PENALIZED-WEIGHT ROBUST OPTIMAL TRANSPORT: FROM HARD CLIPPING TO SOFT COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust optimal transport (OT) methods reduce the influence of outliers, and several produce point-level outlier flags, but we know of no guarantee that these flags recover the contaminated observations. We introduce **PW-ROT** (Penalized-Weight Robust OT), which assigns each transport path a weight $w \in [0, 1]$ and penalizes $1 - w$ linearly. Eliminating the weights in closed form yields ordinary OT with a bounded ground-cost family $\rho_{\lambda,r}$, indexed by a threshold $\lambda > 0$ and a power $r \geq 1$: $r = 1$ is ROBOT's hard-clipped cost (up to its conventional factor of two in the threshold), and $r > 1$ gives smooth soft-compression costs. In a general parametric model, we show that the bounded cost down-weights separated outliers automatically: although no data are discarded and the outlier fraction is not an input, the PW-ROT criterion matches, up to a constant, a criterion that removes that fraction of mass from the *fitted model* (not from the data), exactly for the hard clip ($r = 1$) and up to an explicit error for $r > 1$. We bound the estimation error under arbitrary outliers, prove consistency of the hard clip with a fixed fraction of separated outliers in bounded-support models, and show that the flags recover the outlier set exactly with high probability whenever the estimation error is small relative to a separation margin. For n observations with Gaussian inliers in $\mathbb{R}^d$ of variance σ^2 and a point target, exact recovery holds for every member, including ROBOT, with a fixed fraction of outliers separated at the same scale, when $\lambda_n \asymp \sigma^2(\sqrt{d} + \sqrt{\log n})^2$; slower growth provably flags inliers for any estimator. The power r governs a trade-off: soft powers down-weight rather than discard outliers beyond the threshold, so their weights vary continuously with the cost, but separated outliers then bias the estimate toward themselves, which the hard clip avoids. Empirically, hard clipping is best for one-sided, well-separated outliers; soft compression has lower bias near the threshold and more stable threshold selection on images; and $r = 2$ is the most consistent default.

1 INTRODUCTION

Optimal transport (OT) compares probability distributions by the least costly way to move mass between them. It is used in generative modeling (Arjovsky et al., 2017; Genevay et al., 2018), domain adaptation, and statistics (Peyré & Cuturi, 2019). Because classical OT transports all mass, a small fraction of outlying observations can substantially alter the transport plan and the estimators built on it. Robust OT methods address this sensitivity by down-weighting or discarding mass. ROBOT (Mukherjee et al., 2021) relaxes the marginals in total variation, equivalently yielding the hard-clipped cost $\min(C, 2\lambda)$ for ground cost C and threshold $\lambda > 0$; trimmed OT (Nietert et al., 2022) removes a prescribed mass; and χ^2 reweighting (Balaji et al., 2020) and unbalanced OT (Chizat et al., 2018; Séjourné et al., 2023) penalize marginal deviations.

Although robustness is their primary purpose, several of these methods also yield signals that can be read as outlier flags. This raises a more specific question: *under what conditions do robust-OT flags recover the outlier set?* Existing results study robust transport values under contamination and establish concentration, minimax risk, or consistency (Mukherjee et al., 2021; Nietert et al., 2022; Ma et al., 2026; Staerman et al., 2021); they do not establish that the flags recover the outlier set exactly. The available flagging mechanisms also have limitations: trimming requires the amount of mass to discard, χ^2 weights require a separate cutoff because inlier weights are not fixed at one, and

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.

unbalanced OT does not directly return an outlier set. We give such conditions for a new robust-OT family whose flags need no extra cutoff.

Our approach transfers a familiar idea from robust regression to OT. Penalized mean-shift and penalized-weight procedures attach an observation-level shift or weight with a sparsity-inducing penalty and thereby deliver both a robust estimator and an explicit outlier set (She & Owen, 2011; Kong et al., 2018; Gao & Fang, 2016; Li et al., 2023). PW-ROT instead places a penalized weight on each *transport path*. A linear penalty allows the weight to be eliminated in closed form (a standard half-quadratic step; Black & Rangarajan, 1996; Meng et al., 2017), and the induced bounded cost $\rho_{\lambda,r}$ ranges from hard clipping ($r = 1$) to smooth soft compression ($r > 1$). We study the resulting minimum-distance estimator in a general parametric model, from location families to the Gaussian spread target used by our solver (generative models fit the framework, but we do not verify its identifiability conditions for them). Our analysis yields two main results. First, the bounded cost performs *automatic down-weighting*: for separated outliers, the PW-ROT criterion equals, up to a constant, a partial-transport criterion that removes the true outlier fraction of mass from the fitted model rather than from the data, exactly for $r = 1$ and up to an explicit error for $r > 1$, although this fraction is not an input and no observation is discarded. Second, the flags then recover the outlier set exactly for every $r \geq 1$, with high probability, whenever the estimation error is below a separation margin; with a vanishing outlier fraction, this holds when the threshold λ_n grows with the sample size n at the inliers' extreme-value scale. For Gaussian inliers in $\mathbb{R}^d$ with variance σ^2 and a point target, it holds with a fixed fraction of separated outliers at $\lambda_n \asymp \sigma^2(\sqrt{d} + \sqrt{\log n})^2$, and no slower threshold avoids flagging inliers. The analysis also exposes a trade-off across powers: the weights of soft powers vary continuously across the threshold, but their nonzero tail weights let separated outliers bias the estimate toward themselves, which the hard clip avoids.

Contributions.

1. **Automatic down-weighting and exact outlier recovery in general models (Section 4).** At separated parameters, the PW-ROT criterion equals, up to a constant, that of an estimator that trims the unknown outlier fraction ϵ of mass from the model (not the data), although ϵ is not an input; the equality is exact for $r = 1$ when the model puts no mass near the outliers and holds up to an explicit error for $r > 1$. We bound the estimation error under arbitrary outliers, prove fixed-fraction consistency of the hard clip for bounded-support models, and prove exact recovery for every $r \geq 1$ at a threshold rate that we show to be necessary; for the Gaussian point target, recovery holds with a fixed outlier fraction, and for $r = 1$ the estimator is the oracle inlier mean.
2. **A closed-form robust-OT family containing ROBOT (Section 3).** A per-path penalized weight reduces exactly to ordinary OT with the bounded cost $\rho_{\lambda,r}$, so standard OT solvers apply. Its flags are read off the fitted model at a probability level $\alpha \in (0, 1]$; for a point target, they need no cutoff beyond λ.
3. **The role of r (Section 4.2).** Soft powers incur a bias from separated outliers, with a bound that grows with r (the price of softness), which the hard clip avoids; in exchange, their weights vary continuously across the threshold.
4. **Empirical characterization of the trade-off (Section 5).** Smaller powers are best for one-sided, well-separated contamination, and larger powers near the threshold; on contaminated MNIST, soft powers have a wider stable range of thresholds. These observations motivate, but do not prove, the default $r = 2$.

Organization. Section 2 positions PW-ROT among robust-OT and robust-regression methods, Section 3 derives the cost family, Section 4 develops the theory, Section 5 reports the experiments, and Section 6 discusses limitations.

2 Related Work

We review robust OT, its use for outlier detection, and the robust-regression ideas that PW-ROT carries over to transport.

Robust OT. Unbalanced OT penalizes deviations from the marginals using a divergence (Liero et al., 2018; Chizat et al., 2018; Séjourné et al., 2023). ROBOT uses a total-variation marginal relaxation

and is equivalent to OT with cost $\min(C, 2\lambda)$ (Mukherjee et al., 2021). Other approaches include χ^2-constrained reweighting (Balaji et al., 2020), one-sided KL relaxation (Le et al., 2021), partial OT (Figalli, 2010; Chapel et al., 2020), and trimmed transport (Álvarez-Esteban et al., 2008; Nietert et al., 2022; 2023). Further variants use median-of-means duals (Staerman et al., 2021), minibatch unbalanced OT (Fatras et al., 2021), and E-ROBOT (Ma et al., 2026; La Vecchia & Liu, 2026). PW-ROT instead obtains a bounded ground cost from a weight on each path and transports all mass.

Robust OT and outlier detection. ROBOT provides binary signals tied to its clipping threshold. Trimming identifies the removed mass but requires its amount as input (Nietert et al., 2022 locate it by an elbow under separated supports); χ^2 weights are not exactly one on inliers and require a post-hoc cutoff; unbalanced OT returns no outlier set, whereas Blanchet et al. (2024) identify points moved within a concave-cost OT budget. In PW-ROT, the optimal path weight is exactly one below λ, so a point target flags exactly the points with cost above λ.

Penalized mean shifts and weights. In regression, observation-specific shifts with sparsity penalties yield robust estimation and outlier identification simultaneously. Examples include ℓ_1 shifts (McCann & Welsch, 2007; Gannaz, 2007; Antoniadis, 2007), Θ-IPOD (She & Owen, 2011), and procedures with statistical, efficiency, breakdown, and detection-consistency guarantees (Katayama & Fujisawa, 2017; Kong et al., 2018). Penalized-weight methods also have detection-consistency theory (Gao & Fang, 2016; Gao & Feng, 2018; Li et al., 2023). Eliminating a latent weight is the half-quadratic construction (Geman & Reynolds, 1992; Black & Rangarajan, 1996), also used in self-paced learning (Kumar et al., 2010; Meng et al., 2017). Shape-parameterized robust losses are related in spirit (Huber, 1964; Hampel et al., 1986; Beaton & Tukey, 1974; Barron, 2019); see Appendix A for further discussion, existing robust-OT theory and our positioning. In summary, PW-ROT places the penalized weight on transport paths rather than on observations or marginals; the elimination step is standard, and our contribution is the cost family, its OT interpretation, and its analysis.

3 Penalized-Weight Robust Optimal Transport

Let $\mathbf{a} \in \Delta^n$ and $\mathbf{b} \in \Delta^{n'}$ be discrete probability measures, where $\Delta^n = \{\mathbf{a} \in \mathbb{R}^n_+ : \sum_i a_i = 1\}$ is the probability simplex; let $C = (c_{ij}) \in \mathbb{R}^{n\times n'}_+$ be a cost matrix, with c_{ij} the cost of moving mass from source point i to target point j (the path (i, j)); and let $\mathcal{F}(\mathbf{a}, \mathbf{b}) = \{\mathbf{\Pi} = (\pi_{ij}) \in \mathbb{R}^{n\times n'}_+ : \mathbf{\Pi}\mathbf{1}_{n'} = \mathbf{a},\ \mathbf{\Pi}^\top \mathbf{1}_n = \mathbf{b}\}$ denote the transport polytope, with $\mathbf{1}_{n'}$ the all-ones vector. PW-ROT assigns a weight $w_{ij} \in [0, 1]$ to each path and solves

$$\min_{\mathbf{\Pi}\in\mathcal{F}(\mathbf{a},\mathbf{b}),\, 0\le w_{ij}\le 1} \sum_{i,j} \pi_{ij} \left[\frac{1}{r} w_{ij}^r c_{ij} + \lambda(1 - w_{ij})\right], \qquad r \ge 1, \quad \lambda > 0. \tag{1}$$

The penalty is integrated against the transport plan, so only paths carrying mass are penalized; Appendix B gives the continuous formulation. For a fixed plan the objective separates over paths, and minimizing each term over its weight gives the following closed form.

Theorem 3.1 (Closed-form weights and equivalent cost)**.** *For a scalar $C \ge 0$, the inner problem $\min_{0\le w\le 1}\{r^{-1}w^r C + \lambda(1 - w)\}$ has solution*

$$w^\star(C) = \min\left\{1, \left(\frac{\lambda}{C}\right)^{1/(r-1)}\right\}$$

when $r > 1$. When $r = 1$, it has solution $w^\star(C) = \mathbb{1}\{C \le \lambda\}$. Its value is

$$\rho_{\lambda,r}(C) = \begin{cases} \frac{1}{r}C, & C \le \lambda, \\ \lambda + \frac{1-r}{r}\lambda^{\frac{r}{r-1}} C^{-\frac{1}{r-1}}, & C > \lambda,\ r > 1, \\ \lambda, & C > \lambda,\ r = 1. \end{cases} \tag{2}$$

Consequently, equation 1 is equivalent to the standard OT problem $\min_{\mathbf{\Pi}\in\mathcal{F}(\mathbf{a},\mathbf{b})}\langle \mathbf{\Pi}, \rho_{\lambda,r}(C)\rangle$, where $\rho_{\lambda,r}(C)$ is applied entrywise and $\langle \mathbf{\Pi}, A\rangle = \sum_{i,j} \pi_{ij} A_{ij}$. The same equivalence holds for general measures. The cost $\rho_{\lambda,r}$ is nonnegative, nondecreasing, and bounded by λ.

Theorem 3.1 has computational and statistical consequences. Computationally, PW-ROT is ordinary OT with the entrywise-transformed cost matrix $\rho_{\lambda,r}(c_{ij})$, so any standard solver, including Sinkhorn (Cuturi, 2013), applies with its usual convergence guarantees. Statistically, $w^\star = 1$ exactly when $C \le \lambda$: paths within the threshold keep full weight, and only paths beyond it are down-weighted.

Hard and soft members. Every member (Figure 1) caps the cost at λ and satisfies $w^\star = 1$ exactly when $C \le \lambda$ (Theorem 3.1). For point-target location estimation, where the target is the point mass $\mathbf{b} = \delta_\mu$ at $\mu \in \mathbb{R}^d$ and each data point x_i has one path, of cost $\|x_i - \mu\|^2$, every member therefore flags $\{w^\star < 1\} = \{C > \lambda\}$ with no separate cutoff; see Section 4. When $r > 1$, a flagged path receives a graded weight in $(0, 1)$. As $r \to 1^+$, $\rho_{\lambda,r}(C) \to \min(C, \lambda)$ and $w^\star \to \mathbb{1}\{C \le \lambda\}$, recovering hard rejection.[1] For $r > 1$, the weight beyond the threshold decays as $(\lambda/C)^{1/(r-1)}$, more slowly as r grows. In particular, $r = 2$ gives $w^\star = \min(1, \lambda/C)$ and $\rho_{\lambda,2}(C) = \lambda - \lambda^2/(2C)$ for $C > \lambda$.

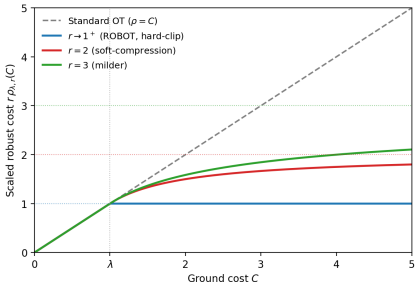

Figure 1: The PW-ROT family at $\lambda = 1$, shown as $r\,\rho_{\lambda,r}(C)$ so that every member coincides with standard OT (dashed) for $C \le \lambda$; the scaling changes neither the weights, the flags nor the estimator. Beyond the threshold, ROBOT's hard clip (blue, $r = 1$) is flat at λ, while $r = 2$ (red) and $r = 3$ (green) compress smoothly toward their caps $r\lambda$ (dotted).

4 Theoretical Analysis

We analyze PW-ROT as a minimum-distance estimator in a general parametric model. We first fix the data model and the flag rule and then show that the bounded cost down-weights outliers automatically (Proposition 4.1). Section 4.1 turns this into estimation and exact-recovery guarantees with explicit rates, and Section 4.2 explains how the power r trades the continuity of the weights against bias. The statements are informal; the proofs are in Appendix C, which points to the full statements in Appendices D (general models) and E (point target, with explicit constants).

Setting. The sample $x_1, \dots, x_n \in \mathbb{R}^d$ splits into inliers, indexed by $\mathcal{I}$ and i.i.d. from a clean law P, and arbitrary, possibly adversarial outliers, indexed by $\mathcal{O}$, with $\mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$ disjoint and $m = |\mathcal{O}| = \epsilon_n n$ (the $\mathcal{O}\cup\mathcal{I}$ model of Ma et al., 2026); ϵ_n is the outlier fraction. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical law of the sample and $P_{\mathcal{I}}$ that of the inliers, where δ_x is the point mass at x. A model is a family $\{\nu_\theta : \theta \in \Theta\}$ of probability measures, $\Theta \subseteq \mathbb{R}^p$; it may be a location or location–scale family, the point target $\nu_\theta = \delta_\theta$, the spread target $\theta + \mathcal{N}(0, I_d)$ of our solver, or, subject to the identifiability conditions of Assumption 2, a generative model $g_\theta \# Z$, the law of $g_\theta(Z)$ for a latent variable Z. For a ground cost $c(x, y) \ge 0$ (squared Euclidean in our experiments; $\|\cdot\|$ is the Euclidean norm throughout), write $\rho = \rho_{\lambda,r}$ and $\mathrm{OT}_\rho(P_1, P_2) = \min_{\pi \in \Pi(P_1,P_2)} \int \rho(c(x, y))\, d\pi(x, y)$ where $\Pi(P_1, P_2)$ is the set of couplings of P_1 and P_2. The estimator is $\hat\theta_n \in \arg\min_\theta \mathrm{OT}_\rho(P_n, \nu_\theta)$. Its target is the *clean* parameter $\theta_0 = \arg\min_\theta \mathrm{OT}_\rho(P, \nu_\theta)$, defined by the uncontaminated law (it may depend on (λ, r)); Ma et al. (2026) instead target the minimizer for the contaminated law. With $A_\lambda(x) = \{y : c(x, y) \le \lambda\}$, the set of targets that x reaches on a full-weight path, and a level $\alpha \in (0, 1]$, the *estimated outlier set*, used throughout Section 4.1, is

$$\hat{\mathcal{O}} = \{i : \nu_{\hat\theta_n}(A_\lambda(x_i)) < \alpha\}.$$

Since $A_\lambda(x) = \{y : w^\star(c(x, y)) = 1\}$, the flags are read off the optimal path weights of Theorem 3.1. For a point target, x_i has a single path, with weight $w_i = w^\star(c(x_i, \hat\theta_n))$, and $\hat{\mathcal{O}} = \{i : w_i < 1\} = \{i : c(x_i, \hat\theta_n) > \lambda\}$. For the Gaussian spread target, with draws $y_1, \dots, y_{n'}$ from $\nu_{\hat\theta_n}$ and path weights $w_{ij} = w^\star(c(x_i, y_j))$, the reachable mass is estimated by the fraction of full-weight paths, and $\hat{\mathcal{O}} = \{i : \frac{1}{n'}\sum_j \mathbb{1}\{w_{ij} = 1\} < \alpha\}$: a point is flagged when fewer

[1] The analytic endpoint is $\min(C, \lambda)$. Experiments use ROBOT in its native parameterization $\min(C, 2\lambda)$; since λ is selected independently for each method by the same rule, this factor is absorbed into the selected threshold.

than a fraction α of its paths keep full weight; $\alpha = 0.05$ gives the 95% rule of our experiments (Lemma E.14).

All guarantees rest on comparing PW-ROT with a criterion that knows the outlier fraction and trims that much mass from the *model*, not from the data; PW-ROT itself transports all data. Let $\hat{Q}_n(\theta)$ be the cost of transporting the inliers into ν_θ after trimming mass ϵ_n from the model, $\hat{Q}_n(\theta) = \min\{\mathrm{OT}_\rho(P_\mathcal{I}, \nu') : \nu' \text{ a probability measure with } (1-\epsilon_n)\nu' \le \nu_\theta \text{ setwise}\}$, so that $(1-\epsilon_n)\nu'$ is the part of ν_θ that is kept and ν' its normalization.

Proposition 4.1 (Sandwich by the trimmed criterion). *Let $\Delta \ge 0$ and $\tau \in [0,1]$. At every θ at which ν_θ puts at most τ of its mass at cost below $\lambda + \Delta$ from the outliers,*

$$(1-\epsilon_n)\hat{Q}_n(\theta) + \epsilon_n\lambda - \gamma_n \;\le\; \mathrm{OT}_\rho(P_n, \nu_\theta) \;\le\; (1-\epsilon_n)\hat{Q}_n(\theta) + \epsilon_n\lambda, \qquad \gamma_n = \epsilon_n\big(\lambda - \rho(\lambda+\Delta)\big) + \lambda\tau.$$

For $r = 1$ and $\tau = 0$ the two sides coincide, so on any region where this separation holds and which contains the minimizers of both criteria, PW-ROT is the model-side trimmed estimator at the true outlier fraction ϵ_n, although ϵ_n is not an input to the estimator.

(Proof: Appendix C.2.) The upper bound states that PW-ROT never costs more than transporting the inliers into a trimmed model and paying λ for each outlier; the lower bound states that, when the model puts little mass near the outliers, it cannot cost much less. For the hard clip, the bounds coincide: the outliers contribute the constant $\epsilon_n\lambda$ and are absorbed by the trimmed mass, so minimizing PW-ROT amounts to minimizing the trimmed criterion at the true fraction ϵ_n. For $r > 1$, the gap γ_n is the price of the soft tail and shrinks as the separation Δ grows. The separation condition holds with $\tau = 0$ when the model has bounded support and every outlier lies at cost at least $\lambda + \Delta$ from it (e.g. uniform location models); for Gaussian models, $\tau > 0$ but is at most $\exp\{-\frac{1}{2}(D - \sqrt{\lambda+\Delta} - \|\theta - \theta_0\| - \sqrt{d})^2\}$ when every outlier is at distance at least $D \ge \sqrt{\lambda+\Delta} + \|\theta - \theta_0\| + \sqrt{d}$ from θ_0 (Corollary D.10). Unlike trimmed OT (Nietert et al., 2022), PW-ROT requires no trimming level, and unlike minimum Hellinger or γ-divergence estimation (Beran, 1977; Fujisawa & Eguchi, 2008), it requires no densities.

4.1 Estimation and exact outlier recovery

Common setting. Throughout, the robust cost is Lipschitz in the data point, $|\rho(c(x,y)) - \rho(c(x',y))| \le L\|x - x'\|$ (for $c(x,y) = \|x-y\|^2$, $L = 2\sqrt{\lambda}/r$), and the model is continuous, identifiable, and coercive (Assumptions 1, 2). The Lipschitz condition holds for the squared-Euclidean cost because the cost is capped; continuity and coercivity hold for location families with a coercive cost, including the point and Gaussian spread targets; and identifiability of θ_0 holds, e.g., for well-specified identifiable models and for the point target with Gaussian inliers (Appendix D, Examples; Lemma E.5). Let $J_0 = \min_\theta \mathrm{OT}_\rho(P, \nu_\theta)$, assumed below λ (automatic for $r > 1$, since then $\rho < \lambda$); $M(\theta) = \mathrm{OT}_\rho(P, \nu_\theta) - J_0$, the excess clean criterion; and $b(s) = \sup\{\|\theta - \theta_0\| : M(\theta) \le s\}$, the largest distance from θ_0 of a parameter whose excess criterion is at most s (the identifiability modulus; $b(s) \to 0$ as $s \to 0$). The inlier fluctuation is $T_n = \mathrm{OT}_{d_\lambda}(P_\mathcal{I}, P)$, the OT value under the cost $d_\lambda(x,x') = \min\{L\|x - x'\|, \lambda\}$; for fixed λ it vanishes almost surely without moment conditions, and by McDiarmid's inequality, for $\delta \in (0,1)$, $T_n \le t_n(\delta) = \mathbb{E}T_n + \lambda\sqrt{\log(1/\delta)/(2(n-m))}$ with probability at least $1-\delta$. Probabilities are over the inlier sample only; the outliers are arbitrary and may depend on it. For a margin $t > 0$, the margin sets $\mathcal{G}_\lambda(t) = \{x : \nu_\theta(A_\lambda(x)) \ge \alpha \text{ whenever } \|\theta - \theta_0\| \le t\}$ and $\mathcal{B}_\lambda(t) = \{x : \nu_\theta(A_\lambda(x)) < \alpha \text{ whenever } \|\theta - \theta_0\| \le t\}$ contain the points that are never, respectively always, flagged while θ stays within t of θ_0; requiring inliers in $\mathcal{G}_\lambda(t)$ and outliers in $\mathcal{B}_\lambda(t)$ is a gap condition (for a point target, an empty annulus between radii $\sqrt{\lambda} \mp t$ around θ_0). For bounded inliers, a fixed threshold suffices: the hard clip is then consistent with a fixed fraction of separated outliers, and the flags recover $\mathcal{O}$ exactly (Theorem D.21, Appendix D). We state the general result, which also covers unbounded inliers; for these, λ must grow with n.

