ADOQ: PREDEFINED ANALYTICAL REGIONS FOR SELECTIVE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective classification usually treats confidence scores as ranking devices and relies on held-out data to determine operating thresholds for specific coverage or risk targets. ADOQ introduces a different construction in which a single learned score organizes the latent decision space into successive predefined analytical shells around fixed class centers, combining a prescribed χ^2_d numerical scale with an ordering driven by discriminative difficulty. Once training and model selection are complete, the encoder, class geometry, score scales, and the entire continuous family of shell boundaries are fixed analytically, without empirical threshold estimation or per-operating-point recalibration, and certification is performed only afterward on these frozen regions. A single unlabeled certification sample provides distribution-free occupancy bounds uniformly over the predefined shell family, labeled observations provide exact reliability bounds for each fixed shell, and a new prediction is thus assigned to a region carrying an independently certified reliability bound rather than only the classifier's global accuracy or an arbitrary confidence score. Experiments on CIFAR-10, SVHN, Imagenette, and CIFAR-100 show that successive analytical shells strongly stratify predictive reliability, with inner and intermediate regions remaining substantially more reliable than the outer tail even when global performance or score-law fidelity is markedly weaker. Thus, a globally imperfect classifier, which would typically be discarded or retrained outright, may still contain substantial predefined regions with independently certified population-level predictive reliability, while less reliable outer regions are explicitly isolated for deferral.

1 INTRODUCTION

Selective classification augments a predictor with the ability to abstain on inputs deemed insufficiently reliable (Chow, 1970; El-Yaniv & Wiener, 2010). This capability is particularly important in high-stakes settings such as medical decision support, autonomous and industrial control, or cybersecurity, where acting on an unreliable prediction may be substantially more costly than deferring a decision. Yet global classification performance alone does not reveal which predictions can still be trusted: a classifier with moderate overall accuracy may contain many highly reliable predictions together with much less reliable ones. Most selective-classification methods address this problem by ranking predictions and choosing confidence thresholds from validation or calibration data. This entangles two distinct questions: which predictions should be trusted? and what evidence supports their actual reliability? The distinction matters whenever operating rules must remain fixed while new evidence is accumulated.

Existing approaches solve important parts of this problem but not their combination. Confidence scores can provide excellent selective rankings, learned selection mechanisms can optimize particular operating points, and calibration-based procedures can attach statistical guarantees to decisions derived from held-out data. Distributional and distance-based representation methods further show that latent geometry can support informative uncertainty scores (Lee et al., 2018; Winkens et al., 2020; Mukhoti et al., 2023; Dinari & Freifeld, 2022; Venkataramanan et al., 2023). However, good ranking alone does not organize the latent space into predefined regions with a stable numerical meaning, while matching a prescribed score distribution alone does not ensure that moving through that space corresponds to increasing predictive difficulty. A score may reproduce its target law while assigning small values to difficult or even misclassified observations; conversely, an excellent con-

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.

fidence ranking may live on an arbitrary numerical scale. Defining meaningful and certifiable latent regions therefore requires the same score to provide both a prescribed coordinate system and an ordering within that system that reflects classification difficulty.

ADOQ (Analytical Difficulty-Ordered Quantiles) is a deterministic representation-learning framework that couples these two properties. ADOQ constructs a fixed class geometry and learns normalized class scores on a prescribed χ^2_d scale while mapping discriminative difficulty onto positions along that scale. Once training is complete, successive analytical intervals define fixed shells around the class centers without empirical threshold estimation. Construction and certification are strictly separated: unlabeled observations provide distribution-free occupancy guarantees, while labeled observations provide exact lower confidence bounds on shell reliability. A new prediction can therefore be associated with a predefined shell carrying an independently certified population-level reliability. This construction raises two complementary research questions. **RQ1** asks whether ADOQ learns the intended analytical coordinate, combining a prescribed numerical scale with an ordering by discriminative difficulty, while retaining useful conventional selective performance. **RQ2** asks whether this learned structure yields predefined analytical shells whose realized occupancy and predictive reliability can be independently certified after training.

Contributions. The main contributions are:

1. ADOQ combines a prescribed χ^2_d numerical scale with a class-conditional ordering driven by discriminative difficulty.
2. A single trained model defines successive analytical shells around fixed class centers without empirical threshold estimation or per-operating-point recalibration.
3. Region construction is separated from certification: unlabeled data provide uniform occupancy bounds, while labeled data provide exact lower confidence bounds on shell reliability.
4. Experiments on CIFAR-10, SVHN, Imagenette, and CIFAR-100 show that these shells strongly stratify predictive reliability, including when global performance or prescribed-law fidelity is limited.

2 Background and Related Work

2.1 Related Work

Selective classification has traditionally focused on the ordering induced by a confidence score. Predictions are ranked and the least confident ones are rejected (Chow, 1970; El-Yaniv & Wiener, 2010), with maximum softmax probability (Hendrycks & Gimpel, 2017) remaining a strong and widely used baseline. Alternative uncertainty scores, including energy-based criteria (Liu et al., 2020b), induce different rankings, while methods such as SelectiveNet (Geifman & El-Yaniv, 2019) and Deep Gamblers (Liu et al., 2019) incorporate rejection directly into the learning objective. More recent work has shown that selective performance also depends strongly on the underlying classifier and representation. Feng et al. (2022) show that appropriate classifier training combined with softmax response can outperform dedicated selection mechanisms, notably through SAT+EM+SR, while Wu et al. (2024) introduce CCL-SC to improve selective classification explicitly at the feature level. Large-scale evaluations further emphasize the importance of both the classifier and confidence function (Galil et al., 2023; Pugnana et al., 2024). These approaches have substantially advanced selective performance, but primarily characterize how well predictions can be ordered and rejected. A strong risk–coverage curve does not by itself organize the latent decision space into predefined regions whose predictive reliability can subsequently be certified.

A complementary line of work provides statistical guarantees for selective operation. Selection with Guaranteed Risk (SGR) (Geifman & El-Yaniv, 2017) uses labeled held-out data, a requested risk level, and a confidence parameter to determine a selection threshold with probabilistic risk control. Conformal prediction similarly uses exchangeable calibration observations to construct data-dependent thresholds or prediction sets with finite-sample guarantees. RAPS (Angelopoulos et al., 2021) adapts this principle to modern image classification, while recent work combines conformal risk control with selective operation (Xu et al., 2025). These methods provide rigorous statistical

guarantees, but calibration observations participate in constructing the deployed rule or its thresholds. ADOQ considers a different regime: the complete family of analytical shell boundaries is fixed before certification observations are examined, and subsequent data are used only to certify the occupancy and population predictive reliability of those already-defined regions.

Structured representations provide another route toward informative uncertainty scores. Mahalanobis-based methods fit class-conditional Gaussian models to deep features (Lee et al., 2018); contrastive representation learning can improve distance-based uncertainty estimates (Winkens et al., 2020); and SNGP (Liu et al., 2020a) explicitly promotes distance awareness in deterministic neural networks. DDU (Mukhoti et al., 2023) combines a regularized feature space with post-hoc Gaussian discriminant analysis. More directly related to Gaussian latent structure, VMDLS (Dinari & Freifeld, 2022) learns separated, approximately isotropic class-conditional latent distributions, while MAPLE (Venkataramanan et al., 2023) encourages approximately Gaussian class representations for Mahalanobis-based uncertainty estimation. These works show that structured latent geometry can support effective distance- and density-based uncertainty scores. However, specifying or encouraging a score distribution does not by itself determine how easy and difficult observations are organized within that space. Conversely, a strong difficulty ranking need not possess a prescribed numerical scale. ADOQ couples these two properties explicitly: discriminative difficulty determines position along a prescribed analytical coordinate, which partitions the latent decision space into successive predefined shells whose occupancy and predictive reliability can be independently certified after training.

2.2 Background

Let (X, Y) denote a random input–label pair with $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{1, \dots, K\}$. A selective classifier combines a predictor $h : \mathcal{X} \to \mathcal{Y}$ with a selection function $g : \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ denotes retention. Its population coverage and selective risk are $C_g = \mathbb{P}(g(X) = 1)$ and $R_g = \mathbb{P}(h(X) \neq Y \mid g(X) = 1)$.

Let $s : \mathcal{X} \to \mathbb{R}$ be a scalar score and $S = s(X)$ the corresponding random score. Thresholding s induces a risk-coverage curve. Any strictly increasing transformation preserves this ordering while changing the numerical thresholds, so ranking quality alone does not give score values an absolute statistical meaning.

For a prescribed continuous cumulative distribution function (CDF) F_0 with quantile function F_0^{-1}, define $\tau_\rho = F_0^{-1}(\rho)$ for any interior level $\rho \in (0, 1)$. If $S \sim F_0$, then $\mathbb{P}(S \leq \tau_\rho) = \rho$.

In particular, for $y \in \mathcal{Y}$ with $C_y \in \mathbb{R}^d$, $\sigma_y > 0$, and I_d the $d \times d$ identity matrix: if $U \sim \mathcal{N}(C_y, \sigma_y^2 I_d)$ and $T_y(U) = \|U - C_y\|_2^2 / \sigma_y^2$, then $T_y(U) \sim \chi_d^2$, the chi-squared distribution with d degrees of freedom.

For CDFs F and G on $\mathbb{R}$, define $\mathrm{KS}(F, G) = \sup_{t \in \mathbb{R}} |F(t) - G(t)|$. The Dvoretzky–Kiefer–Wolfowitz inequality, with the sharp constant of Massart (1990), provides distribution-free uniform control of the difference between an empirical CDF and its population counterpart from independent and identically distributed (i.i.d.) observations. Section 3.4 applies this result to certify the realized occupancy of the predefined analytical shell family, while the population predictive reliability of fixed shells is assessed separately from labeled observations using exact binomial bounds.

3 Method

3.1 Setup and objectives

The objective is to construct a selective classifier whose latent decision space is organized into analytically defined operating regions once training is complete, rather than determined from target data for each operating point.

A geometric representation associates the prescribed statistical scale with explicit class-specific regions. A deterministic encoder maps $x \in \mathcal{X}$ to $z = f_\theta(x) \in \mathbb{R}^d$, where $Z = f_\theta(X)$ denotes the corresponding random latent representation. For each class $y \in \mathcal{Y}$, a fixed center $C_y \in \mathbb{R}^d$ and scale

$\sigma_y > 0$ define

$$T_y(z) = \frac{\|z - C_y\|_2^2}{\sigma_y^2}, \qquad \hat{y}(z) = \arg\min_{y \in \mathcal{Y}} T_y(z). \tag{1}$$

For any $\alpha > 0$, $T_y(z) \leq \alpha$ is precisely the Euclidean ball of radius $\sigma_y\sqrt{\alpha}$ centered at C_y, which motivates referring to C_y and σ_y as the class center and scale.

ADOQ uses the χ^2_d law as the prescribed class-conditional reference for the normalized true-class score, $F_0 = F_{\chi^2_d}$. Analytical levels are indexed by $\rho \in [0, 1)$, with $\tau_0 = 0$ and $\tau_\rho = F^{-1}_{\chi^2_d}(\rho)$ for $\rho \in (0, 1)$, a convention used throughout the paper; $\rho = 1$ is excluded because $\tau_1 = +\infty$ would require τ to take values outside $\mathbb{R}$. Instantiating the ball property above with $\alpha = \tau_\rho$, each nominal level therefore defines the class-specific radial boundary $\|z - C_y\|_2 \leq \sigma_y\sqrt{\tau_\rho}$, and successive levels induce nested analytical regions. This convention is used throughout the remainder of the paper.

3.2 Fixed class geometry

The class centers and scales are fixed before encoder training. Relative scales $\{\sigma_y\}_{y=1}^K$ are estimated once from the training data and normalized to unit mean; the complete procedure is given in Appendix A.

The separation of the class geometry is controlled by a geometry-level reference mass $\rho_{\text{sep}} \in (0, 1)$. The corresponding analytical threshold $\tau_{\rho_{\text{sep}}}$ defines the class-y Euclidean radius r_y^{sep}, while $\sigma_{\max}$ denotes the largest class scale:

$$\tau_{\rho_{\text{sep}}} = F^{-1}_{\chi^2_d}(\rho_{\text{sep}}), \qquad r_y^{\text{sep}} = \sigma_y\sqrt{\tau_{\rho_{\text{sep}}}}, \qquad \sigma_{\max} = \max_{1 \leq y \leq K} \sigma_y. \tag{2}$$

The common pairwise center distance is then set to

$$\alpha = 2\sigma_{\max}\sqrt{\tau_{\rho_{\text{sep}}}}. \tag{3}$$

This choice is sufficient to keep the interiors of all class-specific balls at level ρ_{sep} pairwise disjoint; it does not define a deployment operating level and does not restrict the analytical levels ρ subsequently used to form nested regions and successive shells. Fixing α alone does not yet determine the arrangement of the centers, so three symmetry requirements complete the geometry: a common translation leaves every score unchanged (Appendix A.5), so the configuration's position is fixed by a zero centroid, $\sum_{y=1}^K C_y = 0$; no class should lie closer to this common center than another, so every center is equidistant from it; and since this alone still allows some pairs to sit closer together than others, every pair is additionally required to be equally separated. Together, these requirements leave, up to an orthogonal transformation and class permutation, a single arrangement: the regular simplex of Proposition 1, detailed in Appendix A.

Proposition 1 (Regular simplex class geometry)**.** *For $K \leq d+1$, requiring a centered configuration ($\sum_{y=1}^K C_y = 0$) with equal center norms and pairwise distance α defines, up to an orthogonal transformation and class permutation, a regular $(K-1)$-simplex satisfying*

$$\|C_y\|_2 = \alpha\sqrt{\frac{K-1}{2K}}, \qquad C_y^\top C_{y'} = -\frac{\alpha^2}{2K}, \quad y \neq y'. \tag{4}$$

The separation geometry also constrains how closely the reference score can match its prescribed distribution. Let $\widehat{Y} = \hat{y}(Z)$ and define $\text{KS}_{\text{ref}} = \text{KS}(F_{T_Y(Z)}, F_{\chi^2_d})$, where $F_{T_Y(Z)}$ denotes the CDF of the true-class score.

Proposition 2 (Separation-induced reference KS floor)**.** *Let $a = \mathbb{P}(\widehat{Y} = Y)$ denote classification accuracy. If the interiors of the regions $\mathcal{B}_y^{\text{sep}} = \{z \mid T_y(z) \leq \tau_{\rho_{\text{sep}}}\}$ are pairwise disjoint, then*

$$\text{KS}_{\text{ref}} \geq [\rho_{\text{sep}} - a]_+, \qquad [t]_+ = \max(t, 0). \tag{5}$$

Thus, whenever $a < \rho_{\text{sep}}$, exact agreement of the reference score with the prescribed law is impossible under this geometry. Increasing ρ_{sep} therefore strengthens class separation but also raises the unavoidable reference-distribution mismatch at fixed accuracy. The simplex construction and proofs of both propositions are provided in Appendix A.

3.3 DIFFICULTY-ORDERED QUANTILE LEARNING

For a training observation (x_i, y_i), let $z_i = f_\theta(x_i)$ and define the closest competing-class score and discriminative margin as

$$T_i^{\text{comp}} = \min_{j \neq y_i} T_j(z_i), \qquad m_i = T_i^{\text{comp}} - T_{y_i}(z_i). \tag{6}$$

Ignoring exact ties, $m_i > 0$ corresponds to correct classification, while larger margins indicate greater separation from the closest competitor.

ADOQ maps this relative difficulty onto the prescribed analytical scale. For each ground-truth class y, ADOQ maintains a first-in, first-out (FIFO) memory of recent detached margins from training observations with label y, regardless of their sign. This memory provides a rolling empirical estimate of the class-conditional margin distribution. For a current observation i, its empirical upper-tail percentile $u_i \in (0, 1)$ within the memory of class y_i is used so that larger margins receive smaller percentiles and are therefore assigned closer-to-center analytical positions. The corresponding analytical target is

$$q_i = F^{-1}_{\chi^2_d}(u_i). \tag{7}$$

For a mini-batch of size B, the encoder is trained with the single objective

$$\boxed{\mathcal{L}_{\text{ADOQ}} = \frac{1}{B} \sum_{i=1}^{B} \left(\frac{T_{y_i}(z_i) - q_i}{d} \right)^2} \tag{8}$$

Dividing the residual by d, the mean of the prescribed χ^2_d law, normalizes it by its natural score scale. Since d is fixed for a given model, this does not change the target minimizer but reduces the dependence of the loss magnitude on the chosen latent dimension. Although the loss acts directly only on the true-class score, its target depends on the competing-class configuration through m_i.

The corresponding population construction makes this difficulty organization explicit. For each class y, let M_y denote the model margin conditional on $Y = y$, with continuous CDF G_y that is strictly increasing on its support, and let G_y^{-1} denote its quantile function. Mirroring the empirical construction of q_i above at the population level, define $Q_y = \tau_{1-G_y(M_y)}$.

Proposition 3 (Analytical quantiles encode difficulty)**.** *For every class y, $Q_y \sim \chi^2_d$. Moreover, for every $\rho \in (0, 1)$,*

$$Q_y \leq \tau_\rho \iff M_y \geq G_y^{-1}(1 - \rho). \tag{9}$$

Unlike τ_ρ itself, whose domain excludes only $\rho = 1$, this equivalence additionally excludes $\rho = 0$: its right-hand side $G_y^{-1}(1 - \rho)$ degenerates at the upper edge of the support of M_y when $\rho = 0$. Thus the prescribed law fixes the numerical analytical coordinate, while discriminative difficulty determines how observations are ordered along it. Successive analytical intervals therefore correspond, in the population construction, to successive levels of discriminative difficulty. Moreover, under the fixed geometry of Section 3.2, exact target fit for any $u_i < \rho_{\text{sep}}$ implies $m_i > 0$. Proofs and complete FIFO details are given in Appendix B.

These properties concern the true-class analytical coordinate. At deployment, only the operational score defined in Section 3.4 is available. The next section therefore uses this operational coordinate to define the successive analytical shells whose occupancy and predictive reliability can subsequently be certified.

3.4 PREDEFINED ANALYTICAL SHELLS AND INDEPENDENT CERTIFICATION

Once training and model selection are complete, the encoder, class geometry, and score scales are frozen. Since the true class is unavailable at deployment, ADOQ uses the operational score S_{op},

$$S_{\text{op}}(x) = \min_{1 \leq y \leq K} T_y(f_\theta(x)), \qquad \hat{y}(x) = \arg\min_{1 \leq y \leq K} T_y(f_\theta(x)), \tag{10}$$

with a fixed tie-breaking rule, and let $\widehat{Y} = \hat{y}(X)$.

To express shell boundaries directly on the same normalized coordinate regardless of d, the operational score is transformed onto $R(x) = F_{\chi^2_d}(S_{\mathrm{op}}(x)) \in [0, 1)$.

Any $0 \le \rho_1 < \rho_2 < 1$ fixed before certification defines the analytical shell

$$\mathcal{S}_{\rho_1,\rho_2} = \begin{cases} \{x \mid 0 \le R(x) \le \rho_2\}, & \rho_1 = 0, \\ \{x \mid \rho_1 < R(x) \le \rho_2\}, & 0 < \rho_1 < \rho_2 < 1, \end{cases} \tag{11}$$

or, equivalently, through the corresponding analytical thresholds τ_{ρ_1} and τ_{ρ_2} defined in Section 3.1. Its population occupancy and predictive reliability are

$$C(\rho_1, \rho_2) = \mathbb{P}(X \in \mathcal{S}_{\rho_1,\rho_2}), \qquad A(\rho_1, \rho_2) = \mathbb{P}(\widehat{Y} = Y \mid X \in \mathcal{S}_{\rho_1,\rho_2}), \tag{12}$$

where $A(\rho_1, \rho_2)$ is defined whenever $C(\rho_1, \rho_2) > 0$. The cumulative region is the special case $\mathcal{S}_{0,\rho}$, and its complement $\mathcal{T}_\rho = \{x \mid \rho < R(x)\} = \mathcal{X} \setminus \mathcal{S}_{0,\rho}$ denotes the outer tail beyond level ρ, with occupancy $C(\mathcal{T}_\rho) = 1 - C(0, \rho)$ and reliability $A(\mathcal{T}_\rho)$ defined analogously.

