+2)$ and use the quadratic budgets and $r_n = \lambda_{c,n} = n^{-\kappa}$ of the corollary. Sufficient metric-step conditions are the following:

- For hard ITML, the LogDet objective errors have a uniform deterministic bound, and the sum of the two selected-pool mean violations is $O(n^{-\kappa})$.
- For sum-slack ITML, the returned matrix and positive slacks satisfy its constraints, and the total objective error has a uniform deterministic bound.
- For the averaged-hinge objective, $\lambda_{d,n} = n^{-\kappa}$ and the objective error is $O(\lambda_{d,n})$.

In each case, require the downstream SVM objective error to be $O(n^{-\kappa})$, with deterministic constants. The hard and sum-slack geometric budget is bounded by $C + \sup_n \eta_{d,n}$. The averaged-hinge budget is bounded by $C + \sup_n \eta_{d,n}/\lambda_{d,n}$. Thus all sampling errors remain $O(\sqrt{\log n}/n)$ on the corollary's events. The hard residuals or averaged-hinge losses are $O(n^{-\kappa})$, while the bounded sum-slack budget gives $O(n^{-1})$ selected means. Therefore in all cases the full-pool violation sum is $O(n^{-\kappa})$. The classifier's total objective bound, including its error, is $O(\lambda_{c,n})$. Its norm remains bounded, its metric robustness is $O(r_n)$, and the counting term is $O(n^{-\kappa})$. This proves the same rate and success probability as Corollary 4.2 for these approximate outputs.

C.13 Experimental setting and correspondence with Zhou's example

The experiments were conducted using Ubuntu 20.04.5 LTS and a Tesla A100 80G. The Breast Cancer Wisconsin (Original) data were obtained from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg (Wolberg, 1990). The main plots use 683 observations after missing values are removed, with 450 observations for training and 233 for testing, and report 20 random splits. Pairs are randomly sampled by proportion, with every larger selected set containing the smaller selected sets. Figure 2 reports the SVM and Lasso results with standard-deviation bars.

The metric learner is the LogDet-based ITML method of Davis et al. (2007), whose hard and sum-slack formulations are analyzed above. The downstream learners are a regularized hinge SVM and a linear Lasso score trained with squared loss and an ℓ_1 penalty. The Lasso head is covered by the bounded-loss calculation in Section 3.3. The SVM consistency results concern the linear and fixed-bandwidth Gaussian kernels specified in the theory; the kernel used for the plotted SVM experiment was unspecified in the original experimental record.

The hard LogDet program with $A_0 = I$, a subsequent linear classifier, and the constraint-associated parameter count agree with the mathematical example of Zhou (2021). The standard sum-slack program is given in equation 19. The theoretical results apply under the stated sampling, geometric comparator, and classifier optimization conditions. The empirical curves illustrate the motivating phenomenon; verifying population threshold realizability for the experimental distribution remains a separate question.

D Special Cases and Comparisons

This appendix gives the precise reductions used in Section 5. We first recover finite-sample pairwise robustness bounds from the product-partition argument. We then recover the robustness and generalization bounds for bounded Lipschitz landmark similarities, and derive objective-to-classification-risk guarantees for bilinear similarity learning. In each case, the learned outputs are held fixed when evaluating population expectations.

D.1 Pairwise robustness guarantees

We first show how the product-partition argument used in Corollary 3.6 recovers the finite-sample robustness bounds of Bellet & Habrard (2015). This comparison concerns the pairwise metric-learning loss. Write $\tau_n = \sqrt{(K \log 2 + \log(1/\delta))/(2n)}$, as in Theorem 3.5.