Theorem 4.2 (Growing threshold, any inliers). *Let $\lambda = \lambda_n$, fixed or growing with n, with all quantities of the Common setting, including θ_0, evaluated at λ_n.*

(a) Estimation, any outliers. *If $2\lambda_n\epsilon_n + 2T_n < \lambda_n - J_0$, then, deterministically, $\|\hat{\theta}_n - \theta_0\| \le b(2\lambda_n\epsilon_n + 2T_n)$. If, in addition, $M(\theta) \ge \kappa \min\{\|\theta - \theta_0\|^2, t_0^2\}$ for some $\kappa, t_0 > 0$ (Assumption 3) and $2\lambda_n\epsilon_n + 2T_n < \kappa t_0^2$, then $\|\hat{\theta}_n - \theta_0\| \le ((2\lambda_n\epsilon_n + 2T_n)/\kappa)^{1/2}$.*

(b) Exact recovery. *If $P(X \notin \mathcal{G}_{\lambda_n}(t)) \le \delta/n$ for $X \sim P$, all outliers lie in $\mathcal{B}_{\lambda_n}(t)$, $2\lambda_n\epsilon_n + 2t_n(\delta) < \lambda_n - J_0$ and $b(2\lambda_n\epsilon_n + 2t_n(\delta)) \le t$, then $\mathbb{P}(\hat{\mathcal{O}} = \mathcal{O}) \ge 1 - 2\delta$, whatever the positions of the outliers in $\mathcal{B}_{\lambda_n}(t)$.*

(c) Necessity of growth. *If $\mathbb{P}(\hat{\theta}_n \in \mathcal{U}) \to 1$ for a neighborhood $\mathcal{U}$ of θ_0 and $(n-m)P(X \in S_n) \to \infty$, where S_n is the set of points flagged at every $\theta \in \mathcal{U}$, then some inlier is flagged with probability tending to one. For Gaussian inliers $\mathcal{N}(\theta_0, \sigma^2 I_d)$, the point target and $c(x,y) = \|x-y\|^2$, this holds for every estimator plugged into the flag rule whenever $\lambda_n \le (1-\eta)\sigma^2 \max\{2\log n, d\}$ for some $\eta > 0$.*

What Theorem 4.2 says. (Proof: Appendix C.3.) Part (a) bounds the bias for any outlier configuration: outliers enter only through $\lambda_n\epsilon_n$, because each moves the criterion by at most λ_n/n. Its side condition is a breakdown bound, $\epsilon_n + T_n/\lambda_n < \frac{1}{2}(1 - J_0/\lambda_n)$; the curvature condition holds, e.g., for the Gaussian point target. For the squared-Euclidean cost, $\mathbb{E}T_n \le L\,\mathbb{E}\mathcal{W}_1(P_{\mathcal{I}}, P)$, where $\mathcal{W}_1$ is the 1-Wasserstein distance (OT with cost $\|x-y\|$) and $\mathbb{E}\mathcal{W}_1$ is $O(n^{-1/2})$ for $d=1$, $O(n^{-1/2}\log n)$ for $d=2$ and $O(n^{-1/d})$ for $d \ge 3$ under moment conditions (Fournier & Guillin, 2015); hence, with probability $1-\delta$ and while the bracket stays below $\kappa t_0^2/2$ (true for large n under the conditions below),

$$\|\hat{\theta}_n - \theta_0\| \le \left(\tfrac{2}{\kappa}\right)^{1/2}\Big(\underbrace{\lambda_n\epsilon_n}_{\text{outliers}} + \underbrace{L\,\mathbb{E}\mathcal{W}_1(P_{\mathcal{I}}, P)}_{\text{inlier sampling}} + \underbrace{\lambda_n\sqrt{\log(1/\delta)/n}}_{\text{concentration}}\Big)^{1/2}.$$

Consistency follows (for a target and curvature constants that do not depend on λ_n, e.g. the Gaussian point target; Corollary D.16) if $\lambda_n\epsilon_n \to 0$, $\sqrt{\lambda_n}\,\mathbb{E}\mathcal{W}_1 \to 0$ and $\lambda_n\sqrt{\log n/n} \to 0$; the second condition reads $\lambda_n = o(n)$ for $d=1$, $o(n/\log^2 n)$ for $d=2$ and $o(n^{2/d})$ for $d \ge 3$, and binds for $d \ge 5$, while the third is dimension-free. The confidence level enters only through the concentration term: with $\delta = n^{-a}$, $a > 0$, the bound holds with probability at least $1 - n^{-a}$ and that term is $\lambda_n\sqrt{a\log n/n}$; for fixed λ and $d \ge 3$, $\|\hat{\theta}_n - \theta_0\| = O(\sqrt{\epsilon_n} + n^{-1/(2d)})$ with probability at least $1 - n^{-a}$ (the concentration term is of smaller order). Part (b) turns estimation into detection: recovery requires the error to be below the margin t, not to vanish. Its three conditions read as follows. *Outliers in $\mathcal{B}_{\lambda_n}(t)$*: for the point target, every outlier lies farther than $\sqrt{\lambda_n}+t$ from θ_0. *Error below the margin*, $b(2\lambda_n\epsilon_n + 2t_n(\delta)) \le t$: since $b(s) \to 0$ as $s \to 0$, it holds for large n whenever $\lambda_n\epsilon_n \to 0$ and $t_n(\delta) \to 0$; with a fixed outlier fraction, it requires the bias $b(2\lambda_n\epsilon)$ to stay below t, which is why Corollary 4.3 uses a direct argument. *Inliers in $\mathcal{G}_{\lambda_n}(t)$*: this is where growth enters, since $\mathcal{G}_\lambda(t)$ grows with λ, so λ_n must reach the inliers' extreme-value scale; for Gaussian inliers and a point target this means $\sqrt{\lambda_n} \ge t + \sigma(\sqrt{d} + \sqrt{2\log(n/\delta)})$, i.e. $\lambda_n \asymp \sigma^2(\sqrt{d} + \sqrt{\log n})^2$ (spread target: Remark D.14), and by (c) no slower growth suffices. The conditions of (c) are mild: the first holds whenever $\hat{\theta}_n$ is consistent, e.g. under (a), and the second says that the expected number of inliers flagged at every $\theta \in \mathcal{U}$ diverges; for Gaussian inliers and the point target, it yields the explicit condition $\lambda_n \le (1-\eta)\sigma^2\max\{2\log n, d\}$, for every estimator. Combining (a) and (b), the general theory admits $\sigma^2(\sqrt{d} + \sqrt{\log n})^2 \lesssim \lambda_n \ll \min\{\epsilon_n^{-1}, \sqrt{n/\log n}, (\mathbb{E}\mathcal{W}_1)^{-2}\}$ (Corollary D.16), which requires a vanishing fraction, $\epsilon_n = o(1/\log n)$. With $\delta_n = n^{-a}$, $a > 0$, recovery holds with probability $\ge 1 - 2n^{-a}$ at the same scale, and almost surely for all large n if $a > 1$ (Remark D.17). A fixed fraction with a growing threshold requires a model-specific argument:

Corollary 4.3 (Gaussian point target, fixed fraction)**.** *Let the inliers be $\mathcal{N}(\mu_0, \sigma^2 I_d)$ and $\nu_\theta = \delta_\theta$ (we write $\mu, \hat{\mu}, \mu_0$ for $\theta, \hat{\theta}_n, \theta_0$), fix $\epsilon_n = \epsilon < 1/2$ and $r \ge 1$, let $c(x,y) = \|x-y\|^2$, and let $u_n = \sigma(\sqrt{d} + \sqrt{2\log(2n^2)})$, a radius. For suitable numerical constants $c_\lambda > 2$ and C_D depending on (r, ϵ), if $\lambda_n = c_\lambda^2 u_n^2$ and every outlier lies at distance at least D_n from μ_0, where $D_n \ge C_D u_n$, then with probability at least $1 - 1/n$, $\hat{\mathcal{O}} = \mathcal{O}$ and*

$$\|\hat{\mu} - \mu_0\| \le \sigma\frac{\sqrt{d} + \sqrt{2\log(2n)}}{\sqrt{n-m}} + \mathbb{1}\{r > 1\}\,\frac{\epsilon}{1-\epsilon}\sqrt{\lambda_n}\Big(\frac{\sqrt{\lambda_n}}{D_n - b_n}\Big)^{\frac{r+1}{r-1}},$$

with $b_n \asymp \sqrt{\lambda_n}$ a localization radius (Corollary E.9; not the modulus b); for $r = 1$, $\hat{\mu}$ is exactly the inlier mean on the same event.

What Corollary 4.3 says. (Proof: Appendix C.4.) For the hard clip, a fixed fraction of outliers costs nothing: the error attains the oracle rate $O_p(\sigma\sqrt{d/n})$, and the flags are exact. The probability

$1 - 1/n$ corresponds to $\delta = 1/n$; more generally, with $u_n = \sigma(\sqrt{d} + \sqrt{2\log(2n/\delta)})$ the statement holds with probability at least $1 - \delta$ (Theorem E.7), so $\delta = n^{-a}$ gives $1 - n^{-a}$ at the same scale, with only the constants changing. For $r > 1$, the second term—the price of softness, whose order is attained (Proposition E.11)—vanishes if and only if $\lambda_n = o(D_n^{(r+1)/r})$. Recovery is thus guaranteed for outliers at distance of order $\sigma(\sqrt{d} + \sqrt{\log n})$, whereas consistency of a soft power requires $D_n \gg \lambda_n^{r/(r+1)}$, e.g. $(\log n)^{2/3}$ for $r = 2$ (d fixed): soft powers need outliers to recede faster than detection alone requires.

4.2 The role of r

The power r acts through the tail of the weight $w^\star$, which is continuous in the cost for $r > 1$ and jumps from 1 to 0 at $C = \lambda$ for $r = 1$. *The price of softness*: by Theorem D.21(a) and Corollary 4.3, the soft tail lets outliers beyond the threshold retain some pull, producing a bias whose bound grows with r (explicitly for the point target); this bias vanishes for the hard clip and, for the point target, is attained by one-sided outliers (Proposition E.11). Each outlier pulls toward itself, and the pulls add in the exact error decomposition of Theorem E.7(ii): they accumulate when the outliers lie on one side and can cancel when they surround θ_0. *Continuity*: the same soft tail retains part of the restoring pull of inliers just beyond the threshold, which the hard clip discards, and makes the flag set change gradually with λ. Our theory does not quantify this second effect; Section 5 examines both.

5 Experiments

We examine the two effects of Section 4.2 on synthetic mean estimation, where detection can be scored exactly, and on contaminated MNIST, where the threshold must be selected from data.

Mean-estimation setup. Inliers follow $\mathcal{N}(\mu, I_5)$, with $n = 300$, 100 trials per cell, and outlier proportion $\epsilon_n = m/n \in \{0, 0.1, 0.2, 0.3\}$ (with n fixed we write ϵ). Outliers are shifted by k in each coordinate, either in a common direction, $\mathcal{N}(\mu - k\mathbf{1}_5, I_5)$ with $\mathbf{1}_5$ the all-ones vector, or with independent random coordinate signs; the latter surround μ and tend to cancel in the sample mean. We report $k \in \{0.5, 2, 4\}$; $k = 1$ appears in Appendix F. For each power $r \in \{1, 1.5, 2, 3\}$, λ is selected by trimmed validation: over the thresholds of a data-derived grid that flag fewer than 50% of points, we minimize the held-out squared error after discarding its largest $\zeta = 35\%$ (Algorithm 2). The only contamination information used is that ζ exceeds the largest ϵ. Baselines are Nietert's trimmed OT, ROBOT-χ^2 (the χ^2-budget reweighting of Balaji et al., 2020), coordinate-wise and geometric medians, median-of-means (MoM), and the sample mean (BASE). *Masking* is the fraction of outliers missed; *swamping*, the fraction of inliers flagged.

Clean data. Without outliers, robustness costs little: at $\epsilon = 0$, the four powers have errors 0.121–0.126 versus 0.121 for BASE (Nietert: 0.161; ROBOT-χ^2: 0.166; for both, the selector chooses zero trimming in only a minority of trials).

One-sided, well-separated contamination. The top row of Figure 2 shows the regime in which the price of softness (Section 4.2) favors the hard clip. At $k = 4$, the error increases with r: 0.146–0.152 for ROBOT, 0.159–0.160 for $r = 1.5$, 0.166–0.174 for $r = 2$, and 0.200–0.384 for $r = 3$. At $k = 2$ and $\epsilon = 0.3$, it increases from 0.284 at $r = 1$ to 0.481, 0.658, and 0.810 at $r = 1.5, 2, 3$; Nietert's method (0.230) and ROBOT-χ^2 (0.221) perform best in that cell. This ordering matches, heuristically, the bias mechanism of Theorem D.21(a), Corollary 4.3 and Proposition E.11, whose separation conditions are not met at these shifts.

Symmetric contamination. Placing the outliers around μ removes most of their pull on the mean: in the bottom row of Figure 2, all errors lie between 0.12 and 0.29, and the four PW-ROT powers are within 0.016 of one another. Across all 12 contaminated cells (including $k = 1$, Appendix F) and the clean cell, $r = 2$ has the lowest or second-lowest mean error among the powers in every cell; by paired t-tests on shared data sets ($p < 0.05$, unadjusted for multiplicity; Appendix F) it is never significantly worse than another power and is significantly better than ROBOT in five cells. The geometric median has lower mean error than every power at $k \geq 2$ in this design (significantly in 4 of 6 cells). By contrast, under one-sided contamination $r = 2$ ranks third of four powers in 12

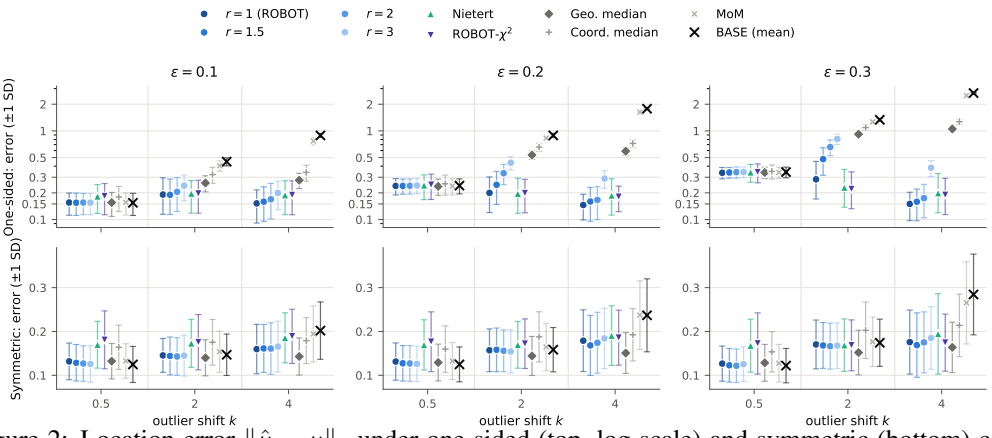

Figure 2: Location error $\|\hat{\mu} - \mu\|_2$ under one-sided (top, log scale) and symmetric (bottom) contamination: mean $\pm$ one standard deviation over 100 trials, by contamination fraction ϵ (columns) and outlier shift k. The PW-ROT family is in blue, from $r = 1$ (dark) to $r = 3$ (light); within each k, methods are offset horizontally (no lines join different k). Trials are paired across methods, so differences can be significant despite overlapping bars. Masking and swamping for all methods are in Figures 5–6 (Appendix F).

Table 1: Outlier detection by the PW-ROT family at $\epsilon = 0.2$: masking / swamping (means over 100 trials; bold, lowest per column). All powers, including ROBOT ($r = 1$), use the same flag rule and threshold selector. Detectors that we constructed from the trimming and χ^2 baselines, whose papers do not propose outlier flags, are compared in Table 4.

	One-sided		Symmetric	
power	k=2	k=4	k=2	k=4
r=1 (ROBOT)	0.04 / **0.23**	**0.00** / **0.04**	0.44 / 0.08	0.29 / **0.01**
r=1.5	**0.02** / 0.27	**0.00** / 0.11	0.40 / 0.11	**0.25** / 0.04
r=2	**0.02** / 0.30	**0.00** / 0.14	**0.39** / 0.10	0.27 / 0.03
r=3	**0.02** / 0.31	**0.00** / 0.06	0.51 / **0.07**	0.32 / **0.01**

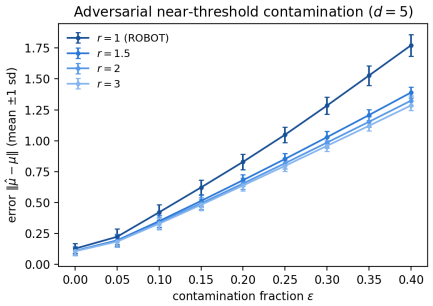

Figure 3: Near-threshold adversarial contamination ($d = 5$, $n = 400$, 300 trials, fixed λ; mean $\pm$ one standard deviation). The hard clip ($r = 1$) has the largest observed bias for every $\epsilon > 0$.

of 13 cells and is significantly worse than ROBOT in 10. The evidence for $r = 2$ is thus consistency across geometries rather than a large gain in any one.

Detection. Table 1 compares only methods whose papers propose outlier detection: PW-ROT and ROBOT. Under one-sided contamination, all powers detect the far outliers ($k = 4$, masking 0.00), and the hard clip swamps least. Under symmetric contamination, every power misses a quarter or more of the outliers, plausibly because cancelling outliers exert little pressure toward a small threshold. No power dominates; for $r = 2$ the table shows the same consistency as its error—the lowest or tied-lowest masking in three of four columns and never the worst—at the price of more swamping than the hard clip on one-sided outliers (0.30, 0.14 vs. 0.23, 0.04).

5.1 Near-threshold adversarial contamination

Outliers placed at the threshold probe the second effect of Section 4.2, the continuity of the weights. We place outliers tightly around squared distance λ, where a single point exerts the largest pull (Lemma E.1(e)), fix λ at the 0.85 quantile of χ^2_5, and compute the point-target estimator by iteratively

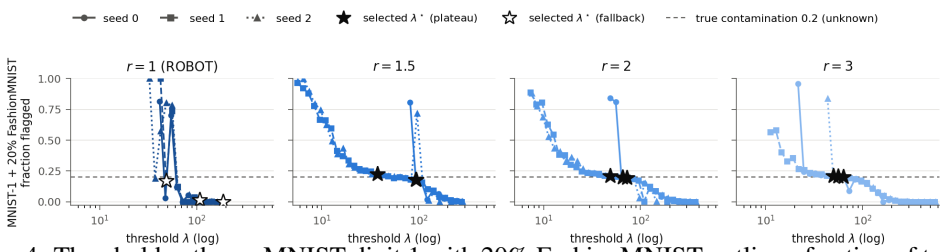

Figure 4: Threshold paths on MNIST digit 1 with 20% FashionMNIST outliers: fraction of training images flagged as λ varies, one line per seed. Stars mark selected thresholds; hollow stars indicate fallback selection. Soft powers have a plateau near the true fraction 0.2, whereas the hard clip is unstable.

reweighted means with weights $w^\star(\|x_i-\mu\|^2)^r$, initialized at the coordinate-wise median ($n = 400$, 300 trials). The ordering reverses: at every $\epsilon > 0$, the bias decreases in r, with errors 1.77, 1.39, 1.32 and 1.29 for $r = 1, 1.5, 2, 3$ at $\epsilon = 0.4$ (Figure 3). Only the soft powers retain part of the inliers' pull beyond the threshold, an effect our theory does not cover.

5.2 Contaminated images: soft compression and threshold selection

We next ask whether the soft tail helps on images. We consider two contaminated-MNIST settings: digit-0 inliers with 20% all-white images, and digit-1 inliers with 20% FashionMNIST class-9 images. For each $r \in \{1, 1.5, 2, 3\}$, we fit a robust semi-discrete OT solver with a neural generator to $n = 6000$ images, with three random seeds. Thresholds are selected by a plateau-gated selector: among thresholds whose flag sets are stable across neighboring values (adjacent Jaccard overlap ≥ 0.9), it chooses the one minimizing a trimmed Sinkhorn-$\mathcal{W}_2$ criterion (entropic 2-Wasserstein), and otherwise minimizes the criterion alone (Appendix H). We report detection only and make no claim about image quality or robust generation; there is no non-robust baseline (Appendix G).

Figure 4 shows the FashionMNIST setting, where the outliers are closest to the inliers. For $r > 1$, the flagged fraction reaches the true contamination level 0.2 over a broad range of λ; across both settings, the selector found this plateau and chose a threshold within it in 16 of 18 soft-power runs. For the hard clip on FashionMNIST, the flagged fraction instead changes sharply between adjacent grid values: its adjacent flag sets have median Jaccard overlap ($|A \cap B|/|A \cup B|$) 0.00 in every seed, versus 0.49–0.93 for the soft powers (Table 5); the trimmed $\mathcal{W}_2$ criterion shows the same contrast. White images are easier: every power has a flat criterion (Figure 8).

All methods have low swamping in these runs (at most 0.03). On the far white outliers, $r \in \{1.5, 2\}$ detects every outlier in every seed, whereas ROBOT misses all outliers in one seed and $r = 3$ in another. On the closer FashionMNIST outliers, masking decreases with r: 0.692 for ROBOT and 0.079, 0.025, and 0.009 for $r = 1.5, 2, 3$. Only $r = 2$ yields a plateau in all six image runs, with masking at most 0.05 in each.

Which power to use? No power is universally best. For unknown contamination geometry, we recommend $r = 2$: it is never the worst on point error, the most consistent under symmetric contamination, and the only power with a plateau in every image run.

6 Limitations and Conclusion

Limitations. The theory covers the exact minimizer at a deterministic threshold, not the stochastic solver or data-selected thresholds, and requires more separation than the experiments use. Constructed trimming detectors often achieve higher F_1 (harmonic mean of precision and recall; Table 4); the experiments cover $d = 5$ and MNIST with three seeds; and r is not selected from data. Because λ is a threshold rather than a scale, PW-ROT targets separated contamination rather than gap-free heavy tails; generation is outside our scope (Appendix G).

Conclusion. PW-ROT connects ROBOT's hard clip to soft compression through penalized path weights. Its bounded cost down-weights separated outliers automatically, its flags recover the outlier set exactly under the stated conditions, and the power r trades weight continuity against bias.

AI USE STATEMENT

We used generative AI tools, including coding assistants, to write and debug experiment code, run and aggregate experiments, search and summarize related literature, assist with mathematical derivations, and draft and edit the manuscript. The research question, methods, mathematical formulations, and final claims were determined by the authors. All AI-assisted code was tested; reported results were verified against experiment logs and code; mathematical derivations and proofs were independently checked by the authors; citations were checked against their sources; and all text was reviewed by the authors. The authors take responsibility for the final content, including the text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Algorithms 1 and 2 in Appendix H specify the estimator and selector. Appendix H also lists settings, data generation, and baselines; Appendices C, D and E contain all assumptions and proofs. Anonymized supplementary code includes the per-trial log used to generate the mean-estimation figures and Tables 1–4, together with the corresponding plotting scripts.

REFERENCES

Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. Trimmed comparison of distributions. *Journal of the American Statistical Association*, 103(482): 697–704, 2008.

Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(2):358–375, 2011.

Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. Similarity of samples and trimming. *Bernoulli*, 18(2):606–634, 2012.

T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955.

Anestis Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 214–223, 2017.

François Bachoc, Alberto González-Sanz, Jean-Michel Loubes, and Yisha Yao. Wasserstein spatial depth. *arXiv preprint arXiv:2411.10646*, 2024.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jonathan T. Barron. A general and adaptive robust loss function. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334, 2019. doi: 10.1109/CVPR.2019.00446.

Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76(12):1298–1302, 2006.

Albert E. Beaton and John W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, 1977.

Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19 (1):57–91, 1996.

Jose Blanchet, Jiajin Li, Markus Pelger, and Greg Zanotti. Automatic outlier rectification via optimal transport. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Eustasio del Barrio and Carlos Matrán. Rates of convergence for partial mass problems. *Probability Theory and Related Fields*, 155(3–4):521–542, 2013. doi: 10.1007/s00440-011-0406-z.

Eustasio del Barrio, Hristo Inouzhe, and Carlos Matrán. On approximate validation of models: A Kolmogorov–Smirnov-based approach. *TEST*, 29:938–965, 2020. doi: 10.1007/s11749-019-00691-1.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

David L. Donoho and Richard C. Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

Kilian Fatras. On making optimal transport robust to all outliers. *arXiv preprint arXiv:2206.11988*, 2022.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 3186–3197, 2021.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010. doi: 10.1007/s00205-008-0212-7.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Irène Gannaz. Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, 17(4):293–310, 2007.

Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2019.

Xiaoli Gao and Yixin Fang. Penalized weighted least squares for outlier detection and robust regression. *arXiv preprint arXiv:1603.07427*, 2016.

Xiaoli Gao and Yang Feng. Penalized weighted least absolute deviation regression. *Statistics and Its Interface*, 11(1):79–89, 2018.

Donald Geman and George Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):367–383, 1992.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *PMLR*, pp. 1608–1617, 2018.

Alberto González-Sanz, Shunan Sheng, Bohan Wu, and Marco Avella Medina. The influence function of transport-based quantiles. *arXiv preprint arXiv:2607.19080*, 2026.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J. Huber. Finite sample breakdown of M- and P-estimators. *The Annals of Statistics*, 12(1): 119–126, 1984.

Shota Katayama and Hironori Fujisawa. Sparse and robust linear regression: An optimization algorithm and its statistical properties. *Statistica Sinica*, 27:1243–1264, 2017. doi: 10.5705/ss.202015.0179.

Dehan Kong, Howard D. Bondell, and Yichao Wu. Fully efficient robust estimation, outlier detection and variable selection via penalized regression. *Statistica Sinica*, 28(2):1031–1052, 2018.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Davide La Vecchia and Hang Liu. E-ROBOT: A dimension-free method for robust statistics and machine learning via Schrödinger bridge. *Statistics and Computing*, 36(4):157, 2026.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Khang Le, Huy Nguyen, Quang M. Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiaqi Li, Linglong Kong, Bei Jiang, and Wei Tu. High-dimensional outlier detection and variable selection via adaptive weighted mean regression. *arXiv preprint arXiv:2306.13850*, 2023.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211 (3):969–1117, 2018.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Yiming Ma, Hang Liu, Davide La Vecchia, and Matthieu Lerasle. Inference via robust optimal transportation: Theory and methods. *International Statistical Review*, 94(2):412–444, 2026.

Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. *Robust Statistics: Theory and Methods (with R)*. Wiley, 2nd edition, 2019.

Lauren McCann and Roy E. Welsch. Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis*, 52(1):249–257, 2007.

