BEYOND IMPURITY: TRANSPORT MARGINS FOR GEOMETRY-AWARE DECISION TREES

Anonymous authors
Paper under double-blind review

ABSTRACT

Decision trees score a split by how many samples of each class it routes to each side. Margin classifiers score a boundary by its distance to the closest point. We show both are functionals of one object. Coupling the class-conditional feature distributions at a node by one-dimensional monotone optimal transport, we define the transport-margin profile $\Phi_t(r)$: the coupled class-to-class mass that crosses a candidate threshold t with both endpoints at distance at least r. Its height at $r = 0$ is exactly the Kolmogorov–Smirnov (KS) split criterion and, after balance-normalized squaring, exactly the Gini gain. Its full-survival radius recovers the hard margin, its value at every r equals the certified robust balanced-accuracy advantage of the split, and its moments $\mathcal{M}_p$ integrate over thresholds to Wasserstein distances, with $\int \mathcal{M}_1(t)\, dt = \frac{1}{4} W_2^2$. We split on the radius-capped area $\mathcal{M}_1^R$, which interpolates between KS and the full transport margin. Sparsity of monotone transport yields an exact linear-time scan of all thresholds after sorting, at near-CART cost, with uniform consistency and contamination-stability guarantees. In forests matched in everything but the criterion, the swap costs under one point of average clean tabular accuracy, significantly improves noise robustness, exact worst-case robustness, and partition stability on real tabular data, and, on geometry-driven families, turns collapses of worst-case and shift robustness into flat curves, with no attack budget committed in advance and a polynomial-time certificate that tracks exact worst-case evaluation.

1 INTRODUCTION

Tree ensembles remain the strongest models for tabular data (Grinsztajn et al., 2022), and the theory of their basic mechanisms is still being consolidated (Scornet & Hooker, 2026). At every node a tree scores candidate splits by the contingency table they induce, the number of samples of each class routed left and right (Breiman et al., 1984; Quinlan, 1986). Figure 1 shows two features on which every such score ties: the same table at the optimal threshold, the same Kolmogorov–Smirnov (KS) separation, the same Gini gain. The boundaries are very different. Feature A separates the classes across a corridor that both crowd, and feature B across a wide empty gap, so small input perturbations flip many predictions at A and none at B. No counting criterion can tell them apart, because counting happens after the feature has been collapsed to two routing states (Proposition 1). KS splitting (Friedman, 1977; Rounds, 1980) and distance-, ranking-, and skew-aware criteria (López de Mántaras, 1991; Ferri et al., 2002; Cieslak et al., 2012) score the same table.

Methods that do add geometry each see one slice of it. A hard-margin separator (Cortes & Vapnik, 1995; Bennett et al., 2000) sees only the closest point and ignores how the remaining mass is arranged. Robust tree learners optimize worst-case impurity at a committed attack budget (Chen et al., 2019a; Vos & Verwer, 2021; 2022; Justin et al., 2023), and certified stumps (Andriushchenko & Hein, 2019) evaluate a split at one radius. A budget that is too small misses the margin, and one that is too large refuses fine-scale structure that carries real signal. The margin distributions that explain boosting (Schapire et al., 1998; Gao & Zhou, 2013) live in a model's output space, and transport enters forests and classifiers through response distributions (Du et al., 2021; Ćevid et al., 2022) or the loss (Kolouri et al., 2016; Li et al., 2020), not through the class-conditional feature distributions at a candidate boundary (Appendix A).

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.

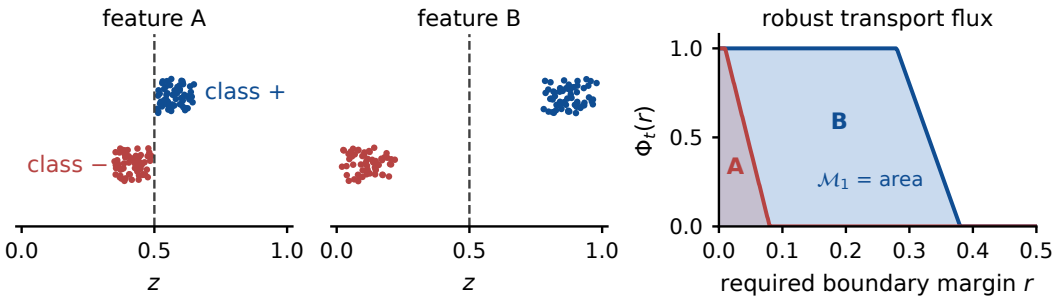

Figure 1: **Same counts, same KS, same Gini: different geometry.** Two features with identical contingency tables at their optimal thresholds (dashed): A separates the classes across a narrow corridor, B across a wide gap. Every contingency-based criterion ties, while the robust transport flux $\Phi_t(r)$ (right) separates them and its area $\mathcal{M}_1$ prefers B.

This paper introduces one object that holds all of these slices. For a node with class-conditional feature distributions μ_+, μ_-, we couple them by one-dimensional monotone optimal transport and ask, for every candidate threshold t and radius $r \geq 0$: *how much coupled opposite-class mass crosses t with both endpoints at distance at least r?* The answer is the *transport-margin profile* $\Phi_t(r)$. Its height at $r = 0$ is the KS split criterion and, after balance-normalized squaring, the Gini gain. Its value at a single r is the certified robust balanced-accuracy advantage of the split, which is what a fixed-budget method reads, and its full-survival radius is the hard margin. Transport-margin splitting (TMS) splits on its area up to a cap R, the *transport margin* $\mathcal{M}_1^R$, which integrates the certified advantage over every radius up to R and so commits to no single budget. Over thresholds its moments $\mathcal{M}_p$ decompose Wasserstein distances, with $\int \mathcal{M}_1(t)\, dt = \frac{1}{4} W_2^2$, so the criterion splits where the W_2 separation of the classes is concentrated.

The criterion costs little more than Gini. Monotone transport on the line is sparse and its margin contributions are piecewise linear, so one linear-time sweep after sorting scores every threshold exactly. In forests matched in everything but the criterion, the swap costs under one point of mean clean accuracy on nineteen OpenML datasets, raises the area under its accuracy-versus-noise curve on fourteen, and stabilizes the selected partitions. On geometry-driven families it turns a collapse from 0.99 to 0.55 into a flat curve, and no fixed attack budget does well on both families we test. Under exact worst-case evaluation on six real datasets it raises robust accuracy over the matched Gini forest on every one at the largest budget, from under one point on wdbc to 38 on spambase. A pre-training test marks the interaction-dominated data on which counting should be kept.

Contributions. **(1)** The transport-margin profile, with classical criteria and fixed-budget certificates as its functionals, and results that exist only for the whole curve: no orientation switching, the layer-cake identity, the hard-margin endpoint, and the capped family. **(2)** A local decomposition of Wasserstein geometry, $\int \mathcal{M}_p(t)\, dt = W_{p+1}^{p+1}/(2^p(p+1))$ (Theorem 4). **(3)** An exact $O(n)$-after-sorting split scan with consistency and contamination guarantees, and matched-forest experiments with cheap-fix and fixed-budget controls.

2 Transport-Margin Splitting

2.1 Tree Splitting as Collapsed Transport

Consider a tree node with binary labels $Y \in \{+1, -1\}$, class prior $\rho = P(Y{=}+1)$, and a scalar feature normalized by a monotone map to $Z = \psi_j(X_j) \in [0, 1]$. Write $\mu_\pm = \mathcal{L}(Z \mid Y{=}\pm 1)$ for the class conditionals, $F_\pm$ for their cumulative distribution functions (CDFs), $Q_\pm$ for their quantile functions, and W_p for the p-Wasserstein distance. A candidate split sends z to $\pi_t(z) = \mathsf{L}$ if $z \leq t$ and to R otherwise. Set $a = F_+(t)$, $b = F_-(t)$, and $q = \rho a + (1-\rho)b$ for the mass routed left.

Theorem 1 (Gini gain is a collapsed transport flux)**.** *Let $\nu_\pm^t = (\pi_t)_{\#}\mu_\pm$ be the routed class conditionals, regarded as measures on the two-point space $\{\mathsf{L}, \mathsf{R}\}$ with unit transport cost. The Gini gain of the split (j, t) equals*

$$\Delta_G(t) = \frac{2\rho^2(1-\rho)^2}{q(1-q)}\big(F_+(t) - F_-(t)\big)^2 = \frac{2\rho^2(1-\rho)^2}{q(1-q)}\, W_1^2\big(\nu_+^t, \nu_-^t\big). \tag{1}$$

CART therefore measures an optimal-transport separation, but only after the feature space has been collapsed to the two routing states. A sample at distance 10^{-3} from the threshold and a sample at the far end of the feature's range occupy the same point of the collapsed space. The balance factor $q(1-q)$ depends on t, so Gini and the flux $|F_+ - F_-|$ have different maximizers in general, but both criteria see only the collapsed measures.

Corollary 1 (KS splitting is the zeroth-order flux). *Let $\gamma^{\uparrow} = (Q_+, Q_-)_{\#}\,\mathrm{Unif}[0,1]$ be the monotone coupling of (μ_+, μ_-) and $\mathcal{C}_t = \{(x,y) : (x-t)(y-t) < 0\}$ the set of coupled pairs separated by t. For atomless distributions, $\gamma^{\uparrow}(\mathcal{C}_t) = |F_+(t) - F_-(t)|$: maximizing the crossing mass of the class-to-class optimal transport is exactly the KS split rule (Friedman, 1977; Rounds, 1980).*

Proposition 1 (Geometry blindness). *Let S be any split criterion that is a function of the contingency table $(n_{L,+}, n_{L,-}, n_{R,+}, n_{R,-})$ alone. Then S is invariant under every side-preserving transformation of the feature values, including maps that place all samples within distance ε of the threshold for arbitrary $\varepsilon > 0$. Consequently no such criterion determines a nontrivial lower bound on the boundary stability of the split in the feature metric.*

Gini, entropy, KS, and threshold-routed Hellinger criteria (Cieslak et al., 2012) are all contingency-only, hence all subject to Proposition 1. A better criterion therefore has to score the split with a quantity that survives the collapse, rather than with yet another function of the same table. Proofs are in Appendix B.

2.2 Transport-margin profiles

Definition 1 (Robust transport flux). *For a threshold t and margin requirement $r \geq 0$, define*

$$\Phi_t(r) = \big[F_+(t-r) - F_-(t+r)\big]_+ + \big[F_-(t-r) - F_+(t+r)\big]_+. \tag{2}$$

By monotonicity of CDFs at most one bracket is positive, so the sum in (2) is also a maximum, and (2) applies to arbitrary (including empirical) distributions. For atomless distributions $\Phi_t(r) = \gamma^{\uparrow}\{(x,y) \in \mathcal{C}_t : \min(|x-t|, |y-t|) \geq r\}$: the transported class-to-class mass that still crosses the boundary when both endpoints must clear it by r. In particular $\Phi_t(0) = |F_+(t) - F_-(t)|$ is the KS split criterion, whose maximum over t is the two-sample KS statistic, and the transport-margin profile $\Phi_t(\cdot)$ is nonincreasing.

Theorem 2 (Robust balanced accuracy). *Let $\mathrm{BA}_r^{\star}(t)$ denote the balanced accuracy of the better-oriented threshold classifier at t under worst-case feature perturbations of magnitude at most r, with a chance-level fallback. Then, for arbitrary distributions,*

$$\mathrm{BA}_r^{\star}(t) = \tfrac{1}{2}\big(1 + \Phi_t(r)\big). \tag{3}$$

Equation (3) is an identity rather than a bound, so $\Phi_t(r) = 2\mathrm{BA}_r^{\star}(t) - 1$ is exactly the certified balanced-accuracy *advantage* of the split. The better orientation in it never changes with r:

Lemma 1 (No orientation switching). *Both brackets of* (2) *are nonincreasing in r, and the bracket that vanishes at $r = 0$ vanishes for every r. Hence the orientation attaining $\mathrm{BA}_r^{\star}$ is, at every radius, the one selected at $r = 0$ (the orientation training assigns to the split), and for that trained split*

$$\int_0^R \Phi_t(r)\,dr = 2\int_0^R \Big[\mathrm{BA}_r^{\mathrm{trained}}(t) - \tfrac{1}{2}\Big]_+ dr, \tag{4}$$

so the area under the profile up to R integrates the positive part of the robust advantage of the split as trained. The chance-level fallback in Theorem 2 affects only radii whose advantage is nonpositive, which contribute nothing.

At a single budget r, quantities of the form $F_{\pm}(t \pm r)$ appear in the certified stump loss of Andriushchenko & Hein (2019). The object here is the entire curve.

Theorem 3 (Layer cake). *Let $m_t(x,y) = \min(|x-t|, |y-t|)$ on $\mathcal{C}_t$ and define the* transport margin *$\mathcal{M}_1(t) = \int_{\mathcal{C}_t} m_t\,d\gamma^{\uparrow}$. Then*

$$\mathcal{M}_1(t) = \int_0^{\infty} \Phi_t(r)\,dr = 2\int_0^{\infty} \Big(\mathrm{BA}_r^{\star}(t) - \tfrac{1}{2}\Big)dr. \tag{5}$$

Theorem 4 (Local decomposition of Wasserstein distances). *For $p \geq 0$ let $\mathcal{M}_p(t) = \int_{\mathcal{C}_t} m_t^p \, d\gamma^{\uparrow}$ (with $m_t^0 = \mathbf{1}[m_t > 0]$, so $\mathcal{M}_0$ is the crossing mass). Then*

$$\int_{\mathbb{R}} \mathcal{M}_p(t)\, dt = \frac{1}{2^p(p+1)}\, W_{p+1}^{p+1}(\mu_+, \mu_-). \tag{6}$$

In particular $\int \mathcal{M}_0(t)\, dt = W_1(\mu_+, \mu_-)$ recovers the classical CDF formula, and $\int \mathcal{M}_1(t)\, dt = \frac{1}{4} W_2^2(\mu_+, \mu_-)$.