Proposition D.1 (Pairwise local variation)**.** *Let $S = (Z_1, \dots, Z_n) \sim \mu^n$, and fix a measurable partition $\Pi = \{C_1, \dots, C_K\}$ before observing S. For any measurable learned loss $h_S : \mathcal{Z}^2 \to [0, B]$, suppose that numbers $c_{ij} \geq 0$ satisfy*

$$|h_S(z, z') - h_S(Z_i, Z_j)| \leq c_{ij} \qquad \text{for all } z \in C(Z_i),\ z' \in C(Z_j).$$

On an event of probability at least $1 - \delta$, simultaneously for every such loss and local bounds,

$$\left| \mathbb{E}_{(Z,Z') \sim \mu \otimes \mu} h_S(Z, Z') - \frac{1}{n^2} \sum_{i,j=1}^{n} h_S(Z_i, Z_j) \right| \leq \frac{1}{n^2} \sum_{i,j=1}^{n} \min\{B, c_{ij}\} + 2B\tau_n. \tag{50}$$

The expectation holds the learned loss fixed and uses independent test observations.

Proof. The assertion is immediate when $B = 0$, so assume $B > 0$. Let $p_k = \mu(C_k)$, $n_k = \sum_i \mathbf{1}\{Z_i \in C_k\}$, and $\widehat{p}_k = n_k/n$. The count event established in Appendix B.3 satisfies

$$T := \frac{1}{2} \sum_{k=1}^{K} |p_k - \widehat{p}_k| \leq \tau_n,$$

and every occupied cell has positive population mass. This event has probability at least $1 - \delta$.

Fix a sample on that event. For $p_k p_l > 0$, set

$$\bar{h}_{kl} = \frac{1}{p_k p_l} \int_{C_k \times C_l} h_S(z, z')\, d\mu(z)\, d\mu(z'),$$

and set $\bar{h}_{kl} = 0$ when $p_k p_l = 0$. In an occupied product cell, define

$$\widehat{h}_{kl} = \frac{1}{n_k n_l} \sum_{i : Z_i \in C_k} \sum_{j : Z_j \in C_l} h_S(Z_i, Z_j),$$

and set it to zero if $n_k n_l = 0$. The risks decompose as

$$\begin{aligned} \mathbb{E}h_S - \frac{1}{n^2} \sum_{i,j} h_S(Z_i, Z_j) = & \sum_{k,l} (p_k p_l - \widehat{p}_k \widehat{p}_l) \bar{h}_{kl} \\ & + \sum_{k,l} \widehat{p}_k \widehat{p}_l (\bar{h}_{kl} - \widehat{h}_{kl}). \end{aligned}$$

Since both product-mass vectors sum to one and $0 \leq \bar{h}_{kl} \leq B$, the first sum has absolute value at most BT_2, where

$$\begin{aligned} T_2 &:= \frac{1}{2} \sum_{k,l} |p_k p_l - \widehat{p}_k \widehat{p}_l| \\ &\leq \frac{1}{2} \sum_{k,l} \big(|p_k - \widehat{p}_k| p_l + \widehat{p}_k |p_l - \widehat{p}_l| \big) = 2T. \end{aligned}$$

For each pair of training indices in an occupied product cell, integrating the assumed local inequality gives

$$|\bar{h}_{C(Z_i), C(Z_j)} - h_S(Z_i, Z_j)| \leq \min\{B, c_{ij}\}.$$

Here the notation on the left denotes the conditional mean of the product cell containing the two observations. Thus the second sum has absolute value at most $n^{-2} \sum_{i,j} \min\{B, c_{ij}\}$. Combining the two estimates and using $T \leq \tau_n$ proves the claim.

The argument is deterministic once the count event is fixed, so that event is common to every measurable learned loss satisfying the local condition. Only the original observations are used in the concentration step. In particular, the argument does not treat the n^2 pairs as independent. □

Robust and pseudo-robust algorithms. For a $(K, \epsilon(S))$-robust metric learner in the sense of Bellet & Habrard (2015, Definition 2), take $h_S(z, z') = l_d(f_S, z, z')$ and $c_{ij} = \epsilon(S)$. Proposition D.1 gives

$$|\mathcal{L}_d(f_S) - \widehat{\mathcal{L}}_d(f_S)| \le \min\{B, \epsilon(S)\} + 2B\tau_n \le \epsilon(S) + 4B\tau_n.$$

The last expression is their Theorem 1 bound. Thus the same hypotheses imply that guarantee through our product-partition analysis. The smaller sampling coefficient comes from bounding integrals of a nonnegative bounded loss by total variation instead of the full ℓ_1 difference of the cell-mass vectors.