Deyu Meng, Qian Zhao, and Lu Jiang. A theoretical understanding of self-paced learning. *Information Sciences*, 414:319–328, 2017.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015.

Debarghya Mukherjee, Aritra Guha, Justin Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 7850–7860, 2021.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *PMLR*, pp. 11691–11719, 2022.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Robust estimation under the Wasserstein distance. *arXiv preprint arXiv:2302.01237*, 2023.

Davy Paindaveine and Riccardo Passeggeri. On the robustness of semi-discrete optimal transport. *The Annals of Applied Probability*, 36(4):2970–2997, 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Elvezio Ronchetti. Robustness aspects of optimal transport. In Yan Liu, Junichi Hirukawa, and Yoshihide Kakizawa (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, pp. 445–453. Springer, 2023.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. In Emmanuel Trélat and Enrique Zuazua (eds.), *Numerical Control: Part B*, volume 24 of *Handbook of Numerical Analysis*, pp. 407–471. Elsevier, 2023. doi: 10.1016/bs.hna.2022.11.003.

Yiyuan She and Art B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. When OT meets MoM: Robust estimation of Wasserstein distance. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *PMLR*, pp. 136–144, 2021.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Yehuda Vardi and Cun-Hui Zhang. The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43:355–386, 1937.

Yikai Xu, Zhao Chen, and Jian Huang. Wasserstein filtering: A sample selection method for robust distribution learning. *arXiv preprint arXiv:2608.13418*, 2026.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283, 2022.

A Extended Related Work

This appendix extends Section 2: it covers unbalanced and partial transport, robust OT in minibatches and generation, existing robust-OT theory, latent-weight constructions, and the penalized mean-shift literature whose recovery theory we adapt, and it closes with our positioning.

Unbalanced and partial transport. The Hellinger–Kantorovich / entropy-transport theory (Liero et al., 2018) and the scaling algorithms of Chizat et al. (2018) are the basis of unbalanced OT; Séjourné et al. (2019) add debiased Sinkhorn divergences and Séjourné et al. (2023) survey the area. For partial transport, Caffarelli & McCann (2010) and Figalli (2010) study existence, uniqueness and free-boundary regularity, and Chapel et al. (2020) give exact solvers and a positive-unlabeled application. Álvarez-Esteban et al. (2011) establish uniqueness and computation of optimal incomplete (trimmed) plans. In each case the mass to discard, or its penalty, sets how outliers are treated. In PW-ROT that role is played by the threshold λ and the shape r of the ground cost.

Robust OT under minibatches, near outliers and in generation. Fatras et al. (2021) show that the unbalanced OT value saturates in a Dirac outlier while the balanced value grows with it. Fatras (2022) observe that cost truncation fully transports outliers lying close to the target, and propose classifier-based reweighting. Xu et al. (2026) discard a fraction of samples chosen by a Wasserstein criterion and prove minimax optimality under bounded covariance. On the generative side, WGAN (Arjovsky et al., 2017) and Sinkhorn-divergence models (Genevay et al., 2018) are sensitive to contaminated training data. Balaji et al. (2020) robustify them by marginal reweighting, and Gao et al. (2019) connect GANs with depth-based robust estimation under Huber contamination. Our MNIST study concerns detection. We make no image-quality claim, and, as noted in Section 6, a bounded ground cost alone cannot robustify a flexible generator.

Existing robust-OT theory. Robust-OT theory has bounded robust values (Mukherjee et al., 2021; Balaji et al., 2020; Fatras et al., 2021), and established concentration, consistency, or minimax guarantees (Ma et al., 2026; La Vecchia & Liu, 2026; Staerman et al., 2021; Nietert et al., 2023). Nietert et al. (2022) recover a clean transport distance under separated supports. Influence functions and breakdown analyses are available for some OT-based quantiles and depths (Paindaveine & Passeggeri, 2026; González-Sanz et al., 2026; Bachoc et al., 2024); a general robustness theory for OT remains open (Ronchetti, 2023). We are not aware of influence-function, gross-error-sensitivity, or breakdown results for minimum-Kantorovich location estimators with robust costs of this kind, nor of a stated threshold rate for exact recovery of robust-OT flags.

Positioning. PW-ROT places the penalized weight on transport paths rather than observations or marginals. The elimination step itself is standard; the contribution is the cost family, its OT interpretation, and its analysis. Our experimental baselines are the coordinate-wise median, geometric median (Weiszfeld, 1937; Vardi & Zhang, 2000; Minsker, 2015), median-of-means (Nemirovsky & Yudin, 1983; Lugosi & Mendelson, 2019), Nietert's robust OT, and ROBOT-χ^2; on MNIST the powers are compared with ROBOT ($r = 1$). We do not pursue high-dimensional minimax optimality (Diakonikolas et al., 2019).

Robustness theory for OT-based estimators. Minimum-Kantorovich estimators are consistent, also under misspecification (Bassetti et al., 2006; Bernton et al., 2019). For the ROBOT cost, Ma et al. (2026) prove concentration with an additive contamination term and consistency without moments, and La Vecchia & Liu (2026) give dimension-free rates for the entropic version. None of these works derives an influence function or breakdown point for the estimator. Classical tools for global minimizers of bounded losses, namely the automatic robustness of minimum-distance functionals (Donoho & Liu, 1988), finite-sample breakdown of bounded-ρ M-estimators (Huber, 1984) and the influence function (Hampel, 1974; Hampel et al., 1986; Maronna et al., 2019), are what our breakdown-type bound (Theorem D.8(c)) and influence function (Proposition E.3) adapt to the OT objective. Breakdown and influence results now exist for OT-based quantiles and medians (Paindaveine & Passeggeri, 2026; González-Sanz et al., 2026) and Wasserstein depth (Bachoc et al., 2024). These concern transport maps and depth functionals, not minimum-distance estimation with a robust cost.

Latent weights and robust losses with a shape parameter. In the half-quadratic construction (Geman & Reynolds, 1992; Black & Rangarajan, 1996), a robust loss is the minimum over an auxiliary weight of weight$\times$loss plus a penalty on the weight. The truncated quadratic (skipped

mean) is the standard example. Self-paced learning writes the same step as a latent loss $F_\lambda(\ell) = \min_v v\ell + f(v,\lambda)$, for a loss $\ell \geq 0$, a latent weight $v \in [0,1]$ and a regularizer f (Kumar et al., 2010; Meng et al., 2017). With $u = w^r$, PW-ROT's inner problem is of this form with regularizer $f(u;\lambda) = \lambda(1-u^{1/r})$; the hard regularizer of Kumar et al. (2010) gives $\min(\ell,\lambda)$, the $r = 1$ member. For $r > 1$ the tail of $\rho_{\lambda,r}(t^2)$ satisfies $\lambda - \rho_{\lambda,r}(t^2) \asymp t^{-2/(r-1)}$, which matches the tail of Barron's general loss (Barron, 2019) with shape parameter $\alpha_{\mathrm{B}} = -2/(r-1)$. Thus $r = 2$ has the Geman–McClure tail, $r \to \infty$ approaches the Cauchy (logarithmic) end, and $r \to 1^+$ the hard clip. Barron's losses give weight below 1 to every nonzero residual, whereas $\rho_{\lambda,r}$ is exactly quadratic, with weight 1, on $C \leq \lambda$. That quadratic core is what yields cutoff-free flags. Unlike Huber's loss (Huber, 1964), whose score (the derivative of the loss) is monotone, and Tukey's biweight (Beaton & Tukey, 1974), which is redescending with a finite rejection point (a residual size beyond which the score is zero), the score of $\rho_{\lambda,r}$ for $r > 1$ is bounded and decays to 0 without vanishing; only $r = 1$ has a finite rejection point (Hampel et al., 1986).

Related penalized mean-shift theory. She & Owen (2011) tune the threshold by BIC and leave nonasymptotic recovery bounds open; Kong et al. (2018) obtain full efficiency ($\mathbb{P}(\hat{\gamma} = 0) \to 1$ on clean data, for the estimated mean-shift vector $\hat{\gamma}$), breakdown $1/2$ and detection consistency, using adaptive weights from a robust initial fit; Li et al. (2023), building on the PWLAD weights of Gao & Feng (2018), prove $\mathbb{P}(\hat{\mathcal{O}} = \mathcal{O}) \to 1$ for their estimated outlier set $\hat{\mathcal{O}}$ under sub-Gaussian noise with $\sqrt{\log n}$-scale tuning. Our recovery theory for location follows the same structure. It adds an explicit nonasymptotic probability and adversarial outliers, and it shows that the $\log n$ growth of λ_n is necessary.

B Continuous Formulation

We record the PW-ROT problem for general probability measures; Theorem 3.1 covers this case.

Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$. For $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and a cost $c : \mathcal{X} \times \mathcal{Y} \to [0,\infty)$, PW-ROT is

$$V_1 = \inf_{\pi \in \Pi(\mu,\nu),\ 0 \leq w(x,y) \leq 1} \iint \Big[\tfrac{1}{r} w(x,y)^r c(x,y) + \lambda(1 - w(x,y)) \Big] d\pi(x,y), \tag{3}$$

over measurable w, and Theorem 3.1 states $V_1 = V_2 := \inf_{\pi \in \Pi(\mu,\nu)} \iint \rho_{\lambda,r}(c(x,y))\, d\pi(x,y)$. The bounded cost also caps the distance under ϵ-contamination: for $\tilde{\mu} = (1-\epsilon)\mu + \epsilon\mu_c$ with an arbitrary contaminating law μ_c, a mixture-plan argument gives $\mathrm{OT}_\rho(\tilde{\mu},\nu) \leq \min\{(1-\epsilon)\mathrm{OT}_\rho(\mu,\nu) + \epsilon\lambda,\ \lambda\}$.

C Proofs of the Main-Text Results

This appendix proves each result of Sections 3 and 4. The proofs of the results of Section 4 assemble the full statements of Appendices D (general models) and E (point target), to which each subsection points for the details.

C.1 Proof of Theorem 3.1

The proof eliminates the weights path by path: with the plan fixed, the objective separates, and each one-dimensional problem has the closed form stated in Theorem 3.1.

Proof of Theorem 3.1. The proof follows the same structure for both the discrete and continuous cases by analyzing the inner minimization over the weights.

Discrete Case: Consider the objective function in equation 1 (the discrete version from the main text). For a fixed transport plan $\mathbf{\Pi}$, the objective is separable with respect to the weights w_{ij}. For any path (i,j) where $\pi_{ij} > 0$, we need to find the optimal weight $w_{ij}^\star$ by solving the inner minimization problem:

$$\min_{0 \leq w \leq 1} \ell(w;C) = \left[\frac{1}{r} w^r C + \lambda(1-w) \right] \tag{4}$$

where $C = c_{ij}$.

- **Case 1:** $r > 1$**.** For $C > 0$, $\ell(\cdot; C)$ is strictly convex (for $C = 0$ it is linear and minimized at $w = 1$). The unconstrained minimum is $w^\star_{\text{unc}} = (\lambda/C)^{1/(r-1)}$, the minimizer without the constraint $w \leq 1$. Applying the constraint $w \leq 1$, the optimal weight is $w^\star_{ij} = \min(1, (\lambda/C)^{1/(r-1)})$. Substituting $w^\star_{ij}$ back into $\ell(w; C)$ yields $\rho_{\lambda,r}(C)$ as defined in Equation equation 2 for $r > 1$.
- **Case 2:** $r = 1$**.** The problem is linear: $\min_{0\leq w\leq 1}[w(C - \lambda) + \lambda]$. The minimum is C if $C \leq \lambda$ (at $w = 1$), and λ if $C > \lambda$ (at $w = 0$). This is $\min(C, \lambda)$, matching $\rho_{\lambda,1}(C)$.

Since the optimal value of the inner minimization for $\pi_{ij}\ell(w_{ij}; c_{ij})$ is $\pi_{ij}\rho_{\lambda,r}(c_{ij})$, substituting this back into the objective of equation 1 (value V_1) yields that of the OT problem with cost $\rho_{\lambda,r}$ (value V_2). The constraints on $\mathbf{\Pi}$ remain the same. Therefore, the optimal values V_1 and V_2 are equal for the discrete case.

Continuous Case: Consider the objective function in equation 3. For a fixed transport plan $\pi \in \Pi(\mu, \nu)$, we want to find the optimal measurable weight function $w^\star(x, y)$. The objective can be written as:

$$\iint_{\mathcal{X}\times\mathcal{Y}} \ell(w(x, y); c(x, y))d\pi(x, y)$$

where $\ell(w; C)$ is the same inner objective function as defined in equation 4. To minimize the integral, we can minimize the integrand $\ell(w(x, y); c(x, y))$ pointwise for almost every (x, y) with respect to the measure π.

For a fixed (x, y), let $C = c(x, y)$. The minimization problem $\min_{0\leq w\leq 1} \ell(w; C)$ is identical to the one solved in the discrete case. The optimal value is $\rho_{\lambda,r}(C) = \rho_{\lambda,r}(c(x, y))$. The pointwise minimizer $w^\star(c(x, y))$ is a continuous function of the measurable cost, hence measurable, so it is admissible.

Therefore, the infimum of the objective in equation 3 is achieved when we substitute the pointwise optimal value back into the integral:

$$V_1 = \inf_{\pi\in\Pi(\mu,\nu)} \iint_{\mathcal{X}\times\mathcal{Y}} \rho_{\lambda,r}(c(x, y))d\pi(x, y)$$

This is exactly the OT problem with cost $\rho_{\lambda,r}$ (V_2 in Appendix B). Thus, $V_1 = V_2$ also holds for the continuous case. Finally, $\rho_{\lambda,r}$ is non-negative, non-decreasing and bounded by λ by Lemma E.1(a) (Appendix E). □

C.2 Proof of Proposition 4.1

Proposition 4.1 is Proposition D.7 at a single θ: its separation condition is $\nu_\theta(U) \leq \tau$ for $U = \bigcup_{i\in\mathcal{O}}\{y : c(x_i, y) < \lambda + \Delta\}$, and $\mathrm{OT}_\rho(P_n, \nu_\theta) = F_n(\theta)$. The upper bound needs no separation. For $r = 1$, $\rho(\lambda + \Delta) = \lambda$, so with $\tau = 0$ the two bounds coincide and $F_n = (1 - \epsilon_n)\hat{Q}_n + \epsilon_n\lambda$ on the separated region; minimizing either side over that region gives the same minimizers.

C.3 Proof of Theorem 4.2

All results below are stated for a fixed threshold and are applied at each n with $\lambda = \lambda_n$, which is why every quantity of the Common setting is evaluated at λ_n. *Part (a).* Lemma D.5(i) gives $\sup_\Theta |F_n - J| \leq \lambda_n\epsilon_n + T_n$ for every outlier configuration, and Lemma D.5(iii) gives existence of the minimizer when $2\lambda_n\epsilon_n + 2T_n < \lambda_n - J_0$. Theorem D.8(a) (with $\xi_n = 0$) then gives $M(\hat{\theta}_n) \leq 2\lambda_n\epsilon_n + 2T_n$, hence the bound through b, and Theorem D.8(b) gives the curvature rate. The almost-sure convergence $T_n \to 0$ (for fixed λ) and the McDiarmid tail $T_n \leq t_n(\delta)$ are Lemma D.5(iv). *Part (b).* This is Theorem D.15(iii) with $\xi_n = 0$: on the event $\{T_n \leq t_n(\delta)\}$ part (a) gives $\|\hat{\theta}_n - \theta_0\| \leq t$, the inlier condition and a union bound give that every inlier lies in $\mathcal{G}_{\lambda_n}(t)$ with probability at least $1 - \delta$ (Theorem D.15(ii)), and Theorem D.15(i) turns these two events into $\hat{\mathcal{O}} = \mathcal{O}$. *Part (c).* The first statement is Proposition D.19; the Gaussian point-target statement, which needs no localization of the estimator, is Proposition E.10. *Rates and the threshold window* discussed after the theorem are Corollary D.16, Remark D.14 (Gaussian margin sets) and Remark D.17 (polynomial confidence levels); the 1-Wasserstein rates are those of Fournier & Guillin (2015).

C.4 Proof of Corollary 4.3

Take $\delta = 1/n$ in Lemma E.6, so that $\tau = u_n$ there. Corollary E.9 shows that $\lambda_n = c_\lambda^2 u_n^2$ and $D_n \geq C_D u_n$ imply the conditions (L) and (S) of Theorem E.7; that theorem gives the error decomposition and bound (and $\hat{\mu} = \bar{x}_I$ for $r = 1$) on the event of Lemma E.6, and Corollary E.8(b) gives $\hat{\mathcal{O}} = \mathcal{O}$ on the same event. For a general δ, the same argument with $u_n = \sigma(\sqrt{d} + \sqrt{2\log(2n/\delta)})$ gives probability at least $1 - \delta$.

D General Framework: Full Statements and Proofs

This appendix gives the precise statements and proofs behind Section 4. Section D.1 fixes the cost, the data model, the target and the flags; Section D.2 bounds the robust OT value; Section D.3 proves the estimation results, including automatic down-weighting and fixed-fraction consistency; and Section D.4 proves exact recovery and the rate window for λ_n.

D.1 Setting

Cost and robust OT Let $\mathcal{X} \subseteq \mathbb{R}^d$ be closed, and let $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a continuous, symmetric ground cost with $c(x, y) = 0$ if and only if $x = y$; the main examples are $c(x, y) = \|x - y\|^2$ and $c(x, y) = \|x - y\|$. For $r \geq 1$ and $\lambda > 0$ let $\rho = \rho_{\lambda,r}$ be the PW-ROT cost ($\rho(C) = C/r$ for $C \leq \lambda$, $\lambda + \frac{1-r}{r}\lambda^{r/(r-1)}C^{-1/(r-1)}$ for $C > \lambda$ and $r > 1$, $\min(C, \lambda)$ for $r = 1$), and set

$$h(x, y) = \rho\big(c(x, y)\big), \qquad 0 \leq h \leq \lambda, \quad h(x, x) = 0.$$

For probability measures μ, ν on $\mathcal{X}$, $\mathrm{OT}_\rho(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int h \, d\pi$ is the PW-ROT value (Theorem 3.1), as in Section 4; h is only a shorthand for the pointwise cost. For any other cost k on $\mathcal{X} \times \mathcal{X}$ we write $\mathrm{OT}_k(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int k \, d\pi$ (e.g. OT_c, OT_{d_λ}). $\mathrm{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$.

Assumption 1 (Lipschitz cost)**.** *There is $L = L_\lambda < \infty$ with $|h(x, y) - h(x', y)| \leq L\|x - x'\|$ for all x, x', y (by symmetry also in the second argument). Put $d_\lambda(x, x') = \min\{L\|x - x'\|, \lambda\}$ and $T_\lambda = \mathrm{OT}_{d_\lambda}$.*

For $c = \|\cdot\|^2$, $L = 2\sqrt{\lambda}/r$; for $c = \|\cdot\|$, $L = 1/r$ (since $\rho' \leq 1/r$, with ρ' the a.e. derivative; Lemma E.1(b)).

Data: the $\mathcal{O} \cup \mathcal{I}$ contamination model Following Ma et al. (2026), the sample $x_1, \ldots, x_n$ splits into inliers $\mathcal{I}$ and outliers $\mathcal{O}$, $|\mathcal{O}| = m$, $\epsilon_n = m/n < 1/2$; the partition is fixed. The inliers $(x_i)_{i \in \mathcal{I}}$ are i.i.d. from the *clean law* P; the outliers are arbitrary deterministic points, possibly chosen after seeing the inliers. Write $P_n = \frac{1}{n}\sum_i \delta_{x_i}$, $P_{\mathcal{I}} = \frac{1}{n-m}\sum_{i \in \mathcal{I}} \delta_{x_i}$ and $P_{\mathcal{O}} = \frac{1}{m}\sum_{i \in \mathcal{O}} \delta_{x_i}$. Huber contamination $(1 - \epsilon)P + \epsilon\mu_c$, with a contaminating law μ_c, is the case of i.i.d. draws with random labels.

Model, target, estimator and flags Let $\{\nu_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$ closed, be probability measures on $\mathcal{X}$. Define the *clean criterion* and its excess

$$J(\theta) = \mathrm{OT}_\rho(P, \nu_\theta), \qquad J_0 = \inf_\Theta J, \qquad M(\theta) = J(\theta) - J_0 \geq 0.$$

We assume $J_0 < \lambda$ (automatic for $r > 1$, since $h < \lambda$).

Assumption 2 (Model)**.** *(M1)* Clean target: *J has a unique minimizer $\theta_0 \in \Theta$, the* clean parameter*. (M2)* Continuity: *$\theta \mapsto \nu_\theta$ is continuous for weak convergence. (M3)* Coercivity: *$\{\theta : M(\theta) \leq t\}$ is compact for every $t < \lambda - J_0$.*

Interpretation. (M1) makes the target well defined, (M2) makes the criterion continuous in θ, and (M3) prevents near-minimizers from escaping to the boundary of Θ or to infinity; together they give existence and a finite modulus b.

Remark D.1 (The target is the clean parameter)**.** *θ_0 is defined by the* uncontaminated *law P; outliers never enter its definition. This differs from Ma et al. (2026), whose target* $\arg\min_\theta \mathrm{OT}_\rho(\mu_\star, \nu_\theta)$ *(with their $r = 1$ cost) is defined by the (possibly contaminated) data law $\mu_\star$. If the model is*

well specified, $P = \nu_{\theta^\dagger}$, then $J_0 = 0$ and, by Lemma D.5(ii), $\theta_0 = \theta^\dagger$ exactly when the model is identifiable. The definition also covers working models: for the point target $\nu_\theta = \delta_\theta$ and $P = \mathcal{N}(\mu_0, \sigma^2 I_d)$, $\theta_0 = \mu_0$ for every λ and r (Lemma E.5; Anderson, 1955). In general $\theta_0 = \theta_0(\lambda, r)$ depends on (λ, r) unless the model is well specified and identifiable, or symmetric as in this example; results with $\lambda = \lambda_n$ assume $\theta_0(\lambda_n) = \theta_0$.

Definition D.2 (Estimator)**.** *$\hat{\theta}_n$ is any measurable point with $F_n(\hat{\theta}_n) \le \inf_\Theta F_n + \xi_n$, where $F_n(\theta) = \mathrm{OT}_\rho(P_n, \nu_\theta)$ and $\xi_n \ge 0$ is an optimization tolerance.*

Definition D.3 (Flags)**.** *For $x \in \mathcal{X}$ let $A_\lambda(x) = \{y : c(x, y) \le \lambda\}$, the targets that x reaches on a full-weight path ($w^\star(c(x, y)) = 1$). For $\alpha \in (0, 1]$,*

$$\hat{\mathcal{O}}_\alpha = \big\{\, i : \ \nu_{\hat{\theta}_n}\big(A_\lambda(x_i)\big) < \alpha \,\big\},$$

the set $\hat{\mathcal{O}}$ of Section 4.

For $\nu_\theta = \delta_\theta$ (any $\alpha \in (0, 1]$) this is the point-target flag $\mathbb{1}\{c(x_i, \hat{\theta}_n) > \lambda\}$; for $\nu_\theta = \theta + \mathcal{N}(0, I_d)$ and $\alpha = 0.05$ it is the 95% rule of the experiments (Lemma E.14), up to Monte-Carlo target samples and the convention at ties. (The experiments' ROBOT clips at $\min(C, 2\lambda_{\mathrm{code}})$ (`pwrot/costs.py`), i.e. $\lambda = 2\lambda_{\mathrm{code}}$ here for $r = 1$, where λ_{code} is the code's threshold argument.) The rule uses the path weights but not the optimal plan.

Examples *Location $\nu_\theta = \nu_0(\cdot - \theta)$* for a fixed base law ν_0, $\Theta = \mathbb{R}^d$, with a coercive cost ($\inf_{\|x-y\| \ge R} c(x, y) \to \infty$ as $R \to \infty$): (M2) holds, and (M3) holds because $J(\theta) \to \lambda$ as $\|\theta\| \to \infty$ (Theorem E.13(i)). *Location–scale $\nu_{\mu,\sigma} = \nu_0((\cdot - \mu)/\sigma)$*: use $\Theta = \mathbb{R}^d \times [\underline{\sigma}, \infty)$ with $\underline{\sigma} > 0$ (or include $\sigma = 0$) and an atomless base law, since (M3) can fail as $\sigma \to 0$. *Mixtures and generative models $g_\theta \# Z$*: (M1) and (M3) generally require restrictions (label switching, vanishing components, non-identifiable generators).

D.2 Bounds on the robust OT value

The estimation results rest on three properties of the bounded cost: contamination changes the value by at most λ times the contaminated mass, the value is Lipschitz in total variation, and it is Lipschitz in a bounded transport metric that metrizes weak convergence.

Lemma D.4 (Value bounds)**.** *For all probability measures μ, μ_c, ν, P, P' on $\mathcal{X}$ and $\epsilon \in [0, 1]$, with $\tilde{\mu} = (1 - \epsilon)\mu + \epsilon\mu_c$ and OT_c the OT value with the raw cost c:*

(a) $\mathrm{OT}_\rho(\tilde{\mu}, \nu) \le \min\big\{(1 - \epsilon)\mathrm{OT}_\rho(\mu, \nu) + \epsilon\lambda,\ \ \lambda\,\mathrm{TV}(\tilde{\mu}, \nu),\ \ \tfrac{1}{r}\mathrm{OT}_c(\tilde{\mu}, \nu)\big\}$*;*

(b) $|\mathrm{OT}_\rho(P, \nu) - \mathrm{OT}_\rho(P', \nu)| \le \lambda\,\mathrm{TV}(P, P')$*;*

(c) under Assumption 1, $|\mathrm{OT}_\rho(P, \nu) - \mathrm{OT}_\rho(P', \nu)| \le T_\lambda(P, P')$, and likewise in the second argument; T_λ is a metric, bounded by $\lambda\,\mathrm{TV}$, that metrizes weak convergence with no moment condition;

(d) in the $\mathcal{O} \cup \mathcal{I}$ model, $\sup_\nu |\mathrm{OT}_\rho(P_n, \nu) - \mathrm{OT}_\rho(P_{\mathcal{I}}, \nu)| \le \lambda\,\epsilon_n$.