The pair (ρ_1, ρ_2) fixes where the shell lies on the prescribed analytical coordinate, but does not prescribe its realized population occupancy. If $S_{\mathrm{op}}(X) \sim \chi^2_d$, then $R(X)$ follows the uniform distribution on $[0, 1]$, denoted $\mathrm{Uniform}(0, 1)$, and the shell has nominal mass $\rho_2 - \rho_1$.

Let F_{op} and $\widehat{F}_n$ denote respectively the population and empirical CDFs of $S_{\mathrm{op}}(X)$, based on n independent certification observations, and let $\epsilon_n(\delta) = \sqrt{\log(2/\delta)/(2n)}$. The Dvoretzky–Kiefer–Wolfowitz inequality gives, with probability at least $1 - \delta$,

$$\sup_t \left| F_{\mathrm{op}}(t) - \widehat{F}_n(t) \right| \le \epsilon_n(\delta). \tag{13}$$

Writing $\widehat{C}(\rho_1, \rho_2)$ for the corresponding empirical shell occupancy, it follows on the same event that, for all interior shells $0 < \rho_1 < \rho_2 < 1$,

$$\left| C(\rho_1, \rho_2) - \widehat{C}(\rho_1, \rho_2) \right| \le 2\epsilon_n(\delta). \tag{14}$$

For the endpoint-anchored region $\mathcal{S}_{0,\rho}$ and its complement $\mathcal{T}_\rho$, the corresponding bound is $\epsilon_n(\delta)$.

Theorem 1 (Independent certification of a predefined analytical shell)**.** *Fix an analytical shell $I = \mathcal{S}_{\rho_1,\rho_2}$ before observing an independent i.i.d. certification sample $\{(X_i, Y_i)\}_{i=1}^n$, and let N_I and K_I denote, respectively, the number of certification observations falling inside I and the number of correct predictions among them:*

$$N_I = \sum_{i=1}^n \mathbf{1}\{X_i \in I\}, \qquad K_I = \sum_{i=1}^n \mathbf{1}\{X_i \in I, \hat{y}(X_i) = Y_i\}. \tag{15}$$

Its population occupancy $C(\rho_1, \rho_2)$ is controlled by the uniform DKW bound above. If $N_I > 0$, then

$$K_I \mid N_I \sim \mathrm{Binomial}\left(N_I, A(\rho_1, \rho_2)\right), \tag{16}$$

so an exact one-sided Clopper–Pearson lower confidence bound at any prescribed confidence level, computed from (K_I, N_I), provides a valid lower confidence bound for the population reliability $A(\rho_1, \rho_2)$. These guarantees remain valid without requiring $S_{\mathrm{op}}(X) \sim \chi^2_d$.

The prescribed law fixes the shell boundaries, while certification determines their realized population occupancy and predictive reliability. A single uniform DKW event controls all analytical shells without an additional multiplicity correction. For any finite family of shells fixed before certification, simultaneous reliability bounds are obtained by correcting the individual confidence levels, as detailed in Appendix C.

4 Experiments

ADOQ is evaluated along the two research questions introduced in Section 1.

Table 1: Selective performance and analytical score structure. Accuracy and native AURC are reported with reference/operational KS and Spearman diagnostics. Lower AURC and KS are better; more negative Spearman indicates stronger alignment with discriminative difficulty. Structural diagnostics use matched-dimensional distance scores when applicable; "–" denotes a non-applicable diagnostic.

	CIFAR-10			SVHN		
Method	Acc./AURC	$\mathrm{KS}_{\mathrm{ref/op}}$	$\rho_S^{\mathrm{ref/op}}$	Acc./AURC	$\mathrm{KS}_{\mathrm{ref/op}}$	$\rho_S^{\mathrm{ref/op}}$
MSP	**96.20**/.00366	–	–	**97.40**/.00525	–	–
SelectiveNet	96.14/.00947	–	–	97.23/.01020	–	–
Mahalanobis	96.15/**.00357**	.2741/.2726	-.6751/-.6698	97.37/**.00426**	.2654/.2667	-.3956/-.3862
SAT+EM+SR	95.61/.00395	.3326/.3328	-.5294/-.5216	97.04/.00516	.4498/.4516	-.3437/-.3330
CCL-SC	95.91/.00427	.2917/.2920	-.4056/-.3973	97.05/.00626	.3423/.3437	-.3167/-.3055
MAPLE	93.36/.01062	.2559/.2578	-.4812/-.4608	96.50/.00616	.2676/.2698	-.4284/-.4125
VMDLS	93.48/.01177	.2837/.2869	-.2852/-.2552	96.22/.00542	.4389/.4418	-.2100/-.1925
ADOQ	95.15/.01000	**.0977/.0999**	**-.9637/-.9462**	96.88/.00835	**.0469/.0322**	**-.9893/-.9745**

	Imagenette			CIFAR-100		
Method	Acc./AURC	$\mathrm{KS}_{\mathrm{ref/op}}$	$\rho_S^{\mathrm{ref/op}}$	Acc./AURC	$\mathrm{KS}_{\mathrm{ref/op}}$	$\rho_S^{\mathrm{ref/op}}$
MSP	85.20/.02751	–	–	**80.12/.04912**	–	–
SelectiveNet	83.80/.03950	–	–	80.02/.10736	–	–
Mahalanobis	85.07/.02777	.6809/.6804	-.9355/-.8999	79.84/.05120	.6514/.6474	-.9384/**-.8644**
SAT+EM+SR	82.55/.03911	.4952/.4885	-.8884/-.8225	79.79/.05164	.6726/.6664	-.8632/-.7614
CCL-SC	82.55/.03721	.5785/.5752	-.8562/-.7850	79.65/.05228	.6683/.6612	-.8372/-.7241
MAPLE	76.18/.06941	.5401/.5276	-.9191/-.8056	73.83/.08337	.5319/.5071	-.9140/-.7812
VMDLS	83.21/.03749	.5161/.5114	-.7329/-.6694	61.43/.23355	.5706/.5301	-.5838/-.3969
ADOQ	**88.59/.02448**	**.1772/.1537**	**-.9862/-.9469**	77.90/.07293	**.3900/.3574**	**-.9645**/-.8529

4.1 Experimental setup

CIFAR-10, SVHN, Imagenette, and CIFAR-100 are evaluated under a common data-split and evaluation protocol. ADOQ uses WideResNet-28-10 on CIFAR-10, SVHN, and CIFAR-100, and ResNet-18 on Imagenette, with $d = 16$ for the ten-class datasets and $d = 100$ for CIFAR-100. The value $\rho_{\mathrm{sep}} = .999$ is used throughout.

Model selection uses validation data only: among checkpoints within 0.5 percentage point of the best validation accuracy, the checkpoint minimizing $\widehat{\mathrm{KS}}_{\mathrm{op}}$ is retained. The selected model is then frozen before test evaluation and certification; the analytical thresholds remain fixed by the prescribed reference law. Experimental details and controls are provided in Appendices D and G; code and configurations are available at `https://anonymous.4open.science/r/adoq-selective-classification-CD03`.

4.2 RQ1: Selective performance and analytical difficulty structure

RQ1 tests the two structural properties targeted by ADOQ: prescribed-scale fidelity and ordering by discriminative difficulty, alongside conventional selective performance.

Table 1 reports native accuracy and AURC together with matched structural diagnostics. For competing methods with a class-conditional latent representation, KS and Spearman diagnostics are computed from matched-dimensional Mahalanobis-type coordinates, where ρ_S^{ref} and ρ_S^{op} denote the Spearman correlation between the true-class discriminative margin $M = \min_{j \neq Y} T_j - T_Y$ and, respectively, the reference and operational scores. Accuracy and AURC are instead evaluated using each method's native predictor and selective score. Methods without an intrinsic class-conditional distance coordinate are marked as not applicable. Additional diagnostics are reported in Appendices E and G.

ADOQ does not uniformly optimize selective ranking alone: several standard scores achieve lower AURC on CIFAR-10, SVHN, and CIFAR-100, whereas ADOQ has both the highest accuracy and

lowest AURC on Imagenette. The structural diagnostics show a different pattern. ADOQ has the smallest reference and operational KS and strongest corresponding difficulty ordering on CIFAR-10, SVHN, and Imagenette. On CIFAR-100, Mahalanobis has a slightly stronger operational correlation ($-.8644$ versus $-.8529$), while ADOQ retains a smaller operational KS ($.3574$ versus $.6474$) and stronger reference ordering ($-.9645$ versus $-.9384$). Thus selective ranking, prescribed-scale fidelity, and difficulty ordering capture distinct properties.

CIFAR-100 is particularly informative: at 77.90% accuracy, $\widehat{\mathrm{KS}}_{\mathrm{ref}} = .3900$ and $\widehat{\mathrm{KS}}_{\mathrm{op}} = .3574$, while difficulty ordering remains strong ($\rho_S^{\mathrm{ref}} = -.9645$, $\rho_S^{\mathrm{op}} = -.8529$). The empirical analogue of the geometric argument in Proposition 2 gives

$$\widehat{\mathrm{KS}}_{\mathrm{ref}} \geq [\rho_{\mathrm{sep}} - \widehat{a}]_+,$$

where $\widehat{a}$ denotes empirical test accuracy, so $\rho_{\mathrm{sep}} = .999$ and the observed accuracy already impose $\widehat{\mathrm{KS}}_{\mathrm{ref}} \geq .2200$. Part of the observed mismatch is therefore structurally constrained by the separation geometry.

These results motivate RQ2: whether the predefined analytical regions retain differentiated predictive reliability and remain independently certifiable when global accuracy and prescribed-law fidelity deteriorate. Dimensional controls are reported in Appendix G.

4.3 RQ2: Certified analytical shells and local reliability

The learned analytical coordinate is examined through its central operational consequence: whether it defines local regions of the latent decision space whose occupancy and predictive reliability can be independently certified after training. This structure is evaluated through fixed-width analytical windows $\mathcal{W}_\rho = \mathcal{S}_{\rho-\Delta\rho,\rho}$ for $\rho \in \{.10, .11, \ldots, .99\}$, together with the outermost window $\mathcal{T}_{.9}$, with $\Delta\rho = 0.1$ throughout: 91 predefined overlapping windows in total. Under exact realization of the prescribed score law, every such window would contain nominal population mass $\Delta\rho = 0.1$.

For each predefined window, empirical occupancy $\widehat{C}(\rho_1, \rho_2) = N_I/n$ and local predictive reliability $\widehat{A}(\rho_1, \rho_2) = K_I/N_I$ ($N_I > 0$) are reported together with the simultaneous (but separate, not joint) DKW and Clopper–Pearson lower bounds, C_{LB} and A_{LB}, defined in Section 3.4 and derived in Appendix C. Figure 1 therefore shows both how population mass is distributed along the analytical coordinate and which predefined regions carry independently certified population-level predictive reliability. The marker at $\rho = .99$ additionally reports the empirical cumulative occupancy $\widehat{C}(0, .99)$ before the extreme outer tail, not the area under the local-occupancy curve.

Across datasets, the analytical windows reveal a clear stratification of predictive reliability along the latent coordinate. SVHN is closest to the prescribed regime: local occupancy remains near the nominal 10% level over most of the coordinate, while reliability stays high until the outer tail. CIFAR-10 and Imagenette exhibit progressively stronger redistribution of population mass, yet their inner and intermediate regions remain substantially more reliable than their outer regions. The analytical coordinate therefore does more than provide a global ranking: it defines a predefined family of local regions with markedly different certified reliability.

CIFAR-100 provides the strongest stress test. Despite a global accuracy of only 77.90% and the largest discrepancy from the prescribed score law, the innermost window $\mathcal{W}_{.10} = \mathcal{S}_{0,.1}$ reaches 96.04% empirical reliability, with a simultaneous 95% reliability lower bound of 94.08%, whereas the outermost window $\mathcal{T}_{.9} = \{x \mid .9 < R(x)\}$ falls to 56.30%, with a corresponding lower bound of 53.86%. Moreover, the cumulative region $\mathcal{S}_{0,.99}$ has an empirical occupancy of 64.73% and a simultaneous 95% reliability lower bound of 93.23%. Thus, imperfect global performance and score-law fidelity do not make the classifier uniformly unreliable: ADOQ still identifies substantial predefined regions in which predictions remain markedly more reliable than the classifier as a whole.

Operationally, once a predefined non-overlapping analytical partition is chosen, each new input is assigned a predicted class and an analytical position, and therefore to exactly one region carrying an independently certified population-level reliability bound. Predictions in regions whose certificate meets the required reliability level may be retained, while predictions in less reliable regions can be deferred. This certificate characterizes the region rather than the individual prediction. Complete statistics for non-overlapping analytical regions induced by the predefined grid, together with complementary cumulative regions $\mathcal{S}_{0,\rho}$, are reported in Appendix F.

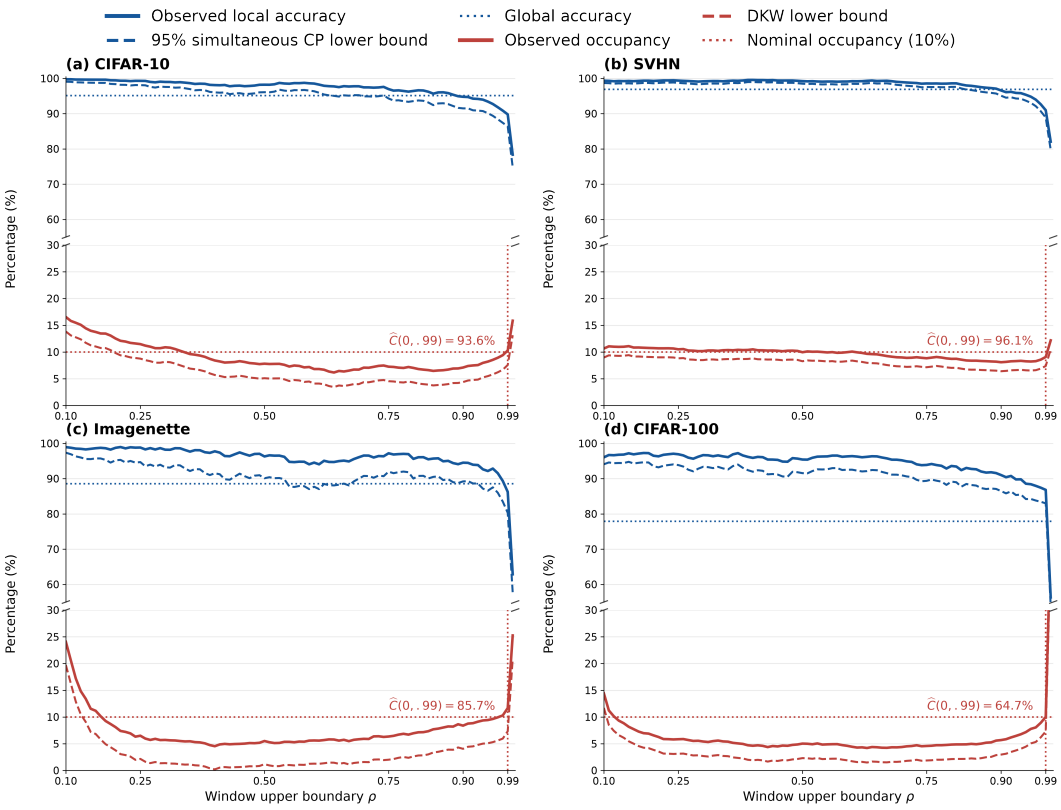


Figure 1: Local analytical reliability and occupancy for ADOQ across the 91 predefined regions: 90 interior sliding windows $\mathcal{W}_\rho = \mathcal{S}_{\rho-0.1,\rho}$ plus the outermost tail $\mathcal{T}_{.9}$. Solid lines are empirical values, dashed lines the simultaneous 95% lower bounds, dotted lines global accuracy and nominal 10% occupancy.

5 Discussion and Conclusion

ADOQ organizes the frozen latent decision space into predefined analytical shells whose realized occupancy and population predictive reliability are certified afterward. Across CIFAR-10, SVHN, Imagenette, and CIFAR-100, these regions strongly stratify reliability, including when global classification performance and prescribed-law fidelity deteriorate; nominal analytical mass need not match realized occupancy, but this does not prevent shell-level certification. Three considerations nonetheless qualify this construction. First, the geometry-level mass ρ_{sep} controls a separation–fidelity trade-off, formalized in Proposition 2 and illustrated empirically in Section 4.2, motivating adaptive or less conservative choices of ρ_{sep} as future work. Second, classification errors introduce wrong-center intrusion, which can reduce operational reliability and prevent strict monotonicity across shells (Appendix C.6). Third, the exact regular-simplex construction requires $d \geq K - 1$, which may become restrictive for very large class spaces.

Despite these limitations, the experiments show that a single learned score can simultaneously support a prescribed analytical coordinate, a strong ordering by discriminative difficulty, and independently certifiable operating regions whose guarantees are region-level rather than individual, depending only on independent certification data from the target population. A model with markedly imperfect global accuracy is typically treated as unreliable overall, leaving little recourse beyond retraining or discarding it; ADOQ instead extracts a certified, reliably usable outcome from it. On CIFAR-100, for instance, a classifier at only 77.90% global accuracy still yields the predefined cumulative region $\mathcal{S}_{0,.99}$ with 64.73% empirical occupancy and a simultaneous 95% reliability lower bound of 93.23% (Section 4.3). ADOQ therefore identifies reliably usable regions of an imperfect classifier while explicitly isolating less reliable regions for deferral, with code available as noted in Section 4.1.

LLM USAGE DISCLOSURE

In this work, generative AI tools were used to aid and polish the writing of the manuscript (drafting and editing text for clarity and readability), to help verify the completeness of literature coverage (checking for potentially missing or relevant related work), and to accelerate the implementation of code based on theoretical constructions and design choices developed by the authors. Generative AI tools were not used to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, or design research methodology or experiments. All AI-assisted work has been reviewed and verified by the authors: literature suggestions were checked against the original sources before inclusion, and all generated code was reviewed, tested, and validated against the theoretical results and experimental protocol described in the paper. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In International Conference on Learning Representations (ICLR), 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

C. K. Chow. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Or Dinari and Oren Freifeld. Variational-and metric-based deep latent space for out-of-distribution detection. In The 38th Conference on Uncertainty in Artificial Intelligence, 2022.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir Abdi. Towards better selective classification. arXiv preprint arXiv:2206.09034, 2022.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 ImageNet classifiers? In International Conference on Learning Representations (ICLR), 2023. URL https://openreview.net/forum?id=p66AzKi6Xim. arXiv:2302.11874.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Proceedings of the 36th International Conference on Machine Learning (ICML), volume 97 of Proceedings of Machine Learning Research, pp. 2151–2159, 2019. URL http://proceedings.mlr.press/v97/geifman19a.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations (ICLR), 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Advances in Neural Information Processing Systems (NeurIPS), volume 31, 2018.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pp. 7498–7512, 2020a.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020b.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, 2019.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. The annals of Probability, pp. 1269–1283, 1990.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24384–24394, 2023.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. Journal of Data-centric Machine Learning Research, 1:1–58, 2024. URL https://arxiv.org/abs/2401.12708. arXiv:2401.12708.

Aishwarya Venkataramanan, Assia Benbihi, Martin Laviale, and Cédric Pradalier. Gaussian latent representations for uncertainty estimation using mahalanobis distance in deep classifiers. In 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 4490–4499. IEEE, 2023.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection, 2020. URL https://arxiv.org/abs/2007.05566.

Yu-Chang Wu, Shen-Huan Lyu, Haopu Shang, Xiangyu Wang, and Chao Qian. Confidence-aware contrastive learning for selective classification. arXiv preprint arXiv:2406.04745, 2024.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control, 2025. URL https://arxiv.org/abs/2512.12844.

A Fixed class geometry

This appendix provides the complete construction of the fixed class geometry summarized in Section 3.2. The construction follows three principles. First, class-specific scales preserve coarse differences in within-class variability. Second, class regions are separated at a fixed geometry-level reference mass ρ_{sep} under the prescribed χ^2_d reference law. Third, once the class-specific scales are fixed, a class-symmetric center geometry is imposed, which leads to a regular simplex whenever $K \le d+1$.

The resulting centers $\{C_y\}_{y=1}^K$ and scales $\{\sigma_y\}_{y=1}^K$ are computed once before encoder training and remain fixed throughout optimization.