Maximizing $\mathcal{M}_1$ therefore splits where the squared W_2 separation of the class conditionals is most concentrated.

Radius cap. The integral in (5) weights all perturbation radii equally. On interleaved multimodal classes this lets a wide plateau of small advantage outweigh a narrow region of large advantage: the midpoint of a long transport ray can carry moderate flux for a very large range of r while a cleaner but narrower gap carries higher flux for a shorter range (Figure 4, left). The *radius-capped* transport margin

$$\mathcal{M}_1^R(t) = \int_0^R \Phi_t(r)\, dr = \int_{\mathcal{C}_t} \min\{m_t, R\}\, d\gamma^{\uparrow} \tag{7}$$

resolves this with one interpretable parameter, the largest perturbation radius of interest. As $R \to 0$, $\mathcal{M}_1^R / R \to \Phi_t(0)$ recovers the KS ordering, as $R \to \infty$ it recovers $\mathcal{M}_1$, and integrating over thresholds Huberizes the transport cost: $\int \mathcal{M}_1^R(t)\, dt = \sum_e w_e\, h_R(d_e)$ for coupling edges e (coupled pairs) with weights w_e and lengths d_e, where $h_R(d) = d^2/4$ for $d \leq 2R$ and $R(d - R)$ otherwise, which truncates long transport rays in the spirit of outlier-robust transport (Mukherjee et al., 2021).

Choosing R. R is the largest perturbation radius of interest, expressed in the unit fixed by the normalization. When a deployment perturbation scale is known, setting R to it is an assumption of the method, not a tuned parameter. When it is not known, standardizing each feature by its training standard deviation (std) and setting $R = 0.1$ (a tenth of a std per feature) is an automatic, validation-free default. Any per-feature normalization already realizes per-feature raw radii $R_j = R \cdot \mathrm{scale}_j$, so the scalar cap is a per-feature budget in raw units. The sensitivity analysis of Section 3.4 shows a flat optimum over $R \in [0.05, 0.2]$ on both synthetic failure regimes, and the same window is optimal or within noise of optimal on real data (Table 20). For scale-free prediction with rank margins we use the $R \to \infty$ endpoint: quantile margins are bounded by construction, capping them trades a known protection (the plateau regime above) against a small average accuracy cost. The rule is to use rank normalization with $R \to \infty$ for scale-free clean prediction, the known unit (or std) with the deployment scale (or $R = 0.1$) for metric perturbations, and the impurity criterion where the signal-density test of Section 3.3 fails (Table 22).

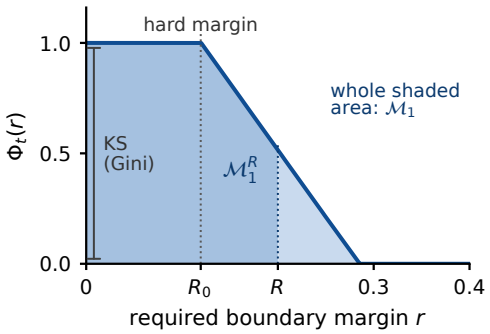

Figure 2: **One profile, four functionals.** $\Phi_t(r)$ for two uniform class conditionals separated by a gap, with its KS/Gini height, the capped area $\mathcal{M}_1^R$ (darker) inside the full transport-margin area $\mathcal{M}_1$, and the hard-margin radius R_0 marked.

Theorem 5 (Hard-margin endpoint). *Let $R_0(t) = \sup\{r : \Phi_t(r) = \Phi_t(0)\}$. For linearly separable data x_i and directions θ with $\|\theta\|_2 = 1$, let $\Phi_{\theta,t}$ and $R_0(\theta, t)$ be computed on the projections $\theta^\top x_i$. Then*

$$\max_{\theta,\, t}\ R_0(\theta, t) \quad \textit{subject to} \quad \Phi_{\theta,t}(0) = 1$$

is the geometric hard-margin problem $\max_{\theta,t} \min_i |\theta^\top x_i - t|$. The constraint is necessary: without it any t outside the data support attains $\Phi_t \equiv 0$ and $R_0 = \infty$.

Why transport? Pointwise, $\Phi_t(r)$ needs only the two CDFs in (2), and at one threshold the collapsed flux of Theorem 1 is elementary (a two-point W_1 is a total-variation distance). The coupling supplies what the CDF form does not: the layer-cake object $\mathcal{M}_1$, one coupling optimal for every

convex cost, the hard-margin endpoint, the trimmed variant, and the sparse edge decomposition the linear-time sweep exploits. One identity becomes a family of criteria with a shared computation. Figure 2 shows these quantities on one curve.

Margin semantics. The normalization ψ_j fixes the meaning of r. With the pooled empirical CDF (**TM-Rank**) margins are probability-mass distances, scale-free and monotone-invariant, and the adversary of (3) moves each point past at most rn of the n samples. With winsorized affine scaling (**TM-Metric**) margins are raw-metric distances and (3) holds literally for additive perturbations. Use TM-Metric, in the perturbation's units, when a metric robustness statement is the goal, and TM-Rank when no canonical metric exists. Measuring margins in units that do not match the perturbation forfeits the geometric advantage, as Table 7 shows. Every robustness statement in Section 3 names its normalization.

2.3 Exact linear-time splitting

Evaluating an optimal-transport quantity at every candidate threshold of every feature at every node sounds expensive. On the line it costs little more than a sort.

Sparse exact coupling. For sorted class samples $z_1^+ \le \cdots \le z^+{}_{n_+}$ and $z_1^- \le \cdots \le z^-{}_{n_-}$ with uniform weights, the monotone coupling is the north-west-corner plan on the merged integer grid $\{i\, n_-\}_{i \le n_+} \cup \{j\, n_+\}_{j \le n_-}$. It is exact in integer arithmetic and has at most $n_+ + n_- - 1$ edges (Peyré & Cuturi, 2019). It is optimal simultaneously for every convex cost, which is what makes Theorem 4 a statement about a single coupling.

Event sweep. A coupling edge with endpoints $\ell < h$ and weight w contributes $w\,[\min(t-\ell,\, h-t)]_+$ to $\mathcal{M}_1(t)$: a tent, encoded by three slope events ($+w$ at ℓ, $-2w$ at the midpoint, $+w$ at h). Under the radius cap the tent becomes a trapezoid with inner breakpoints $\min(\ell+R, (\ell+h)/2)$ and $\max(h-R, (\ell+h)/2)$, which merge at the midpoint into the tent when $h - \ell \le 2R$, and for $\mathcal{M}_2$ the pieces are quadratic. Monotonicity of the coupling makes each event list already sorted, so one cumulative-sum pass evaluates the criterion at all distinct-value midpoints simultaneously.

Proposition 2 (Complexity)**.** *For one feature at a node with n samples, all candidate thresholds are scored exactly in $O(n)$ time after sorting ($O(n \log n)$ from raw values). Binary problems use a single coupling, and K-class problems use the one-vs-rest sum $S(t) = \sum_k \rho_k\, \mathcal{M}_1^{(k)}(t)$ with class priors ρ_k, at total cost $O(Kn)$, which reduces to the binary criterion when $K = 2$.*

Corollary 2 (Classwise certificate semantics)**.** *For each class k, Theorem 2 applied to $(\mu_k, \mu_{\neg k})$ makes $\frac{1}{2}(1 + \Phi_t^{(k)}(r))$ the exact worst-case balanced accuracy of the class-k-versus-rest stump, so the one-vs-rest criterion is the prior-weighted sum of integrated certified one-vs-rest advantages: the certificate semantics of the binary criterion is inherited classwise. This does not certify the multiclass decision itself: the summands bound each one-vs-rest stump separately, not the argmax over classes.*

Training never calls a general transport solver, never forms a quadratic-size plan, and has no regularization parameter to tune.

Trees and forests. A transport-margin tree (TM-tree) is CART with the threshold scan replaced, and leaves keep class proportions. A TM-forest keeps every random forest mechanism (bootstrap resampling, per-node feature subsets, and probability averaging, Breiman, 2001), so setting the criterion back to Gini yields an exactly matched baseline, which is how all forest comparisons in Section 3 are constructed.

Variants at the same complexity. Transport trimming (dropping the most expensive fraction τ of the coupling mass before measuring margins), sparse oblique directions, and prior-corrected leaves all reuse the one-dimensional kernel. Appendix C gives the definitions, and Appendices D and E the supporting comparisons.

2.4 STATISTICAL GUARANTEES

The integrand $g_t(x, y) = \mathbf{1}[(x - t)(y - t) < 0]\ m_t(x, y)$ vanishes on the boundary of the crossing set and is therefore 1-Lipschitz in each argument. With the quantile representation of the coupling this gives dimension-one rates with no further assumptions.

Theorem 6 (Uniform consistency). *For a fixed normalization map ψ, let $\widehat{\mu}_\pm$ be the empirical class conditionals of n samples, $\widehat{\mathcal{M}}_1$ the criterion they induce, and $\hat{t}$ its maximizer. On bounded support, with the class prior bounded away from 0 and 1 (so that $\min(n_+, n_-) = \Theta_p(n)$),*

$$\sup_t \left|\widehat{\mathcal{M}}_1(t) - \mathcal{M}_1(t)\right| \ \le\ W_1(\widehat{\mu}_+, \mu_+) + W_1(\widehat{\mu}_-, \mu_-) \ =\ O_p(n^{-1/2}),$$

and the same bound holds for every $\mathcal{M}_1^R$. If the population criterion has a unique maximizer $t^\star$ then $\hat{t} \to t^\star$ in probability, and under quadratic growth of the population criterion at $t^\star$, $|\hat{t} - t^\star| = O_p(n^{-1/4})$.

Theorem 7 (Contamination stability). *For a fixed normalization map, if $\widetilde{\mu}_\pm = (1 - \epsilon)\mu_\pm + \epsilon\kappa_\pm$ for arbitrary contaminating distributions $\kappa_\pm$ on the normalized feature space, then $\sup_t |\mathcal{M}_1(t; \widetilde{\mu}_\pm) - \mathcal{M}_1(t; \mu_\pm)| \le 2\epsilon$.*

Corollary 3 (Split stability). *If additionally $\mathcal{M}_1(t^\star) - \mathcal{M}_1(t)\ \ge\ c\,|t - t^\star|^2$, the contaminated maximizer satisfies $|\tilde{t}^\star - t^\star| \le 2\sqrt{\epsilon/c}$.*

When ψ is estimated, rank normalization keeps a 4ϵ bound, winsorized scaling needs stable anchoring quantiles, and standardization inherits the guarantee only under moment assumptions (Remark 1). Both theorems hold verbatim for the capped criterion (the cap is 1-Lipschitz) and transfer to training, where the scan maximizes over a finite candidate set. Theorem 2 certifies one split, and a tree composes splits along paths.

Proposition 3 (Tree-level certificate). *Let T be a tree with axis-aligned splits and let x reach a leaf labeled y. If $|x_{j_v} - t_v| > r$ at every split (j_v, t_v) on its path, every perturbation δ with $\|\delta\|_\infty \le r$ leaves the prediction unchanged. Hence the certified accuracy of T at radius r is at least $P(\text{correct leaf}) - \sum_v P\big(|X_{j_v} - t_v| \le r,\ v \text{ on the path of } X\big)$.*

The bound is loose but directional: it grows when every selected split keeps mass away from its threshold, which is what the per-node margin criterion promotes. Forest-level robustness is established empirically in Section 3 and certified from below by the following computable bound.

Proposition 4 (Ensemble box certificate). *Let a forest average the leaf class-probability vectors p_l of trees $T_1, \dots, T_m$ with axis-aligned splits, and let $\mathcal{R}_k(x, \varepsilon)$ be the leaves of T_k whose cells lie within ℓ_∞ distance ε of x. If*

$$\sum_{k=1}^{m} \min_{l \in \mathcal{R}_k(x,\varepsilon)} \Big(p_l(y) - \max_{c \ne y} p_l(c)\Big) \ >\ 0,$$

then every perturbation δ with $\|\delta\|_\infty \le \varepsilon$ leaves the prediction at y (per-tree adversaries relax the shared perturbation, as in the tree-ensemble verification of Chen et al., 2019b).

Proposition 4 supplies the certified values of Section 3. Relaxing instead the leaf integrality of the exact leaf-selection program, while keeping every cross-tree consistency constraint, gives a linear-programming (LP) certificate that is never weaker than the box. On real data it is within 0.011 of the exact value at $\varepsilon \le 0.05$ for every forest tested, and within 0.01 at every budget for TMS forests (Table 18). Appendix B gives both programs and all proofs.