Interpretation. The bounded cost makes the robust OT value insensitive to contamination: moving an ϵ fraction of mass changes it by at most $\epsilon\lambda$, whatever the outliers, and it is continuous for weak convergence without moments. Part (d) is the only place where the outliers enter the analysis.

Proof. (a) *First term.* Fix $\xi > 0$, $\pi_1 \in \Pi(\mu, \nu)$ with $\int h\, d\pi_1 \le \mathrm{OT}_\rho(\mu, \nu) + \xi$ and any $\pi_2 \in \Pi(\mu_c, \nu)$; $(1 - \epsilon)\pi_1 + \epsilon\pi_2 \in \Pi(\tilde{\mu}, \nu)$ costs at most $(1 - \epsilon)(\mathrm{OT}_\rho(\mu, \nu) + \xi) + \epsilon\lambda$. *Second term.* Let $t = \mathrm{TV}(\tilde{\mu}, \nu)$ and $\beta = \tilde{\mu} \wedge \nu$ (the setwise minimum of the two measures, of mass $1 - t$); below, $\#$ is the push-forward and $\otimes$ the product measure. If $t = 0$ the diagonal coupling costs 0; otherwise $\pi = (\mathrm{id}, \mathrm{id})_\# \beta + t^{-1}(\tilde{\mu} - \beta) \otimes (\nu - \beta) \in \Pi(\tilde{\mu}, \nu)$ costs at most λt. *Third term.* $\rho(0) = 0$ and $\rho' \le 1/r$ (for $r = 1$, $\min(C, \lambda) \le C$), so $h \le c/r$. (b) Let $t = \mathrm{TV}(P, P') > 0$, $\beta = P \wedge P'$, $a = P - \beta$, $a' = P' - \beta$. Fix $\xi > 0$ and $\pi \in \Pi(P, \nu)$ with $\int h\, d\pi \le \mathrm{OT}_\rho(P, \nu) + \xi$. With the Radon–Nikodym derivative $\varphi = d\beta/dP \in [0, 1]$, $\pi_\beta = \varphi(x)\pi(dx, dy)$ has first marginal β; $\pi - \pi_\beta \ge 0$ has first marginal a and second marginal some ν_a of mass t. Then $\pi' = \pi_\beta +$

$t^{-1}a' \otimes \nu_a \in \Pi(P', \nu)$ and $\int h\, d\pi' \le \mathrm{OT}_\rho(P, \nu) + \xi + \lambda t$. Let $\xi \downarrow 0$ and exchange P, P'. (c) $h(x, y) \le h(x', y) + d_\lambda(x, x')$ as $h \in [0, \lambda]$. Fix $\xi > 0$, $\gamma \in \Pi(P, P')$ with $\int d_\lambda\, d\gamma \le T_\lambda(P, P') + \xi$ and $\pi' \in \Pi(P', \nu)$ with $\int h\, d\pi' \le \mathrm{OT}_\rho(P', \nu) + \xi$. The gluing lemma (Villani, 2009, Ch. 1) gives (X, X', Y) with $(X, X') \sim \gamma$, $(X', Y) \sim \pi'$ ($\mathcal{X}$ is Polish), so $\mathrm{OT}_\rho(P, \nu) \le \mathbb{E}h(X, Y) \le \mathrm{OT}_\rho(P', \nu) + T_\lambda(P, P') + 2\xi$. The second argument is identical by symmetry. d_λ is a bounded metric ($\min(a + b, k) \le \min(a, k) + \min(b, k)$) inducing the Euclidean topology, so T_λ metrizes weak convergence (Villani, 2009, Thm. 6.9, Cor. 6.13); $T_\lambda \le \lambda \mathrm{TV}$ by the diagonal argument of (a). (d) $P_n - P_{\mathcal{I}} = \epsilon_n(P_{\mathcal{O}} - P_{\mathcal{I}})$, so $\mathrm{TV}(P_n, P_{\mathcal{I}}) \le \epsilon_n$; apply (b). □

These bounds are essentially known: (a) extends Mukherjee et al. (2021, Thm. 2.1) from the clip to every $r \ge 1$, and for $r = 1$ the metric property in (c) and the two-sided $\mathcal{O} \cup \mathcal{I}$ bound (d) are in Ma et al. (2026, Thms. 3.2, 3.5 and eq. (11)). We include short proofs because the estimation results use them for every r and every target.

D.3 Estimation of the clean parameter

We first approximate the empirical criterion uniformly by the clean criterion (Lemma D.5), which gives a bias bound for arbitrary outliers, and then use automatic down-weighting (Proposition D.7) to remove the bias for separated outliers (Theorem D.8(d)).

Put $T_n = T_\lambda(P_{\mathcal{I}}, P)$ and $e_n = \lambda\epsilon_n + T_n$.

Lemma D.5 (Uniform approximation, existence, modulus)**.** *Under Assumptions 1 and 2:*

- *(i)* $\sup_\Theta |F_n - J| \le e_n$ *for every outlier configuration.*
- *(ii)* F_n *and* J *are continuous;* $M(\theta) = 0$ *iff* $\theta = \theta_0$*. If* $P = \nu_{\theta^\dagger}$*, then* $J(\theta) = 0$ *iff* $\nu_\theta = P$*.*
- *(iii) Whenever* $2e_n < \lambda - J_0$*,* F_n *attains its minimum.*
- *(iv) For fixed* λ*,* $\mathbb{E}T_n \to 0$ *and* $T_n \to 0$ *almost surely, with* $\mathbb{P}(T_n \ge \mathbb{E}T_n + s) \le \exp\{-2(n - m)s^2/\lambda^2\}$*, with no moment condition.*
- *(v)* $b(t) = \sup\{\|\theta - \theta_0\| : M(\theta) \le t\}$ *is finite, nondecreasing and right-continuous on* $[0, \lambda - J_0)$*, and* $b(t) \downarrow 0$ *as* $t \downarrow 0$*.*

Interpretation. Uniformly in θ, the empirical criterion is within $e_n = \lambda\epsilon_n + T_n$ of the clean criterion, so outliers act only through $\lambda\epsilon_n$, and the inlier fluctuation T_n vanishes without moment conditions; the modulus b translates a criterion gap into a parameter error.

Proof. (i) Lemma D.4(d) and (c): $|F_n(\theta) - \mathrm{OT}_\rho(P_{\mathcal{I}}, \nu_\theta)| \le \lambda\epsilon_n$ and $|\mathrm{OT}_\rho(P_{\mathcal{I}}, \nu_\theta) - J(\theta)| \le T_n$. (ii) Continuity: $\theta_k \to \theta$ gives $T_\lambda(\nu_{\theta_k}, \nu_\theta) \to 0$ by (M2), and Lemma D.4(c) in the second argument. The zero set of M is $\{\theta_0\}$ by (M1). If $P = \nu_{\theta^\dagger}$ and $J(\theta) = 0$: an optimal plan exists (h bounded continuous, $\Pi(P, \nu_\theta)$ tight and weakly closed), so an optimal $\pi \in \Pi(P, \nu_\theta)$ has $h = 0$, hence $c = 0$ and $x = y$, π-a.e. ($\rho(C) = 0$ only at $C = 0$), i.e. $\nu_\theta = P$. (iii) $\inf F_n \le F_n(\theta_0) \le J_0 + e_n$. For $0 < \xi < \lambda - J_0 - 2e_n$, (i) gives $\{F_n \le \inf F_n + \xi\} \subseteq \{M \le 2e_n + \xi\}$, compact by (M3); the set is closed and nonempty, so the continuous F_n attains its infimum. (iv) T_n is a function of $n - m$ i.i.d. inliers; replacing one changes $P_{\mathcal{I}}$ by total variation $1/(n - m)$ and, since T_λ is a metric bounded by $\lambda\mathrm{TV}$, changes T_n by at most $\lambda/(n - m)$. McDiarmid gives the tail. $\mathbb{E}T_n = f(n - m)$, with $f(k) = \mathbb{E}T_\lambda(P_k, P)$ and P_k the empirical law of k i.i.d. draws from P, depends only on the inlier sample size; $T_\lambda(P_k, P) \to 0$ a.s. (Varadarajan and (c)) and $T_\lambda \le \lambda$, so $f(k) \to 0$. Since $n - m > n/2$, the tails are summable in n for each $s > 0$, and Borel–Cantelli gives $T_n \to 0$ a.s. for any sampling scheme. (v) Finiteness from (M3). $K_t = \{M \le t\}$ are compact and decrease to $K_s = \bigcap_{t>s} K_t$ as $t \downarrow s$; taking maximizers $\theta_t \in K_t$ of $\|\theta - \theta_0\|$, a subsequence converges in K_{t_1}, for any fixed $t_1 > s$, to a point of K_s, so $\sup_{K_t} \|\theta - \theta_0\| \downarrow \sup_{K_s} \|\theta - \theta_0\|$ (right-continuity); $K_0 = \{\theta_0\}$. □

Assumption 3 (Curvature)**.** *There are* $\kappa, t_0 > 0$ *with* $M(\theta) \ge \kappa \min\{\|\theta - \theta_0\|^2, t_0^2\}$ *for all* θ*.*

Interpretation. The clean criterion grows at least quadratically near θ_0; since the cost is bounded ($M \le \lambda - J_0$), the growth can only be required locally, up to the level κt_0^2. It holds if (a) M

is twice differentiable at θ_0 with positive-definite Hessian and (b) the minimum is well separated, $\inf_{\|\theta-\theta_0\|\geq t_0} M(\theta) > 0$, which follows from (M1)–(M3). For the Gaussian point target and every $r \geq 1$, the Hessian is $\frac{2}{r} m_r I$ with $m_r > 0$ (Lemma E.4), the minimizer is unique (Lemma E.5) and $M \to \lambda - J_0$ at infinity, so Assumption 3 holds; for the spread target and generative models it must be checked.

Remark D.6 (Dependence on λ)**.** *$\rho_{\lambda,r}(C)$ is nondecreasing in λ for every C and $r \geq 1$ (for $C > \lambda$ and $r > 1$, $\partial_\lambda \rho = 1 - (\lambda/C)^{1/(r-1)} \geq 0$; ρ is continuous across $\lambda = C$). Write M_λ, b_λ, L_λ for M, b, L at threshold λ. Hence in the well-specified case ($J_0 = 0$) $M_\lambda \geq M_{\lambda_1}$ for $\lambda \geq \lambda_1$, Assumption 3 at λ_1 implies it at every $\lambda \geq \lambda_1$ with the same (κ, t_0), and $b_\lambda(t) \leq b_{\lambda_1}(t)$ for $t < \lambda_1$. (M3) does not transfer automatically. The Lipschitz constant grows, $L_\lambda = 2\sqrt{\lambda}/r$ for $c = \|\cdot\|^2$.*

For $\epsilon \in [0, 1)$ let $\mathcal{C}_\epsilon(\theta) = \{\nu' \in \mathcal{P}(\mathcal{X}) : (1-\epsilon)\nu' \leq \nu_\theta\}$, the probability measures obtained from ν_θ by trimming mass ϵ, and let $Q_\epsilon(\theta) = \inf_{\nu' \in \mathcal{C}_\epsilon(\theta)} \mathrm{OT}_\rho(P, \nu')$ be the model-side partial-transport criterion (the one-sided partial mass problem of del Barrio & Matrán (2013); Álvarez-Esteban et al. (2011)). Its inlier version is $\hat{Q}_n(\theta) = \inf_{\nu' \in \mathcal{C}_{\epsilon_n}(\theta)} \mathrm{OT}_\rho(P_\mathcal{I}, \nu')$.

Proposition D.7 (PW-ROT is sandwiched by the trimmed criterion at the true outlier fraction)**.** *Let $\bar{F}_n(\theta) = (1-\epsilon_n)\hat{Q}_n(\theta) + \epsilon_n \lambda$, and let $\Delta \geq 0$, $\tau \in [0, 1]$. For every θ, $F_n(\theta) \leq \bar{F}_n(\theta)$. If moreover $\nu_\theta(U) \leq \tau$ for $U = \bigcup_{i \in \mathcal{O}} \{y : c(x_i, y) < \lambda + \Delta\}$, then*

$$\bar{F}_n(\theta) - \gamma_n \leq F_n(\theta) \leq \bar{F}_n(\theta), \qquad \gamma_n = \epsilon_n\big(\lambda - \rho(\lambda + \Delta)\big) + \lambda\tau.$$

In particular, for $r = 1$ and $\tau = 0$, $F_n(\theta) = (1-\epsilon_n)\hat{Q}_n(\theta) + \epsilon_n\lambda$: at such θ the PW-ROT criterion equals, up to a constant, the criterion of the model-side trimmed estimator at the trimming level ϵ_n, although ϵ_n is not an input to the estimator.

Interpretation. For the hard clip, minimizing PW-ROT over a separated region is the same as minimizing the model-side trimmed criterion at the true outlier fraction: the outliers cost exactly λ each, whatever θ is. For $r > 1$ they cost slightly less than λ, by an amount that depends on θ; γ_n bounds this dependence.

Proof. Splitting a coupling. $P_n = (1-\epsilon_n)P_\mathcal{I} + \epsilon_n P_\mathcal{O}$. Given $\pi \in \Pi(P_n, \nu_\theta)$, let $\varphi = d((1-\epsilon_n)P_\mathcal{I})/dP_n \in [0, 1]$ and put $\pi_\mathcal{I} = \varphi(x)\pi(dx, dy)$, $\pi_\mathcal{O} = \pi - \pi_\mathcal{I}$. Their first marginals are $(1-\epsilon_n)P_\mathcal{I}$ and $\epsilon_n P_\mathcal{O}$, and their second marginals $(1-\epsilon_n)\nu_\mathcal{I}$, $\epsilon_n\nu_\mathcal{O}$ satisfy $(1-\epsilon_n)\nu_\mathcal{I} + \epsilon_n\nu_\mathcal{O} = \nu_\theta$, so $\nu_\mathcal{I} \in \mathcal{C}_{\epsilon_n}(\theta)$. Conversely every such split and pair of couplings gives an element of $\Pi(P_n, \nu_\theta)$. *Upper bound.* For any $\nu_\mathcal{I} \in \mathcal{C}_{\epsilon_n}(\theta)$, couple $P_\mathcal{I}$ with $\nu_\mathcal{I}$ near-optimally and $P_\mathcal{O}$ with the probability $\nu_\mathcal{O} = (\nu_\theta - (1-\epsilon_n)\nu_\mathcal{I})/\epsilon_n$ arbitrarily (cost $\leq \lambda$). *Lower bound.* In any split, $\pi_\mathcal{O}$ has mass ϵ_n, first marginal on $\{x_i : i \in \mathcal{O}\}$ and second marginal $\epsilon_n\nu_\mathcal{O} \leq \nu_\theta$; its mass on $\{(x, y) : c(x, y) < \lambda + \Delta\}$ is at most $\epsilon_n\nu_\mathcal{O}(U) \leq \nu_\theta(U) \leq \tau$. Since ρ is nondecreasing, $\int h\, d\pi_\mathcal{O} \geq \rho(\lambda + \Delta)(\epsilon_n - \tau) \geq \epsilon_n\rho(\lambda + \Delta) - \lambda\tau$, and the inlier part costs at least $(1-\epsilon_n)\hat{Q}_n(\theta)$. For $r = 1$, $\rho(\lambda + \Delta) = \lambda$. □

Unlike the trimmed Wasserstein distance $\mathcal{W}_p^\epsilon$ of Nietert et al. (2022), no trimming level is supplied; the bounded cost selects it.

Theorem D.8 (Estimation of θ_0)**.** *Let Assumptions 1 and 2 hold.* Arbitrary outliers:

(a) Bias bound. *For any near-minimizer (one exists by Lemma D.5(iii) when $\xi_n = 0$ and $2e_n < \lambda - J_0$), deterministically and uniformly over outlier configurations, $M(\hat{\theta}_n) \leq 2\lambda\epsilon_n + 2T_n + \xi_n$; hence $\|\hat{\theta}_n - \theta_0\| \leq b(2\lambda\epsilon_n + 2T_n + \xi_n)$, which is finite when the argument is $< \lambda - J_0$.*

(b) Rate. *Under Assumption 3, whenever $2\lambda\epsilon_n + 2T_n + \xi_n < \kappa t_0^2$, $\|\hat{\theta}_n - \theta_0\| \leq \big((2\lambda\epsilon_n + 2T_n + \xi_n)/\kappa\big)^{1/2}$; and with probability at least $1 - \delta$, $T_n \leq t_n(\delta) := \mathbb{E}T_n + \lambda\sqrt{\log(1/\delta)/(2(n-m))}$, where $\mathbb{E}T_n \leq \min\{L\,\mathbb{E}\mathcal{W}_1(P_\mathcal{I}, P), \lambda\}$ and $\mathcal{W}_1$ is the 1-Wasserstein distance.*

(c) Consistency. *If $\epsilon_n \to 0$ and $\xi_n \to 0$, then $\hat{\theta}_n \to \theta_0$ almost surely, for any outliers. If $\epsilon_n \to \epsilon$ with $2\lambda\epsilon < \lambda - J_0$ and $\xi_n \to 0$, then $\limsup_n \|\hat{\theta}_n - \theta_0\| \leq b(2\lambda\epsilon)$ almost surely, uniformly over outlier configurations.*

Separated outliers:

(d) Suppose $P = \nu_{\theta_0}$ (so $J_0 = 0$), $\epsilon_n \to \epsilon \in (0, 1/2)$ and $\xi_n \to 0$. Let $\Theta_R = \Theta \cap \bar{B}(\theta_0, R)$, where $\bar{B}(x, R)$ (resp. $B(x, R)$) is the closed (resp. open) Euclidean ball, with $R > b(2\lambda\epsilon)$, and assume, for all large n and with Δ, τ not depending on n (for every realization, if the outliers depend on the inliers):

- (Δ, τ)-separation: $\sup_{\theta \in \Theta_R} \nu_\theta\big(\bigcup_{i \in \mathcal{O}}\{y : c(x_i, y) < \lambda + \Delta\}\big) \le \tau$ *for some* $\Delta \ge 0$, $\tau \ge 0$;
- trimmed identifiability *at some* $\bar{\epsilon} \in (\epsilon, 1)$*: for* $\theta \in \Theta_R$, $(1 - \bar{\epsilon})\nu_{\theta_0} \le \nu_\theta$ *(as measures) only if* $\theta = \theta_0$.[2]

With Q_ϵ as in Proposition D.7, let $b_{\mathrm{POT}}(s) = \sup\{\|\theta - \theta_0\| : \theta \in \Theta_R,\ Q_{\bar{\epsilon}}(\theta) \le s\}$. Then $b_{\mathrm{POT}}(s) \to 0$ as $s \downarrow 0$, and for all large n, on $\{\hat{\theta}_n \in \Theta_R\}$ (which holds eventually, almost surely),

$$\|\hat{\theta}_n - \theta_0\| \le b_{\mathrm{POT}}\Big(2T_n + \frac{\gamma_n + \xi_n}{1 - \epsilon_n}\Big), \qquad \gamma_n = \epsilon_n\big(\lambda - \rho(\lambda + \Delta)\big) + \lambda\tau,$$

where $\lambda - \rho(\lambda + \Delta) = \frac{r-1}{r}\lambda^{\frac{r}{r-1}}(\lambda + \Delta)^{-\frac{1}{r-1}}$ for $r > 1$ and $= 0$ for $r = 1$. Hence $\limsup_n \|\hat{\theta}_n - \theta_0\| \le \inf_{s > \gamma/(1-\epsilon)} b_{\mathrm{POT}}(s)$ a.s., with $\gamma = \epsilon(\lambda - \rho(\lambda + \Delta)) + \lambda\tau$. In particular, for $r = 1$ and $\tau = 0$, $\hat{\theta}_n \to \theta_0$ almost surely despite a fixed fraction of outliers; for $r > 1$ the bound vanishes as $\Delta \to \infty$ and $\tau \to 0$, more slowly for larger r.

Interpretation. Parts (a)–(c) hold for any outliers and give an error of order $b(2\lambda\epsilon_n)$, which vanishes only with the outlier fraction. Part (d) shows that separation removes this term: the error is then controlled by the partial-transport modulus b_{POT} and the price of softness γ, which is zero for the hard clip.

Proof of (a)–(c). (a) By Lemma D.5(i), $J(\hat{\theta}_n) \le F_n(\hat{\theta}_n) + e_n \le F_n(\theta_0) + \xi_n + e_n \le J_0 + 2e_n + \xi_n$. The bound on $\|\hat{\theta}_n - \theta_0\|$ is the definition of b. Outliers enter only through $\lambda\epsilon_n$; no triangle inequality for OT_ρ is used. This is the modulus-of-continuity argument of Donoho & Liu (1988), in the form of Zhu et al. (2022); what is specific here is the target, the clean θ_0 rather than the minimizer for the contaminated law of Ma et al. (2026), and the uniformity over outlier configurations. (b) If $M(\hat{\theta}_n) < \kappa t_0^2$, Assumption 3 forces $\|\hat{\theta}_n - \theta_0\|^2 \le M(\hat{\theta}_n)/\kappa$. The tail of T_n is Lemma D.5(iv); $T_\lambda \le \mathrm{OT}_{L\|\cdot\|} = L\mathcal{W}_1$ and $T_\lambda \le \lambda$. (c) $T_n \to 0$ a.s. (Lemma D.5(iv)). If $\epsilon_n \to 0$, the argument of b tends to 0 and $b(t) \downarrow 0$. Otherwise the argument has $\limsup$ equal to $2\lambda\epsilon < \lambda - J_0$, and b is nondecreasing and right-continuous. The null set depends only on the inliers, which gives uniformity over outliers. □

The admissible fixed fraction in (c), $\epsilon < (1 - J_0/\lambda)/2$, is a breakdown-type bound; it degrades when λ is small relative to the inlier spread, since then $J_0/\lambda \to 1$.

Proof of (d). Steps 1–2. By Proposition D.7, $\bar{F}_n(\theta) - \gamma_n \le F_n(\theta) \le \bar{F}_n(\theta)$ for $\theta \in \Theta_R$.

Step 3 (comparison with θ_0). On $\{\hat{\theta}_n \in \Theta_R\}$, $\bar{F}_n(\hat{\theta}_n) - \gamma_n \le F_n(\hat{\theta}_n) \le F_n(\theta_0) + \xi_n \le \bar{F}_n(\theta_0) + \xi_n$, so $(1-\epsilon_n)\hat{Q}_n(\hat{\theta}_n) \le (1-\epsilon_n)\hat{Q}_n(\theta_0) + \gamma_n + \xi_n$. Since $\nu_{\theta_0} \in \mathcal{C}_{\epsilon_n}(\theta_0)$, $\hat{Q}_n(\theta_0) \le \mathrm{OT}_\rho(P_{\mathcal{I}}, \nu_{\theta_0}) \le T_n$ (Lemma D.4(c) and $\mathrm{OT}_\rho(P, P) = 0$).

Step 4 (to the population partial criterion). For every ν', $|\mathrm{OT}_\rho(P_{\mathcal{I}}, \nu') - \mathrm{OT}_\rho(P, \nu')| \le T_n$, so $Q_{\epsilon_n}(\theta) \le \hat{Q}_n(\theta) + T_n$ uniformly in θ. If $\epsilon_n \le \bar{\epsilon}$ (true for large n), then $(1 - \bar{\epsilon})\nu' \le (1 - \epsilon_n)\nu'$, so $\mathcal{C}_{\epsilon_n}(\theta) \subseteq \mathcal{C}_{\bar{\epsilon}}(\theta)$ and $Q_{\bar{\epsilon}} \le Q_{\epsilon_n}$. Therefore

$$Q_{\bar{\epsilon}}(\hat{\theta}_n) \le \hat{Q}_n(\hat{\theta}_n) + T_n \le 2T_n + \frac{\gamma_n + \xi_n}{1 - \epsilon_n},$$

[2] Equivalently $\nu_\theta \notin \mathcal{V}_{\bar{\epsilon}}(\nu_{\theta_0}) = \{(1-\bar{\epsilon})\nu_{\theta_0} + \bar{\epsilon}Q :\ Q \in \mathcal{P}(\mathcal{X})\}$ for $\theta \ne \theta_0$, the contamination neighborhood of del Barrio et al. (2020); cf. the trimmed similarity of Álvarez-Esteban et al. (2012). Only its use as an identification condition for a parametric OT estimator is new.

which is the displayed bound by definition of b_{POT}.