A.1 Estimation of relative class scales

For each class y, let

$$X_y = \{x_i \mid y_i = y\} \tag{17}$$

denote its training observations after the same input normalization used by the encoder. After vectorization, let

$$\bar{x}_y = \frac{1}{|X_y|} \sum_{x \in X_y} x, \tag{18}$$

and let $\widetilde{\mathbf{X}}_y$ denote the matrix whose rows are the centered vectors $x - \bar{x}_y$, $x \in X_y$.

Let

$$v_{y,1}, \ldots, v_{y,d} \tag{19}$$

denote the first d right singular vectors of $\widetilde{\mathbf{X}}_y$, equivalently the first d principal directions of the class covariance, and for a scalar random variable Z observed over $x \in X_y$, let

$$\mathrm{Std}[Z \mid x \in X_y] = \sqrt{\frac{1}{|X_y| - 1} \sum_{x \in X_y} \big(Z(x) - \bar{Z}\big)^2} \tag{20}$$

denote its sample standard deviation, with $\bar{Z}$ the corresponding sample mean. The raw scale $\widetilde{\sigma}_y$ is defined as the mean standard deviation across the retained components:

$$\widetilde{\sigma}_y = \frac{1}{d}\sum_{\ell=1}^{d} \mathrm{Std}\left[(x-\bar{x}_y)^\top v_{y,\ell} \mid x \in X_y\right]. \tag{21}$$

The raw scales are subsequently normalized to unit mean,

$$\sigma_y = \frac{\widetilde{\sigma}_y}{\frac{1}{K}\sum_{j=1}^{K}\widetilde{\sigma}_j}, \tag{22}$$

so that

$$\frac{1}{K}\sum_{y=1}^{K}\sigma_y = 1. \tag{23}$$

The PCA step is used only to obtain relative class scales before encoder training. No PCA projection is subsequently used by the classifier. These quantities are used only as relative geometric scale factors; they are not intended as estimates of the learned latent variance or as an assumption on the final class-conditional latent distribution.

A.2 Why the common scale can be normalized

The ADOQ score depends only on distances normalized by the corresponding class scale:

$$T_y(z) = \frac{\|z - C_y\|_2^2}{\sigma_y^2}. \tag{24}$$

Consequently, the absolute common scale of the latent geometry is arbitrary.

For any scalar $\kappa > 0$, consider the transformation

$$z' = \kappa z, \qquad C'_y = \kappa C_y, \qquad \sigma'_y = \kappa \sigma_y. \tag{25}$$

Then

$$T'_y(z') = \frac{\|z' - C'_y\|_2^2}{(\sigma'_y)^2} \tag{26}$$

$$= \frac{\|\kappa(z - C_y)\|_2^2}{\kappa^2\sigma_y^2} \tag{27}$$

$$= \frac{\|z - C_y\|_2^2}{\sigma_y^2} \tag{28}$$

$$= T_y(z). \tag{29}$$

Thus a common rescaling leaves all class scores, predictions, analytical thresholds expressed in score space, and the training objective unchanged under the corresponding latent rescaling. Normalizing the mean class scale to one therefore fixes an otherwise arbitrary global unit while preserving the estimated relative class variability.

A.3 Geometry-level class regions

The fixed separation between class centers is controlled by a geometry-level reference mass $\rho_{\mathrm{sep}} \in (0, 1)$. The corresponding analytical threshold is

$$\tau_{\rho_{\mathrm{sep}}} = F^{-1}_{\chi^2_d}(\rho_{\mathrm{sep}}). \tag{30}$$

For class y, consider the corresponding region

$$\mathcal{B}_y^{\mathrm{sep}} = \left\{ z \in \mathbb{R}^d \mid T_y(z) \le \tau_{\rho_{\mathrm{sep}}} \right\}. \tag{31}$$

Using the definition of T_y,

$$T_y(z) \le \tau_{\rho_{\text{sep}}} \iff \frac{\|z - C_y\|_2^2}{\sigma_y^2} \le \tau_{\rho_{\text{sep}}} \tag{32}$$

$$\iff \|z - C_y\|_2 \le \sigma_y \sqrt{\tau_{\rho_{\text{sep}}}}. \tag{33}$$

Hence $\mathcal{B}_y^{\text{sep}}$ is a Euclidean ball centered at C_y with radius

$$r_y^{\text{sep}} = \sigma_y \sqrt{\tau_{\rho_{\text{sep}}}}. \tag{34}$$

The role of ρ_{sep} is purely geometric: it determines how far apart the fixed class centers are placed before encoder training. It is not a deployment operating level and does not constrain the shell boundaries $\rho_1, \rho_2 \in [0, 1)$ subsequently used to define analytical regions (Section 3.4).

A.4 Pairwise separation of class regions

Consider two classes $y \neq y'$. At the geometry level ρ_{sep}, their regions have radii

$$r_y^{\text{sep}} = \sigma_y \sqrt{\tau_{\rho_{\text{sep}}}}, \qquad r_{y'}^{\text{sep}} = \sigma_{y'} \sqrt{\tau_{\rho_{\text{sep}}}}. \tag{35}$$

A sufficient condition for the two closed balls not to overlap in their interiors is

$$\|C_y - C_{y'}\|_2 \ge r_y^{\text{sep}} + r_{y'}^{\text{sep}}, \tag{36}$$

which gives

$$\|C_y - C_{y'}\|_2 \ge \sqrt{\tau_{\rho_{\text{sep}}}} \left(\sigma_y + \sigma_{y'}\right). \tag{37}$$

Let

$$\sigma_{\max} = \max_{1 \le y \le K} \sigma_y. \tag{38}$$

For every class pair,

$$\sigma_y + \sigma_{y'} \le 2\sigma_{\max}. \tag{39}$$

Therefore, choosing the common pairwise distance

$$\boxed{\alpha = 2\sigma_{\max} \sqrt{\tau_{\rho_{\text{sep}}}}} \tag{40}$$

is sufficient to satisfy Equation (37) simultaneously for every class pair.

The use of a single distance α is deliberately conservative for class pairs with $\sigma_y + \sigma_{y'} < 2\sigma_{\max}$, but it avoids class-dependent pairwise separations and therefore preserves a symmetric center configuration.

A.5 Translation invariance and choice of origin

The absolute position of the latent geometry is also arbitrary. For any $t \in \mathbb{R}^d$, define

$$z' = z + t, \qquad C'_y = C_y + t. \tag{41}$$

Then

$$T'_y(z') = \frac{\|(z + t) - (C_y + t)\|_2^2}{\sigma_y^2} \tag{42}$$

$$= \frac{\|z - C_y\|_2^2}{\sigma_y^2} \tag{43}$$

$$= T_y(z). \tag{44}$$

A common translation therefore leaves every score and prediction unchanged. This free degree of freedom can consequently be fixed by requiring the center configuration to have zero centroid:

$$\frac{1}{K} \sum_{y=1}^{K} C_y = 0, \tag{45}$$

or equivalently

$$\sum_{y=1}^{K} C_y = 0. \tag{46}$$

The origin is therefore the centroid of the center configuration and is not assigned to any class. Under the regular-simplex construction derived below, all class centers additionally have the same norm and therefore lie on a sphere centered at the origin.

A.6 Symmetry requirements

Equation (40) specifies how far apart class centers should be but does not determine their relative arrangement. Infinitely many configurations satisfy only a lower bound on pairwise distance.

Since the class labels themselves contain no geometric ordering, the construction avoids introducing an arbitrary preference for any particular class or pair of classes. After fixing the common origin through Equation (46), the centers are therefore required, for some $R > 0$, to satisfy

$$\|C_y\|_2 = R, \qquad \forall y, \tag{47}$$

and, as the final symmetry constraint,

$$\|C_y - C_{y'}\|_2 = \alpha, \qquad \forall y \neq y'. \tag{48}$$

The first condition places every class center at equal distance from the common origin. The second places every pair of class centers at equal distance from one another. Together with the zero-centroid condition, these constraints eliminate any class-dependent geometric preference.

For $K \leq d + 1$, this configuration is precisely a regular $(K - 1)$-simplex.

A.7 Proof of Proposition 1

Assume

$$\|C_y\|_2 = R \tag{49}$$

for every class and

$$\|C_y - C_{y'}\|_2 = \alpha \tag{50}$$

for every $y \neq y'$.

Expanding the squared pairwise distance gives

$$\alpha^2 = \|C_y - C_{y'}\|_2^2 \tag{51}$$

$$= \|C_y\|_2^2 + \|C_{y'}\|_2^2 - 2C_y^\top C_{y'} \tag{52}$$

$$= 2R^2 - 2C_y^\top C_{y'}. \tag{53}$$

Hence all off-diagonal inner products have the same value,

$$C_y^\top C_{y'} = c = R^2 - \frac{\alpha^2}{2}, \qquad y \neq y'. \tag{54}$$

Using the zero-centroid condition,

$$\sum_{j=1}^{K} C_j = 0, \tag{55}$$

and taking the inner product with C_y gives

$$0 = C_y^\top \sum_{j=1}^{K} C_j \tag{56}$$

$$= \|C_y\|_2^2 + \sum_{j \neq y} C_y^\top C_j \tag{57}$$

$$= R^2 + (K - 1)c. \tag{58}$$

Therefore

$$c = -\frac{R^2}{K-1}. \tag{59}$$

Combining Equations (54) and (59),

$$R^2 - \frac{\alpha^2}{2} = -\frac{R^2}{K-1}. \tag{60}$$

Multiplying by $K-1$ gives

$$(K-1)R^2 - \frac{K-1}{2}\alpha^2 = -R^2, \tag{61}$$

and hence

$$KR^2 = \frac{K-1}{2}\alpha^2. \tag{62}$$

Thus

$$R^2 = \alpha^2 \frac{K-1}{2K}, \tag{63}$$

or

$$R = \alpha\sqrt{\frac{K-1}{2K}}. \tag{64}$$

Substituting this expression into Equation (59) yields

$$c = -\frac{\alpha^2}{2K}. \tag{65}$$

Therefore

$$C_y^\top C_{y'} = -\frac{\alpha^2}{2K}, \qquad y \neq y', \tag{66}$$

which establishes Equation (4). The Gram-matrix characterization below establishes the required latent dimension and completes the uniqueness statement of Proposition 1.

A.8 Gram matrix and required latent dimension

The center Gram matrix $G \in \mathbb{R}^{K\times K}$ is defined by

$$G_{yy'} = C_y^\top C_{y'}. \tag{67}$$

From the previous derivation,

$$G_{yy} = R^2 = \alpha^2 \frac{K-1}{2K}, \tag{68}$$

while

$$G_{yy'} = -\frac{\alpha^2}{2K}, \qquad y \neq y'. \tag{69}$$

Let I_K denote the $K \times K$ identity matrix and $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^K$. Then

$$G = \frac{\alpha^2}{2}\left(I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right). \tag{70}$$

The matrix in parentheses is the orthogonal projector onto the hyperplane

$$\mathcal{H} = \left\{v \in \mathbb{R}^K \mid \mathbf{1}^\top v = 0\right\}, \tag{71}$$

which has dimension $K-1$. Consequently,

$$\operatorname{rank}(G) = K - 1. \tag{72}$$

Although there are K class centers, they are linearly dependent because

$$\sum_{y=1}^{K} C_y = 0. \tag{73}$$

They therefore span a $(K-1)$-dimensional subspace. An exact realization of the center geometry consequently requires

$$d \geq K-1, \tag{74}$$

equivalently

$$K \leq d+1. \tag{75}$$

Moreover, the Gram matrix fixes all inner products between class centers. Hence any two realizations of this geometry are related by an orthogonal transformation on their $(K-1)$-dimensional spans, which can be extended to an orthogonal transformation of $\mathbb{R}^d$. A permutation of the class labels simply permutes the corresponding centers. This completes the uniqueness statement in Proposition 1.

A.9 Explicit regular simplex construction

Let $e_1, \ldots, e_K$ denote the canonical basis of $\mathbb{R}^K$ and define

$$\widetilde{C}_y = \frac{\alpha}{\sqrt{2}}\left(e_y - \frac{1}{K}\mathbf{1}\right). \tag{76}$$

The centroid is

$$\sum_{y=1}^{K} \widetilde{C}_y = \frac{\alpha}{\sqrt{2}}\left(\sum_{y=1}^{K} e_y - \sum_{y=1}^{K} \frac{1}{K}\mathbf{1}\right) \tag{77}$$

$$= \frac{\alpha}{\sqrt{2}}\left(\mathbf{1} - \mathbf{1}\right) \tag{78}$$

$$= 0. \tag{79}$$

For $y \neq y'$,

$$\widetilde{C}_y - \widetilde{C}_{y'} = \frac{\alpha}{\sqrt{2}}(e_y - e_{y'}), \tag{80}$$

so

$$\|\widetilde{C}_y - \widetilde{C}_{y'}\|_2^2 = \frac{\alpha^2}{2}\|e_y - e_{y'}\|_2^2 \tag{81}$$

$$= \frac{\alpha^2}{2}(2) \tag{82}$$

$$= \alpha^2. \tag{83}$$

Hence

$$\|\widetilde{C}_y - \widetilde{C}_{y'}\|_2 = \alpha. \tag{84}$$

The squared center norm is

$$\|\widetilde{C}_y\|_2^2 = \frac{\alpha^2}{2}\left\|e_y - \frac{1}{K}\mathbf{1}\right\|_2^2 \tag{85}$$

$$= \frac{\alpha^2}{2}\left(1 - \frac{1}{K}\right) \tag{86}$$

$$= \alpha^2 \frac{K-1}{2K}, \tag{87}$$

recovering Equation (64).

Furthermore, every $\widetilde{C}_y$ belongs to the $(K-1)$-dimensional hyperplane $\mathcal{H}$. Choosing any orthonormal basis of $\mathcal{H}$ therefore gives coordinates in $\mathbb{R}^{K-1}$ that preserve all inner products and pairwise distances. If $d > K-1$, the remaining coordinates can simply be set to zero.

Thus the simplex can be embedded exactly in $\mathbb{R}^d$ for every $d \geq K-1$.

For geometric intuition, when $K=3$ and $d=2$, the construction reduces to an equilateral triangle of side length α whose centroid is the origin. The three class centers are the triangle vertices: none lies at the origin, and all are at the same distance R from it.

A.10 SEPARATION-INDUCED REFERENCE-DISTRIBUTION MISMATCH

To prove Proposition 2, for each class y, let $a_y = \mathbb{P}(\widehat{Y} = Y \mid Y = y)$ denote its classification accuracy, and define the corresponding class-conditional reference-score CDF by

$$F_{\mathrm{ref},y}(t) = \mathbb{P}(T_y(Z) \le t \mid Y = y)\,. \tag{88}$$

Consider an observation with $Y = y$ and $T_y(Z) < \tau_{\rho_{\mathrm{sep}}}$. If $\widehat{Y} \neq y$, then by definition of the classifier there exists some $j \neq y$ such that

$$T_j(Z) \le T_y(Z) < \tau_{\rho_{\mathrm{sep}}}. \tag{89}$$

Hence Z belongs to the interiors of both $\mathcal{B}_y^{\mathrm{sep}}$ and $\mathcal{B}_j^{\mathrm{sep}}$, contradicting their pairwise disjointness. Therefore

$$T_y(Z) < \tau_{\rho_{\mathrm{sep}}} \quad \Longrightarrow \quad \widehat{Y} = y \tag{90}$$

whenever $Y = y$. Consequently, for every $t < \tau_{\rho_{\mathrm{sep}}}$,

$$F_{\mathrm{ref},y}(t) \le a_y. \tag{91}$$

Let $\pi_y = \mathbb{P}(Y = y)$. Averaging over classes gives

$$F_{\mathrm{ref}}(t) = \sum_{y=1}^{K} \pi_y F_{\mathrm{ref},y}(t) \le \sum_{y=1}^{K} \pi_y a_y = \mathbb{P}(\widehat{Y} = Y) = a. \tag{92}$$

Hence, for every $t < \tau_{\rho_{\mathrm{sep}}}$,

$$\mathrm{KS}_{\mathrm{ref}} \ge F_{\chi^2_d}(t) - F_{\mathrm{ref}}(t) \ge F_{\chi^2_d}(t) - a. \tag{93}$$

Since $F_{\chi^2_d}(\tau_{\rho_{\mathrm{sep}}}) = \rho_{\mathrm{sep}}$ and $F_{\chi^2_d}$ is continuous, taking the limit as t increases to $\tau_{\rho_{\mathrm{sep}}}$ from below (written $t \uparrow \tau_{\rho_{\mathrm{sep}}}$) gives

$$\mathrm{KS}_{\mathrm{ref}} \ge \rho_{\mathrm{sep}} - a. \tag{94}$$

Because $\mathrm{KS}_{\mathrm{ref}} \ge 0$, it follows that

$$\mathrm{KS}_{\mathrm{ref}} \ge [\rho_{\mathrm{sep}} - a]_+\,, \tag{95}$$

which proves Proposition 2.

A.11 UNIQUENESS OF THE SYMMETRIC GEOMETRY

Once α is fixed, Equations (46), (47), and (48) determine the center Gram matrix uniquely through Equation (70).

Any two center configurations having the same Gram matrix are related by an orthogonal transformation on their $(K-1)$-dimensional spans, which can be extended to an orthogonal transformation of $\mathbb{R}^d$. Relabeling the simplex vertices additionally corresponds to a permutation of the class indices.

Therefore, for fixed α, the resulting center geometry is unique up to

1. an orthogonal transformation of latent space;
2. a permutation of the class indices.

No numerical repulsion, iterative center optimization, or post-hoc geometric adjustment is required.

B DIFFICULTY-ORDERED QUANTILE LEARNING

This appendix details the training mechanism summarized in Section 3.3. The central objective is to convert the current class-conditional discriminative ordering of samples into target positions on the prescribed χ^2_d analytical scale. Margins are ranked within each class, transformed into empirical upper-tail percentiles, and then mapped through the inverse χ^2_d CDF. The resulting targets therefore encode discriminative difficulty through their ordering rather than through the raw margin magnitude.

B.1 DISCRIMINATIVE SCORE MARGIN

For a training sample (x_i, y_i), let

$$z_i = f_\theta(x_i). \tag{96}$$

Its score relative to class j is

$$T_{ij} = T_j(z_i) = \frac{\|z_i - C_j\|_2^2}{\sigma_j^2}. \tag{97}$$

The true-class score is

$$T_{i,y_i} = T_{y_i}(z_i), \tag{98}$$

and the closest competing-class score is

$$T_i^{\text{comp}} = \min_{j \neq y_i} T_{ij}. \tag{99}$$

The discriminative score margin is

$$m_i = T_i^{\text{comp}} - T_{i,y_i}. \tag{100}$$

Let

$$\hat{y}_i = \hat{y}(z_i). \tag{101}$$

Ignoring exact ties, the sign of m_i has the direct classification interpretation

$$m_i > 0 \quad \Longleftrightarrow \quad T_{i,y_i} < T_{ij} \quad \forall j \neq y_i \quad \Longleftrightarrow \quad \hat{y}_i = y_i. \tag{102}$$

A larger positive margin additionally indicates greater separation between the true class and the nearest competitor in normalized score space.

The margin is used only to determine the target rank assigned to the current sample. It does not directly prescribe a radial distance. Gradients are not propagated through either the margin or the empirical ranking when the current optimization target is constructed.

B.2 PER-CLASS FIFO MEMORIES

Margin distributions are not assumed to be identical across ground-truth classes. Ranking is therefore performed independently within each class.

For every ground-truth class y, an ordered FIFO memory is maintained:

$$\mathcal{Q}_y^{(t)} = \left(m_{y,1}^{(t)}, \ldots, m_{y,N_y^{(t)}}^{(t)} \right), \qquad N_y^{(t)} \leq M, \tag{103}$$

where M is the maximum FIFO capacity and t denotes the current optimization step. The memory contains recent detached margins from training observations whose ground-truth label is y, irrespective of their predicted class or margin sign. Thus, ignoring exact ties, both correctly classified ($m_i > 0$) and misclassified ($m_i < 0$) observations contribute to the same class-conditional memory.

The FIFO truncation is purely temporal. Once the memory reaches capacity, new margins are appended and the oldest entries are removed so that at most M margins are retained; no entry is filtered according to its margin value or classification outcome.