3 EXPERIMENTS

All forest comparisons are *matched*: bootstrap, feature subsets, tree depth, and leaf rules are identical, and only the split criterion changes. Synthetic and exact worst-case experiments consider all d features at every node, and the benchmark experiments use the random-forest default of $\sqrt{d}$ (Appendix F repeats the former at $\sqrt{d}$). TMS denotes the capped criterion $\mathcal{M}_1^R$ with $R = 0.1$ in the stated unit in every perturbation-facing experiment, and robustness AUC is the area under

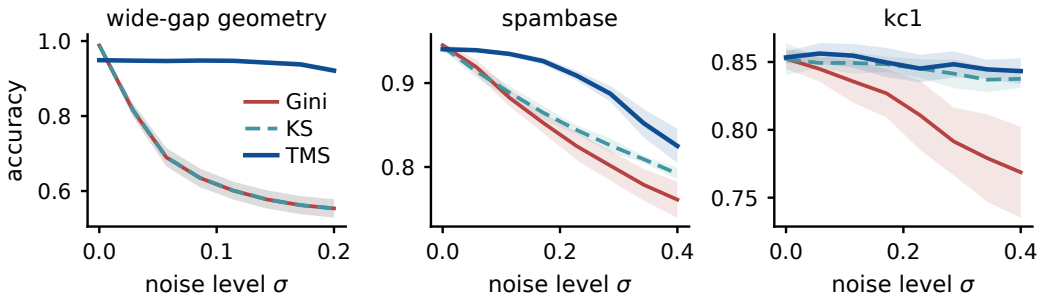

Figure 3: **Matched forests under additive test noise** (mean ± sd over seeds). (a) Wide-gap family. (b,c) spambase and kc1, per-feature-std-scaled noise, TM-Metric.

test accuracy as a function of the noise level, divided by the width of the noise range. The clean benchmarks of Table 12 use TM-Rank at the $R \to \infty$ endpoint, following the rule of Section 2.2. Significance uses paired Wilcoxon tests with Holm correction and mean ranks (Demšar, 2006; García & Herrera, 2008). Synthetic experiments use the two-feature family of Figure 1 extended with an overlap parameter (a fraction of each class placed on the wrong side of feature B's gap, making B count-suboptimal) and eight noise features. All features share one raw scale and splits use that scale directly. Full protocols and additional tables are in Appendices D–G. In method comparisons the best value is bold and the second best underlined, and ties share the mark.

3.1 Mechanism and theory match

With any positive overlap, count-based criteria choose the fragile feature A in 100% of runs (its contingency table is strictly better) and their stumps drop to 0.55–0.61 accuracy under feature noise of $\sigma = 0.1$. The transport margin chooses B in 100% of runs and holds 0.90–0.95, unchanged under $1{:}5$ class imbalance. Within the family the chosen split's $\mathcal{M}_1^R$ tracks its robustness with Spearman correlation 0.96 over 120 geometries (Figure 5), as Theorem 3 predicts.

3.2 Robustness

Figure 3 traces accuracy under additive test noise for matched forests, and Table 8 summarizes the synthetic family: the criterion swap turns the Gini/KS collapse ($0.99 \to 0.55$ over $\sigma \in [0, 0.2]$) into a flat curve ($0.95 \to 0.92$), holds under boundary-targeted noise, location shift, and 30% label flips, and stabilizes the selected split. On nineteen OpenML datasets with per-feature-std-scaled noise, the automatic std-unit forest, the prespecified default of Section 2.2, attains the higher robustness AUC on 14, with mean gain $+1.6$ points (Wilcoxon over datasets, $p = 0.006$, surviving Holm correction over the two normalization variants at $p = 0.012$), largest on eeg-eye-state ($+13.8$) and MagicTelescope ($+5.1$). The winsorized metric variant gives $+1.2$ ($p = 0.016$), largest on spambase and kc1 ($+6$), whose curves Figure 3 (b,c) shows, and Table 13 has the full table. The same configuration carries the clean benchmark (mean clean accuracy within 0.3 points of the matched Gini forest, $p = 0.42$), and a variant normalized by the median absolute deviation (MAD) that restores the estimated-normalization guarantee is indistinguishable ($p = 0.98$). Cheap fixes do not reproduce the gain: impurity-guided threshold re-location trails by 1.6 AUC points ($p = 0.026$), and Gaussian training augmentation at the perturbation scale matches the average-case AUC ($p = 0.77$) at $2.8\times$ the training cost but stays at the Gini level on both geometric families (Table 11) and behind on partition stability. Over 100 bootstrap replicates per dataset the TMS root split induces significantly more stable validation partitions (mean adjusted Rand index, ARI, 0.70 vs. 0.61, $p = 0.029$, with feature agreement 0.75 vs. 0.69, $p = 0.22$). On five multiclass datasets the one-vs-rest criterion carries the robustness over, with $+1.7$ AUC points on pendigits, $+2.6$ on segment, and ties elsewhere.

Fixed-budget robust baselines. The methods in this section differ in what they must be told. Impurity criteria need nothing, TMS needs a unit for margins (supplied automatically by standardization), and fixed-budget methods need the unit *and* a committed attack radius: the certified-advantage objective $\Phi_t(r_0)$ (the robust-stump criterion of Andriushchenko & Hein, 2019, expressible inside our framework and run with the identical architecture) and GROOT forests (Vos & Verwer, 2021) at attack radius ε. Table 1 evaluates both across two families. On the wide-gap family every sufficiently

Table 1: Fixed budgets (in parentheses) vs. the integrated margin: robustness AUC on each family and the minimum over both. Full grids: Table 6.

	wide-gap	multiscale	min
Gini	0.67	0.85	0.67
$\Phi(0.02)$	0.78	**0.97**	0.78
$\Phi(0.1)$	**0.94**	0.81	0.81
GROOT (0.02)	<u>0.80</u>	<u>0.95</u>	0.80
GROOT (0.05)	**0.94**	0.88	<u>0.88</u>
GROOT (0.1)	**0.94**	0.81	0.81
TMS (R=0.1)	**0.94**	**0.97**	**0.94**

Table 2: Exact robust accuracy at $\varepsilon = 0.1$ on six real datasets. Fixed: worst and post-hoc best of Φ and GROOT at 0.02, 0.05, 0.1. The post-hoc best is an oracle and is not marked.

	Gini	TMS	worst fixed	best fixed
diabetes	0.62	**0.64**	<u>0.63</u>	0.70
banknote	0.94	**0.97**	<u>0.95</u>	0.98
wdbc	<u>0.89</u>	**0.90**	0.88	0.90
kc1	0.78	**0.86**	<u>0.81</u>	0.87
phoneme	0.74	**0.78**	<u>0.76</u>	0.81
spambase	<u>0.44</u>	**0.82**	0.23	0.88
mean	<u>0.74</u>	**0.83**	0.71	0.86

large budget matches TMS, but the small budget loses the margin preference. On a multiscale family, where structure at margins $\sim$0.3 carries 60% of the mass and structure at margins ≤ 0.08 carries the rest, the ordering reverses: budgets of 0.05 and above cannot certify the fine-scale splits and stall at 0.79–0.88 AUC, while the small budget survives. Every fixed-budget configuration fails one of the two families. The integrated criterion is the only configuration in the top group of both, and its worst-case AUC across the families (0.94) exceeds the best fixed-budget configuration (0.88) by six points. Training times are of the same order as GROOT's.

Certified robustness. Theorem 2 speaks about the worst case. Table 9 reports *exact* robust accuracy under worst-case ℓ_∞ perturbations from a mixed-integer linear program (MILP) in the style of Kantchelian et al. (2016) on depth-limited matched forests. The Gini forest collapses to 0.04 at $\varepsilon = 0.05$ and to zero at $\varepsilon = 0.1$ on the wide-gap family while the geometry-aware methods hold 0.94–0.95. On the multiscale family the fixed-budget methods pay 18 clean points for their committed budget while TMS keeps clean accuracy at 1.0 and is best at $\varepsilon \leq 0.05$. At $\varepsilon = 0.1$ GROOT is ahead (0.71 vs. 0.59): the fine-band samples are indefensible at that budget by any split that classifies them, so the two methods sit at different points of a data-level Pareto frontier. The box certificate of Proposition 4 is already within 0.033 of these exact values.

Worst-case robustness on real data. Table 2 applies the same exact evaluation to six OpenML datasets (every budget in Table 19), with features standardized by the training standard deviation so that a scalar budget ε means $\varepsilon \cdot \mathrm{std}_j$ per feature. Against impurity splitting, TMS raises exact robust accuracy on all six datasets at $\varepsilon = 0.05$ and $\varepsilon = 0.1$, by 9.0 points on average and 3.2 at the median at $\varepsilon = 0.1$ (Wilcoxon over the 30 splits, $p < 0.001$), from 0.44 to 0.82 on spambase and from 0.78 to 0.86 on kc1. Fixed budgets are fragile here as on the synthetic families. On spambase at $\varepsilon = 0.1$, Φ and GROOT trained at 0.02 or 0.05 fall to 0.23–0.36 while TMS holds 0.82. A budget matched to the attack wins at its own radius: $\Phi(0.1)$ and GROOT at 0.1 lead TMS at $\varepsilon = 0.1$ by 1.6 and 2.8 points on average (Table 16), and each dataset's post-hoc best fixed budget leads by 0.8, 1.4, and 3.0 points on average at $\varepsilon = 0.02$, 0.05, and 0.1. Large budgets pay for this wherever the informative structure is finer than the budget: $\Phi(0.1)$ falls 21 clean points below the matched Gini forest on eeg-eye-state, 8 on mozilla4, and 2 on phoneme (Table 15), the refused-splits failure of Table 1, where GROOT at 0.1 fails the same way. Small budgets collapse in the worst case, large budgets in clean accuracy on fine-scale data, and the integrated criterion collapses on neither axis. Gaussian augmentation ties TMS here ($p = 0.23$).

3.3 Clean tabular benchmarks

Table 12 compares TM-forests with the matched Gini forest and library-default baselines (Breiman, 2001; Geurts et al., 2006; Chen & Guestrin, 2016; Ke et al., 2017) on nineteen OpenML datasets: the sixteen OpenML-CC18 datasets (Bischl et al., 2021) passing a mechanical numeric-binary filter plus three more, evaluated by five-fold cross-validation with 100 trees. Averaged over datasets, the criterion swap costs 0.6 points against the matched Gini forest (95% CI $[-1.4, 0.0]$), 0.7 against random forests, and about 1 point against default boosting, with no paired difference significant (all $p > 0.06$) and equivalence within ± 0.013 against the matched Gini forest (± 0.01 against extremely randomized trees). TMS is the best of all seven models on four datasets. More than half of the mean

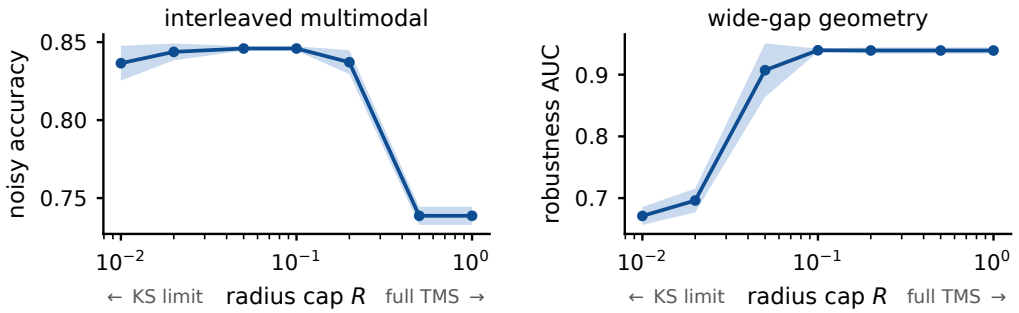

Figure 4: **The radius cap interpolates KS $\leftrightarrow$ TMS** (mean $\pm$ sd over 6 seeds). Small R is KS-like, accurate on interleaved classes (left) but blind to margins (right), and large R, the full margin, does the reverse. $R \in [0.05, 0.2]$ achieves both.

cost is one dataset, madelon (-6.5 points: 480 of its 500 features are noise probes and the signal is XOR-like), and the case is identifiable before training: the marginal-signal density (fraction of features with Bonferroni-significant two-sample tests) is 0.02 there and at least 0.22 (median 0.86) elsewhere, so the rule "impurity below density 0.1", with the cutoff fixed on this benchmark, repairs madelon (clean 0.66, robustness AUC 0.63) and, in sample, raises the robustness gain to $+1.8$ AUC points ($p = 0.002$) at 0.1 points mean clean cost. Applied unchanged to eight datasets outside the benchmark, it flags the only one on which TMS clearly trails (dexter, -3.4 AUC points), and TMS stays within 1.7 points of Gini on every dataset above the cutoff. Appendix F gives these results and constructed interaction families, and adds a second diagnostic, the worst-case headroom of the impurity forest, which bounds the swap's worst-case gain. On five multiclass datasets (up to $K = 10$) the one-vs-rest criterion stays within half a point of the matched Gini forest and 0.6 of scikit-learn's random forest (Table 21).

3.4 Ablations

Margin order. $\mathcal{M}_0$ (the KS flux) reproduces Gini-level robustness (AUC 0.66), $\mathcal{M}_1$ and $\mathcal{M}_2$ agree (0.94), and the partial and oblique variants preserve it (Table 7). **Radius cap.** Figure 4 traces both regimes of Section 2.2: on interleaved multimodal classes the uncapped margin falls into the wide-plateau trap (noisy accuracy 0.74) while every $R \leq 0.2$ holds 0.84–0.85. On the wide-gap family small caps lose the margin preference (AUC 0.67 at $R = 0.01$) while $R \geq 0.1$ attains the full 0.94. The two curves overlap on a wide sweet spot, $R \in [0.05, 0.2]$ in normalized units. **Pairing.** At equal crossing mass, random pairings and the $O(n^2)$ all-pairs average blur the margin structure of interleaved classes, and the capped transport margin matches the best alternative at 0.1 ms per scan, over $100\times$ faster than all-pairs (Table 4).