More generally, suppose the local condition with bound $\epsilon(S)$ holds on an index set $G_S \subseteq \{1, \dots, n\}^2$ of cardinality $q(S)$. Boundedness gives the local bound B on all other pairs. Consequently,

$$|\mathcal{L}_d(f_S) - \widehat{\mathcal{L}}_d(f_S)| \le \frac{q(S)}{n^2} \min\{B, \epsilon(S)\} + B\left(1 - \frac{q(S)}{n^2}\right) + 2B\tau_n.$$

Increasing the sampling term to $4B\tau_n$ and removing the truncation recovers their Theorem 2. Both comparisons use a fixed measurable partition, as required by the quantifiers in the original robustness definitions.

A norm-regularized Mahalanobis example. The recovery can also be checked directly in a concrete metric-learning problem. Suppose $\|x\|_2 \le R$ on $\mathcal{X}$ and consider

$$\min_{A \succeq 0} \lambda_d \|A\|_F + \frac{1}{n^2} \sum_{i,j=1}^{n} \varphi\big(y_{ij}[1 - (X_i - X_j)^\top A(X_i - X_j)]\big),$$

where $y_{ij} = 1$ when $Y_i = Y_j$ and -1 otherwise, $\lambda_d > 0$, and $\varphi \ge 0$ is U-Lipschitz. Let $g_0 = \max\{\varphi(1), \varphi(-1)\}$. Comparison with $A = 0$ gives $\|A\|_{\mathrm{op}} \le \|A\|_F \le g_0/\lambda_d$ for every minimizer. For the representation $f(x) = A^{1/2}x$, therefore,

$$a \le r\sqrt{g_0/\lambda_d}, \qquad D \le 2R\sqrt{g_0/\lambda_d}.$$

Here Π is the fixed label-pure partition of input diameter at most r obtained from a radius-$r/2$ cover, with $K \le |\mathcal{Y}|\mathcal{N}(r/2, \mathcal{X}, \|\cdot\|_2)$. The loss is U-Lipschitz in squared distance, so Corollary 3.6 yields

$$|\mathcal{L}_d(f) - \widehat{\mathcal{L}}_d(f)| \le \min\left\{B_d, \frac{8URrg_0}{\lambda_d}\right\} + 2B_d\tau_n.$$

A deterministic valid loss bound is $B_d = g_0 + 4UR^2 g_0/\lambda_d$: squared distances are at most $4R^2 g_0/\lambda_d$, and comparison with $\varphi(y_{ij})$ using Lipschitz continuity proves the bound. If a smaller common loss bound is known, it may be used instead. The factor $8URrg_0/\lambda_d$ is precisely the local robustness factor in Example 1 of Bellet & Habrard (2015), with r and λ_d corresponding to their γ and c. The learned-radius calculation in Example 3.7, combined with the squared-distance version of Corollary 3.6, therefore recovers this metric-loss guarantee as well as furnishing the geometry used for our downstream classification analysis.

D.2 Recovery of landmark-classifier bounds

We compare our result with Theorems 1 and 2 of Nicolae et al. (2015). Their analysis concerns a jointly learned similarity and a linear classifier on landmark features. The following proposition recovers their local robustness constant and implies their classification generalization bound.

Proposition D.2 (Landmark specialization). *Let $S = ((X_i, Y_i))_{i=1}^n$ consist of independent observations from μ on $\mathcal{X} \times \{-1, +1\}$, and let Π be a fixed measurable partition into K label-pure cells whose input diameters are at most r. Suppose a measurable learning procedure returns a similarity s, landmarks $\zeta_1, \dots, \zeta_N$, and coefficients $\alpha \in \mathbb{R}^N$ satisfying, for all $x, x' \in \mathcal{X}$ and j,*

$$|s(x, \zeta_j)| \le 1, \qquad |s(x, \zeta_j) - s(x', \zeta_j)| \le L_s\|x - x'\|_{\mathcal{X}}, \qquad \|\alpha\|_1 \le 1/\gamma,$$