Step 5 ($b_{\mathrm{POT}}(s) \to 0$). Suppose not: there are $\eta > 0$ and $\theta_k \in \Theta_R$ with $Q_{\bar\epsilon}(\theta_k) \to 0$ and $\|\theta_k - \theta_0\| \geq \eta/2$. Choose $\nu'_k \in \mathcal{C}_{\bar\epsilon}(\theta_k)$ with $\mathrm{OT}_\rho(P, \nu'_k) \to 0$. Θ_R is compact, so along a subsequence $\theta_k \to \bar\theta \in \Theta_R$, $\|\bar\theta - \theta_0\| \geq \eta/2$, and $\nu_{\theta_k} \to \nu_{\bar\theta}$ weakly (M2); hence $\{\nu_{\theta_k}\}$ is tight, and so is $\{\nu'_k\}$ because $\nu'_k \leq \nu_{\theta_k}/(1-\bar\epsilon)$. Along a further subsequence $\nu'_k \to \nu^*$ weakly. $\mathrm{OT}_\rho(P, \cdot)$ is weakly continuous (Lemma D.4(c)), so $\mathrm{OT}_\rho(P, \nu^*) = 0$, whence $\nu^* = P = \nu_{\theta_0}$ (by the argument of Lemma D.5(ii)). For every bounded continuous $f \geq 0$, $(1-\bar\epsilon)\int f\, d\nu'_k \leq \int f\, d\nu_{\theta_k}$ passes to the limit, so $(1-\bar\epsilon)\nu_{\theta_0} \leq \nu_{\bar\theta}$ (on closed sets F via continuous $f_j \downarrow \mathbb{1}_F$, then on all Borel sets by inner regularity) and $\bar\theta = \theta_0$ by trimmed identifiability, a contradiction.

Step 6 (localization and limits). By (c), $\limsup_n \|\hat\theta_n - \theta_0\| \leq b(2\lambda\epsilon)$, which is smaller than the radius of Θ_R, so $\hat\theta_n \in \Theta_R$ for all large n almost surely. $T_n \to 0$ a.s. (Lemma D.5(iv)), $\xi_n \to 0$ and $\gamma_n \to \gamma$, so the argument of b_{POT} is eventually below any $s > \gamma/(1-\epsilon)$. For $r = 1$, $\rho(\lambda+\Delta) = \lambda$; with $\tau = 0$, $\gamma = 0$ and Step 5 gives $\hat\theta_n \to \theta_0$. For $r > 1$, $\lambda - \rho(\lambda+\Delta)$ is the displayed expression by the formula for ρ, and the bound (not necessarily the bias) tends to 0 as $\Delta \to \infty$, at the rate $\Delta^{-1/(r-1)}$. Nothing requires ϵ_n to be constant: only $\epsilon_n \to \epsilon < 1/2$ and $\epsilon_n \leq \bar\epsilon$ eventually are used. □

Corollary D.9 (The price of softness)**.** *In the setting of Theorem D.8(d) with $\tau = 0$: for $r = 1$, $\hat\theta_n \to \theta_0$ almost surely; for $r > 1$, $\limsup_n \|\hat\theta_n - \theta_0\| \leq \inf_{s > s_r} b_{\mathrm{POT}}(s)$ with*

$$s_r = \frac{\epsilon}{1-\epsilon} \cdot \frac{r-1}{r}\, \lambda \Big(\frac{\lambda}{\lambda+\Delta} \Big)^{1/(r-1)}.$$

s_r is nondecreasing in r and tends to 0 as $\Delta \to \infty$, like $\Delta^{-1/(r-1)}$: the argument of the bias bound grows with r.

Proof. Theorem D.8(d) with $\tau = 0$ and $\gamma = \epsilon(\lambda - \rho(\lambda+\Delta))$, where $\lambda - \rho(\lambda+\Delta) = \frac{r-1}{r}\lambda^{r/(r-1)}(\lambda+\Delta)^{-1/(r-1)} = \frac{r-1}{r}\lambda(\lambda/(\lambda+\Delta))^{1/(r-1)}$. Both factors are nondecreasing in r, since $\lambda/(\lambda+\Delta) \leq 1$ and $1/(r-1)$ decreases. □

Corollary D.10 (Gaussian location at finite separation)**.** *Let $\nu_\theta = \mathcal{N}(\theta, I_d)$, $P = \nu_{\theta_0}$, $c = \|\cdot\|^2$, $\Theta_R = \Theta \cap \bar B(\theta_0, R)$, and suppose every outlier satisfies $\|x_i - \theta_0\| \geq D$ with $D \geq \sqrt{\lambda+\Delta} + R + \sqrt{d}$. Then the (Δ, τ)-separation of Theorem D.8(d) holds with*

$$\tau = \exp\Big\{ -\tfrac{1}{2}\big(D - \sqrt{\lambda+\Delta} - R - \sqrt{d}\big)^2 \Big\},$$

trimmed identifiability holds for every $\bar\epsilon < 1$, and therefore $\limsup_n \|\hat\theta_n - \theta_0\| \leq \inf_{s > \gamma/(1-\epsilon)} b_{\mathrm{POT}}(s)$ with $\gamma = \epsilon(\lambda - \rho(\lambda+\Delta)) + \lambda\tau$, a finite-$D$ bound, explicit in τ, that vanishes as $D \to \infty$ for $r = 1$ (take $\Delta = 0$) and, for $r > 1$, as $D \to \infty$ with $\Delta \to \infty$ and $D - \sqrt{\lambda+\Delta} \to \infty$.

Interpretation. Gaussian models have unbounded support, so the separation holds only up to the Gaussian tail τ; the resulting bias bound is explicit and vanishes as the outliers move away.

Proof. If $c(x_i, y) < \lambda + \Delta$ then $\|y - \theta_0\| > D - \sqrt{\lambda+\Delta}$. For $\theta \in \Theta_R$ and $Z \sim \mathcal{N}(0, I_d)$, $\nu_\theta(U) \leq \mathbb{P}(\|Z + \theta - \theta_0\| > D - \sqrt{\lambda+\Delta}) \leq \mathbb{P}(\|Z\| > D - \sqrt{\lambda+\Delta} - R)$, and Gaussian concentration ($\mathbb{E}\|Z\| \leq \sqrt{d}$, $\|\cdot\|$ 1-Lipschitz) gives the bound. If $(1-\bar\epsilon)\nu_{\theta_0} \leq \nu_\theta$, the density ratio f_{θ_0}/f_θ (f_θ the density of ν_θ), $f_{\theta_0}/f_\theta = \exp\{\langle y - \theta_0, \theta_0 - \theta\rangle + \|\theta - \theta_0\|^2/2\}$ is bounded by $1/(1-\bar\epsilon)$, which forces $\theta = \theta_0$. □

Remark D.11 (Scope of (d))**.** Examples. *Uniform location families $\nu_\theta = \mathrm{Unif}(\theta + [-1,1]^d)$ are trimmed-identifiable (a support inclusion forces $\theta = \theta_0$) and give $\tau = 0$ once the outliers lie at cost $\geq \lambda + \Delta$ from $\theta + [-1,1]^d$ for all $\theta \in \Theta_R$; with $r = 1$ this is exact consistency for θ_0 with a fixed fraction of outliers. Gaussian location families are trimmed-identifiable (with f the $\mathcal{N}(0, I_d)$ density, the ratio $f(\cdot - \theta_0)/f(\cdot - \theta) = \exp\{\langle y, \theta_0 - \theta\rangle + \mathrm{const}\}$ is unbounded for $\theta \neq \theta_0$) and have $\tau > 0$, small when the outliers are far from θ_0. Heavy-tailed location families with bounded density ratios (e.g. Cauchy) are* not *trimmed-identifiable, and (d) does not apply to them. For the experiments' ROBOT cost $\min(C, 2\lambda_{\mathrm{code}})$ the result holds with threshold $2\lambda_{\mathrm{code}}$.* Not covered: *the*

point target is misspecified ($P \neq \delta_{\theta_0}$) and is handled by the direct argument of Theorem E.7, where the corresponding bias is, with $\epsilon' = \epsilon/(1-\epsilon)$, $\epsilon'\sqrt{\lambda}(\sqrt{\lambda}/D)^{(r+1)/(r-1)}$. Mechanism. *For $r = 1$ every path from an outlier to the model beyond the threshold costs exactly λ, so the outliers add a constant and are absorbed by an ϵ-fraction of the model; the inliers are then fitted by partial transport, which is identifying. For $r > 1$ the soft tail makes the outlier cost depend on θ through $\rho(c) \in [\rho(\lambda+\Delta), \lambda)$, and this dependence is the bias $\epsilon(\lambda - \rho(\lambda+\Delta))$.* Relation to prior work. *That a fixed fraction of far outliers can leave the clean parameter unbiased is classical: minimum Hellinger distance estimates recover θ_0 exactly as the contamination moves to infinity (Beran, 1977, Thm. 7), and γ-divergence estimates have small bias when outliers lie where the model has little mass (Fujisawa & Eguchi, 2008). Both require densities (and smoothing for empirical measures) and give limiting or approximate statements. Nietert et al. (2022, Prop. 11) recover the clean $\mathcal{W}_p$ distance exactly under separated compact supports with ϵ known. At the population level, minimum total variation shares the exact-consistency property under the hypotheses of (d), but $\mathrm{TV}(P_n, \nu_\theta) = 1$ for continuous models, so it cannot be used on empirical measures. Theorem D.8(d), via Proposition D.7, gives the property for an estimator computed on empirical measures, without densities and without knowledge of ϵ; it is exact at finite separation for $r = 1$ and $\tau = 0$, gives the finite-D Gaussian bound of Corollary D.10, and quantifies the price of softness for $r > 1$ (Corollary D.9).*

Proposition D.12 (Rates under separated contamination)**.** *Assume the setting of Theorem D.8(d) with λ fixed, and in addition:*

- Growth of the partial criterion: $Q_{\bar\epsilon}(\theta) \geq \omega_Q(\|\theta - \theta_0\|)$ *for $\theta \in \Theta_R$, with $\omega_Q : [0,\infty) \to [0,\infty)$ nondecreasing, $\omega_Q(0) = 0$ and $\omega_Q(u) > 0$ for $u > 0$; write $\omega_Q^{-1}(v) = \sup\{u : \omega_Q(u) \leq v\}$, finite for $v < \sup \omega_Q$;*
- Finite-sample localization: $2\lambda\epsilon_n + 2t_n(\delta) + \xi_n < \lambda$ *and* $b(2\lambda\epsilon_n + 2t_n(\delta) + \xi_n) < R$ *(true for large n, since $R > b(2\lambda\epsilon)$ and b is right-continuous).*

(Trimmed identifiability on Θ_R is implied by the growth condition; the limits $\epsilon_n \to \epsilon$, $\xi_n \to 0$ are not used.) Then, for n with $\epsilon_n \leq \bar\epsilon$ and the separation in force, with probability at least $1 - \delta$,

$$\|\hat\theta_n - \theta_0\| \leq \omega_Q^{-1}\Big(2t_n(\delta) + \frac{\gamma_n + \xi_n}{1 - \epsilon_n}\Big).$$

For $r = 1$ and $\tau = 0$ we have $\gamma_n = 0$, so the error is controlled by the inlier fluctuation $t_n(\delta)$ alone: the fixed outlier fraction contributes no term*, in contrast with the $2\lambda\epsilon_n$ term of Theorem D.8(b).*

Interpretation. Under separation, a fixed fraction of outliers contributes no error term for the hard clip; the rate is then set by how fast the partial criterion grows, ω_Q, and by the inlier fluctuation.

Proof. On $\{T_n \leq t_n(\delta)\}$, which has probability at least $1 - \delta$ (Lemma D.5(iv)), Theorem D.8(a) gives $\|\hat\theta_n - \theta_0\| \leq b(2\lambda\epsilon_n + 2t_n(\delta) + \xi_n) < R$, so $\hat\theta_n \in \Theta_R$; no curvature condition is needed, so any fixed $\epsilon < 1/2$ is allowed. Steps 1–4 of the proof of (d) are deterministic given $\hat\theta_n \in \Theta_R$ and $\epsilon_n \leq \bar\epsilon$, and give $\omega_Q(\|\hat\theta_n - \theta_0\|) \leq Q_{\bar\epsilon}(\hat\theta_n) \leq 2T_n + (\gamma_n + \xi_n)/(1 - \epsilon_n)$. Apply ω_Q^{-1}. □

Remark D.13 (Two contrasting examples)**.** Uniform location, $d = 1$ (bounded support, polynomial rate). *Let $\nu_\theta = \mathrm{Unif}[\theta - 1, \theta + 1]$, $P = \nu_{\theta_0}$, $c(x, y) = |x - y|^2$, and $s = |\theta - \theta_0|$. Every $\nu' \in \mathcal{C}_{\bar\epsilon}(\theta)$ is supported in $[\theta - 1, \theta + 1]$, while P puts density $1/2$ on the length-$\min(s, 2)$ interval outside it; a point at distance u from that support moves at least u and costs at least $\rho(u^2) \geq \min(u^2, \lambda)/r$. Hence*

$$Q_{\bar\epsilon}(\theta) \geq \frac{1}{2r}\int_0^{\min(s,2,\sqrt{\lambda})} u^2\, du = \frac{\min(s, 2, \sqrt{\lambda})^3}{6r} =: \omega_Q(s),$$

and for $r = 1$, $\tau = 0$, $\xi_n = 0$: $\|\hat\theta_n - \theta_0\| \leq (12\, t_n(\delta))^{1/3}$ whenever $12\, t_n(\delta) < \min(2, \sqrt{\lambda})^3$, a polynomial rate for the clean parameter despite a fixed fraction of outliers. The cubic order is tight: an explicit ν' gives $Q_{\bar\epsilon}(\theta) \leq s^3(k_{\bar\epsilon}^2 + k_{\bar\epsilon} + 1)/(6r)$ with $k_{\bar\epsilon} = (1 - \bar\epsilon)/\bar\epsilon$ for small s; with $t_n \asymp n^{-1/2}$ in $d = 1$ the rate is $n^{-1/6}$.

Gaussian location (full support, logarithmic guarantee). *Let $\nu_\theta = \mathcal{N}(\theta, I_d)$, $P = \nu_{\theta_0}$, $s = \|\theta - \theta_0\|$, $\bar a = \log \frac{1}{1-\bar\epsilon}$. The partial criterion is extremely flat:*

$$Q_{\bar\epsilon}(\theta) \leq \lambda\, \bar\Phi\Big(\frac{\bar a}{s} - \frac{s}{2}\Big),$$

where $\bar{\Phi}$ is the standard normal upper tail. Proof. *With p, f_θ the densities of P, ν_θ, put $\tilde{p} = \min\{p, f_\theta/(1-\bar{\epsilon})\}$; its deficit is $\int(p-\tilde{p}) = \int(p - f_\theta/(1-\bar{\epsilon}))_+ \le P\{p/f_\theta > 1/(1-\bar{\epsilon})\}$. Since $p/f_\theta = \exp\{-s\,Z + s^2/2\}$ with $Z = \langle y-\theta_0, (\theta-\theta_0)/s\rangle \sim \mathcal{N}(0,1)$ under P, this probability is $\bar{\Phi}(\bar{a}/s - s/2)$. Completing $\tilde{p}$ to a probability ν' with $(1-\bar{\epsilon})\nu' \le \nu_\theta$ is possible because the room $\int(f_\theta/(1-\bar{\epsilon}) - \tilde{p}) = \bar{\epsilon}/(1-\bar{\epsilon}) + \int(p-\tilde{p})$ exceeds the deficit $\int(p-\tilde{p})$; the added mass sits where $\tilde{p} = p$, so $\mathrm{TV}(P,\nu') = \int(p-\tilde{p})$ and $\mathrm{OT}_\rho(P,\nu') \le \lambda \mathrm{TV}(P,\nu')$ by Lemma D.4(a).* □
Consequently any valid ω_Q satisfies $\omega_Q(s) \le \lambda\bar{\Phi}(\bar{a}/s - s/2)$, and the bound of Proposition D.12 is never smaller than $(1+o(1))\,\bar{a}/\sqrt{2\log(\lambda/(2t_n))}$, i.e. of order $1/\sqrt{\log n}$. Since $\tau > 0$ for Gaussians, the bias floor is likewise of order $\bar{a}/\sqrt{2\log(1/\tau)}$. The reason is structural: the model may discard its spare ϵ of mass where it is most convenient, which makes small shifts of θ nearly invisible in the tails. (For Gaussians $\tau > 0$ in any case, so (d) gives only a bias bound.)

Reading. *Model-side trimming gives fixed-fraction consistency with a polynomial rate for bounded-support models, but only a weak (logarithmic) guarantee for smooth full-support models. Whether the estimator itself, and not only this bound, is slow in the Gaussian case is open.*

D.4 Outlier-detection consistency

Detection follows from estimation through margin sets: points whose flag cannot change while the parameter stays within a margin t of θ_0. We define these sets, prove exact recovery when the estimation error is below the margin, and derive the admissible growth of λ_n.

For a margin $t > 0$ define the *margin sets* at threshold λ,

$$\mathcal{G}_\lambda(t) = \Big\{x : \inf_{\theta\in\Theta,\ \|\theta-\theta_0\|\le t} \nu_\theta\big(A_\lambda(x)\big) \ge \alpha\Big\}, \qquad \mathcal{B}_\lambda(t) = \Big\{x : \sup_{\theta\in\Theta,\ \|\theta-\theta_0\|\le t} \nu_\theta\big(A_\lambda(x)\big) < \alpha\Big\},$$

the points whose flag cannot change while θ stays within t of θ_0. For each θ, $x \mapsto \nu_\theta(A_\lambda(x))$ is upper semicontinuous ($A_\lambda(x)$ is closed and c continuous; reverse Fatou), so $\mathcal{G}_\lambda(t)$, an infimum, is closed. Under (M2), $(\theta, x) \mapsto \nu_\theta(A_\lambda(x))$ is jointly upper semicontinuous: if $(\theta_k, x_k) \to (\theta, x)$ then $\{\nu_{\theta_k}\}$ is tight, and for compact K and $\eta > 0$, uniform continuity of c on $K \times K$ gives $A_\lambda(x_k) \cap K \subseteq A_{\lambda+\eta}(x)$ for large k, so $\limsup_k \nu_{\theta_k}(A_\lambda(x_k)) \le \nu_\theta(A_{\lambda+\eta}(x)) + \sup_k \nu_{\theta_k}(K^c)$ by the Portmanteau theorem for the closed set $A_{\lambda+\eta}(x)$; let $K \uparrow \mathcal{X}$ and $\eta \downarrow 0$. Hence $\mathcal{B}_\lambda(t)$ is open and the flags are measurable.

Remark D.14 (Examples of margin sets)**.** Point target, $\nu_\theta = \delta_\theta$, $c = \|\cdot\|^2$: *$\mathcal{G}_\lambda(t) = \{\|x-\theta_0\| \le \sqrt{\lambda} - t\}$ and $\mathcal{B}_\lambda(t) = \{\|x-\theta_0\| > \sqrt{\lambda} + t\}$ (with $\Theta = \mathbb{R}^d$), the conditions of Theorem E.7.* Gaussian spread target, $\nu_\theta = \theta + \mathcal{N}(0, I_d)$: *$\nu_\theta(A_\lambda(x)) = \Phi_{d,\|x-\theta\|^2}(\lambda)$, where $\Phi_{d,t}$ is the CDF of the noncentral χ^2_d law with noncentrality t, is strictly decreasing in $\|x-\theta\|$, so for $\lambda \ge \Phi^{-1}_{d,0}(\alpha)$ (the α-quantile of χ^2_d), with $R_\alpha(\lambda)$ solving $\Phi_{d,R^2}(\lambda) = \alpha$, $\mathcal{G}_\lambda(t) = \{\|x-\theta_0\| \le R_\alpha(\lambda) - t\}$ and $\mathcal{B}_\lambda(t) = \{\|x-\theta_0\| > R_\alpha(\lambda) + t\}$ (Lemma E.14). Since $\|x-Y\| \le \|x-\theta\| + \|Y-\theta\|$ for $Y \sim \nu_\theta$, $\nu_\theta(A_\lambda(x)) \ge \alpha$ whenever $\|x-\theta\| \le \sqrt{\lambda} - \sqrt{\chi^2_{d,\alpha}}$, with $\chi^2_{d,\alpha}$ the α-quantile of χ^2_d; hence $R_\alpha(\lambda) \ge \sqrt{\lambda} - \sqrt{\chi^2_{d,\alpha}}$, and for inliers $\mathcal{N}(\theta_0, \sigma^2 I_d)$ the condition $\mathbb{P}(X \notin \mathcal{G}_\lambda(t)) \le \delta/n$ holds once $\sqrt{\lambda} \ge t + \sqrt{\chi^2_{d,\alpha}} + \sigma(\sqrt{d} + \sqrt{2\log(n/\delta)})$, by Gaussian concentration as in Lemma E.6.* Models continuous in total variation, $\mathrm{TV}(\nu_\theta, \nu_{\theta_0}) \le \omega(\|\theta-\theta_0\|)$ *with ω nondecreasing: $\mathcal{G}_\lambda(t) \supseteq \{x : \nu_{\theta_0}(A_\lambda(x)) \ge \alpha + \omega(t)\}$ and $\mathcal{B}_\lambda(t) \supseteq \{x : \nu_{\theta_0}(A_\lambda(x)) < \alpha - \omega(t)\}$.*

Theorem D.15 (Exact recovery of the outlier set)**.**

(i) On $\{\|\hat{\theta}_n - \theta_0\| \le t\} \cap \{x_i \in \mathcal{G}_\lambda(t)\ \forall i \in \mathcal{I}\}$, if $x_i \in \mathcal{B}_\lambda(t)$ for all $i \in \mathcal{O}$, then $\hat{\mathcal{O}}_\alpha = \mathcal{O}$: no masking and no swamping.

(ii) If $P(X \notin \mathcal{G}_\lambda(t)) \le \delta/n$ for $X \sim P$, then $\mathbb{P}(x_i \in \mathcal{G}_\lambda(t)\ \forall i \in \mathcal{I}) \ge 1-\delta$; for a fixed target θ_0 the condition is monotone in λ (it persists for larger λ).

(iii) (Fixed fraction, arbitrary positions in $\mathcal{B}_\lambda(t)$.) Under Assumptions 1–2, if $P(X \notin \mathcal{G}_\lambda(t)) \le \delta/n$, $x_i \in \mathcal{B}_\lambda(t)$ for all $i \in \mathcal{O}$, $2\lambda\epsilon_n + 2t_n(\delta) + \xi_n < \lambda - J_0$ and $b(2\lambda\epsilon_n + 2t_n(\delta) + \xi_n) \le t$, then $\mathbb{P}(\hat{\mathcal{O}}_\alpha = \mathcal{O}) \ge 1 - 2\delta$.

Interpretation. Exact recovery is a deterministic consequence of two events, a small estimation error and inliers inside their margin sets; the first is controlled by Theorem D.8 and the second by the tail of P.

Proof. (i) For $i \in \mathcal{I}$, $\nu_{\hat{\theta}_n}(A_\lambda(x_i)) \geq \inf_{\|\theta-\theta_0\|\leq t} \nu_\theta(A_\lambda(x_i)) \geq \alpha$, so i is not flagged; for $i \in \mathcal{O}$, $\nu_{\hat{\theta}_n}(A_\lambda(x_i)) \leq \sup_{\|\theta-\theta_0\|\leq t} \nu_\theta(A_\lambda(x_i)) < \alpha$, so i is flagged. (ii) Union bound over at most n i.i.d. inliers. $A_\lambda(x)$ increases with λ, hence so does $\mathcal{G}_\lambda(t)$. (iii) By Theorem D.8(a)–(b), with probability at least $1-\delta$, $\|\hat{\theta}_n - \theta_0\| \leq b(2\lambda\epsilon_n + 2t_n(\delta) + \xi_n) \leq t$. Combine with (ii) and (i). □

Recovery needs a bias that is small relative to the margin t, not a vanishing one, and holds with a fixed fraction of outliers placed anywhere in $\mathcal{B}_\lambda(t)$. Detection consistency with a fixed outlier fraction is known for penalized mean-shift regression (Kong et al., 2018, Thm. 3); to our knowledge Theorem D.15 is the first such guarantee for robust OT, and the model-based flag makes it applicable to general, including generative, models.

Corollary D.16 (A rate window for λ_n)**.** *Let $\lambda = \lambda_n$, let $\delta_n \in (0,1)$ with $\delta_n \to 0$, $\xi_n \to 0$, fix a margin $t > 0$, assume $\theta_0(\lambda_n) = \theta_0$ for all n (Remark D.1), and let Assumptions 1–3 hold with (κ, t_0) uniform in n (in the well-specified case it suffices that Assumption 3 holds at some $\lambda_1 \leq \inf_n \lambda_n$, Remark D.6). Suppose*

(W1) (no swamping) $n\,P(X \notin \mathcal{G}_{\lambda_n}(t)) \leq \delta_n$;

(W2) (separation at the chosen threshold) $x_i \in \mathcal{B}_{\lambda_n}(t)$ *for all* $i \in \mathcal{O}$;

(W3) (bias small relative to the margin) *either*

- *(i) arbitrary positions in $\mathcal{B}_{\lambda_n}(t)$: $\lambda_n\epsilon_n + t_n(\delta_n) \to 0$; for $c = \|\cdot\|^2$ and $\mathbb{E}\mathcal{W}_1(P_\mathcal{I}, P) < \infty$, $t_n(\delta) \leq \frac{2\sqrt{\lambda_n}}{r}\mathbb{E}\mathcal{W}_1(P_\mathcal{I}, P) + \lambda_n\sqrt{\log(1/\delta)/(2(n-m))}$ (note that $\mathbb{E}T_{\lambda_n} \to 0$ is not automatic when $\lambda_n \to \infty$); or*
- *(ii) separated outliers with a fixed fraction and a* fixed *threshold $\lambda_n \equiv \lambda$ (compatible with (W1) for bounded-support inliers): the conditions of Theorem D.8(d), and $\inf_{s>\gamma/(1-\epsilon)} b_{\mathrm{POT}}(s) < t$.*

Then $\mathbb{P}(\hat{\mathcal{O}}_\alpha = \mathcal{O}) \to 1$, and under (W3)(i) also $\hat{\theta}_n \to \theta_0$ in probability.