4 Limitations

The method handles threshold splits on numeric features, not categorical features or missing values. Under heavy class imbalance the criterion is skew-insensitive but majority voting is not, so balanced-accuracy use needs prior-corrected leaves (Appendix D). Theorems 6 and 7 control the criterion and the selected threshold, not the risk of the greedy tree, which is as open as for impurity-based CART (Scornet et al., 2015). When the signal lives in feature interactions, margin-seeking degrades more than counting, and the signal-density test flags that regime before training.

5 Conclusion

Gini measures separation without geometry, and the hard margin measures geometry without distribution. The transport-margin profile measures both and holds the two as its extreme functionals. Its capped area is scored exactly by one sort and one linear sweep, and in matched forests it costs under one point of clean accuracy while adding 0.2–0.4 of accuracy under perturbation where boundary geometry matters, with two computable diagnostics bounding where that can appear. Next steps are boosting-style reweighting of the profile, multivariate couplings, and exact partial matchings.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the theory, including formulating and refining claims and writing proofs, to design and give feedback on experiments, to implement the methods, experiments, and data processing, and to interpret results. We did not use them for translation, the synthetic data are drawn from the distributions of Appendix G rather than generated by AI models, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to create figures, write and edit code and text, search and summarize literature, and format references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments use public OpenML benchmarks or synthetic data. The work involves no human subjects or personal data, and we are not aware of ethical concerns beyond those of decision-tree classifiers in general.

REPRODUCIBILITY STATEMENT

Appendix B proves every result under its stated assumptions, and Appendix G specifies the data, protocols, and hyperparameters. The code, which regenerates every figure and table from seeded scripts and verifies its kernel against brute-force oracles to 10^{-9}, will be released upon acceptance.

REFERENCES

Maksym Andriushchenko and Matthias Hein. Provably robust boosted decision stumps and trees against adversarial attacks. In *NeurIPS*, 2019.

Sebastian Angrick, Jacobus Conradi, Mónika Csikós, Niko Hastrich, Danny Mittal, André Nusser, Krzysztof Onak, and Sharath Raghvendra. Computing all optimal partial p-Wasserstein matchings on the line. *arXiv preprint arXiv:2608.18875*, 2026.

Kristin P. Bennett, Nello Cristianini, John Shawe-Taylor, and Donghui Wu. Enlarging the margins in perceptron decision trees. *Machine Learning*, 41:295–313, 2000.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *NeurIPS Datasets and Benchmarks*, 2021.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge–Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010.

Emilio Carrizosa, Thomas Halskov, and Dolores Romero Morales. Wasserstein support vector machine: Support vector machines made fair. *European Journal of Operational Research*, 329(2): 641–652, 2026.

Domagoj Ćevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Laetitia Chapel and Romain Tavenard. One for all and all for one: Efficient computation of partial Wasserstein distances on the line. In *ICLR*, 2025.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. Robust decision trees against adversarial examples. In *ICML*, 2019a.

Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane Boning, and Cho-Jui Hsieh. Robustness verification of tree-based models. In *NeurIPS*, 2019b.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *KDD*, 2016.

David A. Cieslak, T. Ryan Hoens, Nitesh V. Chawla, and W. Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24:136–158, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

André F. Cruz, Catarina Belém, Sérgio Jesus, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient boosting with fairness constraints. In *ICLR*, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):267–277, 1974.

Federico D'Onofrio, Giorgio Grani, Marta Monaci, and Laura Palagi. Margin optimal classification trees. *Computers & Operations Research*, 161:106441, 2024.

Qiming Du, Gérard Biau, François Petit, and Raphaël Porcher. Wasserstein random forests and applications in heterogeneous treatment effects. In *AISTATS*, 2021.

Cèsar Ferri, Peter Flach, and José Hernández-Orallo. Learning decision trees using the area under the ROC curve. In *ICML*, 2002.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195:533–560, 2010.

Jerome H. Friedman. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, C-26(4):404–408, 1977.

Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

Salvador García and Francisco Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS Datasets and Benchmarks*, 2022.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *UAI*, 2020.

Nathan Justin, Sina Aghaei, Andrés Gómez, and Phebe Vayanos. Learning optimal classification trees robust to distribution shifts. *arXiv preprint arXiv:2310.17772*, 2023.

Alex Kantchelian, J. D. Tygar, and Anthony D. Joseph. Evasion and hardening of tree ensemble classifiers. In *ICML*, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS*, 2017.

Silas Koemen. Distributional split criteria for random forests: Extensions, shrinkage, and the robustness of mean splitting. *arXiv preprint arXiv:2607.23721*, 2026.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *CVPR*, 2016.

Jiajin Li, Caihua Chen, and Anthony Man-Cho So. Fast epigraphical projection-based incremental algorithms for Wasserstein distributionally robust support vector machine. In *NeurIPS*, 2020.

Ramón López de Mántaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.

Hoora Moradian, Denis Larocque, and François Bellavance. L_1 splitting rules in survival forests. *Lifetime Data Analysis*, 23:671–691, 2017.

Debarghya Mukherjee, Aritra Guha, Justin Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *ICML*, 2021.

Sreerama K. Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Lev Reyzin and Robert E. Schapire. How boosting the margin can also boost classifier complexity. In *ICML*, 2006.

E. M. Rounds. A combined nonparametric approach to feature selection and binary decision tree design. *Pattern Recognition*, 12(5):313–317, 1980.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

Erwan Scornet and Giles Hooker. Theory of random forests. *Annual Review of Statistics and Its Application*, 13(1):99–121, 2026.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Robert Tibshirani and Trevor Hastie. Margin trees for high-dimensional classification. *Journal of Machine Learning Research*, 8:637–652, 2007.

Tyler M. Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L. Patsolic, Benjamin Falk, Carey E. Priebe, Jason Yim, Randal Burns, Mauro Maggioni, and Joshua T. Vogelstein. Sparse projection oblique randomer forests. *Journal of Machine Learning Research*, 21(104):1–39, 2020.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2014.

Alexander Vargo, Fan Zhang, Mikhail Yurochkin, and Yuekai Sun. Individually fair gradient boosting. In *ICLR*, 2021.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Daniël Vos and Sicco Verwer. Efficient training of robust decision trees against adversarial examples. In *ICML*, 2021.

Daniël Vos and Sicco Verwer. Robust optimal classification trees against adversarial examples. In *AAAI*, 2022.

Teng Zhang and Zhi-Hua Zhou. Large margin distribution machine. In *KDD*, 2014.

Teng Zhang and Zhi-Hua Zhou. Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1143–1156, 2020.

A Related work

Split criteria. KS-based splitting originates with Friedman (1977) and Rounds (1980), impurity criteria with CART and ID3 (Breiman et al., 1984; Quinlan, 1986), and distance- and ranking-based selection with López de Mántaras (1991) and Ferri et al. (2002). Hellinger trees address skew (Cieslak et al., 2012). All are functions of the routing contingency table and hence subject to Proposition 1. Two statistical neighbors are close: the quantile difference $|Q_+(u) - Q_-(u)|$ underlying our coupling is Doksum's shift function (Doksum, 1974), and Ramdas et al. (2017) connect Wasserstein, KS, and ROC statistics for two-sample testing. Those are global statistics of a pair of samples. This paper decomposes the same geometry pointwise along candidate decision boundaries and uses the decomposition as a training signal.

Distributional forests. Wasserstein random forests (Du et al., 2021), distributional random forests (Ćevid et al., 2022), energy-distance and survival splitting (Ishwaran et al., 2008; Moradian et al., 2017), and the recent comparison of maximum-mean-discrepancy and sliced-Wasserstein criteria (Koemen, 2026) compare the *response* distributions of candidate children, targeting conditional distribution estimation. TMS transports the *class-conditional predictor* distributions within a node and localizes the coupling to the candidate boundary, which is a different object for a different question.

Margins. Margin objectives enter trees through per-node separators (Bennett et al., 2000; Tibshirani & Hastie, 2007) and mixed-integer formulations (D'Onofrio et al., 2024), as point margins in the support vector machine (SVM) sense (Cortes & Vapnik, 1995). Margin *distributions* explain boosting (Schapire et al., 1998; Reyzin & Schapire, 2006; Gao & Zhou, 2013) and are optimized directly by large and optimal margin distribution machines (Zhang & Zhou, 2014; 2020). Those margins live in a model's output space. The transport margin lives in the input feature space and describes how the coupled class-to-class mass clears a candidate split, with the decomposition of Theorem 4 and the identity (3) that this affords.

Robust trees. Adversarially robust tree learning optimizes worst-case impurity at a fixed budget (Chen et al., 2019a; Vos & Verwer, 2021) or solves the minimax problem exactly (Vos & Verwer, 2022; Justin et al., 2023). Andriushchenko & Hein (2019) certify stump losses at one budget through the same $F_{\pm}(t \pm r)$ as (2). At one threshold and one budget our identities are transport readings of known quantities. The curve is what has no fixed-budget counterpart (Lemma 1, Theorems 3–5, the coupling-native trim, and one $O(n)$ sweep for every order, cap, and trim level), and TMS commits to no single adversary. Exact verification of tree ensembles solves a mixed-integer program (Kantchelian et al., 2016), and polynomial verification relaxes it (Chen et al., 2019b). Proposition 4 and the LP relaxation of Appendix B are certificates of that kind, and the LP tracks the exact program closely (Table 18).

Optimal transport and classification. Sliced-Wasserstein kernels (Kolouri et al., 2016), Wasserstein-distributionally-robust SVMs (Li et al., 2020), and Wasserstein fairness penalties (Jiang et al., 2020; Vargo et al., 2021; Cruz et al., 2023; Carrizosa et al., 2026) combine transport with classifiers at the model or loss level. Here transport acts inside the split search, and on the line it needs only the sorted monotone coupling (Peyré & Cuturi, 2019), not an entropic solver (Cuturi, 2013).

Outlier-robust and exact partial transport (Mukherjee et al., 2021; Chapel & Tavenard, 2025; Angrick et al., 2026) enter through the radius cap (Section 2.2) and the trimmed variant (Appendices C and E).

B PROOFS

Theorem 1 (Gini gain is a collapsed transport flux). With $p_L = \rho a/q$, $p_R = \rho(1-a)/(1-q)$ and $G(p) = 2p(1-p)$, expand $\Delta_G = G(\rho) - qG(p_L) - (1-q)G(p_R)$. Using $q - \rho a = (1-\rho)b$ and $1 - q - \rho(1-a) = (1-\rho)(1-b)$, the numerator over $q(1-q)$ collapses to $2\rho^2(1-\rho)^2\big[a(1-b) + b(1-a) - a(1-a) - b(1-b)\big] = 2\rho^2(1-\rho)^2(a-b)^2$, which gives the first equality. On the two-point routing space, $\nu_+^t = (a, 1-a)$ and $\nu_-^t = (b, 1-b)$, and the transport cost of moving one Bernoulli to the other with unit cost is $|a-b|$. □

Corollary 1 (KS splitting is the zeroth-order flux). Under the quantile representation, the pair $(Q_+(u), Q_-(u))$ crosses t iff exactly one of $Q_\pm(u) \le t$ holds, i.e. iff u lies between $F_+(t)$ and $F_-(t)$, up to a Lebesgue-null set for atomless distributions. □

Proposition 1 (geometry blindness). A side-preserving transformation fixes the four cell counts by definition, hence fixes S. Choosing the transformation that maps each side's samples into an ε-neighborhood of t realizes arbitrarily small boundary distances at unchanged criterion value. □

Theorem 2 (robust balanced accuracy). A positive point survives every perturbation of size $\le r$ on the "+ right" orientation iff $Z > t + r$ (probability $1 - F_+(t+r)$), and a negative point survives iff $Z \le t - r$ (probability $F_-(t-r)$). The two orientations give balanced accuracies $\frac{1}{2}[1 \pm (F_\mp(t-r) - F_\pm(t+r))]$-type expressions, and by monotonicity of CDFs at most one bracket in (2) is positive, so $\max\big(\mathrm{BA}^{(+)}, \mathrm{BA}^{(-)}, \frac{1}{2}\big) = \frac{1}{2}\big(1 + \Phi_t(r)\big)$ exactly, for arbitrary distributions. □

Lemma 1 (no orientation switching). Each bracket of (2) has a first term nonincreasing and a subtracted term nondecreasing in r, so both brackets are nonincreasing. At $r = 0$ the brackets are $\pm(F_+(t) - F_-(t))$, so at most one is positive and the other is nonpositive, hence nonpositive for all r. For the orientation positive at $r = 0$, $\mathrm{BA}_r^{\mathrm{trained}} = \frac{1}{2}(1 + B(r))$ with B its bracket, so $[\mathrm{BA}_r^{\mathrm{trained}} - \frac{1}{2}]_+ = \frac{1}{2}[B(r)]_+ = \frac{1}{2}\Phi_t(r)$ for all r. Integrating over $[0, R]$ gives (4). □

Theorem 3 (layer cake). Layer-cake for the nonnegative m_t on $\mathcal{C}_t$: $\int_{\mathcal{C}_t} m_t \, d\gamma^\uparrow = \int_0^\infty \gamma^\uparrow(\mathcal{C}_t \cap \{m_t \ge r\}) \, dr$, and $\mathcal{C}_t \cap \{m_t \ge r\}$ is the union of the two rectangles whose measures are the brackets of (2). The second identity is Theorem 2. □