Let

$$\mathcal{M}_y^{(t)} = \{ m_i \mid y_i = y,\ i \text{ belongs to the current minibatch} \} \tag{104}$$

denote the multiset of current class-y batch margins. Before updating the FIFO memory, the reference multiset used for percentile estimation is

$$\mathcal{R}_y^{(t)} = \mathcal{Q}_y^{(t)} \uplus \mathcal{M}_y^{(t)}, \tag{105}$$

where $\uplus$ denotes multiset union, with the entries of $\mathcal{Q}_y^{(t)}$ viewed as a multiset for percentile computation. Hence,

$$|\mathcal{R}_y^{(t)}| = N_y^{(t)} + |\mathcal{M}_y^{(t)}|. \tag{106}$$

Including the complete current minibatch has two practical consequences. First, percentile estimation is possible from the first minibatch even when the FIFO memories are initially empty. Second, current samples are ranked relative to both recent historical margins and their same-ground-truth-class peers in the current update.

B.3 Empirical upper-tail percentile

For sample i with ground-truth class y_i, define

$$N_i^{>} = \sum_{m \in \mathcal{R}_{y_i}^{(t)}} \mathbf{1}\{m > m_i\}. \tag{107}$$

Its raw empirical upper-tail percentile is

$$\widetilde{u}_i = \frac{N_i^{>} + \frac{1}{2}}{|\mathcal{R}_{y_i}^{(t)}|}. \tag{108}$$

Because the current sample itself belongs to $\mathcal{R}_{y_i}^{(t)}$, the reference multiset is nonempty and $N_i^{>} \leq |\mathcal{R}_{y_i}^{(t)}| - 1$. Hence the half-count correction ensures $\widetilde{u}_i \in (0, 1)$ for every finite reference multiset. The strict comparison also ensures that identical margin values receive the same upper-tail count.

For numerical stability when evaluating the inverse χ^2_d CDF, the value actually used to construct the analytical target is

$$u_i = \mathrm{clip}\left(\widetilde{u}_i, \varepsilon, 1 - \varepsilon\right) := \min(\max(\widetilde{u}_i, \varepsilon),\, 1 - \varepsilon)\,, \qquad \varepsilon = 10^{-4}. \tag{109}$$

B.4 Why the upper-tail convention reverses difficulty

Consider two samples i and j from the same ground-truth class and ranked against the same reference multiset. If $m_i > m_j$, every reference margin exceeding m_i also exceeds m_j. Moreover, since m_i itself belongs to the common reference multiset, it contributes to $N_j^{>}$ but not to $N_i^{>}$. Therefore

$$N_i^{>} < N_j^{>}, \tag{110}$$

and hence

$$\widetilde{u}_i < \widetilde{u}_j. \tag{111}$$

Clipping through Equation (109) is monotone nondecreasing, so

$$u_i \leq u_j. \tag{112}$$

Since the χ^2_d quantile function is monotonically increasing,

$$q_i = F^{-1}_{\chi^2_d}(u_i) \leq F^{-1}_{\chi^2_d}(u_j) = q_j. \tag{113}$$

Hence

$$m_i > m_j \quad \Longrightarrow \quad \widetilde{u}_i < \widetilde{u}_j \quad \Longrightarrow \quad u_i \leq u_j \quad \Longrightarrow \quad q_i \leq q_j. \tag{114}$$

The upper-tail convention therefore maps larger discriminative margins to no larger target scores. Before clipping, the empirical percentile ordering is strict; clipping may merge only values falling within the same clipped tail. This is the mechanism by which discriminative difficulty is injected into the prescribed analytical scale.

B.5 Population quantile construction

The following derivation justifies the idealized population construction underlying Proposition 3.

Proof of Proposition 3. For a fixed class y, let M_y denote the discriminative margin conditional on $Y = y$, and assume that its CDF G_y is continuous and strictly increasing on its support. Define

$$V_y = 1 - G_y(M_y). \tag{115}$$

By the probability integral transform,

$$G_y(M_y) \sim \mathrm{Uniform}(0, 1), \tag{116}$$

and therefore

$$V_y \sim \mathrm{Uniform}(0, 1). \tag{117}$$

Defining

$$Q_y = F^{-1}_{\chi^2_d}(V_y), \tag{118}$$

the inverse-transform construction gives

$$Q_y \sim \chi^2_d. \tag{119}$$

Moreover, for any $\rho \in (0, 1)$, since the χ^2_d quantile function is strictly increasing,

$$Q_y \le \tau_\rho \iff F^{-1}_{\chi^2_d}(V_y) \le F^{-1}_{\chi^2_d}(\rho) \tag{120}$$

$$\iff V_y \le \rho \tag{121}$$

$$\iff 1 - G_y(M_y) \le \rho \tag{122}$$

$$\iff G_y(M_y) \ge 1 - \rho. \tag{123}$$

Let

$$G^{-1}_y(p) = \inf\{m \in \mathbb{R} \mid G_y(m) \ge p\} \tag{124}$$

denote the generalized quantile function. Since G_y is continuous and strictly increasing on its support,

$$G_y(M_y) \ge 1 - \rho \quad \iff \quad M_y \ge G^{-1}_y(1 - \rho). \tag{125}$$

Thus,

$$Q_y \le \tau_\rho \quad \iff \quad M_y \ge G^{-1}_y(1 - \rho) \tag{126}$$

which establishes the second statement of the proposition. □

This population argument explains the role of the empirical FIFO construction: the raw percentile $\widetilde{u}_i$ approximates the upper-tail transform $1 - G_y(M_y)$ while allowing the class-conditional margin distribution itself to evolve during training. The clipped value u_i is then used for numerical stability when constructing the target.

B.6 Target quantile

The clipped empirical percentile is converted into a prescribed analytical target by

$$q_i = F^{-1}_{\chi^2_d}(u_i). \tag{127}$$

The FIFO therefore does not directly prescribe a Euclidean distance. It estimates the empirical rank of the sample within its evolving class-conditional difficulty distribution, and the fixed reference law then converts this rank into a score-space position.

For sample i, the corresponding target Euclidean radius relative to its ground-truth class center is

$$r_i = \sigma_{y_i}\sqrt{q_i}. \tag{128}$$

Consequently, the same percentile u_i represents the same normalized analytical position across classes even when their Euclidean scales differ.

B.7 Update order and absence of within-batch ordering effects

All values $\{u_i, q_i\}$ for the current minibatch are computed using the FIFO states that existed before the minibatch was processed, augmented by the complete multiset of current same-class margins.

Only after all current targets have been computed are the current margins appended to their respective FIFO memories:

$$\mathcal{Q}_y^{(t+1)} \leftarrow \mathrm{FIFO}_M\left(\mathcal{Q}_y^{(t)} \parallel \mathcal{M}_y^{(t)}\right), \tag{129}$$

where $\parallel$ denotes ordered concatenation and FIFO_M retains the M most recent elements.

This ordering is important. If samples were appended one by one before subsequent percentiles were evaluated, otherwise identical minibatches could produce different targets depending on the arbitrary processing order of their elements.

The delayed update makes the target construction permutation invariant with respect to the ordering of samples within the minibatch.

B.8 FINITE-MEMORY INTERPRETATION

The FIFO estimator approximates the evolving class-conditional margin distribution rather than the distribution accumulated over the complete training trajectory.

This is intentional. As the encoder changes, margins computed many epochs earlier may no longer represent the current geometry. A finite queue therefore provides a moving empirical approximation to the contemporary margin distribution.

The FIFO capacity M controls an adaptation–stability trade-off:

- a smaller M adapts more rapidly to changes in the representation but yields a noisier empirical rank estimate;
- a larger M provides a smoother rank estimate but retains margins generated by older encoder states for longer.

The FIFO memories persist across epoch boundaries and are not reset between epochs; entries are removed only through the capacity-M FIFO mechanism. Optimization is performed at the minibatch level: after the analytical targets for the current minibatch have been constructed, the minibatch loss is backpropagated and the encoder parameters are updated before processing the next minibatch.

At inference time the FIFO is not used. It is purely a training-time mechanism for constructing analytical targets.

B.9 ADOQ TRAINING OBJECTIVE

Once q_i has been constructed, it is treated as fixed during the current gradient update. The per-sample objective is

$$\ell_{\mathrm{ADOQ},i} = \left(\frac{T_{i,y_i} - q_i}{d} \right)^2, \tag{130}$$

and the minibatch objective is

$$\mathcal{L}_{\mathrm{ADOQ}} = \frac{1}{B} \sum_{i=1}^{B} \ell_{\mathrm{ADOQ},i}. \tag{131}$$

Since d is fixed during training, this normalization does not change the target minimizer:

$$T_{i,y_i} = q_i. \tag{132}$$

Its role is instead to normalize optimization scale across latent dimensions. Because

$$\mathbb{E}[\chi^2_d] = d, \tag{133}$$

both the prescribed targets and their corresponding scores naturally scale with d. Dividing their difference by d expresses the residual relative to this natural score scale and reduces the dependence of loss and gradient magnitudes on the chosen latent dimension.

Although Equation (131) directly penalizes only the true-class score, its target is not defined independently of the competing classes. The current competing scores determine the margin m_i, which determines the percentile u_i and therefore the target q_i. As the encoder changes, this construction is repeated and the analytical target assigned to a sample evolves with its discriminative ordering.

There is no cross-entropy term, explicit competing-class penalty, reconstruction term, KL term, warm-up schedule, calibration objective, or adaptive weighting coefficient in the ADOQ objective.

B.10 QUANTITATIVE SEPARATION CONSEQUENCE

The fixed geometry additionally provides a quantitative connection between an analytical score position and discriminative separation.

Let z be an observation of true class y such that

$$t = T_y(z) \leq \tau_{\rho_{\mathrm{sep}}}. \tag{134}$$

By definition of the normalized score,

$$\|z - C_y\|_2 = \sigma_y \sqrt{t}. \tag{135}$$

For any competing class $j \neq y$, the triangle inequality gives

$$\|z - C_j\|_2 \geq \|C_j - C_y\|_2 - \|z - C_y\|_2 \tag{136}$$

$$= 2\sigma_{\max}\sqrt{\tau_{\rho_{\text{sep}}}} - \sigma_y\sqrt{t} \tag{137}$$

$$\geq \sigma_{\max}\left(2\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{t}\right), \tag{138}$$

where Equation (3) and $\sigma_y \leq \sigma_{\max}$ were used.

Since $t \leq \tau_{\rho_{\text{sep}}}$, the factor in parentheses is positive. Moreover, $\sigma_j \leq \sigma_{\max}$, so

$$T_j(z) = \frac{\|z - C_j\|_2^2}{\sigma_j^2} \tag{139}$$

$$\geq \left(2\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{t}\right)^2. \tag{140}$$

Defining

$$T^{\text{comp}}(z) = \min_{j \neq y} T_j(z), \tag{141}$$

it follows that

$$T^{\text{comp}}(z) \geq \left(2\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{t}\right)^2. \tag{142}$$

Hence the discriminative margin satisfies

$$m(z) = T^{\text{comp}}(z) - T_y(z) \tag{143}$$

$$\geq \left(2\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{t}\right)^2 - t \tag{144}$$

$$= 4\sqrt{\tau_{\rho_{\text{sep}}}}\left(\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{t}\right). \tag{145}$$

Therefore, every observation satisfying

$$T_y(z) < \tau_{\rho_{\text{sep}}} \tag{146}$$

has $m(z) > 0$ and is thus strictly correctly classified under the fixed geometry.

Now consider a training sample i for which the per-sample ADOQ objective vanishes at the current model state. Equation (132) then gives

$$T_{y_i}(z_i) = q_i. \tag{147}$$

If $u_i < \rho_{\text{sep}}$, monotonicity of the χ^2_d quantile function gives

$$q_i = F^{-1}_{\chi^2_d}(u_i) < F^{-1}_{\chi^2_d}(\rho_{\text{sep}}) = \tau_{\rho_{\text{sep}}}. \tag{148}$$

Applying Equation (145) with $t = q_i$ therefore gives

$$m_i \geq 4\sqrt{\tau_{\rho_{\text{sep}}}}\left(\sqrt{\tau_{\rho_{\text{sep}}}} - \sqrt{q_i}\right) \tag{149}$$

$$> 0. \tag{150}$$

Consequently,

$$T_{y_i}(z_i) < T_{ij} \qquad \forall j \neq y_i, \tag{151}$$

so the sample is strictly correctly classified. This establishes the quantitative separation consequence stated in Section 3.3.

This consequence concerns the true-class analytical coordinate used during training. It does not imply that an observation satisfying $S_{\text{op}}(z) < \tau_{\rho_{\text{sep}}}$ is necessarily correctly classified, because the operational minimizer may differ from the unknown true class. Predictive reliability of regions defined from S_{op} is therefore assessed independently as described in Section 3.4.

B.11 TRAINING PROCEDURE

For completeness, one minibatch update can be summarized as follows.

1. Encode every current input:
$$z_i = f_\theta(x_i). \tag{152}$$
2. Compute all normalized class scores:
$$T_{ij} = \frac{\|z_i - C_j\|_2^2}{\sigma_j^2}. \tag{153}$$
3. From detached copies of the current scores, compute the discriminative margins used to construct the targets:
$$m_i = \min_{j \neq y_i} T_{ij} - T_{i,y_i}. \tag{154}$$
4. For each class represented in the minibatch, form
$$\mathcal{R}_y^{(t)} = \mathcal{Q}_y^{(t)} \uplus \mathcal{M}_y^{(t)}. \tag{155}$$
5. Compute the raw upper-tail percentile $\widetilde{u}_i$ of every current margin using Equation (108).
6. Clip $\widetilde{u}_i$ according to Equation (109) to obtain u_i, and compute
$$q_i = F_{\chi_d^2}^{-1}(u_i). \tag{156}$$
7. After all current targets have been computed, append the minibatch margins to their corresponding FIFO memories according to Equation (129).
8. Reuse the original differentiable true-class scores to compute
$$\mathcal{L}_{\mathrm{ADOQ}} = \frac{1}{B} \sum_{i=1}^{B} \left(\frac{T_{i,y_i} - q_i}{d} \right)^2. \tag{157}$$
9. Backpropagate the minibatch loss and update the encoder parameters before processing the next minibatch.

At test time, Steps 3–9 disappear entirely. Inference requires only the deterministic encoder, the fixed geometry, and the class scores.

C ANALYTICAL REGIONS AND INDEPENDENT CERTIFICATION

This appendix develops the certification results summarized in Section 3.4. Once training and model selection are complete, the model, analytical coordinate, and region boundaries are fixed. Independent certification observations are then used only to assess the population occupancy and predictive reliability of these predefined regions. All finite-sample probability statements below are understood conditional on the frozen model.

Two quantities must be kept distinct. Operational distributional fidelity quantifies how closely realized population masses follow the nominal masses induced by the prescribed χ_d^2 scale. Predictive reliability instead measures how frequently predictions are correct within a fixed region. Operational distributional fidelity and population occupancy can be assessed without labels, whereas predictive reliability requires labeled observations. Importantly, the certification results below do not require exact realization of the prescribed score distribution.

C.1 REFERENCE AND OPERATIONAL SCORES

With $Z = f_\theta(X)$, define the true-class reference score and operational deployment score as
$$S_{\mathrm{ref}} = T_Y(Z), \qquad S_{\mathrm{op}} = \min_{1 \leq j \leq K} T_j(Z) = T_{\widehat{Y}}(Z), \tag{158}$$
with respective CDFs F_{ref} and F_{op}.

Since $T_Y(Z)$ is one of the scores entering the minimum,

$$S_{\mathrm{op}} \le S_{\mathrm{ref}}. \tag{159}$$

Consequently,

$$F_{\mathrm{op}}(t) \ge F_{\mathrm{ref}}(t) \qquad \forall t \in \mathbb{R}. \tag{160}$$

More precisely,

$$F_{\mathrm{op}}(t) - F_{\mathrm{ref}}(t) = \mathbb{P}\left(S_{\mathrm{op}} \le t < S_{\mathrm{ref}}\right). \tag{161}$$

Whenever $\widehat{Y} = Y$, the two scores are identical. Hence

$$F_{\mathrm{op}}(t) - F_{\mathrm{ref}}(t) = \mathbb{P}\left(\widehat{Y} \ne Y,\, S_{\mathrm{op}} \le t < S_{\mathrm{ref}}\right), \tag{162}$$

and therefore

$$0 \le F_{\mathrm{op}}(t) - F_{\mathrm{ref}}(t) \le \mathbb{P}(\widehat{Y} \ne Y). \tag{163}$$

Let $F_0 = F_{\chi^2_d}$ and define

$$\mathrm{KS}_{\mathrm{ref}} = \|F_{\mathrm{ref}} - F_0\|_\infty, \qquad \mathrm{KS}_{\mathrm{op}} = \|F_{\mathrm{op}} - F_0\|_\infty. \tag{164}$$

The reverse triangle inequality and Equation (163) give

$$|\mathrm{KS}_{\mathrm{op}} - \mathrm{KS}_{\mathrm{ref}}| \le \|F_{\mathrm{op}} - F_{\mathrm{ref}}\|_\infty \le \mathbb{P}(\widehat{Y} \ne Y). \tag{165}$$

Thus classification error bounds the difference between the reference and operational KS discrepancies. It does not imply that either discrepancy is small: the reference score may itself differ substantially from the prescribed law.

C.2 Population occupancy of analytical shells

For an interior analytical shell $0 < \rho_1 < \rho_2 < 1$,

$$\mathcal{S}_{\rho_1,\rho_2} = \{x \mid \tau_{\rho_1} < S_{\mathrm{op}}(x) \le \tau_{\rho_2}\}, \tag{166}$$

where the analytical thresholds follow the convention defined in Section 3.1.

Its population occupancy is

$$C(\rho_1,\rho_2) = \mathbb{P}\left(\tau_{\rho_1} < S_{\mathrm{op}} \le \tau_{\rho_2}\right) \tag{167}$$

$$= F_{\mathrm{op}}(\tau_{\rho_2}) - F_{\mathrm{op}}(\tau_{\rho_1}). \tag{168}$$

Under exact realization of the prescribed operational score law, $F_{\mathrm{op}} = F_{\chi^2_d}$, and therefore

$$C(\rho_1,\rho_2) = \rho_2 - \rho_1. \tag{169}$$

In general, however, the realized occupancy need not equal this nominal mass. Since $|F_{\mathrm{op}}(t) - F_{\chi^2_d}(t)| \le \mathrm{KS}_{\mathrm{op}}$ for every t, applying the bound at both shell boundaries gives

$$|C(\rho_1,\rho_2) - (\rho_2 - \rho_1)| \le 2\mathrm{KS}_{\mathrm{op}}. \tag{170}$$

For the cumulative region $\mathcal{S}_{0,\rho} = \{x \mid S_{\mathrm{op}}(x) \le \tau_\rho\}$, only one CDF boundary contributes, yielding

$$|C(0,\rho) - \rho| \le \mathrm{KS}_{\mathrm{op}}. \tag{171}$$

Similarly, for the outer tail $\mathcal{T}_\rho$,

$$|C(\mathcal{T}_\rho) - (1-\rho)| \le \mathrm{KS}_{\mathrm{op}}. \tag{172}$$

These relations explain the role of the operational KS diagnostic: it quantifies fidelity of realized population occupancy to the prescribed analytical scale, but does not determine predictive reliability.

C.3 Uniform unlabeled occupancy certification

Let $X_1, \dots, X_n$ be i.i.d. certification observations from the target population, independent of the complete frozen model and model-selection procedure, and let $S_i = S_{\mathrm{op}}(X_i)$. The empirical CDF of the operational scores is

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{S_i \le t\}. \tag{173}$$

The Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\mathbb{P}\left(\sup_t \left| \widehat{F}_n(t) - F_{\mathrm{op}}(t) \right| > \varepsilon \right) \le 2e^{-2n\varepsilon^2}. \tag{174}$$

Defining

$$\epsilon_n(\delta) = \sqrt{\frac{\log(2/\delta)}{2n}}, \tag{175}$$

it follows that, with probability at least $1 - \delta$,

$$\sup_t \left| \widehat{F}_n(t) - F_{\mathrm{op}}(t) \right| \le \epsilon_n(\delta). \tag{176}$$

For an interior analytical shell, define its empirical occupancy as

$$\widehat{C}(\rho_1, \rho_2) = \widehat{F}_n(\tau_{\rho_2}) - \widehat{F}_n(\tau_{\rho_1}). \tag{177}$$

On the single event in Equation (176),

$$\begin{aligned} \left| C(\rho_1, \rho_2) - \widehat{C}(\rho_1, \rho_2) \right| &\le \left| F_{\mathrm{op}}(\tau_{\rho_2}) - \widehat{F}_n(\tau_{\rho_2}) \right| && (178) \\ &\quad + \left| F_{\mathrm{op}}(\tau_{\rho_1}) - \widehat{F}_n(\tau_{\rho_1}) \right| && (179) \\ &\le 2\epsilon_n(\delta). && (180) \end{aligned}$$

Hence

$$\boxed{\left| C(\rho_1, \rho_2) - \widehat{C}(\rho_1, \rho_2) \right| \le 2\epsilon_n(\delta)} \tag{181}$$

simultaneously for all interior shells $0 < \rho_1 < \rho_2 < 1$.