Interpretation. The threshold must be large enough that no inlier leaves its margin set (W1) and small enough that the bias $\lambda_n\epsilon_n$ and the fluctuation stay below the margin (W3)(i); with separated outliers and bounded inliers a fixed threshold suffices (W3)(ii).

Proof under (W3)(i). By Theorem D.8(b), with probability at least $1-\delta_n$, $\|\hat{\theta}_n - \theta_0\| \leq ((2\lambda_n\epsilon_n + 2t_n(\delta_n) + \xi_n)/\kappa)^{1/2} \to 0$ (the side condition holds for large n), so it is $\leq t$ eventually. Theorem D.15(i)–(ii) with (W1)–(W2) give $\mathbb{P}(\hat{\mathcal{O}}_\alpha = \mathcal{O}) \geq 1 - 2\delta_n$ for large n. The bound on t_n uses $L = 2\sqrt{\lambda_n}/r$ and $T_\lambda \leq L\mathcal{W}_1$. □

Proof under (W3)(ii). With λ fixed, Theorem D.8(d) gives, almost surely, $\|\hat{\theta}_n - \theta_0\| \leq b_{\mathrm{POT}}(s) < t$ for some $s > \gamma/(1-\epsilon)$ and all large n, hence in probability; Theorem D.15(i)–(ii) with (W1)–(W2) give $\mathbb{P}(\hat{\mathcal{O}}_\alpha = \mathcal{O}) \geq 1 - \delta_n - o(1)$. □

(W3)(i) requires $\lambda_n = o(1/\epsilon_n)$, $\lambda_n = o(\sqrt{(n-m)/\log(1/\delta_n)})$ and $\sqrt{\lambda_n}\,\mathbb{E}\mathcal{W}_1(P_\mathcal{I}, P) \to 0$.

Remark D.17 (Polynomial confidence levels)**.** *The confidence level δ is ours to choose; letting it shrink with n turns the high-probability statements into rates. Take $\delta_n = n^{-a}$ with $a > 0$. Since $n - m > n/2$, Lemma D.5(iv) gives, with probability at least $1 - n^{-a}$, $T_n \leq \mathbb{E}T_n + \lambda\sqrt{a\log n/n}$, so under Assumption 3 Theorem D.8(b) yields*

$$\|\hat{\theta}_n - \theta_0\| \leq \Big(\big(2\lambda\epsilon_n + 2\mathbb{E}T_n + 2\lambda\sqrt{a\log n/n} + \xi_n\big)/\kappa\Big)^{1/2}$$

with the same probability; only the concentration term depends on a. In Theorem D.15(iii) the inlier condition becomes $\mathbb{P}(X \notin \mathcal{G}_\lambda(t)) \le n^{-(1+a)}$, and then $\mathbb{P}(\hat{\mathcal{O}}_\alpha \neq \mathcal{O}) \le 2n^{-a}$. For Gaussian inliers $\mathcal{N}(\theta_0, \sigma^2 I_d)$ and the point target this holds when $\sqrt{\lambda_n} \ge t + \sigma\big(\sqrt{d} + \sqrt{2(1+a)\log n}\big)$, because $\mathbb{P}(\|Z\| \ge \sqrt{d} + s) \le e^{-s^2/2}$ for $Z \sim \mathcal{N}(0, I_d)$ (Lemma E.6); the scale $\lambda_n \asymp \sigma^2(\sqrt{d} + \sqrt{\log n})^2$ is unchanged, and a enters only through the constant. If $a > 1$, then $\sum_n 2n^{-a} < \infty$, and the first Borel–Cantelli lemma, which requires no independence across n, gives $\mathbb{P}(\hat{\mathcal{O}}_\alpha \neq \mathcal{O}$ for infinitely many $n) = 0$: exact recovery holds almost surely for all large n, provided the conditions of Theorem D.15(iii) hold for all large n.

Interpretation. A faster-shrinking δ_n costs only a constant in the threshold and a factor $\sqrt{a}$ in the concentration term, and for $a > 1$ upgrades recovery with high probability to almost-sure recovery for all large n.

Corollary D.18 (Exact recovery with a fixed threshold)**.** *In the setting of Proposition D.12, suppose also that the support of P lies in $\mathcal{G}_\lambda(t)$, that $x_i \in \mathcal{B}_\lambda(t)$ for all $i \in \mathcal{O}$, and that $\omega_Q^{-1}\big(2t_n(\delta) + (\gamma_n + \xi_n)/(1-\epsilon_n)\big) \le t$. Then $\mathbb{P}(\hat{\mathcal{O}}_\alpha = \mathcal{O}) \ge 1 - \delta$.*

Proof. On $\{T_n \le t_n(\delta)\}$, which has probability at least $1-\delta$, Proposition D.12 gives $\|\hat{\theta}_n - \theta_0\| \le t$. Every inlier lies in the support of P, hence in $\mathcal{G}_\lambda(t)$, almost surely, and every outlier lies in $\mathcal{B}_\lambda(t)$; Theorem D.15(i) gives $\hat{\mathcal{O}}_\alpha = \mathcal{O}$. □

Interpretation. For bounded inliers the inlier event holds surely, so the recovery probability is that of the estimation event alone, and no growth of λ is needed.

Proposition D.19 (The inlier-reach scale is necessary)**.** *Let $\mathcal{U}$ be a neighborhood of θ_0 and $S_n = \{x : \sup_{\theta \in \mathcal{U}} \nu_\theta(A_{\lambda_n}(x)) < \alpha\}$, assumed measurable. If $(n-m)P(X \in S_n) \to \infty$ and $\mathbb{P}(\hat{\theta}_n \in \mathcal{U}) \to 1$, then $\mathbb{P}(\hat{\mathcal{O}}_\alpha \cap \mathcal{I} \neq \emptyset) \to 1$.*

Interpretation. The lower end of the window cannot be relaxed: once the estimator is localized, any threshold below the inliers' reach flags some inlier with probability tending to one. For the Gaussian point target, Proposition E.10 removes the localization assumption.

Proof. $\mathbb{P}(\text{no inlier in } S_n) = (1 - P(X \in S_n))^{n-m} \le \exp\{-(n-m)P(X \in S_n)\} \to 0$. On $\{\hat{\theta}_n \in \mathcal{U}\}$ an inlier in S_n has $\nu_{\hat{\theta}_n}(A_{\lambda_n}(x_i)) < \alpha$ and is flagged. □

Remark D.20 (Reading the window)**.** *Estimation consistency alone does not require $\lambda_n \to \infty$ (Theorem D.8(c)). Growth is forced only by exact recovery with unbounded inliers ((W1) and Proposition D.19, whose conditions match up to the margin t and the neighborhood $\mathcal{U}$), and it tightens (W2) and (W3)(i). For bounded inliers a fixed λ suffices. The $\sqrt{\log n}$ scale is that of the universal threshold (Donoho & Johnstone, 1994), and Kong et al. (2018) note informally that outliers must exceed the inlier extreme-value scale. For Gaussian inliers and the point target, (W1) holds once $\sqrt{\lambda_n} \ge t + \sigma(\sqrt{d} + \sqrt{2\log(n/\delta_n)})$, and for the spread target once $\sqrt{\lambda_n} \ge t + \sqrt{\chi^2_{d,\alpha}} + \sigma(\sqrt{d} + \sqrt{2\log(n/\delta_n)})$ (Remark D.14 and Gaussian concentration).*

D.5 Fixed threshold for bounded inliers

Collecting Theorem D.8(d), Proposition D.12 and Corollary D.18 in the notation of Section 4.1 gives the fixed-threshold counterpart of Theorem 4.2, which the main text summarizes in one sentence.

Theorem D.21 (Fixed threshold, bounded inliers)**.** *Let λ be fixed and $\epsilon_n \to \epsilon < 1/2$. Suppose the model is well specified ($P = \nu_{\theta_0}$, so $J_0 = 0$) and* trimmed-identifiable *($(1-\bar{\epsilon})\nu_{\theta_0} \le \nu_\theta$ only at θ_0, for some $\bar{\epsilon} > \epsilon$), and that the separation of Proposition 4.1 holds with some $\Delta \ge 0$ and $\tau = 0$ for all θ within a radius $R > b(2\lambda\epsilon)$ of θ_0 and all large n (in practice, bounded-support models). Let $Q_{\bar{\epsilon}}$ be $\hat{Q}_n$ with P in place of $P_{\mathcal{I}}$ and level $\bar{\epsilon}$, and $b_{\mathrm{POT}}(s) = \sup\{\|\theta - \theta_0\| : \|\theta - \theta_0\| \le R,\ Q_{\bar{\epsilon}}(\theta) \le s\}$.*

- *(a)* Consistency. *If $r = 1$, then $\hat{\theta}_n \to \theta_0$ almost surely despite the fixed fraction of outliers. If $r > 1$, $\limsup_n \|\hat{\theta}_n - \theta_0\| \le \inf_{s > s_r} b_{\mathrm{POT}}(s)$ almost surely, with $s_r = \frac{\epsilon}{1-\epsilon}\frac{r-1}{r}\lambda\big(\frac{\lambda}{\lambda+\Delta}\big)^{1/(r-1)}$, which grows with r and vanishes as $\Delta \to \infty$.*

(b) Rate. *Suppose $Q_{\bar\epsilon}(\theta) \ge \omega_Q(\|\theta-\theta_0\|)$ on that ball, for a nondecreasing ω_Q with $\omega_Q(0) = 0 < \omega_Q(u)$ for $u > 0$, and let $\omega_Q^{-1}(v) = \sup\{u : \omega_Q(u) \le v\}$. Then for every n with $\epsilon_n \le \bar\epsilon$, $2\lambda\epsilon_n + 2t_n(\delta) < \lambda$ and $b(2\lambda\epsilon_n + 2t_n(\delta)) < R$ (all large n), with probability at least $1-\delta$, $\|\hat\theta_n - \theta_0\| \le \omega_Q^{-1}\big(2t_n(\delta) + \gamma_n/(1-\epsilon_n)\big)$, with γ_n of Proposition 4.1 at $\tau = 0$ (zero for $r = 1$).*

(c) Exact recovery with a fixed threshold. *If, moreover, the support of P lies in $\mathcal{G}_\lambda(t)$, all outliers lie in $\mathcal{B}_\lambda(t)$, and the bound in (b) is at most t, then $\mathbb{P}(\hat{\mathcal{O}} = \mathcal{O}) \ge 1-\delta$.*

Proof. (a) is Theorem D.8(d) and Corollary D.9; (b) is Proposition D.12 with $\xi_n = 0$; (c) is Corollary D.18. □

Interpretation. By the bias bound $M(\hat\theta_n) \le 2\lambda\epsilon_n + 2T_n$ (Theorem D.8(a); Theorem 4.2(a)) and $T_n \to 0$, $\limsup_n \|\hat\theta_n - \theta_0\| \le b(2\lambda\epsilon)$ almost surely, so the separation and trimmed identifiability are needed only on that ball. There each outlier costs a constant under the hard clip (Proposition 4.1), so a fixed fraction leaves no bias. Because bounded inliers never leave $\mathcal{G}_\lambda(t)$ once λ covers their support, recovery needs no growth of λ, and its only failure event is a large estimation error. The rate is set by the growth ω_Q of the trimmed criterion: for $r = 1$ the guaranteed bound is $O(n^{-1/6})$ for uniform location in $d = 1$ (Remark D.13). Unbounded inliers, e.g. a well-specified Gaussian model (full support, so $\tau > 0$), are treated next.

E Point-Target Specialization: Full Statements and Proofs

Some symbols have a local meaning in this appendix that differs from Appendix D: ϵ is a fixed outlier fraction, I, O are the inlier and outlier index sets $\mathcal{I}, \mathcal{O}$, τ, η are noise radii, b is a localization radius, D is the outlier separation, and M_r, κ_r are the matrix and constant of the influence function; each is defined where it first appears. This appendix specializes the general framework (Appendix D) to location estimation, mainly with a point target $\nu_\theta = \delta_\theta$, where PW-ROT reduces to an M-estimator. It contains only the results used in the main text and in Appendix D: the calculus of ρ (Lemma E.1); the M-estimator reduction (Proposition E.2); the influence function (Proposition E.3) and the curvature of the criterion at spherically symmetric models (Lemma E.4); uniqueness of the center at the Gaussian model (Lemma E.5); the explicit point-target estimation, recovery and rate results and the sharpness of the soft-power bias (Theorem E.7, Corollaries E.8, E.9, Proposition E.11); moment-free consistency (Theorem E.13); and the 95%-rule lemma for the spread target (Lemma E.14).

E.1 Notation and elementary facts

We collect the calculus of ρ used throughout: its derivative is the weight function W_r, its Hessian is controlled off the threshold sphere, and its score is bounded by $\sqrt{\lambda}$. Fix $\lambda > 0$ and $r \ge 1$; for $r > 1$ put $q = r/(r-1) \in (1, \infty)$. The robust cost is

$$\rho(C) = \rho_{\lambda,r}(C) = \begin{cases} C/r, & C \le \lambda, \\ \lambda + \frac{1-r}{r}\lambda^q C^{-1/(r-1)}, & C > \lambda,\ r > 1, \\ \lambda, & C > \lambda,\ r = 1, \end{cases} \qquad w^\star(C) = \min\{1, (\lambda/C)^{1/(r-1)}\}$$

($w^\star = \mathbb{1}\{C \le \lambda\}$ for $r = 1$), and $W_r(s) = \min\{1, (\lambda/s)^q\}$ ($W_1(s) = \mathbb{1}\{s \le \lambda\}$). For data $x_1, \dots, x_n \in \mathbb{R}^d$ write $s_i(\mu) = \|x_i - \mu\|^2$,

$$G_n(\mu) = \frac{1}{n}\sum_{i=1}^n \rho\big(s_i(\mu)\big), \qquad \psi_r(x;\mu) = W_r(\|x-\mu\|^2)(x-\mu).$$

Point i is flagged at μ when $w^\star(s_i(\mu)) < 1$, i.e. $s_i(\mu) > \lambda$.

Lemma E.1 (Calculus of ρ)**.** *(a) ρ is continuous, non-decreasing, $0 \le \rho < \lambda$ for $r > 1$ and $0 \le \rho \le \lambda$ for $r = 1$, and $\rho(C) \to \lambda$ as $C \to \infty$. (b) For $r > 1$, $\rho \in C^1[0,\infty)$ with $\rho'(C) = \frac{1}{r}W_r(C) = \frac{1}{r}w^\star(C)^r$. (c) For $r > 1$ and $\vartheta \ge 1$, $\lambda - \rho(\vartheta^2\lambda) = \lambda\frac{r-1}{r}\vartheta^{-2/(r-1)}$. (d) Let $f_x(\mu) = \rho(\|x-\mu\|^2)$. For $r > 1$, $\nabla f_x(\mu) = -\frac{2}{r}\psi_r(x;\mu)$; on $\{\mu : \|x-\mu\|^2 \ne \lambda\}$, f_x is C^∞ with*

$$\nabla^2 f_x(\mu) = \tfrac{2}{r}I \ \ (s < \lambda), \qquad \nabla^2 f_x(\mu) = \tfrac{2}{r}\Big(\tfrac{\lambda}{s}\Big)^q\big(I - 2q\,uu^\top\big) \ \ (s > \lambda),$$

$s = \|x-\mu\|^2$, $u = (x-\mu)/\|x-\mu\|$, $I = I_d$. *Hence* $-\frac{2}{r}(2q-1)(\lambda/s)^q I \preceq \nabla^2 f_x(\mu) \preceq \frac{2}{r} I$ *(Loewner order) off the sphere* $\{\mu : \|x-\mu\|^2 = \lambda\}$. *For* $r = 1$, $\nabla^2 f_x = 2I$ *for* $s < \lambda$ *and* 0 *for* $s > \lambda$. *(e) (Bounded score)* $\|\psi_r(x;\mu)\| = W_r(s)\sqrt{s} \le \sqrt{\lambda}$, *with equality iff* $s = \lambda$; *for* $s \ge \lambda$ *and* $r > 1$, $\|\psi_r\| = \lambda^q s^{1/2-q}$, *which is decreasing in* s. *For* $r = 1$, $\psi_1 = 0$ *when* $s > \lambda$. *(f) (Lipschitz cost) For every* y, $x \mapsto \rho(\|x-y\|^2)$ *is* L_r*-Lipschitz with* $L_r = 2\sqrt{\lambda}/r$.

Interpretation. Inside the threshold the cost is the quadratic C/r and the score is the residual itself; outside, the score is bounded by $\sqrt{\lambda}$ and decays like $s^{1/2-q}$, and the cost is Lipschitz with constant $2\sqrt{\lambda}/r$, which is all that the general framework needs.

Proof. (a),(b): on $C > \lambda$, $\frac{d}{dC}\left[\frac{1-r}{r}\lambda^q C^{-1/(r-1)}\right] = \frac{1}{r}\lambda^q C^{-1/(r-1)-1} = \frac{1}{r}(\lambda/C)^q$ because $\frac{1}{r-1} + 1 = q$; at $C = \lambda$ the two branches have value λ/r and one-sided derivatives $1/r$, so $\rho \in C^1$. Since $w^\star(C)^r = (\lambda/C)^{r/(r-1)}$ for $C > \lambda$, $\rho' = \frac{1}{r}W_r = \frac{1}{r}(w^\star)^r$. Monotonicity follows from $\rho' \ge 0$ (and from the formula when $r = 1$); the second branch is $< \lambda$ and tends to λ. (c): $\lambda - \rho(\vartheta^2\lambda) = \frac{r-1}{r}\lambda^q(\vartheta^2\lambda)^{-1/(r-1)} = \frac{r-1}{r}\lambda^{q-1/(r-1)}\vartheta^{-2/(r-1)}$ and $q - \frac{1}{r-1} = 1$. (d): $\nabla_\mu s = -2(x-\mu)$, so $\nabla f_x = \rho'(s)\nabla s = -\frac{2}{r}W_r(s)(x-\mu)$. Differentiating again off the sphere, $\nabla^2 f_x = \frac{2}{r}\big[W_r(s)I + 2W_r'(s)(x-\mu)(x-\mu)^\top\big]$; $W_r' = 0$ for $s < \lambda$ and $W_r'(s) = -\frac{q}{s}W_r(s)$ for $s > \lambda$. The eigenvalues are $\frac{2}{r}W_r$ (multiplicity $d-1$) and $\frac{2}{r}W_r(1-2q)$, and $0 < W_r \le 1$. (e): for $s \le \lambda$, $\|\psi_r\| = \sqrt{s} \le \sqrt{\lambda}$; for $s > \lambda$, $\|\psi_r\| = \lambda^q s^{1/2-q}$ is strictly decreasing since $q > 1/2$, and equals $\sqrt{\lambda}$ at $s = \lambda$. (f): for $r > 1$, $\|\nabla_x \rho(\|x-y\|^2)\| = \frac{2}{r}\|\psi_r(x;y)\| \le \frac{2\sqrt{\lambda}}{r}$ by (e), and a C^1 function with bounded gradient is Lipschitz. For $r = 1$, $\min(\|x-y\|^2, \lambda) = \min(\|x-y\|, \sqrt{\lambda})^2$ is the composition of the 1-Lipschitz map $x \mapsto \min(\|x-y\|, \sqrt{\lambda}) \in [0, \sqrt{\lambda}]$ with $t \mapsto t^2$, which is $2\sqrt{\lambda}$-Lipschitz on $[0, \sqrt{\lambda}]$. □

E.2 M-ESTIMATOR REDUCTION

For a point target there is a single coupling, so PW-ROT reduces to an M-estimator with the bounded loss ρ; this links it to classical robust location estimation.

Proposition E.2 (M-estimator reduction)**.** *Let* $c(x,y) = \|x-y\|^2$ *and* $P_n = \frac{1}{n}\sum_i \delta_{x_i}$.

- *(i) For every* μ, $\mathrm{OT}_\rho(P_n, \delta_\mu) = G_n(\mu)$.
- *(ii)* G_n *attains its minimum, and every minimizer* $\hat{\mu}$ *satisfies* $G_n(\hat{\mu}) \le \frac{n-1}{n}\lambda$.
- *(iii) If* $r > 1$, *every local minimizer of* G_n *solves* $\sum_i \psi_r(x_i;\mu) = 0$.
- *(iv) If* $r = 1$, *at every local minimizer* $\hat{\mu}$ *of* G_n *no point lies on the sphere,* $s_i(\hat{\mu}) \ne \lambda$ *for all* i. *If moreover* $G_n(\hat{\mu}) < \lambda$ *(in particular at every global minimizer), then* $\mathcal{S} = \{i : s_i(\hat{\mu}) < \lambda\} \ne \emptyset$ *and* $\hat{\mu} = |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}} x_i$; *equivalently* $\sum_i \psi_1(x_i;\hat{\mu}) = 0$.

Minimizers need not be unique, and roots of $\sum_i \psi_r(x_i;\mu) = 0$ *need not be minimizers.*

Interpretation. For a point target PW-ROT is a redescending M-estimator: for $r > 1$ it solves a weighted-mean equation with weights W_r, and for $r = 1$ it is the mean of the points within the threshold, a skipped mean.

Proof. (i) The only coupling of P_n and δ_μ puts mass $1/n$ on each pair (x_i, μ), so $\mathrm{OT}_\rho(P_n, \delta_\mu) = \frac{1}{n}\sum_i \rho(\|x_i - \mu\|^2)$. No envelope argument is needed.

(ii) G_n is continuous (Lemma E.1(a)). $G_n(x_1) \le \frac{1}{n}\rho(0) + \frac{n-1}{n}\lambda = \frac{n-1}{n}\lambda$. As $\|\mu\| \to \infty$ every $s_i(\mu) \to \infty$, so $G_n(\mu) \to \lambda$. Hence there is R with $G_n > \frac{n-1}{n}\lambda$ outside $B(0,R)$ (for $r = 1$, $G_n = \lambda$ once all $s_i > \lambda$). The minimum over the compact ball $\bar{B}(0,R)$ exists and is global.

(iii) For $r > 1$, $G_n \in C^1$ with $\nabla G_n(\mu) = -\frac{2}{rn}\sum_i \psi_r(x_i;\mu)$ (Lemma E.1(d)), so every local minimizer is stationary.

(iv) Let $\hat{\mu}$ be a local minimizer and suppose $K = \{k : s_k(\hat{\mu}) = \lambda\} \ne \emptyset$. For $i \notin K$, $s_i(\hat{\mu}) \ne \lambda$, so on a neighborhood V of $\hat{\mu}$ each such term $\min(s_i, \lambda)$ equals either $s_i(\mu)$ or λ; call their sum $a_K(\mu)$,

a smooth function on V. Fix $k \in K$ and define on V

$$H_0(\mu) = \tfrac{1}{n}\big[a_K(\mu) + |K|\lambda\big], \qquad H_1(\mu) = \tfrac{1}{n}\big[a_K(\mu) + s_k(\mu) + (|K|-1)\lambda\big].$$

Because $\min(s,\lambda) \le \lambda$ and $\min(s,\lambda) \le s$, we have $G_n \le H_0$ and $G_n \le H_1$ on V, with equality at $\hat\mu$. Hence $H_j(\hat\mu) = G_n(\hat\mu) \le G_n(\mu) \le H_j(\mu)$ near $\hat\mu$, so $\hat\mu$ is a local minimizer of the smooth functions H_0 and H_1. Then $\nabla a_K(\hat\mu) = 0$ and $\nabla a_K(\hat\mu) + \nabla s_k(\hat\mu) = 0$, so $\nabla s_k(\hat\mu) = -2(x_k - \hat\mu) = 0$, i.e. $s_k(\hat\mu) = 0 \ne \lambda$, which is a contradiction. So $K = \emptyset$, and near $\hat\mu$, $G_n(\mu) = \frac{1}{n}[\sum_{i\in\mathcal{S}} s_i(\mu) + (n - |\mathcal{S}|)\lambda]$. If $G_n(\hat\mu) < \lambda$ then $\mathcal{S} \ne \emptyset$, and stationarity of this smooth function gives $\sum_{i\in\mathcal{S}}(x_i - \hat\mu) = 0$. By (ii), a global minimizer has $G_n \le \frac{n-1}{n}\lambda < \lambda$.

Non-uniqueness: take $d = 1$, $x_{1,2} = \pm a$. For $r = 1$ and $a^2 > \lambda$, $G_n(\pm a) = \lambda/2 < \lambda = G_n(0)$, so $\{a, -a\}$ are two global minimizers and the root $\mu = 0$ of $\sum \psi_1 = 0$ is not a minimizer. For $r > 1$, $G_n(\pm a) = \frac{1}{2}\rho(4a^2)$ and $G_n(0) = \rho(a^2)$; as $a \to \infty$ these tend to $\lambda/2$ and λ, so for large a the symmetric root 0 is not a minimizer and, by symmetry, the minimizers come in pairs $\pm\hat\mu$. □

E.3 Influence function

We derive the influence function of the point-target estimator and bound its gross-error sensitivity, and we compute the curvature matrix M_r at spherically symmetric models (Lemma E.4), which Appendix D uses to verify Assumption 3.