Theorem 4 (local decomposition of Wasserstein distances). For a fixed coupled pair $x < y$ with $d = y - x$, $\int_{\mathbb{R}} m_t^p \, dt = \int_x^y \min(t-x, y-t)^p \, dt = 2\int_0^{d/2} s^p \, ds = \frac{d^{p+1}}{2^p(p+1)}$. Fubini over the coupling gives $\int \mathcal{M}_p(t) \, dt = \frac{1}{2^p(p+1)} \int |x-y|^{p+1} d\gamma^\uparrow$, and the monotone quantile coupling is optimal for every convex cost on the real line (Villani, 2009; Santambrogio, 2015), so the integral is W_{p+1}^{p+1}. The capped analogue integrates $\min(m_t, R)$: $d^2/4$ for $d \le 2R$ and $2 \cdot \frac{R^2}{2} + (d - 2R)R = R(d-R)$ otherwise. □

Theorem 5 (hard-margin endpoint). With t in the separation gap, $\Phi_t(0) = 1$ and $\Phi_t(r) = 1$ iff $r \le \min\big(t - \sup \mathrm{supp}\, \mu_-, \ \inf \mathrm{supp}\, \mu_+ - t\big)$, so $R_0(t)$ equals that minimum. Maximizing over t gives the gap midpoint and half-gap value, and maximizing over unit directions gives $\max_{\theta,t} \min_i |\theta^\top x_i - t|$. Degeneracy without the constraint: for t outside the support both brackets of (2) vanish for all r. □

Corollary 2 (classwise certificate semantics). Immediate: Theorem 2 and Lemma 1 applied to the pair $(\mu_k, \mu_{\neg k})$ for each k, then the layer-cake identity of Theorem 3 summed with weights ρ_k. □

Theorem 6 (uniform consistency). Each coordinate section of g_t is piecewise linear with slopes in $[-1, 1]$ and continuous across $\partial\mathcal{C}_t$ (where $m_t = 0$), so $|g_t(x, y) - g_t(x', y')| \le |x - x'| + |y - y'|$. Then $|\widehat{\mathcal{M}}_1(t) - \mathcal{M}_1(t)| = \left|\int_0^1 g_t(\widehat{Q}_+, \widehat{Q}_-) - g_t(Q_+, Q_-)\, du\right| \le \int_0^1 |\widehat{Q}_+ - Q_+| + |\widehat{Q}_- - Q_-|\, du = W_1(\widehat{\mu}_+, \mu_+) + W_1(\widehat{\mu}_-, \mu_-)$, which is $O_p(n^{-1/2})$ on bounded support. The bound is uniform in t and applies verbatim to $\min(m_t, R)$. Argmax consistency and the $n^{-1/4}$ rate under quadratic growth follow from the standard argmax argument. □

Theorem 7 (contamination stability). $W_1\big((1-\epsilon)\mu + \epsilon\kappa, \mu\big) \le \epsilon\, W_1(\kappa, \mu) \le \epsilon$ on a unit-diameter space, and the Lipschitz bound of the previous proof gives the claim. □

Corollary 3 (split stability). Write $\mathcal{M}_1$ and $\widetilde{\mathcal{M}}_1$ for the clean and contaminated criteria. $\mathcal{M}_1(\tilde{t}^\star) \ge \widetilde{\mathcal{M}}_1(\tilde{t}^\star) - 2\epsilon \ge \widetilde{\mathcal{M}}_1(t^\star) - 2\epsilon \ge \mathcal{M}_1(t^\star) - 4\epsilon$, so $c\,|\tilde{t}^\star - t^\star|^2 \le 4\epsilon$. □

Proposition 3 (tree-level certificate). If $|x_{j_v} - t_v| > r$ at every path node, then for any $\|\delta\|_\infty \le r$ the perturbed point satisfies the same inequality at every such node with margin > 0, hence follows the same path and receives the same leaf label. A prediction can therefore change only if the point is within r of some threshold on its own path, and the union bound over path nodes gives the statement. □

Proposition 4 (ensemble box certificate). Any perturbed point $x + \delta$ with $\|\delta\|_\infty \le \varepsilon$ lies, in each tree, in a leaf whose cell is within ε of x, i.e. in $\mathcal{R}_k(x, \varepsilon)$. The forest's probability margin at $x + \delta$ is therefore at least the sum over trees of the minimal leaf margin over $\mathcal{R}_k(x, \varepsilon)$. Minimizing per tree lets each tree choose its own δ, which relaxes the shared-δ adversary. □

Exact program and LP certificate. Fix a test point x, its predicted class y, a budget ε, and a wrong class c. Leaf l of tree T_k has cell $\{z : a_{kl} < z \le b_{kl}\}$ (coordinatewise, infinite ends allowed) and class-probability vector p_{kl}. The adversary chooses $\delta \in [-\varepsilon, \varepsilon]^d$ and leaf indicators $z_{kl} \in \{0, 1\}$:

$$\min_{z,\delta} \ \textstyle\sum_{k,l} z_{kl}\big(p_{kl}(y) - p_{kl}(c)\big) \quad \text{s.t.} \quad \sum_l z_{kl} = 1 \ \text{ for every } k,$$

$$x_j + \delta_j \ge a_{klj} + \eta - M^a_{klj}(1 - z_{kl}), \qquad x_j + \delta_j \le b_{klj} + M^b_{klj}(1 - z_{kl}),$$

with a strictness constant $\eta = 10^{-9}$ for the open lower faces and $M^a_{klj} = a_{klj} + \eta - (x_j - \varepsilon)$, $M^b_{klj} = (x_j + \varepsilon) - b_{klj}$. These are the smallest constants for which the constraints of an unselected leaf follow from the bounds on δ. A constraint with a nonpositive constant holds for every admissible δ and is dropped, and a leaf whose cell lies farther than ε from x is fixed to $z_{kl} = 0$. Every feasible integral point therefore selects, in each tree, the leaf that $x + \delta$ reaches (up to the strictness constant), and every admissible δ yields such a point, so x is robust exactly when the optimum is positive for every $c \ne y$. A single global constant would have to exceed every $|a_{klj} - x_j| + \varepsilon$ and $|b_{klj} - x_j| + \varepsilon$. A smaller one makes the program infeasible, and infeasibility must never be read as robustness.

The LP certificate replaces $z_{kl} \in \{0, 1\}$ by $z_{kl} \in [0, 1]$ and certifies x when every optimum is positive. It is valid because the relaxed feasible set contains every integral one, so the relaxed optimum is a lower bound on the attack margin. It is never weaker than the box certificate. Give each tree its own perturbation: its relaxed indicators then range over the simplex spanned by its reachable leaves, and a linear objective attains its minimum over that simplex at a reachable leaf, which is the tree's term in Proposition 4. Sharing one δ across trees only shrinks the feasible set and raises the optimum, so a point certified by the box is certified by the LP. The LP has polynomially many variables and constraints and is solved in polynomial time. □

Remark 1 (Estimated normalizations). When ψ is itself estimated, rank normalization stays controlled: ϵ-contamination moves the pooled empirical CDF by at most ϵ in sup norm, hence every normalized position by at most ϵ and $\mathcal{M}_1$ by at most an additional 2ϵ, so the bound of Theorem 7 becomes 4ϵ. Winsorized affine scaling additionally requires the anchoring quantiles to be stable, which we state as an assumption rather than prove. Standardization is different: the empirical standard deviation has unbounded sensitivity to arbitrary contamination, so the std normalization

of Section 3.2 is an operational default for scale matching and inherits the guarantee only under moment assumptions on the contaminating distribution.

Proof. With $\widetilde{F} = (1-\epsilon)F + \epsilon F_\kappa$ the pooled contaminated CDF (F_κ the CDF of the contamination), $\|\widetilde{F} - F\|_\infty \le \epsilon$, so the two rank maps differ by at most ϵ at every point. Shifting every coupled position by at most δ changes g_t (1-Lipschitz in each argument) by at most 2δ, hence $\mathcal{M}_1$ by at most 2ϵ. Adding the fixed-coordinate bound of Theorem 7 gives 4ϵ. □

C Algorithmic details

Coupling. Give each positive sample n_- units of mass and each negative sample n_+ units. The merged grid $\{i\, n_-\} \cup \{j\, n_+\}$ is exact in integers, and consecutive grid intervals name the coupled pair (positive index $\lfloor u/n_- \rfloor$, negative index $\lfloor u/n_+ \rfloor$ at interval start u) with weight proportional to the interval length. Support size is at most $n_+ + n_- - 1$.

Sweep. Sort the node's values once. Candidates are the midpoints of consecutive distinct values. Each edge emits its slope events. Because the coupling is monotone, the left endpoints, midpoints, and right endpoints are each nondecreasing along the edge order. Under the cap the two inner breakpoints of edge e are $\min\big(\ell_e + R, (\ell_e + h_e)/2\big)$ and $\max\big(h_e - R, (\ell_e + h_e)/2\big)$, with slope events $-w_e$ each. When $h_e - \ell_e \le 2R$ both sit at the midpoint and add up to the tent's $-2w_e$, and all four breakpoint sequences stay nondecreasing, since minima and maxima of nondecreasing sequences are nondecreasing. A single merged pass then maintains the running value and slope, evaluating the criterion at every candidate. Piecewise-quadratic sweeps for $\mathcal{M}_2$ track one additional curvature accumulator.

Prediction. Leaves store class proportions, and forests average them. With prior correction, proportions are divided by the training priors and renormalized.

Variants. *Transport trimming*: dropping the most expensive fraction τ of the coupling mass before measuring margins prevents outlying or mislabeled mass from defining the split geometry. The truncation is not exact partial transport, which re-optimizes the matched sub-measures. Exact partial matchings on the line are available (Chapel & Tavenard, 2025; Angrick et al., 2026), and Appendix E shows the truncation is empirically split-equivalent to the exact partial plan on the contamination family tested. *Oblique splits*: sparse random directions $\theta \in \{-1, 0, 1\}^d / \|\cdot\|_2$ (cf. Murthy et al., 1994; Tomita et al., 2020) are scored by the same one-dimensional kernel. *Prior-corrected leaves*: dividing leaf class proportions by the training priors carries the criterion's skew-insensitivity (each $\mu_\pm$ is separately normalized) through to prediction. Without the correction, majority voting reintroduces the skew that the split criterion removed (Appendix D).

D Class imbalance and prior-corrected leaves

Table 3 reports balanced accuracy on a one-dimensional Gaussian family with training imbalance up to 1:50 and balanced test sets.

Table 3: Balanced accuracy under training imbalance (one-dimensional Gaussian family, balanced test sets).

train ratio	Gini	TMS	Gini + prior corr.	TMS + prior corr.
1:1	**0.835**	0.834	**0.835**	0.834
1:5	0.802	0.719	**0.824**	0.818
1:10	0.730	0.593	0.817	**0.818**
1:50	0.575	0.502	0.787	**0.800**

The split criterion normalizes μ_+ and μ_- separately and is unaffected by the ratio, but majority voting is not: without correction the skew removed at the split returns at the leaf. With prior-corrected leaves the transport-margin forest attains the best balanced accuracy at every ratio of 1:10

and beyond. Resampling (Chawla et al., 2002) and skew-insensitive criteria (Cieslak et al., 2012) are the standard alternatives, surveyed by He & Garcia (2009).

E Partial transport: truncation vs. exact partial matchings

The variant of Section 2.3 drops the most expensive fraction τ of the full monotone coupling. Exact partial transport instead re-optimizes which sub-measures to couple (Figalli, 2010; Caffarelli & McCann, 2010), and on the line it is computable for all masses (Chapel & Tavenard, 2025; Angrick et al., 2026). Comparing the truncation against exact partial plans obtained by linear programming, on two-cluster instances with 10% of one class placed far outside the support (20 instances per setting): the transport-cost ratio of truncated to exact is at most 1.03 when τ matches the contamination (1.24 at τ twice too large), and in 100% of instances both plans place the selected threshold inside the true separation gap, with mean threshold difference at the candidate-grid resolution. Exact partial matchings improve the transport cost, not the split.

F Additional experiments

Pairing ablation. On interleaved bimodal classes with two equal-flux gaps of different widths (30 bootstraps), the same margin functional is evaluated under different pairings of the positive and negative mass (Table 4). The uncapped margin's failure is the plateau trap of Figure 4.

Table 4: Pairing ablation: noisy-test accuracy ($\sigma = 0.05$) and scan time.

pairing	noisy acc.	scan time
random (8 draws)	0.732	0.6 ms
all-pairs average	0.732	12 ms
minimum margin	**0.845**	0.1 ms
monotone, uncapped	0.743	0.2 ms
monotone, capped (R=0.1)	<u>0.844</u>	0.1 ms

Table 5: madelon: fraction of forest splits on the twenty informative features (sample-weighted; chance rate 4%).

criterion	depth 0–2	depth 3–5	depth 6+
Gini	22.7%	13.8%	7.9%
TMS (R=0.1)	17.5%	7.0%	4.0%

Margin order and normalization. Table 7 ablates the criterion on the wide-gap family, whose features share one raw unit. The zeroth-order flux loses the margin preference, the first- and second-order margins agree, and the partial and oblique variants preserve them. Rank and per-feature range normalization score margins in units that no longer match the perturbation (quantile mass or range-relative distance) and lose the robustness gain, while the shared-unit scan retains it. Margin units and perturbation units must agree, as stated in Section 2.2.