where $L_s < \infty$ and $\gamma > 0$. Define $f(x) = (s(x, \zeta_j))_{j=1}^N$ in $(\mathbb{R}^N, \|\cdot\|_\infty)$, $g(v) = \alpha^\top v$, and the hinge loss $\ell(x, y) = [1 - yg(f(x))]_+$. Set $B = 1 + 1/\gamma$. Then the metric robustness evaluated using ℓ/B satisfies

$$\varepsilon_{\mathrm{mr}}(f, g, S; \Pi) \le \frac{1}{n} \sum_{i=1}^{n} \min\left\{1, \frac{\|\alpha\|_1 a_i}{B}\right\} \le \min\left\{1, \frac{L_s r}{\gamma B}\right\}.$$

In particular, on the event of Theorem 3.5,

$$\left|\mathbb{E}_\mu \ell - \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, Y_i)\right| \le \frac{L_s r}{\gamma} + \left(1 + \frac{1}{\gamma}\right)\sqrt{\frac{K \log 2 + \log(1/\delta)}{2n}}. \tag{51}$$

Moreover, any two same-cell observations, one of which is a training observation, have hinge losses differing by at most $L_s r/\gamma$.

Proof. The similarity bound and coefficient constraint give $|g(f(x))| \le \|\alpha\|_1 \le 1/\gamma$, and hence $0 \le \ell(x, y) \le B$. Thus ℓ/B is an admissible evaluation loss for Theorem 3.5. For any x, x',

$$d_f(x, x') = \max_{1 \le j \le N} |s(x, \zeta_j) - s(x', \zeta_j)| \le L_s \|x - x'\|_{\mathcal{X}}.$$

Since the input diameter of $C(Z_i)$ is at most r, this implies $a_i \le L_s r$. The dual-norm inequality gives

$$|g(f(x)) - g(f(X_i))| \le \sum_{j=1}^{N} |\alpha_j||s(x, \zeta_j) - s(X_i, \zeta_j)| \le \|\alpha\|_1 d_f(x, X_i).$$

For fixed y, the hinge loss is 1-Lipschitz as a function of the score. Consequently, whenever $d_f(x, X_i) \le t$,

$$\left|\frac{\ell(x, Y_i)}{B} - \frac{\ell(X_i, Y_i)}{B}\right| \le \min\left\{1, \frac{\|\alpha\|_1 t}{B}\right\}.$$

Taking the supremum over this ball and averaging at $t = a_i$ proves the metric-robustness bounds. Applying Theorem 3.5 to ℓ/B and multiplying by B proves equation 51. Finally, if $(x, y) \in C(Z_i)$, label purity gives $y = Y_i$ and the definition of a_i gives $d_f(x, X_i) \le a_i \le L_s r$. The preceding score and loss inequalities therefore give $|\ell(x, y) - \ell(X_i, Y_i)| \le L_s r/\gamma$. □

The last statement is exactly the $(K, L_s r/\gamma)$ robustness conclusion of Theorem 1 of Nicolae et al. (2015), under the identifications $K = M$, $L_s = l$, $r = \rho$, and $n = d_l$. Their Theorem 2 has the same local term and the sampling term $(1 + 1/\gamma)\sqrt{(2K \log 2 + 2\log(1/\delta))/n}$. Equation equation 51 implies that bound, with a factor of two improvement in the sampling term. This improvement comes from the total-variation estimate in equation 23.

The landmarks and both learned components can depend on the labeled sample and on additional unlabeled observations. The event in Theorem 3.5 depends only on the labeled sample's counts in Π, and the deterministic inequality holds simultaneously for all measurable returned outputs. Thus no independent training split or additional term involving the number of landmarks is needed. The norm and similarity conditions above are precisely the conditions used for the local and bounded-hinge steps; optimization enters through ensuring that the returned coefficients and similarity meet those conditions.