Let $X \sim F$ for a law F on $\mathbb{R}^d$, $G_F(\mu) = \mathbb{E}_F\rho(\|X-\mu\|^2)$, $\Psi_F(\mu) = \mathbb{E}_F\psi_r(X;\mu)$, and let μ_∞ be a local minimizer of G_F. Write $S = \|X - \mu_\infty\|^2$ and, when $X \ne \mu_\infty$, $U = (X - \mu_\infty)/\|X - \mu_\infty\|$.

Assumption 4 (Regularity of F)**.** *For $r > 1$: $F(\{x : \|x-\mu\|^2 = \lambda\}) = 0$ for all μ in a neighborhood V of μ_∞. For $r = 1$: F has a density f that is C^1 on a neighborhood of $\bar B(\mu_\infty, \sqrt{\lambda})$. (Both hold for any Gaussian or any smooth density.)*

Interpretation. The law puts no mass on threshold spheres, so the score is differentiable in μ after integration; this holds for every law with a density.

Proposition E.3 (Influence function and gross-error sensitivity)**.** *Under Assumption 4, Ψ_F is C^1 near μ_∞. Let $M_r = -D\Psi_F(\mu_\infty)$, where D denotes the Jacobian (unrelated to the separation D of Section E.4), and write $\mathbb{E}[Y;A] = \mathbb{E}[Y\mathbb{1}_A]$:*

$$M_r = \mathbb{E}\big[W_r(S)\big] I - 2q\,\mathbb{E}\big[(\lambda/S)^q\, UU^\top \mathbb{1}\{S > \lambda\}\big] \qquad (r > 1),$$
$$M_1 = \mathbb{P}(S < \lambda)\, I - \lambda^{-1/2} \textstyle\int_{\partial B(\mu_\infty, \sqrt{\lambda})} (x - \mu_\infty)(x - \mu_\infty)^\top f(x)\, d\sigma(x),$$

with $\partial B(\mu_\infty, \sqrt{\lambda})$ the sphere of radius $\sqrt{\lambda}$ and $d\sigma$ its surface measure. Assume M_r is nonsingular. Then for every x with $\|x - \mu_\infty\|^2 \ne \lambda$ there is, for small $t \ge 0$, a unique root $T(F_t)$ of $\Psi_{F_t} = 0$ near μ_∞ (the estimator viewed as a functional T of the law), where $F_t = (1-t)F + t\delta_x$. The map $t \mapsto T(F_t)$ is C^1 and

$$\mathrm{IF}_r(x; F) = \frac{d}{dt} T(F_t)\Big|_{t=0} = M_r^{-1}\psi_r(x; \mu_\infty).$$

If $M_r \succ 0$ (positive definite), $T(F_t)$ is a strict local minimizer of G_{F_t}. The gross-error sensitivity is

$$\gamma_r^\star := \sup_{x:\|x-\mu_\infty\|^2 \ne \lambda} \|\mathrm{IF}_r(x; F)\| = \frac{\sqrt{\lambda}}{\sigma_{\min}(M_r)},$$

where $\sigma_{\min}$ is the smallest singular value. For $r > 1$ and every unit vector v,

$$v^\top M_r v \;\ge\; \kappa_r := \mathbb{P}(S \le \lambda) + (1 - 2q)\,\mathbb{E}\big[(\lambda/S)^q; S > \lambda\big], \tag{5}$$

so $\kappa_r > 0$ implies $M_r \succ 0$ and $\gamma_r^\star \le \sqrt{\lambda}/\kappa_r$. If $M_r \succ 0$, then moreover $\gamma_r^\star \ge \sqrt{\lambda}$, since $\sigma_{\min}(M_r) = \min_{\|v\|=1} v^\top M_r v \le \mathbb{E}W_r(S) \le 1$.

Interpretation. Because the score is bounded by $\sqrt{\lambda}$, the influence of a single point is bounded, and the gross-error sensitivity is $\sqrt{\lambda}$ divided by the smallest singular value of M_r.

Proof. C^1 *for* $r > 1$. Fix x. Along any segment $u \mapsto \mu_u$, $s(u) = \|x - \mu_u\|^2$ is a quadratic in u and crosses λ at most twice. The map $\mu \mapsto \psi_r(x;\mu)$ is continuous, and it is C^1 off the sphere $\{\|x-\mu\|^2 = \lambda\}$ with $D_\mu \psi_r = -[W_r I + 2W_r'(s)(x-\mu)(x-\mu)^\top]$. This matrix has operator norm $\le \max(1, 2q-1) =: \ell$ by the eigenvalue computation of Lemma E.1(d). Hence $\psi_r(x;\cdot)$ is ℓ-Lipschitz, uniformly in x. For $\mu \in V$ and increments $e \to 0$, the quotient $\big(\psi_r(X;\mu+e) - \psi_r(X;\mu) - D_\mu\psi_r(X;\mu)e\big)/\|e\|$ is bounded by 2ℓ and tends to 0 for every X off the sphere around μ, a set of F-measure one. Dominated convergence gives differentiability, with $D\Psi_F(\mu) = \mathbb{E} D_\mu \psi_r(X;\mu)$. The same argument with $\mu_k \to \mu$ shows that $\mu \mapsto \mathbb{E} D_\mu \psi_r(X;\mu)$ is continuous on V. Evaluating at μ_∞ with $W_r'(S) = -\frac{q}{S} W_r(S) \mathbb{1}\{S > \lambda\}$ gives the formula for M_r.

C^1 *for* $r = 1$. $\Psi_F(\mu) = \int_{B(0,\sqrt{\lambda})} z\, f(\mu + z)\, dz$, and $f \in C^1$ near the closed ball, so $D\Psi_F(\mu) = \int_B z\, \nabla f(\mu+z)^\top dz$, which is continuous in μ. By the divergence theorem on $B = B(0,\sqrt{\lambda})$ applied to $z_i f(\mu+z) e_j$, with e_j the j-th basis vector, whose divergence is $\mathbb{1}\{i=j\} f + z_i \partial_j f$, $\int_B z_i \partial_j f\, dz = \int_{\partial B} z_i z_j f\, \lambda^{-1/2} d\sigma - \mathbb{1}\{i=j\} \int_B f\, dz$. This is the formula for M_1 ($\mathbb{P}(S = \lambda) = 0$ under a density).

IF. Put $H(t,\mu) = (1-t)\Psi_F(\mu) + t\psi_r(x;\mu) = \Psi_{F_t}(\mu)$. If $\|x - \mu_\infty\|^2 \neq \lambda$, then $\psi_r(x;\cdot)$ is C^1 near μ_∞, so H is C^1 near $(0,\mu_\infty)$. Also $H(0,\mu_\infty) = \Psi_F(\mu_\infty) = 0$: G_F is differentiable with $\nabla G_F = -\frac{2}{r}\Psi_F$ (dominated differentiation, using the bound $\|\nabla f_x\| \le 2\sqrt{\lambda}/r$ of Lemma E.1(e)), and μ_∞ is a local minimizer. For $r = 1$, $\nabla G_F = -2\Psi_F$ by the same divergence computation applied to $\mathbb{E}\min(\|X-\mu\|^2, \lambda)$. Finally $D_\mu H(0,\mu_\infty) = -M_r$ is invertible. The implicit function theorem yields a unique C^1 curve $\mu_t = T(F_t)$ near μ_∞ with $H(t,\mu_t) = 0$, and $\dot\mu_0 = -D_\mu H^{-1} \partial_t H = M_r^{-1}(\psi_r(x;\mu_\infty) - \Psi_F(\mu_\infty)) = M_r^{-1}\psi_r(x;\mu_\infty)$. If $M_r \succ 0$, the Jacobian $-D_\mu H(t,\mu_t)$ is positive definite for small t by continuity. It equals $\frac{r}{2}$ times the Hessian of G_{F_t} (which exists near μ_t), so μ_t is a strict local minimizer.

Sensitivity. As x ranges over $\{0 < \|x - \mu_\infty\|^2 < \lambda\}$, $\psi_r(x;\mu_\infty) = x - \mu_\infty$ covers the punctured open ball of radius $\sqrt{\lambda}$. Off this set $\|\psi_r\| < \sqrt{\lambda}$ (Lemma E.1(e)). Hence $\gamma_r^\star = \sup_{\|z\|<\sqrt{\lambda}} \|M_r^{-1} z\| = \sqrt{\lambda}\|M_r^{-1}\| = \sqrt{\lambda}/\sigma_{\min}(M_r)$.

Bound (5). $v^\top M_r v = \mathbb{E} W_r(S) - 2q\mathbb{E}[(\lambda/S)^q (v^\top U)^2; S > \lambda] \ge \mathbb{E} W_r(S) - 2q\mathbb{E}[(\lambda/S)^q; S > \lambda]$ because $(v^\top U)^2 \le 1$. Also $\mathbb{E} W_r(S) = \mathbb{P}(S \le \lambda) + \mathbb{E}[(\lambda/S)^q; S > \lambda]$. For the lower bound on $\gamma_r^\star$ when $M_r \succ 0$: the second term of $v^\top M_r v$ is non-positive, so $\sigma_{\min}(M_r) = \min_{\|v\|=1} v^\top M_r v \le \mathbb{E} W_r(S) \le 1$. □

Lemma E.4 (Curvature at spherically symmetric models)**.** *Let $X - \mu_\infty$ have density $x \mapsto g(\|x\|^2)$ on $\mathbb{R}^d$, with radial profile g ($d \ge 1$; here μ_∞ is the center of symmetry, a root of Ψ_F by symmetry), where $g : [0,\infty) \to [0,\infty)$ is non-increasing and locally absolutely continuous on $(0,\infty)$. Put $S = \|X - \mu_\infty\|^2$, $c_d = \pi^{d/2}/\Gamma(d/2)$, so that S has density $f_S(s) = c_d\, s^{d/2-1} g(s)$. Define*

$$m_1 := \mathbb{P}(S \le \lambda) - \tfrac{2}{d}\lambda f_S(\lambda), \qquad m_r := m_1 + \frac{2\lambda c_d}{d} \int_\lambda^\infty \Big(\frac{\lambda}{s}\Big)^{q-1} s^{d/2-1} \big(-g'(s)\big)\, ds \quad (r > 1).$$

Then:

- *(i) For every $r > 1$, $M_r = m_r I$, i.e. $m_r = \mathbb{P}(S \le \lambda) + (1 - \frac{2q}{d})\mathbb{E}[(\lambda/S)^q; S > \lambda]$. If in addition g is C^1 on a neighborhood of $[0,\lambda]$, then $M_1 = m_1 I$.*
- *(ii) $m_1 \ge 0$, with equality if and only if g is constant on $(0,\lambda]$.*
- *(iii) If $\mathbb{P}(S > \lambda) > 0$, then $m_r > 0$ for every $r > 1$, so $M_r \succ 0$ and μ_∞ is a strict local minimizer of G_F; the same holds for $r = 1$ if g is C^1 near $[0,\lambda]$ and not constant on $[0,\lambda]$.*

For $F = \mathcal{N}(\mu_\infty, \sigma^2 I_d)$, $-g' = g/(2\sigma^2)$ and the formula reads

$$m_r = m_1 + \frac{\lambda}{d\sigma^2}\, \mathbb{E}\Big[\Big(\frac{\lambda}{S}\Big)^{q-1}; S > \lambda\Big], \qquad m_1 = \mathbb{P}(S \le \lambda) - \tfrac{2}{d}\lambda f_S(\lambda) > 0,$$

so (iii) holds at every Gaussian, for all $d \ge 1$, $\lambda > 0$ and $r \ge 1$.

Interpretation. At spherically symmetric models the curvature matrix is a positive multiple of the identity, so the clean criterion is locally strictly convex at the center; this is the curvature used for Assumption 3.

Proof. Preliminaries. The polar-coordinate formula $\mathbb{P}(S \le t) = |\mathbb{S}^{d-1}| \int_0^{\sqrt{t}} v^{d-1} g(v^2)\, dv$, with $|\mathbb{S}^{d-1}| = 2\pi^{d/2}/\Gamma(d/2)$ the area of the unit sphere $\mathbb{S}^{d-1}$, and the substitution $s = v^2$ give $f_S = c_d s^{d/2-1} g$. Since g is non-increasing and $\int_0^\infty s^{d/2-1} g(s)\, ds = 1/c_d < \infty$, for $s > 0$

$$\int_{s/2}^{s} t^{d/2-1} g(t)\, dt \;\ge\; g(s) \int_{s/2}^{s} t^{d/2-1}\, dt = \tfrac{2}{d}(1 - 2^{-d/2})\, s^{d/2} g(s),$$

and the left side tends to 0 as $s \to \infty$. Hence $s^{d/2} g(s) \to 0$; in particular $g(s) \to 0$ and $s^a g(s) \to 0$ for every $a \le d/2$.

(i). By Assumption 4 for $r > 1$ (spheres are Lebesgue-null), Proposition E.3 gives M_r. Spherical symmetry makes U uniform on $\mathbb{S}^{d-1}$ and independent of S, so $\mathbb{E}[UU^\top \mid S] = I/d$ and $M_r = \big(\mathbb{E}W_r(S) - \frac{2q}{d}\mathbb{E}[(\lambda/S)^q; S > \lambda]\big) I$; with $\mathbb{E}W_r(S) = \mathbb{P}(S \le \lambda) + \mathbb{E}[(\lambda/S)^q; S > \lambda]$ this is $\big(\mathbb{P}(S \le \lambda) + (1 - \frac{2q}{d})A\big) I$, $A := \mathbb{E}[(\lambda/S)^q; S > \lambda]$. By symmetry $\Psi_F(\mu_\infty) = \mathbb{E}[W_r(S)(X - \mu_\infty)] = 0$, so Proposition E.3 applies at μ_∞ (its proof uses only $\Psi_F(\mu_\infty) = 0$). It remains to identify the scalar with m_r. Put $a = d/2 - q$, so $1 - 2q/d = 2a/d$ and $A = c_d \lambda^q \int_\lambda^\infty s^{a-1} g(s)\, ds$ (absolutely convergent: $s^{a-1} \le \lambda^{-q} s^{d/2-1}$ on $s \ge \lambda$). For $\bar{s} > \lambda$, integration by parts for the absolutely continuous functions s^a and g on $[\lambda, \bar{s}]$ gives

$$\int_\lambda^{\bar{s}} a\, s^{a-1} g(s)\, ds = \bar{s}^a g(\bar{s}) - \lambda^a g(\lambda) + \int_\lambda^{\bar{s}} s^a \big(-g'(s)\big)\, ds.$$

As $\bar{s} \to \infty$ the left side converges, $\bar{s}^a g(\bar{s}) \to 0$ by the preliminaries ($a < d/2$), and the last integral increases (its integrand is ≥ 0 a.e.) to a finite limit. Therefore

$$\big(1 - \tfrac{2q}{d}\big) A = \tfrac{2}{d} c_d \lambda^q \Big[-\lambda^a g(\lambda) + \int_\lambda^\infty s^a (-g')\, ds\Big] = -\tfrac{2}{d}\lambda f_S(\lambda) + \tfrac{2\lambda c_d}{d} \int_\lambda^\infty \Big(\frac{\lambda}{s}\Big)^{q-1} s^{d/2-1} (-g')\, ds,$$

using $c_d \lambda^q \lambda^{d/2-q} g(\lambda) = \lambda\, c_d \lambda^{d/2-1} g(\lambda) = \lambda f_S(\lambda)$ and $\lambda^q s^{d/2-q} = \lambda (\lambda/s)^{q-1} s^{d/2-1}$. This is $m_r - \mathbb{P}(S \le \lambda)$. For $r = 1$ with $g \in C^1$ near $[0, \lambda]$, the density is C^1 near $\bar{B}(\mu_\infty, \sqrt{\lambda})$, and the boundary term of M_1 in Proposition E.3 equals $\frac{\lambda}{d} \lambda^{-1/2} \int_{\partial B} f\, d\sigma\, I = \frac{2\lambda}{d} f_S(\lambda) I$ (since $\mathbb{E}[UU^\top \mid S] = I/d$ on the sphere and $\int_{\partial B(\mu_\infty, \sqrt{\lambda})} f\, d\sigma = 2\sqrt{\lambda} f_S(\lambda)$). Hence $M_1 = m_1 I$.

(ii). $\mathbb{P}(S \le \lambda) = c_d \int_0^\lambda s^{d/2-1} g(s)\, ds \ge c_d g(\lambda) \int_0^\lambda s^{d/2-1} ds = \frac{2}{d} c_d \lambda^{d/2} g(\lambda) = \frac{2}{d} \lambda f_S(\lambda)$, because $g(s) \ge g(\lambda)$ on $[0, \lambda]$. Equality holds iff $g(s) = g(\lambda)$ for a.e. $s \in (0, \lambda)$, i.e. (by monotonicity) iff g is constant on $(0, \lambda]$.

(iii). Write $e(r) = \int_\lambda^\infty (\lambda/s)^{1/(r-1)} s^{d/2-1} (-g'(s))\, ds$ (note $q - 1 = 1/(r-1)$), so that $m_r = m_1 + \frac{2\lambda c_d}{d} e(r)$. First, $\int_\lambda^\infty s^{d/2-1}(-g')\, ds < \infty$: integrating by parts on $[\lambda, \bar{s}]$, $\int_\lambda^{\bar{s}} s^{d/2-1}(-g') = \lambda^{d/2-1} g(\lambda) - \bar{s}^{d/2-1} g(\bar{s}) + (\frac{d}{2} - 1) \int_\lambda^{\bar{s}} s^{d/2-2} g$, and every term has a finite limit (for the last, $s^{d/2-2} \le \lambda^{-1} s^{d/2-1}$ on $s \ge \lambda$); so $e(r)$ is finite, since $0 < (\lambda/s)^{1/(r-1)} < 1$ for $s > \lambda$. Since $\mathbb{P}(S > \lambda) > 0$, $g(s_0) > 0$ for some $s_0 > \lambda$; as $g(s) \to 0$, $g(s_0) = \int_{s_0}^\infty (-g')\, ds > 0$, so $-g' > 0$ on a subset of (λ, ∞) of positive measure, and $e(r) > 0$. With $m_1 \ge 0$ from (ii), $m_r > 0$, so $M_r = m_r I \succ 0$. G_F is C^1 with $\nabla G_F = -\frac{2}{r}\Psi_F$, Ψ_F is C^1 with $D\Psi_F(\mu_\infty) = -M_r$ (Proposition E.3), and $\Psi_F(\mu_\infty) = 0$; so $\nabla^2 G_F(\mu_\infty) = \frac{2}{r} M_r \succ 0$ and μ_∞ is a strict local minimizer. For $r = 1$ with $g \in C^1$ near $[0, \lambda]$ and not constant there, $m_1 > 0$ by (ii) and $M_1 = m_1 I$ by (i), and the same argument applies.

Gaussian case. $g(s) = (2\pi\sigma^2)^{-d/2} e^{-s/(2\sigma^2)}$ is strictly decreasing and smooth, with $-g' = g/(2\sigma^2)$; then $c_d s^{d/2-1}(-g'(s)) = f_S(s)/(2\sigma^2)$ and $\frac{2\lambda}{d} \int_\lambda^\infty (\lambda/s)^{q-1} \frac{f_S(s)}{2\sigma^2} ds = \frac{\lambda}{d\sigma^2} \mathbb{E}[(\lambda/S)^{q-1}; S > \lambda]$. $\mathbb{P}(S > \lambda) > 0$ and g is not constant on $[0, \lambda]$, so $m_1 > 0$ by (ii). □

Lemma E.5 (The center minimizes G_F at the Gaussian model)**.** *If $F = \mathcal{N}(\mu_0, \sigma^2 I_d)$, then μ_0 is the unique global minimizer of G_F for every $r \ge 1$, $\lambda > 0$.*

Interpretation. At the Gaussian model the clean target is the center for every λ and r, so all members estimate the same parameter and can be compared directly.

Proof. By the layer-cake formula and monotonicity of ρ, $G_F(\mu) = \int_0^\lambda \mathbb{P}\big(\rho(\|X-\mu\|^2) > t\big)\,dt = \int_0^\lambda \big[1 - \mathbb{P}(\|X-\mu\| \le R_t)\big]dt$, where $R_t = \sup\{R : \rho(R^2) \le t\} \in (0,\infty)$ for $t \in (0,\lambda)$. For $Z \sim \mathcal{N}(0,\sigma^2 I)$, $\mathbb{P}(\|Z+y\| \le R)$ is a strictly decreasing function of $\|y\|$ for each $R > 0$: $\|Z+y\|^2/\sigma^2$ is noncentral χ^2_d with noncentrality $\|y\|^2/\sigma^2$, and this family is strictly stochastically increasing in the noncentrality. (This is the Gaussian case of Anderson's inequality; the strict version follows from the Poisson-mixture representation of the noncentral χ^2 and the strict ordering $\chi^2_{d+2j} \succ_{\mathrm{st}} \chi^2_{d+2j'}$ for $j > j'$, where $\succ_{\mathrm{st}}$ is the usual stochastic order.) Hence the integrand, and so G_F, is strictly increasing in $\|\mu - \mu_0\|$. □

The same argument with Anderson's inequality (Anderson, 1955) gives the non-strict statement for every spherically symmetric unimodal F, and strictness wherever g is strictly decreasing.

E.4 Nonasymptotic estimation and exact recovery

We first bound $\|\hat{\mu} - \mu_0\|$ with high probability (Theorem E.7), then derive exact recovery from it (Corollary E.8) and rates (Corollary E.9), following the order of the penalized mean-shift literature (She & Owen, 2011; Kong et al., 2018; Li et al., 2023).

Assumption 5 (Contaminated sample)**.** *The index set splits as $\{1,\dots,n\} = I \cup O$ (the sets $\mathcal{I}, \mathcal{O}$ of Section 4) with $|O| = m$ and $\epsilon = m/n < 1/2$; write $\epsilon' = m/(n-m) = \epsilon/(1-\epsilon) < 1$. Inliers are $x_i = \mu_0 + \varepsilon_i$ ($i \in I$) with noise ε_i i.i.d. $\mathcal{N}(0,\sigma^2 I_d)$ (not to be confused with the fraction ϵ). Outliers x_i ($i \in O$) are arbitrary points, possibly chosen adversarially after seeing the inliers (the partition $I \cup O$ itself is fixed, independent of $(\varepsilon_i)_{i\in I}$), with $\min_{i\in O}\|x_i - \mu_0\| \ge D$ (D is irrelevant when $m = 0$, and Proposition E.10 does not use it). (The mean-shift model $x_i = \mu_0 + \gamma_i + \varepsilon_i$, with shift vectors γ_i, is the special case where D bounds $\|\gamma_i + \varepsilon_i\|$ from below.)*

Interpretation. The inliers are Gaussian around μ_0, and the outliers are arbitrary points, possibly adversarial, at distance at least D from μ_0.

For $\delta \in (0,1)$ put

$$\tau = \sigma\big(\sqrt{d} + \sqrt{2\log(2n/\delta)}\big), \qquad \eta = \frac{\sigma\big(\sqrt{d} + \sqrt{2\log(2/\delta)}\big)}{\sqrt{n-m}},$$

$$k_r = \max\Big\{1, \Big[\frac{2(r-1)}{r(1-\epsilon')}\Big]^{(r-1)/2}\Big\}\ (k_1 = 1), \qquad b = \tau + k_r\sqrt{\lambda},$$

$$\beta_r = \begin{cases} \epsilon'\,\dfrac{(b+\tau)^{2q}}{(D-b)^{2q-1}}, & r > 1,\ m > 0, \\ 0, & r = 1 \text{ or } m = 0. \end{cases}$$

Lemma E.6 (Noise event)**.** *Let $\mathcal{E} = \{\max_{i\in I}\|\varepsilon_i\| \le \tau\} \cap \{\|\bar{\varepsilon}_I\| \le \eta\}$, where $\bar{\varepsilon}_I = \frac{1}{n-m}\sum_{i\in I}\varepsilon_i$. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.*

Interpretation. With probability $1-\delta$, every inlier lies within $\tau \asymp \sigma(\sqrt{d} + \sqrt{\log n})$ of μ_0 and the inlier mean within $\eta \asymp \sigma\sqrt{d/n}$; all later statements hold on this event.

Proof. $z \mapsto \|z\|$ is 1-Lipschitz and $\mathbb{E}\|Z\| \le \sqrt{d}$ for $Z \sim \mathcal{N}(0, I_d)$. By Gaussian concentration for Lipschitz functions (Boucheron et al., 2013, Thm. 5.6), $\mathbb{P}(\|Z\| \ge \sqrt{d} + t) \le e^{-t^2/2}$. With $t = \sqrt{2\log(2n/\delta)}$ and a union bound over $|I| \le n$ indices, $\mathbb{P}(\max_I \|\varepsilon_i\| > \tau) \le \delta/2$. Since $\bar{\varepsilon}_I \sim \mathcal{N}(0, \frac{\sigma^2}{n-m} I)$, the same bound with $t = \sqrt{2\log(2/\delta)}$ gives $\mathbb{P}(\|\bar{\varepsilon}_I\| > \eta) \le \delta/2$. □

Theorem E.7 (Nonasymptotic estimation bound)**.** *Let Assumption 5 hold, $r \ge 1$, and suppose*

(L) $\sqrt{\lambda} > 2\tau + \beta_r$ *and* $\lambda(1-2\epsilon) > \frac{2(1-\epsilon)}{r}\tau^2$;

(S) *if $m > 0$:* $D \ge b + \sqrt{\lambda}$ *and, if also $r > 1$,* $(2q-1)\,\epsilon'\Big(\dfrac{\lambda}{(D-b)^2}\Big)^q < 1$.