Fixed-budget baselines, full grids. Table 6 extends Table 1 to the full budget grids. Large budgets shrink the trees on the multiscale family (mean leaves per tree drops from 26 to 15 as r_0 grows past the fine scale), the signature of refused splits. Mean fit times per forest are 0.01–0.2 s for all methods at this size.

Table 6: Fixed-budget objectives over the full budget grids: robustness AUC.

criterion	wide-gap	multiscale
Gini	0.67	0.85
$\Phi(0.02)$	0.78	**0.97**
$\Phi(0.05)$	<u>0.94</u>	0.85
$\Phi(0.1)$	<u>0.94</u>	0.81
$\Phi(0.2)$	**0.95**	0.79
GROOT (0.02)	0.80	<u>0.95</u>
GROOT (0.05)	<u>0.94</u>	0.88
GROOT (0.1)	<u>0.94</u>	0.81
GROOT (0.2)	<u>0.94</u>	0.80
TMS (R=0.1)	<u>0.94</u>	**0.97**

Table 7: Wide-gap family: AUC by margin order (top) and normalization (bottom).

variant	robustness AUC
$\mathcal{M}_0$ (KS flux)	0.66
$\mathcal{M}_1$	0.94
$\mathcal{M}_2$	0.94
partial (τ=0.05)	0.94
oblique (8 directions)	0.94
rank normalization	0.67
per-feature range normalization	0.67
shared raw unit	0.94

Synthetic family, summary. Table 8 collects the robustness and stability results of Section 3.2 for the wide-gap family.

Table 8: Synthetic family: robustness and stability of matched forests. TMS is $\mathcal{M}_1^{R=0.1}$.

	Gini	KS	TMS
clean accuracy	**0.99**	**0.99**	0.95
uniform-noise robustness AUC	0.67	0.67	**0.94**
boundary-targeted noise acc.	0.57	0.57	**0.95**
location shift ± 0.05 acc.	0.53	0.53	**0.95**
label flips 30%, noisy test	0.63	0.63	**0.85**
root-feature agreement	0.98	0.98	**1.00**
root-threshold std ($\times 10^3$)	3	**2**	**2**
validation-partition ARI	0.90	0.92	**1.00**

Exact worst-case robustness on the synthetic families. Table 9 gives the exact MILP robust accuracies discussed in Section 3.2 (3 seeds, 150 points per seed). The multiscale inversion at $\varepsilon = 0.1$ (GROOT 0.71, TMS 0.59) is a property of the data, not a mis-set cap: the fine-band samples (40% of the mass, margins at most 0.08) are indefensible at that budget by any split that classifies them, so defending them is exactly the 18-point clean concession the fixed-budget methods make, and raising the cap does not chase the trade (TMS at $R = 0.2$ trains the identical forest).

Table 9: Exact worst-case robust accuracy (MILP; matched depth-6 forests, GROOT at budget 0.05). Box certificates are within 0.033 of every value.

	wide-gap				multiscale			
	clean	ε=.02	.05	.1	clean	ε=.02	.05	.1
Gini	**0.98**	0.65	0.04	0.00	**1.00**	0.66	0.00	0.00
$\Phi(r_0{=}0.05)$	0.95	**0.95**	**0.95**	0.94	0.82	0.79	0.72	0.64
GROOT (ε=0.05)	0.95	**0.95**	**0.95**	0.94	0.81	0.79	**0.78**	**0.71**
TMS (R=0.1)	0.95	**0.95**	**0.95**	**0.95**	**1.00**	**0.96**	**0.78**	0.59

Feature subsampling. The synthetic families and the exact worst-case evaluations consider every feature at every node ($m_{\text{try}} = d$ candidate features), and the benchmark experiments use $m_{\text{try}} = \sqrt{d}$. Table 10 repeats the former at $\sqrt{d}$. On the six real datasets the worst-case advantage of TMS persists and grows: its exact robust accuracy at $\varepsilon = 0.1$ exceeds the matched Gini forest's on all six (+12.0 points on average, split-level $p < 0.001$), against +9.0 with every feature. On the synthetic families it shrinks, most in the worst case. These families carry their robust structure in one or two of nine or ten features, so a node that samples three features misses it most of the time, and a criterion can only choose among the features it is shown. Gini is unaffected because it ignores that structure even when shown it. Where robust structure is concentrated in a few features and worst-case robustness matters, a larger m_{try} lets the criterion find it.

Table 10: Feature subsampling for the matched forests: every feature per node ($m_{\text{try}} = d$) against the random-forest default ($m_{\text{try}} = \sqrt{d}$). The real-data row averages the six MILP datasets.

	$m_{\text{try}} = d$		$m_{\text{try}} = \sqrt{d}$	
	Gini	TMS	Gini	TMS
wide-gap, robustness AUC	0.67	**0.94**	0.67	0.84
multiscale, robustness AUC	0.85	**0.97**	0.85	0.88
wide-gap, exact at $\varepsilon = 0.1$	0.00	**0.95**	0.00	0.26
multiscale, exact at $\varepsilon = 0.05$	0.00	**0.78**	0.00	0.06
six real datasets, exact at $\varepsilon = 0.1$	0.74	0.83	0.72	**0.84**

Criterion value and robustness within the family. Figure 5 plots, for 120 generated geometries of the wide-gap family (margin scale, overlap, and seed varied), the capped margin $\mathcal{M}_1^R$ of the split each stump selects against that stump's robustness AUC. The two have Spearman correlation 0.96,

as the layer-cake identity predicts. Both quantities are monotone in the generating margin scale, so this is a consistency check within the family rather than a cross-dataset forecast (the cross-dataset comparison is in the interaction paragraph below).

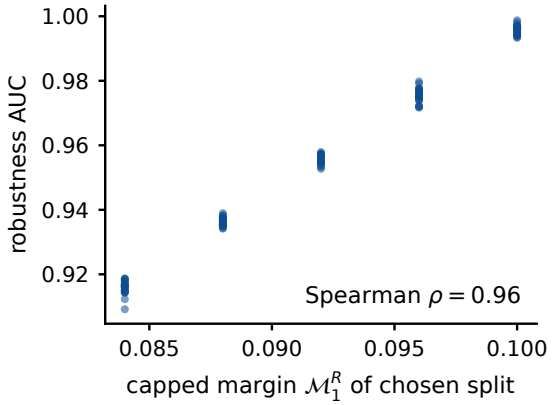

Figure 5: **Within the synthetic family, the chosen split's capped margin tracks its robustness** (120 geometries varying margin scale, overlap, and seed).

Cheap baselines. Table 11 evaluates the two attributions that Proposition 1 does not exclude, selecting by impurity and re-locating the threshold to the widest gap among near-optimal candidates (two tolerances, gaps measured in the perturbation unit) and training a matched Gini forest on Gaussian-augmented data at the perturbation scale (each tree sees its bootstrap sample plus one jittered copy, $\sigma = R$), together with Hellinger splitting (Cieslak et al., 2012), a functional of the cell counts to which Proposition 1 applies. On both geometric families every one of them stays at the Gini level: threshold re-location cannot change which feature is selected, and augmentation smooths the counts but still counts. On the nineteen real datasets (same protocol as Table 13), relocation trails TMS by 1.6 AUC points (Wilcoxon $p = 0.026$). Augmentation is the strongest cheap baseline, matching TMS's average-case AUC ($p = 0.77$) at 2.8× the training cost, but remains behind on partition stability (mean ARI 0.63 vs. 0.70, TMS ahead on 14 of 19) and recovers none of the geometric families' worst-case gains.

Table 11: Cheap baselines on the geometric families: clean accuracy and robustness AUC.

	wide-gap		multiscale	
model	clean	AUC	clean	AUC
Gini	0.986	0.668	1.000	0.845
Hellinger	<u>0.989</u>	0.669	1.000	0.845
two-stage (tol 1%)	0.988	0.666	1.000	0.845
two-stage (tol 5%)	0.987	0.666	1.000	0.845
augmented Gini (σ=0.1)	**0.991**	<u>0.673</u>	1.000	<u>0.846</u>
TMS (R=0.1)	0.948	**0.942**	1.000	**0.969**

Robustness on real data, full table. Table 13 reports clean accuracy and robustness AUC under per-feature-std-scaled noise for all nineteen datasets. The std-unit capped criterion attains the higher robustness AUC on 14 of 19 datasets (mean +1.6 points over the Gini forest, Wilcoxon $p = 0.006$), and the winsorized metric variant gives +1.2 ($p = 0.016$).

Fixed budgets on the same benchmark. Table 14 runs the Φ budget grid under the identical protocol. Every budget matches TMS's mean AUC (all $p \geq 0.20$), and every budget pays for it in clean accuracy where the informative structure is finer than the commitment: relative to the matched Gini forest, $\Phi(0.1)$ loses 21 clean points on eeg-eye-state, 8 on mozilla4, and 2 on phoneme, and even $\Phi(0.02)$ loses 6 on eeg-eye-state, while TMS's mean clean gap is 0.3 points with no collapse on any dataset. The eeg-eye-state dataset is the multiscale family realized in a benchmark: informative structure lives below the committed budgets, and refusing it costs clean accuracy exactly as in Table 1. Table 15 isolates the three fine-scale datasets.

Table 12: Clean accuracy on nineteen OpenML datasets (five-fold cross-validation, mean ± sd; 100 trees, boosters at library defaults).

dataset	Gini (matched)	RandomForest	ExtraTrees	HistGB	XGBoost	LightGBM	TM-forest
diabetes	0.763±.03	0.763±.03	**0.764**±.03	0.736±.03	0.751±.03	0.732±.04	0.757±.03
spambase	0.950±.01	0.950±.01	0.950±.01	0.956±.01	0.953±.01	**0.959**±.01	0.939±.01
kc1	0.861±.01	0.861±.01	0.864±.01	0.859±.01	0.854±.01	0.849±.01	**0.866**±.01
pc1	**0.939**±.01	0.936±.01	0.937±.01	0.931±.02	0.935±.01	0.936±.01	0.938±.01
blood-transf.	0.778±.03	0.782±.03	**0.789**±.03	0.763±.02	0.757±.03	0.758±.03	0.770±.04
ozone-8hr	0.943±.00	0.944±.00	0.942±.00	0.943±.01	0.943±.01	**0.945**±.01	0.942±.00
phoneme	0.907±.01	0.907±.01	0.905±.01	0.897±.01	0.897±.00	0.894±.01	**0.910**±.01
qsar-biodeg	0.865±.03	0.860±.03	0.863±.02	0.865±.03	0.862±.02	**0.868**±.04	0.850±.03
wdbc	0.958±.02	0.960±.01	0.963±.01	0.970±.01	**0.974**±.01	0.972±.01	0.968±.01
climate-crashes	0.920±.01	0.926±.01	0.915±.00	0.939±.01	**0.941**±.01	0.939±.01	0.917±.01
pc4	0.907±.01	0.905±.01	0.898±.01	**0.909**±.01	0.906±.02	0.907±.01	0.895±.01
pc3	0.901±.01	0.899±.01	**0.905**±.01	0.891±.01	0.889±.01	0.891±.01	0.896±.01
kc2	0.828±.04	0.835±.04	**0.843**±.04	0.824±.05	0.820±.05	0.820±.04	0.828±.04
MagicTelescope	0.878±.00	0.879±.00	0.874±.00	**0.881**±.00	**0.881**±.01	**0.881**±.00	0.872±.00
banknote-authentication	0.991±.00	0.990±.01	0.998±.00	0.997±.00	0.997±.00	0.996±.00	**0.999**±.00
eeg-eye-state	0.926±.00	0.926±.01	0.876±.03	0.902±.01	0.932±.01	0.904±.01	**0.934**±.00
mozilla4	**0.953**±.00	0.951±.00	0.941±.01	0.949±.00	0.951±.00	0.951±.00	0.940±.01
madelon	0.695±.01	0.713±.04	0.668±.01	**0.824**±.03	0.807±.03	0.810±.03	0.629±.02
wilt	0.982±.01	0.980±.01	0.968±.01	0.982±.00	**0.985**±.00	0.984±.00	0.984±.00

Table 13: Clean accuracy and robustness AUC under per-feature-std-scaled noise ($\sigma \in [0, 0.4]$); marks refer to AUC.