The corresponding lower certificate is

$$C(\rho_1, \rho_2) \ge \max\left\{ 0,\, \widehat{C}(\rho_1, \rho_2) - 2\epsilon_n(\delta) \right\}. \tag{182}$$

For cumulative regions, $\widehat{C}(0, \rho) = \widehat{F}_n(\tau_\rho)$, so only one CDF boundary contributes. Therefore

$$\boxed{\left| C(0, \rho) - \widehat{C}(0, \rho) \right| \le \epsilon_n(\delta)} \tag{183}$$

simultaneously for every $\rho \in (0, 1)$. The same one-boundary argument applies to the outer tail $\mathcal{T}_\rho$, since $\widehat{C}(\mathcal{T}_\rho) = 1 - \widehat{F}_n(\tau_\rho)$ likewise involves a single empirical-CDF evaluation.

A key point is that the DKW sample size is the total number n of independent certification observations, not the number falling inside a particular shell. Moreover, no multiple-comparison correction is required across analytical shells: the single uniform event in Equation (176) controls every score threshold and therefore every shell simultaneously.

C.4 Predictive reliability certification for a fixed shell

Fix an analytical shell $I = \mathcal{S}_{\rho_1, \rho_2}$ independently of the certification observations. Whenever $\mathbb{P}(X \in I) > 0$, let

$$A_I = A(\rho_1, \rho_2) = \mathbb{P}\left(\widehat{Y} = Y \mid X \in I \right) \tag{184}$$

denote its population predictive reliability.

For an independent i.i.d. labeled certification sample $\{(X_i, Y_i)\}_{i=1}^n$, define

$$H_i = \mathbf{1}\{X_i \in I\}, \qquad J_i = \mathbf{1}\{\hat{y}(X_i) = Y_i\}. \tag{185}$$

Let

$$N_I = \sum_{i=1}^n H_i, \qquad K_I = \sum_{i=1}^n H_i J_i \tag{186}$$

denote respectively the number of certification observations falling inside the shell and the number of correct predictions among them.

Because the shell is fixed independently of the certification sample, conditional on N_I,

$$K_I \mid N_I \sim \mathrm{Binomial}\,(N_I, A_I)\,. \tag{187}$$

For $N_I > 0$ and $K_I > 0$, the exact one-sided Clopper–Pearson lower confidence bound at error level δ is

$$A_{\mathrm{LB}}(I, \delta) = F^{-1}_{\mathrm{Beta}(K_I, N_I - K_I + 1)}(\delta). \tag{188}$$

For $K_I = 0$, including when $N_I = 0$, define $A_{\mathrm{LB}}(I, \delta) = 0$.

For every realized value of N_I,

$$\mathbb{P}\,(A_I \geq A_{\mathrm{LB}}(I, \delta) \mid N_I) \geq 1 - \delta. \tag{189}$$

Averaging over the random value of N_I therefore gives

$$\boxed{\mathbb{P}\,(A_I \geq A_{\mathrm{LB}}(I, \delta)) \geq 1 - \delta.} \tag{190}$$

Unlike the DKW occupancy certificate, whose sample size is the total number n of certification observations, the effective sample size of the reliability certificate is N_I, the number of labeled observations actually falling inside the shell.

The cumulative case follows directly by taking $I = \mathcal{S}_{0,\rho}$, and the outer-tail case by taking $I = \mathcal{T}_\rho$.

Proof of Theorem 1. For a shell fixed independently of the certification sample, Equation (181) gives the population occupancy certificate from the uniform DKW event. For predictive reliability, Equation (187) gives the conditional binomial model, and Equation (190) establishes the validity of the one-sided Clopper–Pearson lower confidence bound.

Neither argument requires $S_{\mathrm{op}}(X) \sim \chi^2_d$. The prescribed law is used only to fix the analytical shell boundaries before certification. □

Thus occupancy certification requires only operational scores, whereas predictive reliability certification additionally requires labels. Without additional assumptions, unlabeled score values alone cannot certify the population probability of correct classification within a shell.

C.5 Simultaneous reliability certification

The DKW occupancy certificate is automatically simultaneous over the complete analytical shell family because it follows from a single uniform empirical-CDF event. Clopper–Pearson reliability bounds are different: the result above certifies one fixed region at a specified confidence level.

Suppose a finite family of J analytical regions $I_1, \ldots, I_J$ is specified before the certification observations are examined. The regions may be disjoint, nested, or overlapping. Assign error level δ/J to each individual Clopper–Pearson bound. Then, for every $j = 1, \ldots, J$,

$$\mathbb{P}\left(A_{I_j} \geq A_{\mathrm{LB}}\left(I_j, \frac{\delta}{J}\right)\right) \geq 1 - \frac{\delta}{J}. \tag{191}$$

By the union bound,

$$\boxed{\mathbb{P}\left(A_{I_j} \geq A_{\mathrm{LB}}\left(I_j, \frac{\delta}{J}\right), \quad \forall j = 1, \ldots, J\right) \geq 1 - \delta.} \tag{192}$$

This Bonferroni construction does not require the regions or their estimates to be independent. It therefore remains valid for the strongly overlapping sliding windows used in Section 4.3, although it may be conservative.

The distinction between the two simultaneous statements is important:

- the DKW event controls all shell occupancies simultaneously and therefore requires no multiplicity correction across analytical shells;
- Clopper–Pearson is pointwise for a fixed shell, so a family-wise reliability statement over a finite collection requires a multiplicity correction such as Bonferroni.

If a single joint confidence statement covering both the DKW occupancy event and the complete finite reliability family is required, the total error probability can additionally be split between the two components, for example by assigning $\delta/2$ to each.

C.6 What the certificates do and do not imply

The occupancy and reliability certificates concern population properties of fixed analytical regions. They do not assign an individual probability of correctness to a new observation. In particular, membership of x in a shell I does not identify an individual conditional correctness probability with the region-level quantity A_I.

The valid interpretation is that the new observation belongs to a predefined region whose population predictive reliability has been lower-bounded by an independent certification procedure.

Likewise, neither DKW nor Clopper–Pearson implies that local reliability must vary monotonically along the analytical coordinate. ADOQ is trained to encode a global discriminative ordering, but finite learned representations and operational wrong-center assignments can produce local deviations from strict shell-by-shell monotonicity.

Finally, certification does not assume that

$$C(\rho_1, \rho_2) = \rho_2 - \rho_1. \tag{193}$$

The nominal interval specifies the analytical location and nominal mass of the region under the prescribed operational score law. Its realized population occupancy is measured and certified afterward.

C.7 Separation between construction, selection, and certification

The finite-sample guarantees above require the certification observations to be independent of all data-dependent choices used to construct and select the frozen model.

Let $\mathcal{D}_{\mathrm{train}}$ denote the data used to construct the class geometry and optimize the encoder, and let $\mathcal{D}_{\mathrm{val}}$ denote the data used for model selection. After these stages, the complete model

$$\widehat{\mathcal{M}} = \left(f_{\widehat{\theta}}, \{C_y\}_{y=1}^{K}, \{\sigma_y\}_{y=1}^{K}\right)$$

is frozen. Consequently, the complete family $\{\mathcal{S}_{\rho_1,\rho_2}\}_{0\le\rho_1<\rho_2<1}$ is also fixed through the operational score and the analytical-threshold convention of Section 3.1.

Only then is an independent certification dataset $\mathcal{D}_{\mathrm{cert}}$ used to estimate and certify $C(\rho_1, \rho_2)$ and $A(\rho_1, \rho_2)$. The three data roles are therefore

$$\mathcal{D}_{\mathrm{train}} \longrightarrow \text{geometry construction and representation learning}, \tag{194}$$

$$\mathcal{D}_{\mathrm{val}} \longrightarrow \text{model selection}, \tag{195}$$

$$\mathcal{D}_{\mathrm{cert}} \longrightarrow \text{independent assessment and certification}. \tag{196}$$

Certification observations do not modify the encoder, class centers, class scales, or analytical thresholds. They quantify how the already-defined analytical family behaves on the target population. Any finite collection of regions for which simultaneous Clopper–Pearson reliability guarantees are reported must likewise be specified before the certification observations are examined.

C.8 Contrast with threshold calibration

The distinction between construction and certification also separates ADOQ from procedures in which held-out observations are used to determine the deployed operating threshold.

In a calibration-based selective procedure, calibration observations may be used to construct a data-dependent threshold $\widehat{\tau} = \widehat{\tau}(\mathcal{D}_{\mathrm{cal}}, \text{target criterion})$, for example to target a specified risk or coverage level. The calibration data therefore participate directly in defining the deployed operating rule.

For ADOQ, by contrast, the analytical thresholds τ_ρ are fixed by the prescribed analytical coordinate according to the convention of Section 3.1, before certification observations are examined. More generally, every shell boundary $(\tau_{\rho_1}, \tau_{\rho_2})$ is already fixed once the model has been selected.

Independent certification data are subsequently used only to estimate or bound the realized population occupancy $C(\rho_1, \rho_2)$ and predictive reliability $A(\rho_1, \rho_2)$. This distinction does not make certification observations unnecessary: finite-sample statements about target-population behavior necessarily require data from that population. Rather, certification assesses an already-fixed family of analytical regions instead of defining that family.

D Experimental Details

This appendix provides the experimental details underlying Section 4. It first describes the datasets, architectures, and ADOQ training configuration, then specifies model selection, baseline evaluation, structural diagnostics, and the certification protocol used in RQ2.

D.1 Datasets, splits, and architectures

CIFAR-10, SVHN, Imagenette, and CIFAR-100 are evaluated. In all cases, the test set is excluded from training and model selection and is accessed only after the final checkpoint has been selected. The untouched test split is therefore used as the independent held-out sample for final evaluation and for the certification experiments reported in Section 4.3 and Appendix F.

For CIFAR-10, the official 50,000-image training set is partitioned into 45,000 training and 5,000 validation observations. The official 10,000-image test set is kept untouched until final evaluation.

For SVHN, the official 73,257-image training split is partitioned into 65,932 training and 7,325 validation observations, using the same fixed partition throughout training and model selection. The official 26,032-image test split is kept untouched until final evaluation. The additional `extra` split is not used.

For Imagenette, the experiments use 8,522 training, 947 validation, and 3,925 test observations. Inputs are processed at spatial resolution 160×160. The ResNet-18 backbone is initialized from scratch using `torchvision.models.resnet18(weights=None)`; no ImageNet-pretrained weights are used.

For CIFAR-100, the official 50,000-image training set is partitioned into 45,000 training and 5,000 validation observations. The official 10,000-image test set is kept untouched until final evaluation.

ADOQ uses WideResNet-28-10 on CIFAR-10, SVHN, and CIFAR-100, and ResNet-18 on Imagenette. The latent dimension is $d = 16$ for CIFAR-10, SVHN, and Imagenette, and $d = 100$ for CIFAR-100. The latter permits an exact regular simplex for $K = 100$ classes since the construction only requires $d \geq K - 1$.

D.2 ADOQ training configuration

The fixed class centers and class-specific scales are constructed before encoder optimization according to Appendix A. The geometry-level reference mass is fixed to

$$\rho_{\mathrm{sep}} = 0.999 \tag{197}$$

for all datasets.

The class-specific empirical margin distributions used to construct the difficulty-ordered targets are represented by FIFO buffers as described in Appendix B. Each class uses a maximum FIFO capacity

Table 2: Datasets and ADOQ configurations. Test observations are excluded from training and model selection.

Dataset	Train	Val.	Test	d	Backbone
CIFAR-10	45,000	5,000	10,000	16	WRN-28-10
SVHN	65,932	7,325	26,032	16	WRN-28-10
Imagenette	8,522	947	3,925	16	ResNet-18
CIFAR-100	45,000	5,000	10,000	100	WRN-28-10

Table 3: ADOQ training hyperparameters. All runs use SGD with momentum 0.9, Nesterov acceleration, cosine annealing with $T_{\max} = 400$, no warm-up, and weight decay 5×10^{-4}.

Dataset	Epochs	Batch	Initial LR	LR schedule	Weight decay	Augmentation
CIFAR-10	400	100	0.10	Cosine	5×10^{-4}	Crop + flip
SVHN	400	100	0.10	Cosine	5×10^{-4}	Crop
Imagenette	400	50	0.05	Cosine	5×10^{-4}	Resized crop + flip
CIFAR-100	400	200	0.10	Cosine	5×10^{-4}	Crop + flip

of

$$M = 1024. \tag{198}$$

For each minibatch, percentile targets are computed from the previous FIFO state augmented by all current same-class margins, and the FIFO memories are updated only after every current target has been determined.

Empirical upper-tail percentiles are clipped to

$$[10^{-4}, 1 - 10^{-4}] \tag{199}$$

before applying the inverse χ^2_d CDF.

Training minimizes only

$$\mathcal{L}_{\mathrm{ADOQ}} = \frac{1}{B} \sum_{i=1}^{B} \left(\frac{T_{y_i}(z_i) - q_i}{d} \right)^2 . \tag{200}$$

No cross-entropy, KL-divergence, reconstruction, calibration, or additional auxiliary objective is used.

The optimization settings used for the reported models are summarized in Table 3. All models are trained for 400 epochs with SGD, momentum 0.9, and Nesterov acceleration. Learning rates follow cosine annealing with $T_{\max} = 400$ and no warm-up. Weight decay is 5×10^{-4} for all datasets.

Training minibatches are class balanced. Consequently, each minibatch contains 10 observations per class on CIFAR-10 and SVHN, 5 per class on Imagenette, and 2 per class on CIFAR-100. This ensures that every class contributes current margins to its percentile reference set at each optimization step.

For CIFAR-10, training augmentation consists of `RandomCrop(32,padding=4)` followed by `RandomHorizontalFlip()`. Inputs are normalized using mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2023, 0.1994, 0.2010)$.

For CIFAR-100, the same crop and horizontal-flip augmentation is used, with normalization mean $(0.5071, 0.4867, 0.4408)$ and standard deviation $(0.2675, 0.2565, 0.2761)$.

For SVHN, training augmentation uses `RandomCrop(32,padding=4)` without horizontal flipping. Inputs are normalized using mean $(0.4377, 0.4438, 0.4728)$ and standard deviation $(0.1980, 0.2010, 0.1970)$.

For Imagenette, training uses `RandomResizedCrop(160, scale=(0.60,1.0), ratio=(0.75,4/3))` followed by `RandomHorizontalFlip()`. Inputs are normalized using the ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

Evaluation uses resizing followed by `CenterCrop(160)` and the same normalization. CIFAR-10, CIFAR-100, and SVHN use no stochastic augmentation at evaluation time.

The WideResNet-28-10 models are trained from scratch. Convolutional layers use Kaiming-normal initialization with `fan_out` mode and ReLU nonlinearity; BatchNorm weights and biases are initialized to one and zero, respectively; and linear layers use Xavier-normal initialization with zero bias. The Imagenette ResNet-18 is likewise trained from scratch with `weights=None`.

All reported ADOQ results correspond to one training run per dataset using random seed 42; no averaging across multiple random seeds is performed. Consequently, the experiments characterize the selected trained models rather than training-run variability.

D.3 Model selection

ADOQ checkpoint selection uses validation data only and follows the same rule for all four datasets. Let e range over the saved training checkpoints, A_e^{val} denote validation classification accuracy at checkpoint e, and $\widehat{\mathrm{KS}}_{\mathrm{op},e}^{\mathrm{val}}$ denote the empirical operational KS discrepancy from the prescribed χ_d^2 law.

First, define

$$A_{\max}^{\mathrm{val}} = \max_e A_e^{\mathrm{val}}, \tag{201}$$

and

$$\mathcal{E}_\eta = \left\{ e \mid A_e^{\mathrm{val}} \geq A_{\max}^{\mathrm{val}} - \eta \right\}, \qquad \eta = 0.005. \tag{202}$$

Thus, only checkpoints within 0.5 percentage point of the best validation accuracy are eligible. Among them, the selected checkpoint is

$$e^\star = \arg\min_{e \in \mathcal{E}_\eta} \widehat{\mathrm{KS}}_{\mathrm{op},e}^{\mathrm{val}}. \tag{203}$$

This rule prioritizes classification performance by first restricting selection to accuracy-competitive checkpoints, and uses operational distributional fidelity only to discriminate within that set.

Ties are broken by higher validation accuracy, then lower validation AURC, and finally by the earliest epoch.

No test-set statistic, analytical-shell occupancy, regional reliability, or certification result enters model selection.

The selected checkpoints are

- CIFAR-10: epoch 386;
- SVHN: epoch 388;
- Imagenette: epoch 344;
- CIFAR-100: epoch 400.

After checkpoint selection, the encoder, class centers, class scales, and analytical coordinate are frozen. The analytical thresholds remain fixed by the convention of Section 3.1, so the complete family of analytical region boundaries is determined without using any test or certification observations. This preserves the separation between model construction, model selection, and independent certification required by Section 3.4.

D.4 Baseline evaluation protocol

The comparison in Table 1 includes maximum softmax probability (MSP), SelectiveNet, Mahalanobis, SAT+EM+SR, CCL-SC, MAPLE, and VMDLS. All methods use the same training, validation, and untouched test partitions. Test observations are never used for checkpoint selection, hyperparameter selection, covariance fitting, dimensionality reduction, or threshold tuning.

For classification accuracy and AURC, every method is evaluated using its native prediction rule and native confidence or uncertainty score. AURC is computed from the complete empirical risk–coverage curve induced by that score. The structural diagnostics reported separately below do not replace these native scores.

Supervised MSP and Mahalanobis backbones. MSP and Mahalanobis use supervised WideResNet-28-10 models on CIFAR-10, SVHN, and CIFAR-100 and ResNet-18 on Imagenette. Training lasts 200 epochs with SGD, momentum 0.9, Nesterov acceleration, initial learning rate 0.1, cosine annealing with $T_{\max} = 200$, and weight decay 5×10^{-4}. Batch size is 128 on CIFAR-10, SVHN, and CIFAR-100 and 64 on Imagenette. No pretrained weights are used.

For CIFAR-10 and CIFAR-100, training augmentation uses `RandomCrop(32,padding=4)` and horizontal flipping. SVHN uses the same crop without horizontal flipping. Imagenette uses resizing to 160×160 and horizontal flipping. Dataset-specific normalization is used for CIFAR-10, SVHN, and CIFAR-100, and ImageNet normalization for Imagenette.

The validation split is a fixed 10% subset of the original training data, constructed with `val_seed=42`. The checkpoint with highest validation accuracy is retained. MSP uses maximum softmax probability. Mahalanobis uses the supervised feature representation with class-specific means and a shared within-class covariance, and predicts the class with minimum Mahalanobis distance.

SelectiveNet. SelectiveNet is trained for 200 epochs using SGD with momentum 0.9, Nesterov acceleration, initial learning rate 0.1, cosine annealing, weight decay 5×10^{-4}, and batch size 128. The same dataset-specific augmentation and normalization protocol as the supervised baselines is used. Models are trained from scratch and selected by highest accuracy on the same fixed 10% validation split. Its native selection-head output $g(x)$ is used as the confidence score for AURC. Since larger values of $g(x)$ indicate greater confidence, the implementation evaluates AURC using $-g(x)$ only to match the internal "lower score = higher confidence" sorting convention; this sign change does not alter the induced ranking.

For MSP, Mahalanobis, and SelectiveNet, `val_seed=42` fixes only the train/validation partition. The original training scripts did not set a global PyTorch random seed, so weight initialization, minibatch ordering, and dropout masks are stochastic across independent retraining runs. Reported values are those obtained from the stored checkpoints used throughout the final evaluation. Accordingly, these results characterize the evaluated stored checkpoints rather than retraining variability.