For a concrete instance from Nicolae et al. (2015), let $\mathcal{X}$ be the Euclidean unit ball, with landmarks $\zeta_j \in \mathcal{X}$, and take $s_A(x, \zeta) = x^\top A \zeta$, with A diagonal and $|A_{kk}| \le 1$. Then $\|A\|_{\mathrm{op}} \le 1$, so

$$|s_A(x, \zeta)| \le \|x\|_2 \|A\|_{\mathrm{op}} \|\zeta\|_2 \le 1, \qquad |s_A(x, \zeta) - s_A(x', \zeta)| \le \|x - x'\|_2.$$

Thus $L_s = 1$ and Proposition D.2 applies to every feasible output of their joint learning problem, recovering the stated local and generalization guarantees for their bilinear similarity example.

D.3 Similarity objectives and classification risk

The landmark representation also gives an explicit comparison between similarity-learning objectives and the empirical classification loss. We state this comparison separately because it supplies the empirical term in a classification-risk bound, whereas the preceding proposition controls the generalization gap.

Proposition D.3 (Similarity objective)**.** *Fix a label-pure partition Π before observing the binary labeled sample S. Let a measurable learning procedure return a similarity s, labeled landmarks $(\zeta_j, b_j)_{j=1}^N$, with $b_j \in \{-1, +1\}$, and weights $p_j \ge 0$ satisfying $\sum_j p_j = 1$. For $\gamma > 0$, define*

$$f(x) = (s(x, \zeta_j))_{j=1}^N \in (\mathbb{R}^N, \|\cdot\|_\infty), \qquad g_\alpha(v) = \alpha^\top v,$$

and the empirical similarity objective

$$\widehat{Q}(s) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{Y_i}{\gamma}\sum_{j=1}^{N} p_j b_j s(X_i, \zeta_j)\right]_+. \tag{52}$$

Suppose $\|\alpha\|_1 \le 1/\gamma$ and its empirical hinge loss is at most $\widehat{Q}(s) + \eta$, for $\eta \ge 0$. Then, on the event of Theorem 3.5,

$$R_{01}(g_\alpha \circ f) \le \widehat{Q}(s) + \eta + \frac{1}{n}\sum_{i=1}^{n} \min\{1, a_i/\gamma\} + \tau_n. \tag{53}$$

The empirical comparison holds whenever α minimizes the empirical hinge loss over the ℓ_1 ball of radius $1/\gamma$ to within additive error η. If the cells have input diameter at most r, and every landmark section $s(\cdot, \zeta_j)$ is L_s-Lipschitz, the average response in equation 53 is at most $\min\{1, L_s r/\gamma\}$.

Proof. Set $\alpha_j^0 = p_j b_j/\gamma$. Since each $|b_j| = 1$,

$$\|\alpha^0\|_1 = \frac{1}{\gamma}\sum_j p_j = \frac{1}{\gamma}.$$

This score is feasible for the constrained classification problem, and substituting α^0 into its empirical hinge loss gives exactly $\widehat{Q}(s)$. An η-optimal classifier therefore has empirical hinge loss at most $\widehat{Q}(s)+\eta$. The feasible coefficient ball is compact and the finite-sample hinge objective is continuous, so an exact minimum exists. As elsewhere, the bound applies to any measurable returned choice.

For an arbitrary coefficient vector in that ball,

$$|g_\alpha(f(x)) - g_\alpha(f(x'))| \le \|\alpha\|_1 \|f(x) - f(x')\|_\infty \le d_f(x, x')/\gamma.$$

The clipped hinge is bounded by one and is 1-Lipschitz in its score. Its local response thus satisfies $\omega_i(a_i) \le \min\{1, a_i/\gamma\}$, with the right-hand side interpreted as one if $a_i = \infty$. Its empirical value is at most the empirical untruncated hinge, and it dominates the misclassification indicator, including score ties. Applying Theorem 3.5 proves equation 53. Finally, the Lipschitz condition gives $d_f(x, x') \le L_s\|x - x'\|_{\mathcal{X}}$, and hence $a_i \le L_s r$ for every cell. This proves the last assertion. All comparisons are pointwise after the learned outputs are fixed; the theorem's common event therefore also covers landmarks, weights, and similarities learned from S. □