	Gini (rank)		TMS (rank, $R\to\infty$)		TMS (metric, R=0.1)		TMS (std, R=0.1)	
dataset	clean	AUC	clean	AUC	clean	AUC	clean	AUC
diabetes	0.792	**0.780**	0.771	0.765	0.779	0.776	0.784	0.776
spambase	0.953	0.855	0.938	0.893	0.939	**0.912**	0.949	0.883
kc1	0.847	0.799	0.847	0.848	0.866	**0.855**	0.853	0.843
pc1	0.940	0.932	0.937	0.931	0.943	**0.936**	0.937	0.933
blood-transf.	0.751	**0.764**	0.756	0.761	0.747	0.760	0.742	0.763
ozone-8hr	0.940	**0.941**	0.941	**0.941**	0.942	0.940	0.943	0.940
phoneme	0.891	0.822	0.893	0.827	0.891	**0.831**	0.899	0.830
qsar-biodeg	0.858	0.818	0.849	0.794	0.864	**0.836**	0.845	0.823
wdbc	0.942	0.938	0.947	0.941	0.947	0.941	0.953	**0.943**
climate-crashes	0.914	0.915	0.920	0.920	0.914	0.914	0.926	**0.925**
pc4	0.897	0.889	0.884	0.879	0.895	**0.890**	0.895	0.886
pc3	0.896	0.895	0.900	**0.898**	0.902	0.896	0.896	0.897
kc2	0.828	0.820	0.834	**0.841**	0.822	0.821	0.847	0.831
MagicTelescope	0.878	0.779	0.870	0.820	0.868	**0.832**	0.873	0.830
banknote	0.988	0.960	0.995	**0.973**	0.993	0.968	0.990	0.967
eeg-eye-state	0.917	0.607	0.925	0.617	0.929	0.611	0.894	**0.745**
mozilla4	0.946	0.781	0.935	0.788	0.932	**0.794**	0.929	0.791
madelon	0.645	**0.631**	0.604	0.574	0.627	0.615	0.617	0.608
wilt	0.978	0.898	0.979	0.898	0.976	**0.926**	0.979	0.917

Worst-case robustness on real data, full table. Table 19 reports exact robust accuracy for all six datasets, the full Φ and GROOT budget grids, and the augmentation baseline, and Table 16 isolates the largest budget against TMS. Against impurity splitting the paired per-split difference at $\varepsilon = 0.1$ is positive on every split for banknote ($+2.3$ points), phoneme ($+4.1$), and spambase ($+38.4$), on four of five for kc1 ($+7.2$, one tie), and on three of five for wdbc ($+0.9$) and diabetes ($+1.3$). GROOT trained at 0.1 is at least as robust as TMS at $\varepsilon = 0.1$ on all six datasets (2.8 points ahead on average, split-level $p < 0.001$) at statistically equal clean accuracy ($p = 0.07$). GROOT trained at 0.02 ties TMS at $\varepsilon = 0.02$ ($p = 0.97$), and GROOT trained at 0.02 or 0.05 collapses on spambase at $\varepsilon = 0.1$ (0.23 and 0.29). The augmentation baseline ties TMS ($p = 0.23$), consistent with its average-case showing, with the geometric families and certificate semantics remaining its gaps.

When can the swap help? Two computable diagnostics. Proposition 5 gives the first, and it operationalizes as marginal-signal density: the fraction of features whose per-feature two-sample KS test is significant at Bonferroni-corrected 5%, computed from the training split in $O(nd \log n)$. On the nineteen datasets the density is 0.02 on madelon and 0.22–1.00 everywhere else, so any cutoff in $[0.05, 0.2]$ isolates the failure case, and we fix 0.1 on this benchmark. Two out-of-sample checks apply it unchanged. The first uses six constructed families (XOR, 3- and 4-parity, a multiplicative-interaction family, and two marginal-signal controls, each padded with noise probes): every family measures density below the cutoff. Falling back costs at most 3.2 accuracy points, on XOR, where both criteria stay near 0.6, and saves 4 points on the heavily diluted control. The second uses eight

Table 14: The Φ budget grid on the nineteen-dataset benchmark: clean accuracy and robustness AUC (protocol of Table 13; marks refer to AUC).

	$\Phi(0.02)$		$\Phi(0.05)$		$\Phi(0.1)$		TMS (std, R=0.1)	
dataset	clean	AUC	clean	AUC	clean	AUC	clean	AUC
diabetes	0.753	0.764	0.775	**0.784**	0.788	0.772	0.784	0.776
spambase	0.952	0.873	0.944	0.876	0.947	0.881	0.949	**0.883**
kc1	0.850	0.849	0.855	0.849	0.850	**0.852**	0.853	0.843
pc1	0.937	0.933	0.931	0.932	0.937	**0.935**	0.937	0.933
blood-transf.	0.773	0.764	0.747	0.763	0.751	**0.770**	0.742	0.763
ozone-8hr	0.942	0.939	0.941	**0.940**	0.938	0.936	0.943	**0.940**
phoneme	0.898	0.829	0.890	**0.830**	0.871	0.826	0.899	**0.830**
qsar-biodeg	0.845	0.819	0.842	0.820	0.845	**0.823**	0.845	**0.823**
wdbc	0.947	0.941	0.965	**0.955**	0.959	0.947	0.953	0.943
climate-crashes	0.926	**0.925**	0.926	**0.925**	0.920	0.922	0.926	**0.925**
pc4	0.890	0.892	0.900	**0.894**	0.890	0.889	0.895	0.886
pc3	0.900	**0.899**	0.893	0.894	0.891	0.895	0.896	0.897
kc2	0.828	0.827	0.828	0.822	0.841	**0.831**	0.847	**0.831**
MagicTelescope	0.874	0.829	0.868	0.829	0.862	**0.832**	0.873	0.830
banknote	0.990	0.965	0.990	**0.967**	0.990	**0.967**	0.990	**0.967**
eeg-eye-state	0.859	0.738	0.833	0.727	0.709	0.672	0.894	**0.745**
mozilla4	0.913	0.790	0.892	0.794	0.870	**0.804**	0.929	0.791
madelon	0.641	**0.628**	0.628	**0.628**	0.629	0.612	0.617	0.608
wilt	0.976	0.900	0.978	0.907	0.974	**0.930**	0.979	0.917

Table 15: Fine-scale benchmark datasets: clean accuracy and robustness AUC for the Φ budget grid and TMS (full grid in Table 14).

	Gini		$\Phi(0.02)$		$\Phi(0.1)$		TMS (R=0.1)	
	clean	AUC	clean	AUC	clean	AUC	clean	AUC
eeg-eye-state	**0.917**	0.607	0.859	0.738	0.709	0.672	0.894	**0.745**
mozilla4	**0.946**	0.781	0.913	0.790	0.870	**0.804**	0.929	0.791
phoneme	0.891	0.822	0.898	0.829	0.871	0.826	**0.899**	**0.830**

binary OpenML datasets outside the benchmark (Table 17). The three near-zero densities belong to the NIPS-2003 datasets arcene and dexter, built like madelon with noise probes, and to the shape-coded hill-valley task. Falling back is right on dexter, where TMS trails by 3.4 AUC points, and costs at most 1.3 points on the other two, where the criteria tie or sit near chance. Gisette, the third probe-padded dataset, keeps enough marginal signal (density 0.16) to stay above the cutoff, and there and on the four ordinary datasets TMS is within 1.7 points of Gini in either direction. Over the eight, the rule beats always using TMS by 0.5 clean and 0.4 AUC points. These are small datasets with three splits each, so the check bounds the rule's cost rather than measuring its gain precisely. Dilution with uninformative features lowers the density even when marginal signal exists, and that is the correct behavior: the noise-gap chasing that hurts the margin criterion grows with the noise-feature count, whether or not the signal is interactional. Falling back to impurity below the cutoff repairs madelon on both axes and strengthens the dataset-level robustness comparison ($+1.8$ AUC points, $p = 0.002$, mean clean cost 0.1 points). A runtime alternative does not work: an out-of-bag (OOB) guardrail cannot see this case, because on madelon the two criteria's OOB scores tie (0.60 vs. 0.59) while the full-ensemble test gap is four points. The counting criterion's advantage there is an ensemble-averaging effect that the smaller out-of-bag sub-ensembles cannot see, so the pre-training test is the right place to catch it. Below the cutoff even the hybrid criterion trails plain impurity (0.65 vs. 0.67 on madelon), so the rule falls back all the way rather than interpolating. The second is worst-case headroom: train the impurity forest and evaluate its exact robust accuracy at the deployment budget (the depth-6 MILP costs under a second per point), and set $\mathrm{headroom}(\varepsilon) = \mathrm{clean} - \mathrm{exact}(\varepsilon)$. Since the swapped model's robust accuracy is at most its own clean accuracy, the swap's worst-case gain is at most the headroom plus its clean-accuracy change. When the swap does not raise clean accuracy, the headroom alone is the bound. On the six datasets at $\varepsilon = 0.1$ the Gini forest's headroom ranges from 0.04 (banknote, wdbc) to 0.49 (spambase), the bound holds on every dataset, and TMS captures 11–83% of the headroom, including 79% of spambase's 0.49. The LP certificate can stand in for the exact evaluation here. It understates the Gini forest's robust accuracy by at most 0.011 at $\varepsilon \leq 0.05$ and by at most 0.11 at $\varepsilon = 0.1$ (kc1, Table 18), and an understated baseline only overstates the headroom, which keeps the gain bound valid.

Table 16: The largest fixed budgets vs. TMS on the six MILP datasets: clean and exact robust accuracy at $\varepsilon = 0.1$ (five-split means). At their own radius the matched budgets lead by up to seven points.

		diabetes	banknote	wdbc	kc1	phoneme	spambase
clean	$\Phi(0.1)$	0.74	0.99	**0.94**	0.87	0.85	0.90
	GROOT (0.1)	**0.75**	0.99	0.93	0.87	0.84	**0.91**
	TMS (R=0.1)	0.74	0.99	**0.94**	**0.88**	**0.86**	0.90
ε=0.1	$\Phi(0.1)$	0.65	**0.98**	0.90	0.86	0.79	**0.88**
	GROOT (0.1)	**0.70**	0.97	0.90	**0.87**	**0.81**	**0.88**
	TMS (R=0.1)	0.64	0.97	0.90	0.86	0.78	0.82

Table 17: The signal-density rule on eight binary OpenML datasets outside the benchmark, cutoff 0.1 unchanged (matched 25-tree forests, std units, three $70/30$ splits). Bold marks the higher robustness AUC.

					Gini		TMS (R=0.1)	
dataset	n	d	density	rule	clean	AUC	clean	AUC
arcene	200	9961	0.020	Gini	0.761	**0.769**	0.767	0.759
dexter	600	11035	0.001	Gini	0.906	**0.876**	0.869	0.841
hill-valley	1212	100	0.000	Gini	0.561	0.516	0.550	**0.529**
gisette	7000	4971	0.158	TMS	0.965	**0.960**	0.963	0.954
ionosphere	351	33	0.677	TMS	0.934	**0.928**	0.928	0.911
sonar	208	60	0.228	TMS	0.799	0.796	0.804	**0.813**
haberman	306	3	0.333	TMS	0.707	0.702	0.728	**0.710**
breast-w	699	8	1.000	TMS	0.967	0.962	0.968	0.962

Covariate shift and split stability on real data. Shift: test features moved by ± 0.1 training standard deviations on the nineteen datasets (matched 25-tree forests, std-unit TMS): mean accuracy 0.864 for TMS vs. 0.847 for Gini, TMS ahead on 14 of 19 (Wilcoxon $p = 0.12$). Stability: 100 bootstrap replicates of each training set, root stumps: mean feature agreement 0.75 vs. 0.69 ($p = 0.22$), mean validation-partition ARI 0.70 vs. 0.61 ($p = 0.029$).

Radius sensitivity on real data. Table 20 varies the cap under std-scaled noise (TM-Metric, 20 trees, 3 splits). The synthetic window $R \in [0.05, 0.2]$ is optimal or within 0.005 of optimal on all four datasets, and the smallest cap is always worst. Across the full nineteen-dataset benchmark (std units, protocol of Table 13), the window is flat: mean robustness AUC $0.849/0.849/0.854$ at $R = 0.05/0.1/0.2$, with the largest per-dataset difference between $R = 0.05$ and the default 0.1 at 0.012 ($p = 1.0$). The default is not the argmax of its own window ($R = 0.2$ is marginally better), which is what a prespecified rather than tuned value looks like.

Interaction-heavy data and a hybrid criterion. The madelon failure mode of Section 3.3 has a population statement:

Proposition 5 (Interaction-only distributions defeat every marginal criterion)**.** *If $\mu_+^{(j)} = \mu_-^{(j)}$ for every feature j, then $\Phi_t^{(j)}(r) = 0$ for all t, r, j, and by Theorem 1 every per-feature Gini gain is zero as well: any split statistic that is a functional of the per-feature class-conditional marginals (impurity, KS, Φ, $\mathcal{M}_p^R$) sits identically at its uninformative value, and threshold selection is driven by sampling noise alone.*

The statement is immediate (both brackets of (2) vanish), and it covers impurity and margin criteria alike. The two differ in how their sampling noise behaves. On madelon (480 of 500 features are noise probes, and the signal is XOR-like, so the informative marginals are nearly equal), a single scan of the full training sample ranks the most informative feature first under both criteria. Table 5 aggregates depths 0–2 across trees and bootstrap draws, where TMS already trails, and as nodes shrink the margin functional is drawn to spuriously wide gaps in the noise features faster than the counting functional. By depth six its selection rate on the twenty informative features reaches the chance rate $20/500$.

The hybrid criterion $\Delta_G + \mathcal{M}_1^R/(2R)$ is the equal-weight average of $2\Delta_G$ and $\mathcal{M}_1^R/R$, where Δ_G is the node's Gini decrease with $G(p) = 2p(1-p)$ and children weighted by their share of the node.

Table 18: Polynomial certificates vs. the exact MILP: the box certificate of Proposition 4, the LP relaxation of Appendix B, and the exact value at $\varepsilon = 0.1$, with the largest LP gap to the exact value at $\varepsilon \in \{0.02, 0.05\}$.