SAT+EM+SR. SAT+EM+SR is trained for 300 epochs with SGD, momentum 0.9, initial learning rate 0.1, and weight decay 5×10^{-4}. The learning rate is multiplied by 0.5 every 25 epochs through epoch 275. Batch size is 128 and the training seed is 1. Entropy minimization uses weight

$$\beta_{\mathrm{EM}} = 0.001.$$

The SAT pretraining period is 150 epochs on CIFAR-10, 200 on CIFAR-100, and zero on SVHN and Imagenette. Imagenette uses input size 160. The reported native selective score is the softmax-response rule obtained after dropping the abstention logit and renormalizing over the class logits. Checkpoints are selected by highest accuracy on the held-out validation split.

CCL-SC. CCL-SC is trained for 300 epochs using SGD with momentum 0.9, initial learning rate 0.1, weight decay 5×10^{-4}, and batch size 64. The learning rate is multiplied by 0.5 every 25 epochs. The training seed is 1 and contrastive pretraining lasts 150 epochs for all datasets.

For CIFAR-10, SVHN, and Imagenette, the configuration uses $m_{\mathrm{MoCo}} = 0.999$, queue size 300, and reward weight 0.5. CIFAR-100 uses $m_{\mathrm{MoCo}} = 0.99$, queue size 3000, and reward weight 1.0. Imagenette uses input size 160. Native AURC uses softmax response from the classification head. The query-encoder checkpoint with highest held-out validation accuracy is retained.

MAPLE. MAPLE is trained for 200 epochs with Adam, learning rate 2×10^{-4}, batch size 64, no learning-rate schedule, weight decay zero, and training seed 1. The triplet margin is 0.7, with `val_epochs=10`, `fnr=0.3`, and at most five clusters per class. Imagenette uses input size 160.

MAPLE predicts from its learned pseudo-cluster representation using Mahalanobis distance, and its native `dvar` uncertainty score is used for AURC. The checkpoint with highest held-out validation accuracy is retained. The evaluated implementation is adapted from the MAPLE implementation of Venkataramanan et al. and corresponds to commit `39262b0`.

Table 4: Summary of baseline implementations and native scores used for the accuracy/AURC comparison. Structural scores used for RQ1 are described separately in Section D.5.

Method	Backbone / representation	Native selective score	Selection
MSP	WRN-28-10 / ResNet-18	maximum softmax probability	best validation accuracy
SelectiveNet	selective architecture	selection-head score	best validation accuracy
Mahalanobis	WRN-28-10 / ResNet-18 features	minimum class Mahalanobis distance	best validation accuracy
SAT+EM+SR	WRN-28-10 / ResNet-18	SAT+SR softmax response	best validation accuracy
CCL-SC	256-d learned embedding	softmax response	best validation accuracy
MAPLE	pseudo-cluster representation	native `dvar`	best validation accuracy
VMDLS	native metric latent space	native class-distance score	final epoch

VMDLS. VMDLS is trained for 200 epochs using Adam with learning rate 10^{-3} and batch size 80, formed as 10 classes with 8 observations per class. A StepLR schedule with step size 100 and factor 0.1 is used. The training seed is 1. The margin is 0.1 and the initial negative margin is 0.15.

The native latent dimension is 32 for CIFAR-10, SVHN, and Imagenette and 100 for CIFAR-100. CIFAR-10, CIFAR-100, and Imagenette use the default `margin_jumps=1.0` and `margins_epoch=0`. SVHN uses `margin_jumps=0.2` and `margins_epoch=50`. Imagenette uses input size 160.

Prediction uses the native latent-space class-distance rule and its associated distance as the native uncertainty score. Following the upstream convention, the final epoch checkpoint is used. The evaluated implementation is adapted from Dinari and Freifeld and corresponds to commit `5add9e2`.

D.5 Matched structural diagnostics

The structural columns of Table 1 answer a different question from native selective performance. They test whether a learned representation supports a distance coordinate exhibiting the two properties explicitly targeted by ADOQ:

1. numerical agreement with the prescribed χ^2_d scale;
2. ordering according to discriminative difficulty.

All class statistics, dimensionality reductions, and covariance estimates used for these diagnostics are fitted on training representations only. Validation and test observations are excluded from this fitting step.

Whenever applicable, diagnostics are reported both in the method's native dimension and at the ADOQ comparison dimension,

$$d_{\mathrm{match}} = \begin{cases} 16, & \text{CIFAR-10, SVHN, Imagenette,} \\ 100, & \text{CIFAR-100.} \end{cases}$$

Dimension matching is performed by PCA fitted on training representations only; the matched representation uses the first d_{match} components of that same fitted basis. The matched-dimensional control reduces confounding from latent dimension when comparing numerical fidelity and difficulty ordering across representations; it does not replace the method's native representation or selective score.

For a class-based representation $h(x)$, the generic diagnostic distance is

$$D_y(x) = (h(x) - \widehat{\mu}_y)^\top \, \widehat{\Sigma}_y^{-1} \, (h(x) - \widehat{\mu}_y) \,, \tag{204}$$

where $\widehat{\Sigma}_y$ denotes a class-specific covariance when used by the corresponding diagnostic. For methods using tied covariance estimation, the same expression is evaluated with a single shared covariance matrix $\widehat{\Sigma}$ in place of $\widehat{\Sigma}_y$. Empirical covariance matrices are regularized by adding the implementation ridge term `cov_eps` to the diagonal before inversion.

For the supervised Mahalanobis baseline, class means are estimated separately, while a single shared covariance matrix is obtained by pooling the within-class centered residuals across all classes.

The native representation has dimension 640 for WideResNet-28-10 and 512 for ResNet-18. The matched controls apply the same tied-covariance Mahalanobis construction after train-fitted PCA reduction to d_{match}.

For SAT+EM+SR and CCL-SC, the diagnostic operates on the learned 256-dimensional feature representation. Full empirical class-specific covariance matrices are fitted separately for each class. These Mahalanobis distances are post-hoc representation diagnostics: they are not the native scores used to obtain the AURC values in Table 1.

VMDLS is evaluated with its native Mahalanobis latent-space geometry, using class-specific full empirical covariance matrices. Its native dimension is 32 on the ten-class datasets and 100 on CIFAR-100. For the ten-class matched controls, the latent representation is reduced to 16 dimensions using the training-fitted PCA basis before recomputing the class-distance model.

MAPLE uses its pseudo-cluster Mahalanobis geometry, with a full empirical covariance matrix estimated separately for each pseudo-cluster. Its native diagnostic dimension is the dimension retained by its explained-variance PCA procedure. For matched controls, the first d_{match} components of the same fitted basis are used. Because multiple pseudo-clusters may represent one semantic class, the class-level distance is the minimum Mahalanobis distance over pseudo-clusters associated with that class.

For any resulting class-level coordinate, the reference diagnostic score is

$$S_{\mathrm{ref}}^{\mathrm{diag}} = D_Y(X), \tag{205}$$

and the operational diagnostic score is

$$S_{\mathrm{op}}^{\mathrm{diag}} = \min_y D_y(X). \tag{206}$$

The associated discriminative margin is

$$M^{\mathrm{diag}} = \min_{j \neq Y} D_j(X) - D_Y(X). \tag{207}$$

The reported structural quantities are the empirical KS discrepancies between the reference and operational diagnostic scores and the corresponding $\chi^2_{d_{\mathrm{diag}}}$ law, where d_{diag} is the dimension of the evaluated native or matched representation, together with their Spearman correlations with M^{diag}.

MSP and SelectiveNet are marked as not applicable for these structural columns. Although both networks contain internal feature representations, attaching a post-hoc class-distance model to them would introduce a new diagnostic not intrinsic to the compared method. Their comparison is therefore restricted to their native selective scores.

These structural scores are diagnostic only. Competing methods are neither assumed nor claimed to have been trained to follow a χ^2_d law; the diagnostics are used solely to compare whether their learned representations exhibit the numerical-scale and difficulty-ordering properties targeted by ADOQ.

D.6 Evaluation metrics

For RQ1, Table 1 reports classification accuracy using each method's native prediction rule and AURC using its native selective score.

For ADOQ, numerical fidelity to the prescribed coordinate is measured by

$$\widehat{\mathrm{KS}}_{\mathrm{ref}} = \sup_t \left| \widehat{F}_{\mathrm{ref}}(t) - F_{\chi^2_d}(t) \right|, \tag{208}$$

and

$$\widehat{\mathrm{KS}}_{\mathrm{op}} = \sup_t \left| \widehat{F}_{\mathrm{op}}(t) - F_{\chi^2_d}(t) \right|, \tag{209}$$

where $\widehat{F}_{\mathrm{ref}}$ is the empirical CDF of $S_{\mathrm{ref}} = T_Y$ and $\widehat{F}_{\mathrm{op}}$ is the empirical CDF of $S_{\mathrm{op}} = T_{\widehat{Y}}$.

Difficulty ordering is measured by Spearman correlation between the true-class margin

$$M = \min_{j \neq Y} T_j - T_Y \tag{210}$$

and, respectively, S_{ref} and S_{op}. A more negative correlation corresponds to stronger alignment between increasing discriminative margin and decreasing analytical score.

D.7 Local analytical-window protocol

RQ2 evaluates the frozen operational analytical coordinate

$$R(X) = F_{\chi^2_d}(S_{\mathrm{op}}(X)) \in [0, 1) \tag{211}$$

through sliding windows

$$\mathcal{W}_\rho = \mathcal{S}_{\rho-0.1,\rho}. \tag{212}$$

The window width is fixed to $\Delta\rho = 0.1$, and the upper boundary is evaluated on $\rho \in \{0.10, 0.11, \ldots, .99\}$, giving 90 interior windows. Together with the outermost window $\mathcal{T}_{.9}$, this gives 91 predefined overlapping windows in total.

For a test / certification sample of size n, the empirical occupancy of a window is

$$\widehat{C}_{\mathcal{W}}(\rho) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[X_i \in \mathcal{W}_\rho\right]. \tag{213}$$

Let

$$N_\rho = \sum_{i=1}^{n} \mathbf{1}\left[X_i \in \mathcal{W}_\rho\right] \tag{214}$$

denote the number of observations in the window, and let

$$K_\rho = \sum_{i=1}^{n} \mathbf{1}\left[X_i \in \mathcal{W}_\rho,\, \widehat{Y}_i = Y_i\right]. \tag{215}$$

The empirical local reliability is

$$\widehat{A}_{\mathcal{W}}(\rho) = \frac{K_\rho}{N_\rho} \tag{216}$$

when $N_\rho > 0$. The formulas above define the statistics for the 90 interior windows. For the outermost tail $\mathcal{T}_{.9}$, the same definitions are used with $\mathcal{W}_\rho$ replaced by $\mathcal{T}_{.9}$.

D.8 Certification settings

The certification results use the held-out test split only after all training and model-selection decisions have been frozen, so the certification sample is independent of model construction and selection. The assumptions and finite-sample guarantees are detailed in Appendix C.

Unless otherwise stated, occupancy certification uses

$$\delta_{\mathrm{occ}} = 0.05. \tag{217}$$

For the sliding local windows, the DKW lower bound displayed in Figure 1 is

$$C_{\mathrm{LB}}(\rho) = \max\left\{0,\, \widehat{C}_{\mathcal{W}}(\rho) - 2\epsilon_n(\delta_{\mathrm{occ}})\right\}, \tag{218}$$

where

$$\epsilon_n(\delta) = \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{219}$$

Because this bound follows from a single uniform empirical-CDF event, the same 95% statement applies simultaneously to the complete 91-region family without a multiplicity correction. The first and final windows, $\mathcal{S}_{0,.1}$ and $\mathcal{T}_{.9}$, admit the tighter one-endpoint DKW bound $\epsilon_n(\delta_{\mathrm{occ}})$. For consistency across the plotted family, Figure 1 uses the conservative $2\epsilon_n(\delta_{\mathrm{occ}})$ bound for these endpoint windows as well.

Predictive-reliability bounds are computed from labeled observations using the one-sided exact Clopper–Pearson construction. For the complete family of $M_{\mathrm{win}} = 91$ analytical regions, simultaneous 95% reliability bounds use Bonferroni correction,

$$\delta_{\mathrm{win}} = \frac{0.05}{91}. \tag{220}$$

No independence between overlapping windows is required for this Bonferroni family-wise guarantee. Thus each plotted lower bound is

$$A_{\mathrm{LB}}(\rho) = F^{-1}_{\mathrm{Beta}(K_\rho, N_\rho - K_\rho + 1)}\left(\frac{0.05}{91}\right) \tag{221}$$

for $K_\rho > 0$, and is set to zero otherwise.

The 95% DKW statement and the 95% family-wise Clopper–Pearson statement refer to the occupancy and reliability families separately. No joint 95% probability statement over both families is claimed.

Appendix F additionally reports statistics for the non-overlapping analytical regions induced by the predefined grid

$$\mathcal{G} = \{0, .01, .10, .25, .50, .75, .90, .95, .98, .99, .995, .999\}, \tag{222}$$

which was fixed as part of the evaluation protocol before computing the region-wise and cumulative certification statistics: eleven interior shells $\mathcal{S}_{\rho_j,\rho_{j+1}}$ for consecutive $\rho_j, \rho_{j+1} \in \mathcal{G}$, together with the outermost tail $\mathcal{T}_{.999}$. The appendix also reports cumulative regions $\mathcal{S}_{0,\rho}$ at the eleven interior grid levels.

E Distributional and Ordering Diagnostics

This appendix provides supplementary structural diagnostics underlying RQ1. Native classification accuracy and AURC are reported in Table 1. The focus here is on the two structural properties targeted by ADOQ: fidelity to the prescribed numerical score scale and ordering according to discriminative difficulty.

For competing representations, the diagnostic construction is described in Appendix D.5. These scores are used only to analyze representation structure and do not replace the native confidence or uncertainty scores used to compute each method's AURC.

E.1 Matched-dimensional structural controls

Good selective ranking does not by itself imply either a prescribed numerical score scale or a strong relationship with discriminative difficulty. Accordingly, whenever applicable, matched-dimensional distance coordinates are evaluated from alternative learned representations.

For CIFAR-10, SVHN, and Imagenette, the comparison dimension is $d = 16$; for CIFAR-100, it is $d = 100$. The reference diagnostic score is the distance to the true class, whereas the operational diagnostic score is the minimum class distance. Numerical fidelity is quantified by KS discrepancy from the corresponding χ^2_d law, and difficulty ordering by Spearman correlation with the true-class discriminative margin. Matching the representation dimension reduces confounding from latent dimension when comparing these structural properties. The ADOQ values correspond to the same selected models used in the main experiments.

The matched-dimensional controls confirm that selective ranking, prescribed-scale fidelity, and difficulty ordering capture distinct properties of a learned representation. On CIFAR-10, SVHN, and Imagenette, ADOQ exhibits both the smallest reference and operational KS discrepancies and the most negative reference and operational margin–score correlations among the evaluated representations.

The distinction is particularly pronounced on SVHN. Mahalanobis obtains $\widehat{\mathrm{KS}}_{\mathrm{ref/op}} = .2654/.2667$ and $\rho_S^{\mathrm{ref/op}} = -.3956/-.3862$, compared with $.0469/.0322$ and $-.9893/-.9745$ for ADOQ. The other evaluated representations likewise remain substantially farther from the prescribed numerical scale and exhibit weaker difficulty ordering, despite several of them achieving competitive native selective performance in Table 1.

CIFAR-100 provides a complementary case. ADOQ retains the smallest reference and operational KS discrepancies and the strongest reference difficulty ordering, whereas Mahalanobis attains a slightly stronger operational Spearman correlation ($-.8644$ versus $-.8529$). Thus strong difficulty

Table 5: Matched structural diagnostics underlying RQ1. Lower KS indicates closer agreement with the prescribed numerical scale; more negative Spearman correlation indicates stronger alignment between smaller scores and larger discriminative margins. Best values within each dataset are shown in bold.

Dataset	Method	$\widehat{\mathrm{KS}}_{\mathrm{ref}}$	$\widehat{\mathrm{KS}}_{\mathrm{op}}$	ρ_S^{ref}	ρ_S^{op}
CIFAR-10	Mahalanobis	.2741	.2726	-.6751	-.6698
	SAT+EM+SR	.3326	.3328	-.5294	-.5216
	CCL-SC	.2917	.2920	-.4056	-.3973
	MAPLE	.2559	.2578	-.4812	-.4608
	VMDLS	.2837	.2869	-.2852	-.2552
	ADOQ	**.0977**	**.0999**	**-.9637**	**-.9462**
SVHN	Mahalanobis	.2654	.2667	-.3956	-.3862
	SAT+EM+SR	.4498	.4516	-.3437	-.3330
	CCL-SC	.3423	.3437	-.3167	-.3055
	MAPLE	.2676	.2698	-.4284	-.4125
	VMDLS	.4389	.4418	-.2100	-.1925
	ADOQ	**.0469**	**.0322**	**-.9893**	**-.9745**
Imagenette	Mahalanobis	.6809	.6804	-.9355	-.8999
	SAT+EM+SR	.4952	.4885	-.8884	-.8225
	CCL-SC	.5785	.5752	-.8562	-.7850
	MAPLE	.5401	.5276	-.9191	-.8056
	VMDLS	.5161	.5114	-.7329	-.6694
	ADOQ	**.1772**	**.1537**	**-.9862**	**-.9469**
CIFAR-100	Mahalanobis	.6514	.6474	-.9384	**-.8644**
	SAT+EM+SR	.6726	.6664	-.8632	-.7614
	CCL-SC	.6683	.6612	-.8372	-.7241
	MAPLE	.5319	.5071	-.9140	-.7812
	VMDLS	.5706	.5301	-.5838	-.3969
	ADOQ	**.3900**	**.3574**	**-.9645**	-.8529

ordering does not imply fidelity to the prescribed numerical coordinate, while numerical fidelity alone would not guarantee difficulty ordering. The two criteria are therefore evaluated separately and considered jointly.

These comparisons are structural diagnostics only. Competing methods are neither assumed nor claimed to have been trained to follow a χ^2_d law; the common diagnostics assess whether their learned representations exhibit the numerical-scale and difficulty-ordering properties explicitly targeted by ADOQ.

E.2 Visual score-law diagnostics

To visualize the distributional discrepancies summarized by the KS statistics, the ADOQ reference and operational scores are transformed through the prescribed χ^2_d CDF:

$$R_{\mathrm{ref}} = F_{\chi^2_d}(S_{\mathrm{ref}}), \qquad R_{\mathrm{op}} = F_{\chi^2_d}(S_{\mathrm{op}}). \tag{223}$$

For finite scores, both transformed coordinates lie in $[0, 1)$. If the corresponding score follows the prescribed χ^2_d law, its transformed coordinate follows a $\mathrm{Uniform}(0, 1)$ distribution, whose CDF is the diagonal on $[0, 1]$. These transformations also place all datasets on the same normalized scale despite their different latent dimensions.

Figure 2 shows that SVHN exhibits the closest agreement with the prescribed scale, followed by CIFAR-10, while Imagenette shows a larger systematic deviation and CIFAR-100 departs substantially from the uniform reference. These visual patterns agree with the corresponding ADOQ KS diagnostics in Table 5.