Bilinear similarities. Let $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \le R\}$, take $s_A(x, \zeta) = x^\top A\zeta$, and suppose the landmarks belong to $\mathcal{X}$. For every real matrix A,

$$|s_A(x, \zeta_j) - s_A(x', \zeta_j)| = |(x - x')^\top A\zeta_j| \le R\|A\|_{\mathrm{op}}\|x - x'\|_2.$$

Consequently, Proposition D.3 yields

$$R_{01}(g_\alpha \circ f_A) \le \widehat{Q}(A) + \eta + \min\{1, rR\|A\|_{\mathrm{op}}/\gamma\} + \tau_n, \tag{54}$$

where $\widehat{Q}(A) = \widehat{Q}(s_A)$ and $f_A(x) = (x^\top A\zeta_j)_j$. The distance in this application is the pseudometric induced by f_A; the matrix in the similarity can be indefinite. The displayed estimate follows from the operator norm inequality for any such matrix.

For the formulation of Guo & Ying (2014), set $N = n$, $\zeta_j = X_j$, $b_j = Y_j$, and $p_j = 1/n$, with A restricted to symmetric matrices and their margin parameter r identified with our γ. Then equation 52 is their empirical similarity objective, and the signed uniform-average score $\alpha_j^0 = Y_j/(n\gamma)$ is the feasible comparator for their constrained classification problem. This identifies exactly the empirical comparison used in their classification analysis. If the similarity learner minimizes $\widehat{Q}(A) + \lambda_d\|A\|$, where $\|\cdot\|$ is a matrix norm and $\lambda_d > 0$, comparison with $A = 0$ gives

$$\widehat{Q}(A) + \lambda_d\|A\| \le 1, \qquad \|A\| \le 1/\lambda_d.$$

For any constant c_{op} such that $\|A\|_{\mathrm{op}} \le c_{\mathrm{op}}\|A\|$, the local term in equation 54 is therefore at most $\min\{1, rRc_{\mathrm{op}}/(\gamma\lambda_d)\}$. Such a finite constant exists for every norm on the finite-dimensional matrix space. For the Frobenius, entrywise ℓ_1, and mixed $(2, 1)$ norms one can take $c_{\mathrm{op}} = 1$: the operator

norm is at most the Frobenius norm, and the Frobenius norm is at most either of the latter two norms. Thus the same matrix-norm choices admit explicit metric-robustness bounds.

For the reasonable-point formulation of Bellet et al. (2012), take the labeled landmarks to be the reasonable points and use uniform weights on this set. Equation equation 52 is then the empirical goodness loss of the associated signed-average score. The score itself satisfies Proposition D.3 with $\eta = 0$. For a head using these landmarks, an ℓ_1 penalty also permits a direct comparison with this score. If its returned coefficients satisfy, for $\lambda_c > 0$,

$$\widehat{\mathcal{L}}_{\mathrm{h}}(\alpha) + \lambda_c \|\alpha\|_1 \le \widehat{Q}(s) + \lambda_c/\gamma + \eta =: E_c,$$

where $\widehat{\mathcal{L}}_{\mathrm{h}}$ is its average hinge loss, then $\widehat{\mathcal{L}}_{\mathrm{h}}(\alpha) \le E_c$ and $\|\alpha\|_1 \le E_c/\lambda_c$. The same dual-norm calculation, now with the actual coefficient norm, gives

$$R_{01}(g_\alpha \circ f) \le E_c + \frac{1}{n} \sum_i \min\{1, E_c a_i/\lambda_c\} + \tau_n.$$

For an exact minimizer the objective comparison holds with $\eta = 0$. Their summed-loss formulation corresponds to $\lambda_c = \lambda/n$ in our average-loss convention. This treats the coefficient penalty explicitly when transferring the similarity objective to the final fitted classifier.

The goodness framework of Balcan et al. (2008) supplies existence and sampling guarantees for landmark separators. The stability and Rademacher analyses of Bellet et al. (2012); Guo & Ying (2014) control the learned similarity through their respective complexity terms. The calculations above identify their empirical comparison scores and give classification-risk bounds through metric robustness. The exact finite-sample bound recoveries established in this appendix are the pairwise robustness bounds in Appendix D.1 and the bounded-Lipschitz landmark bounds in Appendix D.2.