	Gini				TMS (R=0.1)			
	$\varepsilon = 0.1$			gap	$\varepsilon = 0.1$			gap
dataset	box	LP	exact	$\varepsilon \le 0.05$	box	LP	exact	$\varepsilon \le 0.05$
diabetes	0.61	0.62	0.62	0.001	0.63	0.63	0.64	0.001
banknote	0.93	0.94	0.94	0.000	0.97	0.97	0.97	0.000
wdbc	0.89	0.89	0.89	0.000	0.90	0.90	0.90	0.001
kc1	0.64	0.67	0.78	0.011	0.85	0.85	0.86	0.001
phoneme	0.71	0.73	0.74	0.005	0.77	0.77	0.78	0.004
spambase	0.36	0.38	0.44	0.011	0.82	0.82	0.82	0.001

Table 19: Exact robust accuracy (MILP) on real data: means over five splits (per-cell sd ≤ 0.07 except at $\varepsilon = 0.1$ for Gini on kc1 and spambase and for $\Phi(0.02)$ and $\Phi(0.05)$ on spambase, 0.08–0.16). Marks are per dataset and column; fully tied columns are unmarked.

dataset	model	clean	ε=0.02	ε=0.05	ε=0.1
diabetes	Gini	0.74	0.71	0.66	0.62
	$\Phi(0.02)$	0.74	0.71	0.67	0.63
	$\Phi(0.05)$	0.73	0.71	0.68	0.64
	$\Phi(0.1)$	0.74	0.72	0.69	0.65
	GROOT (0.02)	0.73	0.73	0.69	0.67
	GROOT (0.05)	**0.75**	0.74	**0.73**	0.67
	GROOT (0.1)	**0.75**	**0.75**	0.72	**0.70**
	aug. Gini (σ=0.1)	0.74	0.72	0.69	0.65
	TMS (R=0.1)	0.74	0.72	0.68	0.64
banknote	Gini	0.99	0.98	0.97	0.94
	$\Phi(0.02)$	0.99	**0.99**	0.98	0.96
	$\Phi(0.05)$	0.99	**0.99**	**0.99**	0.97
	$\Phi(0.1)$	0.99	**0.99**	**0.99**	**0.98**
	GROOT (0.02)	0.99	0.98	0.97	0.95
	GROOT (0.05)	0.99	0.98	0.97	0.95
	GROOT (0.1)	0.99	0.98	0.98	0.97
	aug. Gini (σ=0.1)	0.99	**0.99**	0.98	0.96
	TMS (R=0.1)	0.99	**0.99**	**0.99**	0.97
wdbc	Gini	0.93	0.92	0.91	0.89
	$\Phi(0.02)$	**0.94**	**0.93**	0.91	0.89
	$\Phi(0.05)$	**0.94**	**0.93**	0.91	0.89
	$\Phi(0.1)$	**0.94**	**0.93**	**0.92**	**0.90**
	GROOT (0.02)	**0.94**	**0.93**	0.91	0.88
	GROOT (0.05)	0.93	0.92	0.91	0.89
	GROOT (0.1)	0.93	0.92	0.91	**0.90**
	aug. Gini (σ=0.1)	0.93	0.92	0.91	0.89
	TMS (R=0.1)	**0.94**	0.92	0.91	**0.90**

dataset	model	clean	ε=0.02	ε=0.05	ε=0.1
kc1	Gini	0.87	0.87	0.86	0.78
	$\Phi(0.02)$	0.87	0.86	0.85	0.81
	$\Phi(0.05)$	0.87	0.87	**0.87**	0.84
	$\Phi(0.1)$	0.87	0.87	0.86	0.86
	GROOT (0.02)	0.87	0.87	0.86	0.86
	GROOT (0.05)	0.87	0.87	**0.87**	0.86
	GROOT (0.1)	0.87	0.87	**0.87**	**0.87**
	aug. Gini (σ=0.1)	0.87	0.87	0.86	0.85
	TMS (R=0.1)	**0.88**	**0.88**	**0.87**	0.86
phoneme	Gini	**0.87**	0.84	0.81	0.74
	$\Phi(0.02)$	**0.87**	**0.85**	**0.83**	0.76
	$\Phi(0.05)$	0.86	**0.85**	**0.83**	0.77
	$\Phi(0.1)$	0.85	0.83	0.81	0.79
	GROOT (0.02)	0.86	0.84	0.82	0.77
	GROOT (0.05)	0.84	0.84	0.82	0.80
	GROOT (0.1)	0.84	0.83	0.82	**0.81**
	aug. Gini (σ=0.1)	**0.87**	**0.85**	0.82	0.76
	TMS (R=0.1)	0.86	**0.85**	0.81	0.78
spambase	Gini	**0.93**	0.90	0.77	0.44
	$\Phi(0.02)$	0.91	0.90	0.70	0.36
	$\Phi(0.05)$	0.91	0.90	**0.89**	0.29
	$\Phi(0.1)$	0.90	0.90	**0.89**	**0.88**
	GROOT (0.02)	0.92	0.90	0.59	0.23
	GROOT (0.05)	0.90	0.90	**0.89**	0.29
	GROOT (0.1)	0.91	0.90	**0.89**	**0.88**
	aug. Gini (σ=0.1)	0.92	**0.91**	**0.89**	0.85
	TMS (R=0.1)	0.90	0.89	0.88	0.82

For binary problems both terms lie in $[0, 1]$, the second being the mean certified advantage over $[0, R]$, so they share a scale at every node without per-scan normalization. The hybrid interpolates the two criteria's behavior: on madelon it recovers most of the interaction-case gap ($0.61 \to 0.65$ against Gini's 0.67) while staying at or between the components elsewhere: 0.945 on spambase, 0.912 on phoneme (the best of the three), and 0.852 on qsar, with 50-tree forests and five-fold cross-validation. The same structure explains why the criterion value does not forecast cross-dataset gains: across the nineteen datasets the margin a Gini tree leaves on the table does not predict the forest-level robustness gain (Spearman correlation 0.23, $p = 0.34$), because a large margin surplus is informative only where the marginals are trustworthy, and madelon-like data produce surplus without signal.

Multiclass benchmark. Table 21: the one-vs-rest criterion stays within half a point of the matched Gini forest and 0.6 of scikit-learn's random forest on every dataset in clean accuracy, and carries the robustness edge of Corollary 2 over ($+1.7$ AUC points on pendigits, $+2.6$ on segment, ties elsewhere).

Runtime. A 20-tree binary forest on $4{,}600 \times 57$ inputs fits in under one second single-threaded, and the transport-margin scan runs within a factor of 1.1 of the matched Gini scan for binary problems and 2.3 for ten classes.

Table 20: Robustness AUC as the radius cap varies (std-scaled noise, TM-Metric).

dataset	R=0.02	0.05	0.1	0.2	0.5	∞
spambase	0.86	0.88	0.91	0.92	0.91	0.91
kc1	0.86	0.86	0.86	0.86	0.86	0.86
qsar-biodeg	0.84	0.85	0.85	0.86	0.85	0.85
wdbc	0.94	0.95	0.95	0.95	0.94	0.94

Table 21: Multiclass: accuracy (five-fold cross-validation, TM-Rank, 50 trees) and robustness AUC under std-scaled noise (std-unit TMS).

	accuracy			robustness AUC	
dataset	Gini (matched)	RandomForest	TM-forest	Gini	TMS
vehicle (K=4)	0.741	**0.747**	<u>0.745</u>	**0.691**	0.689
satimage (K=6)	<u>0.910</u>	**0.911**	0.905	0.871	**0.873**
segment (K=7)	<u>0.978</u>	**0.980**	0.976	0.877	**0.903**
optdigits (K=10)	**0.980**	<u>0.979</u>	0.978	0.962	0.962
pendigits (K=10)	**0.990**	<u>0.989</u>	<u>0.989</u>	0.952	**0.969**

G EXPERIMENTAL PROTOCOLS

Synthetic family. Feature A places each class at distance $U(0.004, c)$ from 0.5 on its own side, so it separates the classes perfectly with margins at most c. Feature B places the classes uniformly on $[0, 0.2]$ and $[0.8, 1]$, with an overlap fraction of each class moved to the other interval. Eight $\mathcal{N}(0.5, 0.25^2)$ noise features and Gaussian jitter of standard deviation 0.005 on A and B complete the design. Unless varied, every forest experiment uses 200 samples per class, $c = 0.05$, and overlap 0.05. The stump experiments of Section 3.1 use 150 samples per class, $c \in \{0.02, 0.05\}$, overlap 0, 0.05, or 0.1, and imbalance 1:1 or 1:5. Figure 5 crosses six values of c from 0.01 to 0.1 with overlaps $0, 0.02, \ldots, 0.08$ and four seeds at 150 samples per class, and the stability experiment uses 150 samples per class with jitter 0.01.

Forests. 15–25 trees for matched comparisons (50 for the multiclass accuracies and the hybrid criterion, 100 for Table 12), minimum leaf size 2–5, and no depth limit except depth 6 for the exact worst-case evaluations. Every feature is considered at every node on the synthetic families and in the exact worst-case evaluations, and $\sqrt{d}$ in all benchmark experiments on the nineteen binary and five multiclass datasets (Table 10 compares the two). Seeds are identical across criteria.

Noise. Uniform: i.i.d. Gaussian added to all features. Boundary-targeted: Gaussian ($\sigma = 0.15$) added only within 0.1 of the true boundary on the two informative features. Real data: Gaussian scaled by each feature's training standard deviation.

Datasets. Binary (Tables 12 and 13): the sixteen OpenML-CC18 datasets passing the mechanical filter (binary target, exclusively numeric features, no missing values, $500 \le n \le 25{,}000$, $d \le 500$) applied to the full suite (IDs 37, 44, 1049, 1050, 1063, 1067, 1068, 1462, 1464, 1485, 1487, 1489, 1494, 1510, 40983, 40994; the filter script will be released with the code), plus three numeric binary datasets outside the suite (1046, 1120, 1471) (Vanschoren et al., 2014; Bischl et al., 2021). Multiclass: IDs 28, 32, 36, 54, 182. Restricting the binary benchmark to the sixteen CC18 datasets gives the same picture as the full nineteen. Columns containing NaN are dropped, and labels are used as provided.

Statistics. The experimental unit is the dataset: fold scores are averaged per dataset, and paired Wilcoxon signed-rank tests, bootstrap confidence intervals (20,000 resamples), and the equivalence check (90% CI within ± 0.01) are computed across datasets, with Holm correction across the six baseline comparisons.

Normalization. All normalizations are dataset-level maps fitted on the training split: TM-Rank uses the pooled empirical CDF per feature; TM-Metric winsorizes to the $(0.01, 0.99)$ quantiles and scales affinely to $[0, 1]$; TM-Std standardizes by the per-feature mean and standard deviation (margins in σ units). A raw additive perturbation of size σ times the per-feature standard deviation corresponds to the per-feature normalized radius $\sigma \, \mathrm{std}_j / (Q_{0.99} - Q_{0.01})_j$ (about $\sigma/4.6$ for a Gaussian-like feature), so the sweep $\sigma \in [0, 0.4]$ probes normalized radii up to roughly 0.09, within the default cap $R = 0.1$. The worst-case experiments standardize features by the training standard deviation so that a scalar budget is exact per feature.

Real-data shift and stability. Shift: all features of the held-out test set moved by $+0.1$ and -0.1 training standard deviations, accuracies averaged. Stability: 100 bootstrap replicates of the training

split; root-stump feature agreement with the modal choice, and mean pairwise adjusted Rand index of the induced validation partitions of the held-out split over 60 random replicate pairs.

Table 22: Recommended configurations (signal density: Section 3.3).

setting	normalization	cap	why
generic tabular	rank	$R \to \infty$	scale-free, monotone-invariant
perturbation scale known	that unit	deployment scale	margins in the perturbation's unit
perturbation scale unknown	std	$R = 0.1$	automatic; flat over $R \in [0.05, 0.2]$
signal density below 0.1	impurity criterion		no marginal signal (Prop. 5)

Criterion defaults. Table 22 lists the recommended configurations. Perturbation-facing experiments: $\mathcal{M}_1^R$, $R = 0.1$ in the experiment's stated unit. Clean benchmarks: TM-Rank, $R \to \infty$. Fixed-budget baselines: $\Phi(r_0)$ shares the TM-forest architecture exactly; GROOT uses `groot-trees` with per-feature attack radius ε, matched tree count and leaf size.

Cheap baselines. Two-stage: Gini scores every candidate of the node's features. Among candidates within a relative tolerance of the best score, the split whose midpoint threshold sits in the widest empty gap is chosen (gaps in std units on real data). Augmentation: each tree trains on its bootstrap sample plus one copy jittered by $\mathcal{N}(0, (0.1\,\mathrm{std}_j)^2)$, fresh noise per tree. Hellinger: the distance between the class-conditional routing distributions, computed in the same scan. MAD normalization: $z = (x - \mathrm{median})/(1.4826\,\mathrm{MAD})$.

Certified evaluation. Leaf boxes are enumerated per tree. The certificate lets each tree independently reach any leaf whose box lies within the budget (a relaxation of the shared perturbation, hence valid) and checks the averaged probability margin. Exact robust accuracy solves the leaf-selection program of Appendix B with HiGHS, one program per wrong class, and the LP certificate solves its relaxation. Each real dataset uses five $70/30$ splits and the first 150 test points of each, and each synthetic family three seeds with 150 test points. All certified forests, GROOT included, have 20 trees of depth 6 with minimum leaf size 5, which keeps the program exact and tractable (≤ 1 s per instance).

Multiscale family. 60% of samples: label determined by a feature with a 0.6 gap; 40%: that feature uniform in the central band and the label determined by a second feature with margins in $[0.01, 0.08]$; one fragile feature separates all samples at margins ≤ 0.05; six noise features.