Importantly, the increasing score-law discrepancy across these datasets does not coincide with a comparable loss of difficulty ordering. In particular, CIFAR-100 exhibits substantial reference and operational distributional mismatch while retaining strongly negative margin–score correlations. To-

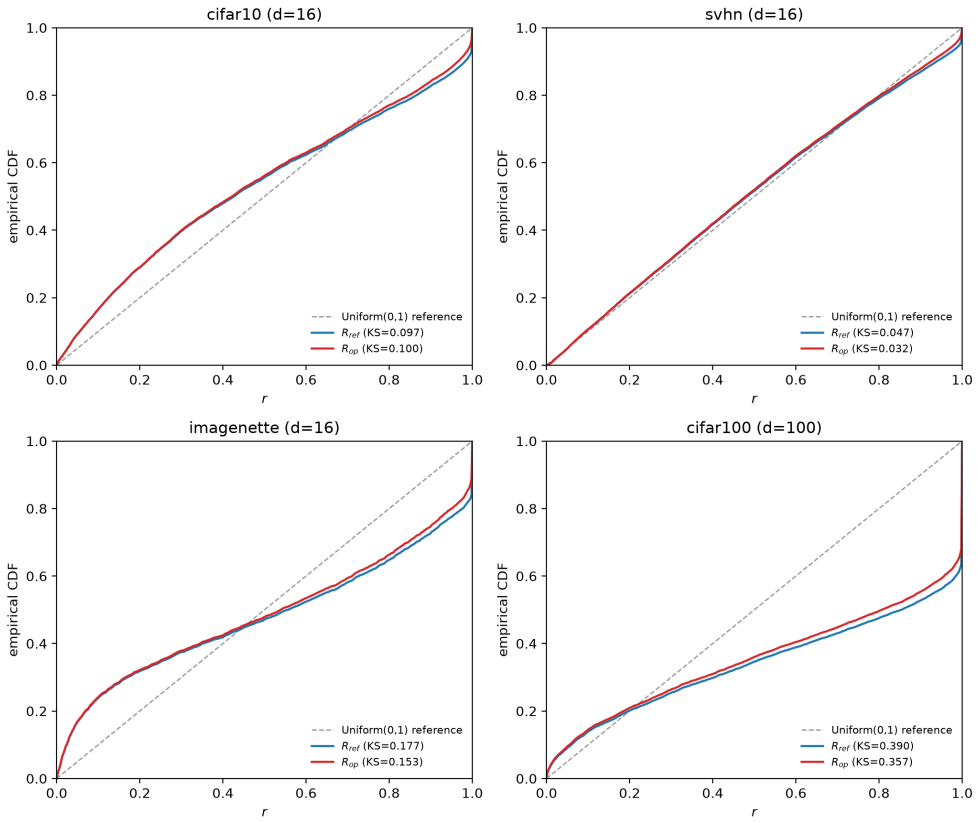


Figure 2: Visual diagnostics of prescribed score-law fidelity. Empirical CDFs of the transformed ADOQ reference and operational coordinates $R_{\mathrm{ref}} = F_{\chi^2_d}(S_{\mathrm{ref}})$ and $R_{\mathrm{op}} = F_{\chi^2_d}(S_{\mathrm{op}})$ are compared with the $\mathrm{Uniform}(0,1)$ reference. Exact realization of the prescribed χ^2_d law for the corresponding score would yield the diagonal. The displayed KS values quantify the maximum vertical discrepancy from this reference.

gether with the matched-dimensional controls, this confirms that numerical fidelity to the prescribed coordinate and organization of discriminative difficulty along that coordinate are related but distinct structural properties.

E.3 Class-conditional fidelity of the ADOQ reference score

ADOQ prescribes the same χ^2_d reference CDF for the normalized true-class score within each class. A pooled empirical KS discrepancy can therefore conceal stronger deviations within individual classes. For each class y, define

$$\widehat{\mathrm{KS}}_y^{\mathrm{ref}} = \sup_t \left| \widehat{F}_{T_y|Y=y}(t) - F_{\chi^2_d}(t) \right|. \tag{224}$$

Table 6 reports the mean and maximum of these classwise empirical discrepancies.

These diagnostics further distinguish exact realization of the prescribed reference scale from approximate empirical fidelity. SVHN exhibits the closest class-conditional agreement, followed by CIFAR-10. Imagenette and CIFAR-100 show substantially greater heterogeneity across classes, including individual classes with large discrepancies from the prescribed reference CDF.

The mean and maximum values are descriptive empirical diagnostics rather than certification bounds. In particular, classwise departures from the prescribed reference law do not alter the analytical region construction, whose boundaries remain fixed by the prescribed coordinate, nor do they invalidate the subsequent certification procedure. They instead show why nominal analytical

Table 6: Class-conditional reference-score fidelity for the selected ADOQ checkpoints. The table reports the mean and maximum empirical KS discrepancy across classes.

Dataset	Mean classwise KS	Maximum classwise KS
CIFAR-10	.1511	.2332
SVHN	.0928	.1892
Imagenette	.4100	.6121
CIFAR-100	.4096	.7538

mass should not be assumed to equal realized operational occupancy. RQ2 therefore measures and independently certifies the realized occupancy and predictive reliability of predefined regions rather than relying on nominal reference behavior.

F Analytical-Region Statistics

This appendix complements the local sliding-window analysis of Section 4.3 with tabulated statistics for two predefined analytical-region families.

First, using the reporting grid fixed in the evaluation protocol (Appendix D.8),

$$\mathcal{G} = \{0, .01, .10, .25, .50, .75, .90, .95, .98, .99, .995, .999\}, \tag{225}$$

the analytical coordinate is partitioned into eleven successive interior shells $\mathcal{S}_{\rho_j,\rho_{j+1}}$ for consecutive $\rho_j, \rho_{j+1} \in \mathcal{G}$, together with the outermost tail $\mathcal{T}_{.999}$, giving twelve regions in total. These shells provide a complementary view to the fixed-width $\Delta\rho = 0.1$ sliding windows shown in Figure 1. The progressively finer partition as $\rho \to 1^-$ provides additional resolution of the extreme outer tail.

Second, cumulative regions

$$\mathcal{S}_{0,\rho} = \{x \mid 0 \le R(x) \le \rho\} \tag{226}$$

are reported at the eleven interior grid levels

$$\rho \in \{.01, .10, .25, .50, .75, .90, .95, .98, .99, .995, .999\}. \tag{227}$$

For every region, $\widehat{C}$ denotes empirical operational occupancy and $\widehat{A}$ empirical predictive reliability. C_{LB} denotes the distribution-free DKW lower bound derived in Appendix C, and A_{LB} the one-sided exact Clopper–Pearson lower confidence bound.

All occupancy bounds use $\delta_{\mathrm{occ}} = 0.05$ and follow from the same uniform DKW event. Interior shells use $\max\{0, \widehat{C} - 2\epsilon_n(\delta_{\mathrm{occ}})\}$. For the two endpoint regions, $\mathcal{S}_{0,.01}$ and $\mathcal{T}_{.999}$, only one empirical-CDF endpoint is uncertain, so the tighter bound $\max\{0, \widehat{C} - \epsilon_n(\delta_{\mathrm{occ}})\}$ is used. Cumulative regions likewise use the one-endpoint DKW bound. A zero occupancy lower bound for a narrow shell therefore reflects finite-sample conservatism of the uniform DKW guarantee, not zero empirical occupancy.

Reliability bounds for these region tables are simultaneous at the 95% family-wise level using Bonferroni correction over the twelve reported regions. For cumulative tables, the correction is applied over the eleven reported levels. These reliability bounds differ from those displayed in Figure 1, where simultaneous reliability is required over all 91 overlapping sliding windows. All analytical boundaries are fixed independently of the certification observations.

F.1 CIFAR-10

Table 7 shows that CIFAR-10 exhibits moderate redistribution of empirical occupancy relative to the prescribed coordinate together with a clear reliability gradient: shellwise empirical reliability remains above 95% up to $\rho = .90$, then decreases progressively through the refined outer-tail shells, reaching 54.06% in $\mathcal{T}_{.999}$.

Cumulatively, Table 8 shows that $\mathcal{S}_{0,.99}$ contains 93.57% of the test observations at 97.40% empirical reliability, with a simultaneous lower reliability bound of 96.94%.

Table 7: CIFAR-10 statistics for the non-overlapping analytical regions induced by the predefined grid $\mathcal{G}$. Reliability bounds are simultaneous over the twelve reported regions.

Region	$\widehat{C}$	C_{LB}	$\widehat{A}$	A_{LB}
$[0, .01]$	2.20%	0.84%	100.00%	97.54%
$(.01, .10]$	14.34%	11.62%	99.72%	99.11%
$(.10, .25]$	18.15%	15.43%	99.39%	98.73%
$(.25, .50]$	21.47%	18.75%	98.28%	97.39%
$(.50, .75]$	17.40%	14.68%	97.59%	96.44%
$(.75, .90]$	10.59%	7.87%	95.47%	93.51%
$(.90, .95]$	4.15%	1.43%	92.29%	88.18%
$(.95, .98]$	3.55%	0.83%	87.61%	82.32%
$(.98, .99]$	1.72%	0.00%	85.47%	77.07%
$(.99, .995]$	1.26%	0.00%	73.02%	61.41%
$(.995, .999]$	1.97%	0.00%	69.04%	59.68%
$(.999, 1)$	3.20%	1.84%	54.06%	46.54%

Table 8: CIFAR-10 cumulative operational statistics. Reliability bounds are simultaneous over the eleven reported analytical levels.

ρ	$\widehat{C}(0, \rho)$	C_{LB}	$\widehat{A}(0, \rho)$	A_{LB}
.010	2.20%	0.84%	100.00%	97.58%
.100	16.54%	15.18%	99.76%	99.23%
.250	34.69%	33.33%	99.57%	99.18%
.500	56.16%	54.80%	99.07%	98.69%
.750	73.56%	72.20%	98.72%	98.34%
.900	84.15%	82.79%	98.31%	97.91%
.950	88.30%	86.94%	98.03%	97.61%
.980	91.85%	90.49%	97.63%	97.18%
.990	93.57%	92.21%	97.40%	96.94%
.995	94.83%	93.47%	97.08%	96.60%
.999	96.80%	95.44%	96.51%	95.99%

F.2 SVHN

Table 9 shows that SVHN remains closest to the prescribed empirical occupancy profile and exhibits very high reliability throughout most of the analytical coordinate: all shells below $\rho = .90$ retain empirical reliability above 97%, and the finer outer-tail partition reveals a progressive degradation beyond $\rho = .90$, with empirical reliability falling to 47.16% in $\mathcal{T}_{.999}$.

At $\rho = .99$, Table 10 shows that the cumulative region contains 96.13% of the test observations at 98.23% empirical reliability, with a simultaneous lower reliability bound of 98.00%.

F.3 IMAGENETTE

Table 11 shows that Imagenette exhibits substantially stronger redistribution of empirical occupancy than CIFAR-10 or SVHN. The refined partition also reveals a clear reliability gradient: empirical shellwise reliability remains above 94% through $\mathcal{S}_{.75,.90}$, then decreases progressively toward the extreme outer tail, reaching 40.10% in $\mathcal{T}_{.999}$. This final shell alone contains 10.29% of the test observations, despite having nominal analytical mass 0.1% under exact realization of the prescribed score law.

Nevertheless, Table 12 shows that the cumulative region $\mathcal{S}_{0,.99}$ still contains 85.66% of the test observations at 95.81% empirical reliability, with a simultaneous reliability lower bound of 94.82%. Thus substantial redistribution of occupancy does not prevent the analytical coordinate from separating a large, highly reliable cumulative region from a markedly less reliable outer tail.

Table 9: SVHN statistics for the non-overlapping analytical regions induced by the predefined grid $\mathcal{G}$. Reliability bounds are simultaneous over the twelve reported regions.

Region	$\widehat{C}$	C_{LB}	$\widehat{A}$	A_{LB}
$[0, .01]$	0.60%	0.00%	99.36%	95.21%
$(.01, .10]$	10.05%	8.36%	99.35%	98.81%
$(.10, .25]$	15.87%	14.19%	99.32%	98.91%
$(.25, .50]$	25.47%	23.79%	99.31%	98.99%
$(.50, .75]$	23.42%	21.74%	98.97%	98.57%
$(.75, .90]$	12.42%	10.74%	97.31%	96.47%
$(.90, .95]$	4.21%	2.52%	94.25%	92.13%
$(.95, .98]$	2.82%	1.14%	88.83%	85.42%
$(.98, .99]$	1.28%	0.00%	82.83%	76.74%
$(.99, .995]$	0.92%	0.00%	76.99%	69.07%
$(.995, .999]$	1.60%	0.00%	69.78%	63.52%
$(.999, 1)$	1.35%	0.51%	47.16%	40.08%

Table 10: SVHN cumulative operational statistics. Reliability bounds are simultaneous over the eleven reported analytical levels.

ρ	$\widehat{C}(0, \rho)$	C_{LB}	$\widehat{A}(0, \rho)$	A_{LB}
.010	0.60%	0.00%	99.36%	95.27%
.100	10.64%	9.80%	99.35%	98.84%
.250	26.51%	25.67%	99.33%	99.03%
.500	51.98%	51.14%	99.32%	99.11%
.750	75.40%	74.56%	99.21%	99.03%
.900	87.83%	86.98%	98.94%	98.75%
.950	92.03%	91.19%	98.73%	98.53%
.980	94.85%	94.01%	98.43%	98.21%
.990	96.13%	95.29%	98.23%	98.00%
.995	97.05%	96.20%	98.02%	97.78%
.999	98.65%	97.81%	97.57%	97.30%

F.4 CIFAR-100

Table 13 shows that CIFAR-100 provides the clearest illustration of the distinction between nominal analytical position, realized occupancy, and predictive reliability: the extreme outer tail $\mathcal{T}_{.999}$, whose nominal mass is only 0.1% under exact realization of the prescribed score law, contains 30.67% of the test observations, and its empirical reliability falls to 43.66%, with a simultaneous lower bound of 41.29%.

Table 14 quantifies the resulting reliably usable portion of the classifier: the cumulative region $\mathcal{S}_{0,.999}$ contains 69.33% of the test observations at 93.05% empirical reliability, with a simultaneous lower bound of 92.21%; at $\rho = .90$, $\mathcal{S}_{0,.90}$ has empirical occupancy 55.35% and reliability 95.32% (lower bound 94.53%); and at $\rho = .99$, the corresponding values are 64.73%, 94.04%, and 93.23%. Thus, despite an empirical global classification accuracy of 77.90%, substantial predefined regions retain independently certified population-level reliability well above the classifier's observed global accuracy.

Across the four datasets, the shell-wise and cumulative statistics reported in this appendix show a consistent pattern: predictive reliability degrades progressively toward the extreme outer tail, most mildly on SVHN and most severely on CIFAR-100, while a large cumulative region below $\rho \approx .99$ remains substantially more reliable than the classifier's global accuracy in every case.

G Additional Experimental Controls

This appendix reports additional controls supporting the experimental analysis. It first contrasts the ADOQ operating mechanism with a representative calibration-based selective procedure. It

Table 11: Imagenette statistics for the non-overlapping analytical regions induced by the predefined grid $\mathcal{G}$. Reliability bounds are simultaneous over the twelve reported regions.

Region	$\widehat{C}$	C_{LB}	$\widehat{A}$	A_{LB}
$[0, .01]$	4.69%	2.52%	99.46%	95.92%
$(.01, .10]$	19.36%	15.03%	98.82%	97.35%
$(.10, .25]$	10.93%	6.59%	98.83%	96.68%
$(.25, .50]$	12.92%	8.58%	97.04%	94.45%
$(.50, .75]$	14.78%	10.44%	95.86%	93.17%
$(.75, .90]$	12.08%	7.74%	94.51%	91.15%
$(.90, .95]$	5.17%	0.84%	92.12%	85.79%
$(.95, .98]$	3.41%	0.00%	83.58%	73.56%
$(.98, .99]$	2.32%	0.00%	74.73%	61.01%
$(.99, .995]$	1.32%	0.00%	63.46%	44.28%
$(.995, .999]$	2.73%	0.00%	57.01%	43.83%
$(.999, 1)$	10.29%	8.13%	40.10%	33.70%

Table 12: Imagenette cumulative operational statistics. Reliability bounds are simultaneous over the eleven reported analytical levels.

ρ	$\widehat{C}(0, \rho)$	C_{LB}	$\widehat{A}(0, \rho)$	A_{LB}
.010	4.69%	2.52%	99.46%	95.98%
.100	24.05%	21.88%	98.94%	97.73%
.250	34.98%	32.81%	98.91%	97.94%
.500	47.90%	45.73%	98.40%	97.49%
.750	62.68%	60.51%	97.80%	96.91%
.900	74.75%	72.58%	97.27%	96.39%
.950	79.92%	77.76%	96.94%	96.05%
.980	83.34%	81.17%	96.39%	95.45%
.990	85.66%	83.49%	95.81%	94.82%
.995	86.98%	84.81%	95.31%	94.29%
.999	89.71%	87.54%	94.15%	93.04%

then verifies numerically that the fixed ADOQ class geometry satisfies its prescribed construction, examines the learned latent representations through complementary visual diagnostics, evaluates the effect of matched latent dimension on competing structural diagnostics, and finally assesses the sensitivity of ADOQ itself to the choice of latent dimension.

G.1 Comparison with calibration-based risk control

ADOQ and calibration-based selective procedures address related but distinct operating scenarios. To make this distinction concrete, Selection with Guaranteed Risk (SGR) is used as a representative calibration-based procedure.

SGR starts from a requested target risk $r^\star$ and uses labeled calibration data to select a data-dependent operating threshold,

$$(r^\star, \mathcal{D}_{\mathrm{cal}}) \longmapsto \widehat{\tau}(r^\star, \mathcal{D}_{\mathrm{cal}}). \tag{228}$$

The resulting retained region therefore depends explicitly on both the target risk and the available calibration evidence.

ADOQ instead starts from a nominal analytical level and fixes

$$\rho \longmapsto \tau_\rho \longmapsto \mathcal{S}_{0,\rho}, \tag{229}$$

where τ_ρ follows the analytical-threshold convention of Section 3.1. This mapping is fixed before certification observations are examined. Independent certification data are subsequently used to quantify the realized occupancy and predictive reliability of this already-defined region without modifying its boundary.

Table 13: CIFAR-100 statistics for the non-overlapping analytical regions induced by the predefined grid $\mathcal{G}$. Reliability bounds are simultaneous over the twelve reported regions.

Region	$\widehat{C}$	C_{LB}	$\widehat{A}$	A_{LB}
$[0, .01]$	4.06%	2.70%	94.33%	90.61%
$(.01, .10]$	10.35%	7.63%	96.71%	94.96%
$(.10, .25]$	9.13%	6.41%	96.60%	94.70%
$(.25, .50]$	12.48%	9.76%	95.83%	94.11%
$(.50, .75]$	11.19%	8.47%	95.35%	93.44%
$(.75, .90]$	8.14%	5.42%	91.77%	88.90%
$(.90, .95]$	3.93%	1.21%	87.02%	81.95%
$(.95, .98]$	3.55%	0.83%	86.76%	81.36%
$(.98, .99]$	1.90%	0.00%	84.74%	76.69%
$(.99, .995]$	1.73%	0.00%	82.08%	73.21%
$(.995, .999]$	2.87%	0.15%	77.35%	70.21%
$(.999, 1)$	30.67%	29.31%	43.66%	41.29%

Table 14: CIFAR-100 cumulative operational statistics. Reliability bounds are simultaneous over the eleven reported analytical levels.

ρ	$\widehat{C}(0, \rho)$	C_{LB}	$\widehat{A}(0, \rho)$	A_{LB}
.010	4.06%	2.70%	94.33%	90.65%
.100	14.41%	13.05%	96.04%	94.51%
.250	23.54%	22.18%	96.26%	95.12%
.500	36.02%	34.66%	96.11%	95.19%
.750	47.21%	45.85%	95.93%	95.12%
.900	55.35%	53.99%	95.32%	94.53%
.950	59.28%	57.92%	94.77%	93.97%
.980	62.83%	61.47%	94.32%	93.51%
.990	64.73%	63.37%	94.04%	93.23%
.995	66.46%	65.10%	93.73%	92.91%
.999	69.33%	67.97%	93.05%	92.21%

For SGR, softmax response is used as the confidence score and $\delta = 0.001$. The purpose of this control is not to rank SGR and ADOQ, but to illustrate the distinction between calibration-dependent operating-region construction and certification of predefined analytical regions.

Table 15 reports representative SGR operating points. For each requested risk, the threshold is selected from the labeled calibration partition and then evaluated unchanged on the untouched test set.

The reported operating points illustrate the expected calibration-dependent trade-off. As the requested risk is relaxed, SGR selects a less restrictive threshold and retains a larger fraction of the observations. For example, on CIFAR-10 the threshold decreases from .9969 at $r^\star = .01$ to .2924 at $r^\star = .05$, while calibration coverage increases from 85.42% to 99.98%. On CIFAR-100, the threshold decreases from .9977 to .9053 as $r^\star$ increases from .02 to .06, with calibration coverage increasing from 33.52% to 66.34%.

This behavior reflects a fundamental difference in operating mechanism. Calibration data participate directly in defining the SGR acceptance region. In ADOQ, by contrast, analytical boundaries are fixed once training and model selection are complete. Certification observations may show that realized occupancy or reliability differs from its nominal interpretation, but they do not redefine the analytical region.

The confidence levels used in the two procedures are intentionally different: SGR uses $\delta = .001$, whereas the ADOQ certification experiments use $\delta = .05$ unless otherwise stated. Confidence-bound widths should therefore not be compared directly. The numerical results in Table 15 are included to illustrate the different operating mechanisms rather than to establish a performance ordering between the two procedures.

Table 15: Representative SGR operating points. $r^\star$ is the requested risk and $b^\star$ the corresponding SGR upper risk bound computed from the labeled calibration sample. The selected threshold is subsequently evaluated unchanged on the test set.

Dataset	$r^\star$	Threshold	Calib. cov.	$b^\star$	Test cov.	Test risk
CIFAR-10	.01	.9969	85.42%	0.97%	84.24%	0.49%
	.03	.8724	96.34%	3.02%	95.73%	2.07%
	.05	.2924	99.98%	4.86%	100.00%	3.80%
SVHN	.02	.9702	95.18%	2.00%	95.89%	1.12%
	.03	.8411	97.56%	2.99%	97.94%	1.64%
	.04	.4352	99.88%	3.99%	99.87%	2.52%
Imagenette	.04	.9942	51.69%	3.94%	52.13%	0.64%
	.05	.9917	56.03%	4.86%	56.08%	0.95%
	.06	.9721	65.22%	5.86%	65.68%	2.17%
CIFAR-100	.02	.9977	33.52%	1.94%	35.11%	1.25%
	.04	.9698	56.08%	3.98%	56.49%	2.99%
	.06	.9053	66.34%	5.97%	65.56%	4.82%

Table 16: Numerical checks of the frozen ADOQ center geometry. Pairwise distances coincide up to numerical precision and $\Delta_{\mathrm{geom}} > 0$ for all four datasets.

Quantity	CIFAR-10	SVHN	Imagenette	CIFAR-100
K	10	10	10	100
d	16	16	16	100
$\sigma_{\min}$	.8723	.9807	.9118	.6971
$\sigma_{\max}$	1.0960	1.0170	1.1527	1.2071
Centroid norm	1.70×10^{-7}	9.41×10^{-8}	2.26×10^{-7}	3.88×10^{-7}
Pairwise dist. min	13.7332	12.7435	14.4434	29.5132
Pairwise dist. max	13.7332	12.7435	14.4434	29.5132
Δ_{geom}	.0259	.0122	.3291	.1329

G.2 Numerical verification of the fixed ADOQ geometry

The class geometry is fixed analytically before encoder optimization. As a sanity check, the geometry stored with each selected ADOQ model is verified numerically against the construction of Appendix A.

For each class pair, define the separation margin

$$\Delta_{\mathrm{geom}} = \min_{y \neq y'} \left[\|C_y - C_{y'}\|_2 - \left(r_y^{\mathrm{sep}} + r_{y'}^{\mathrm{sep}} \right) \right], \tag{230}$$

where

$$r_y^{\mathrm{sep}} = \sigma_y \sqrt{\tau_{\rho_{\mathrm{sep}}}}. \tag{231}$$

A nonnegative value confirms that the interiors of the corresponding geometry-level class balls are pairwise disjoint.

Table 16 shows that, for all four datasets, the centroid is numerically indistinguishable from the origin and the minimum and maximum pairwise center distances coincide up to floating-point precision. The stored centers therefore realize the intended regular-simplex geometry.

Moreover, $\Delta_{\mathrm{geom}} > 0$ in every case, confirming that the geometry-level class-ball interiors are pairwise disjoint at $\rho_{\mathrm{sep}} = .999$.

CIFAR-100 also provides a useful large-class verification. With $K = 100$ and $d = 100$, the condition $d \geq K - 1$ is satisfied, while the stored geometry retains a near-zero centroid, constant pairwise distance, and positive separation margin. Its larger score-law discrepancy is therefore not explained by an implementation failure of the prescribed center geometry.

G.3 LATENT-SPACE VISUAL DIAGNOSTICS

The numerical diagnostics are complemented by two visualizations of the learned latent representations. The first provides a metric-aware view through an origin-preserving orthogonal projection of normalized residual coordinates. The second uses UMAP solely as a qualitative visualization of local class-neighborhood organization and is not interpreted metrically.

Normalized-residual principal-subspace projection. For each observation, the latent representation is recentered on its predicted class center and normalized by the corresponding class scale:

$$w(x) = \frac{f_\theta(x) - C_{\hat{y}(x)}}{\sigma_{\hat{y}(x)}}. \tag{232}$$

By construction,

$$\|w(x)\|_2^2 = S_{\mathrm{op}}(x). \tag{233}$$

All predicted classes therefore share the same analytical radial coordinate in this normalized residual space. For any analytical level $\rho \in [0, 1)$,

$$R(x) \le \rho \iff \|w(x)\|_2 \le \sqrt{\tau_\rho}. \tag{234}$$

To obtain a two-dimensional metric-aware view, the two leading principal directions of the normalized residuals are computed using an origin-preserving decomposition, without subtracting their empirical mean. Let $P_2 \in \mathbb{R}^{d\times 2}$ denote the resulting orthonormal basis. The displayed coordinates are

$$u(x) = P_2^\top w(x). \tag{235}$$

Because the projection is orthogonal and origin preserving,

$$\|u(x)\|_2 \le \|w(x)\|_2. \tag{236}$$

Consequently, every observation belonging to $\mathcal{S}_{0,\rho}$ in the full latent space must appear inside the circle of radius $\sqrt{\tau_\rho}$ in the projected view. The reverse implication does not hold, however: a point that appears inside the projected circle need not truly belong to $\mathcal{S}_{0,\rho}$, since a full-space exterior observation can have its norm reduced by the projection and land inside the circle anyway.

Figure 3 shows this projection for $\rho = .90$. Green and red points denote correct and misclassified predictions, respectively. The displayed circle is a projected visual reference for the full-dimensional analytical boundary and should not be interpreted as a two-dimensional decision boundary.

The visualization is consistent with the reliability stratification reported in RQ2. Correct predictions dominate much of the inner normalized residual space, while misclassified observations become increasingly visible at larger analytical radii. This pattern is especially pronounced for CIFAR-100, consistently with its lower global accuracy and substantially larger empirical occupancy in the extreme outer analytical region. The visualization remains descriptive: all analytical-region membership and certification results in the paper are computed in the original latent dimension rather than from the projected coordinates.

Qualitative neighborhood visualization. Figure 4 provides a complementary UMAP visualization of the learned latent representations. UMAP is used solely for qualitative illustration. Because it distorts metric geometry, analytical distances, shell boundaries, score values, and certification properties cannot be inferred from this embedding.

For the ten-class datasets, all classes are displayed. For CIFAR-100, ten classes are highlighted and labeled for readability, while the remaining classes are shown in gray.

On CIFAR-10, SVHN, and Imagenette, Figure 4 shows classes forming visually distinct, well-separated local neighborhoods. On CIFAR-100, the highlighted classes are comparatively more intermixed, consistent with its higher operational KS and lower accuracy reported in RQ1. The UMAP views provide qualitative visual context for class-neighborhood organization in the learned representations; they are not used to support any metric, structural, or certification claim in the paper.

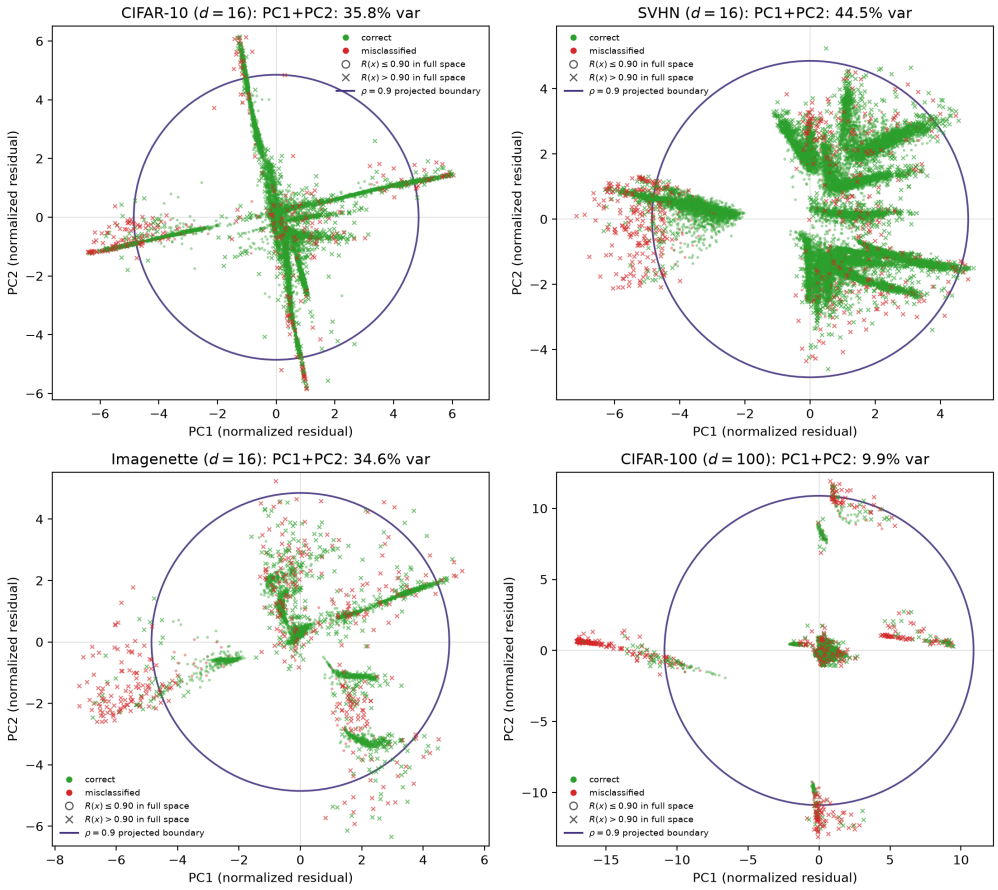


Figure 3: Origin-preserving principal-subspace visualization of the normalized ADOQ operational residuals. Each observation is represented by $w(x) = (f_\theta(x) - C_{\hat{y}(x)})/\sigma_{\hat{y}(x)}$, with $\|w(x)\|_2^2 = S_{\mathrm{op}}(x)$, and projected as $u = P_2^\top w$ onto two orthonormal principal directions. Green and red points denote correct and misclassified predictions, respectively. The displayed $\rho = .90$ circle has radius $\sqrt{\tau_{.90}}$. Since the projection is non-expansive, observations inside the analytical region remain inside the displayed circle, whereas full-space exterior observations may project inside it. The percentage reported in each panel is the fraction of total squared normalized-residual norm captured by the two displayed directions.

G.4 Native versus matched latent dimension

The structural diagnostics in Appendix E use the ADOQ reference dimension whenever the alternative representation permits it. This control matters because the reference distribution itself depends on d, and changing the representation dimension can affect the resulting KS discrepancy.

Table 17 compares the operational KS obtained in each distance-based method's native representation with the corresponding diagnostic after matching the ADOQ reference dimension. For CIFAR-10, SVHN, and Imagenette, the matched dimension is $d = 16$; for CIFAR-100 it is $d = 100$. The dimensionality reduction used for matching is fitted on training representations only, as detailed in Appendix D.5.

Dimension matching has a substantial but strongly method- and dataset-dependent effect. On CIFAR-10 and SVHN, for example, the operational KS of the supervised Mahalanobis representation decreases from .9670 to .2726 and from .9585 to .2667, respectively. SAT+EM+SR and CCL-SC also show substantial reductions after matching on several datasets, whereas MAPLE generally changes more moderately.

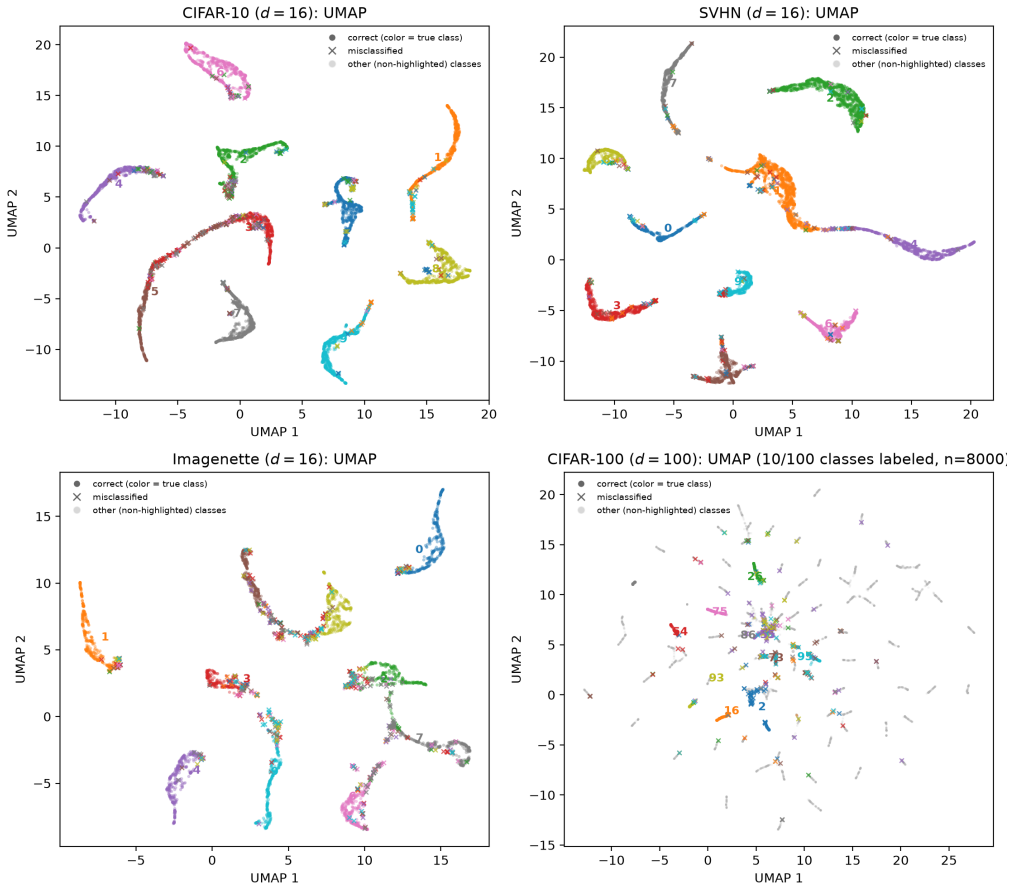


Figure 4: Qualitative UMAP visualization of the learned ADOQ latent representations. Colors denote true classes and crosses indicate misclassified observations. For CIFAR-100, ten classes are highlighted and labeled while the remaining classes are shown in gray for readability. UMAP is used only to visualize local neighborhood organization and is not interpreted metrically; the figure therefore provides no evidence about analytical distances, shell boundaries, score values, or certification.

The effect is not uniformly favorable. VMDLS on SVHN increases from .3315 to .4418 when reduced from $d = 32$ to $d = 16$, while on CIFAR-100 the Mahalanobis diagnostic increases from .5056 to .6474 when reduced from $d = 640$ to $d = 100$. For Imagenette, Mahalanobis is essentially unchanged (.6816 versus .6804). Dimension matching therefore changes the absolute level of discrepancy in a method- and dataset-dependent manner rather than systematically improving agreement with the reference law.

Thus latent dimension can materially affect absolute discrepancy from a χ^2_d reference, but it does not account for the structural differences observed in RQ1. Across all four datasets, the matched competing representations retain larger operational KS discrepancies than ADOQ. These controls therefore rule out latent dimension alone as an explanation for ADOQ's lower operational KS in the matched comparisons.

G.5 ADOQ SENSITIVITY TO LATENT DIMENSION

The main experiments use $d = 16$ for the ten-class datasets and $d = 100$ for CIFAR-100. The sensitivity of the analytical structure learned by ADOQ to alternative latent dimensions is additionally evaluated.

Table 17: Effect of latent dimension on empirical operational KS for the distance-based structural controls. Native and matched diagnostics are derived from the same learned representation; the matched dimension is the dimension used by ADOQ on the corresponding dataset. Lower is better.

Dataset	Method	Native d	Native $\mathrm{KS}_{\mathrm{op}}$	Matched $\mathrm{KS}_{\mathrm{op}}$	ADOQ $\mathrm{KS}_{\mathrm{op}}$
CIFAR-10	Mahalanobis	640	.9670	.2726	.0999
	SAT+EM+SR	256	.5359	.3328	.0999
	CCL-SC	256	.6382	.2920	.0999
	MAPLE	9	.3179	.2578	.0999
	VMDLS	32	.2827	.2869	.0999
SVHN	Mahalanobis	640	.9585	.2667	.0322
	SAT+EM+SR	256	.7560	.4516	.0322
	CCL-SC	256	.8360	.3437	.0322
	MAPLE	9	.3373	.2698	.0322
	VMDLS	32	.3315	.4418	.0322
Imagenette	Mahalanobis	512	.6816	.6804	.1537
	SAT+EM+SR	256	.7540	.4885	.1537
	CCL-SC	256	.7376	.5752	.1537
	MAPLE	11	.5349	.5276	.1537
	VMDLS	32	.5091	.5114	.1537
CIFAR-100	Mahalanobis	640	.5056	.6474	.3574
	SAT+EM+SR	256	.8937	.6664	.3574
	CCL-SC	256	.8966	.6612	.3574
	MAPLE	81	.5039	.5071	.3574
	VMDLS	100	.5301	.5301	.3574

Table 18: ADOQ sensitivity to latent dimension. Accuracy and AURC quantify conventional classification and selective performance, while $\widehat{\mathrm{KS}}_{\mathrm{op}}$ and ρ_S^{op} characterize the operational analytical coordinate. More negative Spearman correlation indicates stronger alignment between smaller analytical scores and larger discriminative margins.

Dataset	d	Accuracy	AURC	$\widehat{\mathrm{KS}}_{\mathrm{op}}$	ρ_S^{op}
CIFAR-10	10	94.91%	.00976	.0695	-.9567
	16	95.15%	.01000	.0999	-.9462
	32	94.97%	.01148	.0726	-.9302
SVHN	10	96.19%	.00880	.0583	-.9752
	16	96.88%	.00835	.0322	-.9745
	32	96.71%	.00936	.0464	-.9725
Imagenette	10	88.13%	.02756	.1613	-.9248
	16	88.59%	.02448	.1537	-.9469
	32	87.29%	.02600	.1617	-.9156
CIFAR-100	100	77.90%	.07293	.3574	-.8529
	128	77.53%	.07656	.3549	-.8385

For CIFAR-10, SVHN, and Imagenette, additional ADOQ models are trained with $d \in \{10, 16, 32\}$. For CIFAR-100, the regular-simplex construction requires $d \geq K - 1 = 99$, so the comparison uses $d = 100$ and $d = 128$.

Table 18 shows that, across the ten-class datasets, changing the latent dimension modifies score-law discrepancy and conventional performance to some extent, while operational difficulty ordering remains strong across all evaluated configurations. In particular, ρ_S^{op} remains below $-.91$ in every case. No single dimension is uniformly optimal across all reported metrics: for example, CIFAR-10 attains its smallest operational KS at $d = 10$, whereas Imagenette attains both its smallest operational KS and strongest operational ordering at $d = 16$.

CIFAR-100 shows a similarly limited change across the two evaluated dimensions. Increasing d from 100 to 128 changes $\widehat{\mathrm{KS}}_{\mathrm{op}}$ only from .3574 to .3549, while the operational Spearman correla-

tion remains strongly negative ($-.8529$ versus $-.8385$). The substantial prescribed-law discrepancy therefore persists at the larger dimension and is not explained by the near-minimal simplex dimension alone.

Together with the matched-dimensional controls of Appendix G.4, which rule out latent dimension as an explanation for ADOQ's advantage over competing representations, this sensitivity analysis rules out latent dimension as an explanation for ADOQ's own structural behavior: the difficulty ordering and (to a lesser extent) the prescribed-law fidelity reported in RQ1 are properties of the ADOQ construction itself, not artifacts of the specific dimension chosen. The value $d = 16$ is retained as the common main configuration for the ten-class datasets, while $d = 100$ provides a near-minimal dimension permitting an exact regular simplex for CIFAR-100.