Then on the event $\mathcal{E}$ (probability $\ge 1-\delta$):

(i) (Localization and uniqueness) G_n has a unique global minimizer $\hat{\mu}$, and $\|\hat{\mu} - \mu_0\| \le a := \tau + \beta_r < b$. For $r > 1$, $\hat{\mu}$ is the only stationary point of G_n in $B(\mu_0, b)$.

(ii) (Exact error decomposition) Every inlier has $s_i(\hat{\mu}) < \lambda$, and

$$\hat{\mu} - \mu_0 = \bar{\varepsilon}_I + \frac{1}{n-m} \sum_{i \in O} \psi_r(x_i; \hat{\mu}).$$

(iii) (Error bound) For $r > 1$,

$$\begin{aligned} \|\hat{\mu} - \mu_0\| \;&\le\; \|\bar{\varepsilon}_I\| + \Big\| \frac{1}{n-m} \sum_{i \in O} \psi_r(x_i; \hat{\mu}) \Big\| \\ &\le\; \eta + \epsilon' \lambda^q (D-b)^{1-2q} = \eta + \epsilon' \sqrt{\lambda} \Big(\frac{\sqrt{\lambda}}{D-b} \Big)^{2q-1}, \end{aligned}$$

and for $r = 1$, $\hat{\mu} = \bar{x}_I$, the inlier (oracle) mean, exactly; so $\|\hat{\mu} - \mu_0\| = \|\bar{\varepsilon}_I\| \le \eta$.

Interpretation. Condition (L) makes the threshold exceed the inliers' reach and (S) keeps the outliers beyond it. Then the estimator is unique and localized, every inlier gets full weight, and the error splits exactly into the oracle noise $\bar{\varepsilon}_I$ and the outliers' residual pull, which is zero for the hard clip and of order $\epsilon' \sqrt{\lambda} (\sqrt{\lambda}/D)^{2q-1}$ for $r > 1$.

Proof. Work on $\mathcal{E}$. Note $\tau^2 < \lambda$ by (L).

Step 1 (localization by values). For $i \in I$, $\rho(s_i(\mu_0)) = \rho(\|\varepsilon_i\|^2) \le \tau^2/r$, and $\rho \le \lambda$ for $i \in O$; so $G_n(\mu_0) \le (1-\epsilon)\tau^2/r + \epsilon\lambda < \lambda/2$, the last inequality being the second part of (L). Now let $\|\mu - \mu_0\| \ge b$. For $i \in I$, $\|x_i - \mu\| \ge b - \tau = k_r \sqrt{\lambda}$, so $\rho(s_i(\mu)) \ge \rho(k_r^2 \lambda)$. For $r > 1$, Lemma E.1(c) and the definition of k_r give $\lambda - \rho(k_r^2 \lambda) \le \lambda(1-\epsilon')/2$; for $r = 1$, $\rho(k_1^2 \lambda) = \lambda$. In both cases $\rho(s_i(\mu)) \ge \lambda(1+\epsilon')/2$. Since outlier terms are ≥ 0,

$$G_n(\mu) \ge (1-\epsilon)\lambda \tfrac{1+\epsilon'}{2} = \tfrac{\lambda}{2} > G_n(\mu_0),$$

using $(1-\epsilon)(1+\epsilon') = 1$. Hence every global minimizer (which exists by Proposition E.2(ii)) lies in the open ball $B(\mu_0, b)$.

Step 2 (outliers are beyond the threshold throughout $B(\mu_0, b)$). For $\|\mu - \mu_0\| < b$ and $i \in O$, $\|x_i - \mu\| > D - b \ge \sqrt{\lambda}$ by (S). By Lemma E.1(e), $\|\psi_r(x_i; \mu)\| \le \lambda^q (D-b)^{1-2q}$ for $r > 1$ and $= 0$ for $r = 1$.

Step 3a ($r > 1$: every stationary point in $B(\mu_0, b)$ is close to μ_0). Let $\mu \in B(\mu_0, b)$ with $\sum_i \psi_r(x_i; \mu) = 0$, and write $W_i = W_r(s_i(\mu))$, $t = \|\mu - \mu_0\|$. Inserting $x_i - \mu = \varepsilon_i - (\mu - \mu_0)$ for $i \in I$,

$$(\mu - \mu_0) \sum_{i \in I} W_i = \sum_{i \in I} W_i \varepsilon_i + \sum_{i \in O} \psi_r(x_i; \mu). \tag{6}$$

For $i \in I$, $s_i \le (t + \tau)^2 < (b + \tau)^2$, so $W_i \ge W_r((b+\tau)^2) = (\lambda/(b+\tau)^2)^q$ (as $b + \tau > \sqrt{\lambda}$). Taking norms in equation 6 and using Step 2,

$$t \le \tau + \frac{m \lambda^q (D-b)^{1-2q}}{(n-m)\lambda^q (b+\tau)^{-2q}} = \tau + \beta_r = a.$$

(If $m = 0$ the outlier sum is empty and $t \le \tau = a$.)

Step 3b ($r > 1$: strong convexity on a neighborhood of $\bar{B}(\mu_0, a)$). By (L), $a + \tau < \sqrt{\lambda}$, and $a < b$ because $\beta_r < \sqrt{\lambda} \le k_r \sqrt{\lambda}$. Choose $a' \in (a, b)$ with $a' + \tau < \sqrt{\lambda}$. On $B(\mu_0, a')$ every inlier has $s_i < \lambda$ and every outlier has $s_i > (D - a')^2 > (D-b)^2 \ge \lambda$. So no point lies on its sphere, G_n is C^∞ there, and by Lemma E.1(d)

$$\begin{aligned} \nabla^2 G_n(\mu) &\succeq \frac{2}{rn} \Big[(n-m) - m(2q-1) \Big(\frac{\lambda}{(D-b)^2} \Big)^q \Big] I \\ &= \frac{2(n-m)}{rn} \Big[1 - (2q-1)\epsilon' \Big(\frac{\lambda}{(D-b)^2} \Big)^q \Big] I \succ 0 \end{aligned}$$

by (S) (for $m = 0$ the bracket is 1). A strictly convex C^1 function on a convex open set has at most one stationary point. By Step 3a all stationary points in $B(\mu_0, b)$ lie in $\bar{B}(\mu_0, a) \subset B(\mu_0, a')$, so there is at most one. By Step 1 and Proposition E.2(iii) every global minimizer is such a point. Hence the global minimizer $\hat{\mu}$ is unique, $\|\hat{\mu}-\mu_0\| \le a$, and it is the only stationary point in $B(\mu_0, b)$.

Step 3c ($r = 1$). Let $\hat{\mu}$ be a global minimizer; $\|\hat{\mu} - \mu_0\| < b$ by Step 1. By Proposition E.2(iv), $\hat{\mu}$ is the mean of $\mathcal{S} = \{i : s_i(\hat{\mu}) < \lambda\} \neq \emptyset$. By Step 2, $\mathcal{S} \subset I$, so $\hat{\mu} - \mu_0 = |\mathcal{S}|^{-1} \sum_{i\in\mathcal{S}} \varepsilon_i$ and $\|\hat{\mu} - \mu_0\| \le \tau$. Then every inlier has $\|x_i - \hat{\mu}\| \le 2\tau < \sqrt{\lambda}$ by (L), so $\mathcal{S} = I$ and $\hat{\mu} = \bar{x}_I$. Since every global minimizer equals $\bar{x}_I$, it is unique, and (i) holds with $\beta_1 = 0$.

(ii) For $i \in I$, $\|x_i-\hat{\mu}\| \le \tau+a < \sqrt{\lambda}$ by (L), so $W_i = 1$, and equation 6 at $\mu = \hat{\mu}$ (a stationary point; for $r = 1$ use $\psi_1 = 0$ on outliers and $\hat{\mu} = \bar{x}_I$) becomes $(n-m)(\hat{\mu} - \mu_0) = \sum_I \varepsilon_i + \sum_O \psi_r(x_i; \hat{\mu})$.

(iii) Take norms in (ii), use $\|\bar{\varepsilon}_I\| \le \eta$ on $\mathcal{E}$ and Step 2 for each outlier score, and $\lambda^q(D-b)^{1-2q} = \sqrt{\lambda}(\sqrt{\lambda}/(D-b))^{2q-1}$. □

Corollary E.8 (Exact recovery from the estimation bound)**.** *(a) (Deterministic.) Let $\tilde{\mu}$ be any point with $e := \|\tilde{\mu} - \mu_0\| < \sqrt{\lambda} - \tau$ and $D - e > \sqrt{\lambda}$. On $\mathcal{E}$ the flag set (at the point $\tilde{\mu}$) $\hat{O}(\tilde{\mu}) = \{i : w^\star(\|x_i-\tilde{\mu}\|^2) < 1\}$ equals O. (b) Under the conditions of Theorem E.7, on $\mathcal{E}$ (probability $\ge 1-\delta$), $\hat{O}(\hat{\mu}) = O$: no masking and no swamping, for every $r \ge 1$. For $r > 1$ the same holds for every stationary point of G_n in $B(\mu_0, b)$, since there is only one.*

Interpretation. Any estimate within $\sqrt{\lambda} - \tau$ of μ_0 gives the correct flags, so recovery needs only a bounded error, not a vanishing one.

Proof. (a) For $i \in I$, $\|x_i - \tilde{\mu}\| \le \|\varepsilon_i\| + e \le \tau + e < \sqrt{\lambda}$, so $w^\star = 1$. For $i \in O$, $\|x_i - \tilde{\mu}\| \ge D - e > \sqrt{\lambda}$, so $w^\star < 1$. (b) By Theorem E.7(i), $e \le a = \tau + \beta_r$, and $a < \sqrt{\lambda} - \tau$ by (L); and $D - a > D - b \ge \sqrt{\lambda}$ by (S) and $a < b$. Apply (a). □

Corollary E.9 (Rates)**.** *Take $\delta = \delta_n = 1/n$, fix $\epsilon < 1/2$, $r \ge 1$ and $c_\lambda > 2$ with $c_\lambda^2 > 2(1-\epsilon)/(r(1-2\epsilon))$, and set $\lambda_n = c_\lambda^2\tau_n^2$ with $\tau_n = \sigma(\sqrt{d} + \sqrt{2\log(2n^2)}) \asymp \sigma(\sqrt{d} + \sqrt{\log n})$. Write D_n for the separation D at sample size n and $b_n = \tau_n(1 + k_r c_\lambda)$ for the radius b. There is a constant $C_D = C_D(r, \epsilon, c_\lambda)$ such that $D_n \ge C_D\tau_n$ implies (L) and (S). In that case, with probability $\ge 1 - 1/n$, $\hat{O} = O$ and*

$$\|\hat{\mu} - \mu_0\| \le \sigma\frac{\sqrt{d} + \sqrt{2\log(2n)}}{\sqrt{n-m}} + \mathbb{1}\{r > 1\}\,\epsilon' c_\lambda \tau_n \Big(\frac{c_\lambda\tau_n}{D_n - b_n}\Big)^{2q-1}.$$

Consequently: (a) $\mathbb{P}(\hat{O} = O) \to 1$ for every $r \ge 1$, including ROBOT; recovery needs only a bounded estimation error, not a vanishing one. (b) For $r = 1$, $\hat{\mu} = \bar{x}_I$ on the event, so $\|\hat{\mu} - \mu_0\| = O_p(\sigma\sqrt{d/n})$, the oracle rate (the bound above is its high-probability version). (c) For $r > 1$ and fixed $\epsilon > 0$, the bias bound tends to 0 if and only if $\lambda_n^q/D_n^{2q-1} \to 0$, i.e. $\lambda_n = o(D_n^{(r+1)/r})$; it is of the oracle order $\sigma\sqrt{(d + \log n)/n}$ if $(\sqrt{\lambda_n}/D_n)^{2q-1}\sqrt{\lambda_n} = O(\sqrt{(d + \log n)/n})$.

Interpretation. Choosing λ_n at the inliers' extreme-value scale yields exact recovery with a fixed outlier fraction and, for the hard clip, the oracle error; soft powers add a bias that vanishes when the separation grows faster than $\lambda_n^{r/(r+1)}$.

Proof. If $D_n - b_n \ge K_0\tau_n$ then $\beta_r \le \epsilon'(2 + k_r c_\lambda)^{2q}\tau_n/K_0^{2q-1}$ and $(\lambda_n/(D_n - b_n)^2)^q \le (c_\lambda/K_0)^{2q}$. Choose K_0 large enough that $\epsilon'(2 + k_r c_\lambda)^{2q}K_0^{1-2q} < c_\lambda - 2$ (so $\sqrt{\lambda_n} = c_\lambda\tau_n > 2\tau_n + \beta_r$) and $(2q-1)\epsilon'(c_\lambda/K_0)^{2q} < 1$, and take $C_D = 1 + k_r c_\lambda + \max(K_0, c_\lambda)$. This gives (S). The second part of (L) holds by the choice of c_λ. The bound is Theorem E.7(iii) with η at $\delta = 1/n$, $\lambda^q(D-b)^{1-2q} = \sqrt{\lambda}(\sqrt{\lambda}/(D-b))^{2q-1}$, $\sqrt{\lambda_n} = c_\lambda\tau_n$; recovery is Corollary E.8(b). In (c), $\tau_n(c_\lambda\tau_n/(D_n - b_n))^{2q-1} \to 0$ iff $\tau_n^{2q}/(D_n - b_n)^{2q-1} \to 0$; since $D_n \ge C_D\tau_n$ with $C_D > 1 + k_r c_\lambda$, $D_n - b_n \asymp D_n$, and $\tau_n \asymp \sqrt{\lambda_n}$, this is $\lambda_n^q/D_n^{2q-1} \to 0$. □

Proposition E.10 (The threshold scale is necessary for every estimator)**.** *Let Assumption 5 hold with $N = n - m \ge 2$ inliers, let $\eta \in (0, 1)$, and let $\tilde{\mu} \in \mathbb{R}^d$ be any point, possibly depending on all the data (in particular any estimator, whatever the outliers). Let $\bar{\Phi}$ be the standard normal upper tail.*

(a) If $\lambda \le (1-\eta)\,2\sigma^2 \log N$*, then with probability at least* $1 - 2\exp\{-N\bar{\Phi}(a_N)\}$*,* $a_N = \sqrt{(2-\eta)\log N}$*, some inlier has* $\|x_i - \tilde{\mu}\|^2 > \lambda$*.*

(b) If $\lambda \le (1-\eta)\,\sigma^2 d(N-1)/N$*, then with probability at least* $1 - \exp\{-\eta^2 d(N-1)/16\}$*, some inlier has* $\|x_i - \tilde{\mu}\|^2 > \lambda$*.*

Both probabilities tend to 1 *as* $N \to \infty$*, since* $N\bar{\Phi}(a_N) \ge (2\pi)^{-1/2}\frac{a_N}{1+a_N^2}N^{\eta/2} \to \infty$*. Consequently, for the point target, exact recovery (no swamping) with probability bounded away from* 0 *requires* $\lambda_n \ge (1-o(1))\,\sigma^2 \max\{2\log n, d\} \ge (1-o(1))\frac{\sigma^2}{4}(\sqrt{d}+\sqrt{2\log n})^2$*, for every estimator, every outlier fraction* $\epsilon < 1/2$ *and every outlier configuration: if* $\liminf_n \mathbb{P}($*no inlier is flagged*$) > 0$*, then* $\liminf_n \lambda_n/(\sigma^2 \max\{2\log n, d\}) \ge 1$*.*

Proof. (a) Write $x_i = \mu_0 + \sigma Z_i$ for $i \in I$, with Z_i i.i.d. $\mathcal{N}(0, I_d)$, and let Z_{i1} be the first coordinate. $\mathbb{P}(\max_{i\in I} Z_{i1} \le a_N) = (1-\bar{\Phi}(a_N))^N \le \exp\{-N\bar{\Phi}(a_N)\}$, and likewise for $\min_{i\in I} Z_{i1} \ge -a_N$. On the complement of both events, for every $\mu \in \mathbb{R}^d$,

$$\max_{i\in I}\|x_i - \mu\| \;\ge\; \max_{i\in I}|x_{i1} - \mu_1| \;\ge\; \tfrac{1}{2}\big(\max_{i\in I} x_{i1} - \min_{j\in I} x_{j1}\big) \;>\; \sigma a_N,$$

so $\max_{i\in I}\|x_i-\mu\|^2 > (2-\eta)\sigma^2\log N \ge \lambda$; take $\mu = \tilde{\mu}$. The tail bound is the Mills-ratio inequality $\bar{\Phi}(a) \ge \frac{a}{1+a^2}\varphi(a)$ with $\varphi(a_N) = (2\pi)^{-1/2}N^{-1+\eta/2}$. (b) For every μ, $\max_{i\in I}\|x_i - \mu\|^2 \ge \frac{1}{N}\sum_{i\in I}\|x_i - \mu\|^2 \ge \frac{1}{N}\sum_{i\in I}\|x_i - \bar{x}_I\|^2 = \frac{\sigma^2}{N}\chi^2_k$ with $k = d(N-1)$. By the lower-tail bound $\mathbb{P}(\chi^2_k \le k - 2\sqrt{kx}) \le e^{-x}$ (Laurent & Massart, 2000, Lemma 1) with $x = \eta^2 k/16$, $\chi^2_k > (1-\eta/2)k$ with probability at least $1 - e^{-\eta^2 k/16}$, and then $\max_{i\in I}\|x_i - \mu\|^2 > (1-\eta/2)\sigma^2 k/N > \lambda$. For the consequence, suppose $\lambda_n \le (1-\eta)\sigma^2\max\{2\log n, d\}$ along a subsequence. Since $N > n/2$, $\log n - \log 2 < \log N \le \log n$ and $(N-1)/N \to 1$, so for $n \ge n_0(\eta)$ either $\lambda_n \le (1-\eta/2)\,2\sigma^2\log N$ or $\lambda_n \le (1-\eta/2)\sigma^2 d(N-1)/N$; (a) or (b) with $\eta/2$ then gives a flagged inlier with probability tending to one, uniformly in d. Finally $\max\{2\log n, d\} \ge \frac{1}{4}(\sqrt{d} + \sqrt{2\log n})^2$ because $(a+b)^2 \le 2(a^2+b^2)$. □

The necessary scale $\sigma^2\max\{2\log n, d\}$ and the sufficient scale $c_\lambda^2\sigma^2(\sqrt{d}+\sqrt{2\log(2n^2)})^2$ of Corollary E.9 have the same order; only the constants differ. Unlike Proposition D.19, no localization of the estimator is assumed: no center whatsoever keeps all Gaussian inliers within a smaller radius.

Proposition E.11 (The soft-power bias is attained by one-sided outliers)**.** *Let* $r > 1$*,* $m \ge 1$*, and let all outliers sit at one point,* $x_i = \mu_0 + Du$ *(*$i \in O$*) for a unit vector* u*, with (L) and (S) holding. Then on* $\mathcal{E}$*,*

$$\epsilon'\lambda^q\,\frac{D-a}{(D+a)^{2q}} - \|\bar{\varepsilon}_I\| \;\le\; u^\top(\hat{\mu}-\mu_0) \;\le\; \|\hat{\mu}-\mu_0\| \;\le\; \epsilon'\lambda^q(D-b)^{1-2q} + \|\bar{\varepsilon}_I\|.$$

In particular, if $D/b \to \infty$ *the two bias terms have ratio* $\to 1$*: the bias bound of Theorem E.7 is sharp up to a factor* $1 + o(1)$*, and for* $r > 1$ *it is of exact order* $\epsilon'\sqrt{\lambda}(\sqrt{\lambda}/D)^{2q-1}$*, while for* $r = 1$ *it is* 0*.*

Proof. By Theorem E.7(ii), $\hat{\mu} - \mu_0 = \bar{\varepsilon}_I + \epsilon'\psi_r(x^*;\hat{\mu})$ with $x^* = \mu_0 + Du$. Here $\psi_r(x^*;\hat{\mu}) = W_r(s^*)(x^* - \hat{\mu})$, $s^* = \|x^* - \hat{\mu}\|^2$, and $\|\hat{\mu} - \mu_0\| \le a$ gives $D - a \le \sqrt{s^*} \le D + a$ and $u^\top(x^* - \hat{\mu}) = D - u^\top(\hat{\mu} - \mu_0) \ge D - a$. Since $s^* > \lambda$ (Step 2 of Theorem E.7) and W_r is non-increasing, $W_r(s^*) = (\lambda/s^*)^q \ge \lambda^q(D+a)^{-2q}$. Taking the inner product with u: $u^\top(\hat{\mu} - \mu_0) \ge -\|\bar{\varepsilon}_I\| + \epsilon'\lambda^q(D+a)^{-2q}(D-a)$. The upper bound is Theorem E.7(iii). The ratio of the two bias terms is $\frac{(D-a)(D-b)^{2q-1}}{(D+a)^{2q}} \to 1$ as $D/b \to \infty$ (recall $a < b$). □

E.5 Consistency without moment conditions

Because the cost is bounded, the minimum-OT_ρ estimator is consistent for the population minimizer under an arbitrary data law, without moment conditions; we prove this for location families.

As in Appendix D, let $h(x,y) = \rho(\|x-y\|^2)$ and, for probability measures P, ν on $\mathbb{R}^d$, $\mathrm{OT}_\rho(P,\nu) = \inf_{\pi\in\Pi(P,\nu)}\int h\,d\pi$. Let $d_\lambda(x,x') = \min\{L_r\|x-x'\|, \lambda\}$ with $L_r = 2\sqrt{\lambda}/r$, and let $T_\lambda(P,P') =$

$\mathrm{OT}_{d_\lambda}(P, P')$. Since d_λ is a bounded metric inducing the Euclidean topology, T_λ metrizes weak convergence of probability measures on $\mathbb{R}^d$ (Villani, 2009, Thm. 6.9 and Cor. 6.13); no moment condition is needed because $d_\lambda \le \lambda$.

Lemma E.12 (Uniform stability of OT_ρ in the data measure)**.** *For all probability measures P, P', ν: $|\mathrm{OT}_\rho(P, \nu) - \mathrm{OT}_\rho(P', \nu)| \le T_\lambda(P, P')$. In particular $\sup_\nu |\mathrm{OT}_\rho(P_n, \nu) - \mathrm{OT}_\rho(P, \nu)| \le T_\lambda(P_n, P)$.*

Interpretation. The robust OT value moves by at most the bounded transport distance between the data laws, uniformly over the target; this is the only property of the empirical measure that consistency uses.

Proof. Fix $\xi > 0$, a coupling $\gamma \in \Pi(P, P')$ with $\int d_\lambda \, d\gamma \le T_\lambda(P, P') + \xi$, and $\pi' \in \Pi(P', \nu)$ with $\int h \, d\pi' \le \mathrm{OT}_\rho(P', \nu) + \xi$. By the gluing lemma (Villani, 2009, Ch. 1) there are random variables (X, X', Y) with $(X, X') \sim \gamma$ and $(X', Y) \sim \pi'$. Then (X, Y) couples (P, ν). Since $h \in [0, \lambda]$ and $h(\cdot, y)$ is L_r-Lipschitz (Lemma E.1(f)), $h(X, Y) \le h(X', Y) + d_\lambda(X, X')$. Taking expectations, $\mathrm{OT}_\rho(P, \nu) \le \mathrm{OT}_\rho(P', \nu) + T_\lambda(P, P') + 2\xi$. Now let $\xi \downarrow 0$ and exchange P and P'. □

Theorem E.13 (Consistency of minimum-OT_ρ estimation)**.** *Let $X_1, X_2, \ldots$ be i.i.d. from an* arbitrary *law P on $\mathbb{R}^d$ (clean or contaminated; no moments assumed), P_n the empirical measure, and $\{\nu_\theta : \theta \in \mathbb{R}^d\}$ a location family $\nu_\theta = \nu_0(\cdot - \theta)$ (e.g. $\nu_0 = \delta_0$ for the point target, $\nu_0 = \mathcal{N}(0, I_d)$ for the spread target). Let $J(\theta) = \mathrm{OT}_\rho(P, \nu_\theta)$ and $F_n(\theta) = \mathrm{OT}_\rho(P_n, \nu_\theta)$ (as in Appendix D, but with P arbitrary). Then:*

- *(i) J is L_r-Lipschitz and $J(\theta) \to \lambda$ as $\|\theta\| \to \infty$. If $\inf J < \lambda$ (automatic for $r > 1$; for $r = 1$ and the point target it holds for every P), then J attains its minimum, and the set of minimizers is compact.*
- *(ii) $\sup_\theta |F_n(\theta) - J(\theta)| \le T_\lambda(P_n, P) \to 0$ almost surely.*
- *(iii) (Identifiability, the only assumption) If J has a unique minimizer $\theta^\star$, then every sequence of near-minimizers, $F_n(\hat\theta_n) \le \inf F_n + o_p(1)$, satisfies $\hat\theta_n \to \theta^\star$ in probability. The same holds almost surely if $o_p(1)$ is replaced by $o(1)$ a.s.*

Interpretation. Consistency needs only identifiability of the population minimizer; no moments are required because the cost is bounded. Under contamination the limit is the minimizer for the contaminated law, which is why the clean-parameter results of Appendix D need the separation arguments.

Proof. (i) Coupling ν_θ and $\nu_{\theta'}$ by translation and using Lemma E.1(f) in the y argument, as in Lemma E.12, gives $|J(\theta) - J(\theta')| \le \min(L_r \|\theta - \theta'\|, \lambda)$.

For the limit, fix $\eta_0 > 0$ and choose K with $P(\|X\| > K) < \eta_0$ and $\nu_0(\|Y\| > K) < \eta_0$, and fix $R > 0$. For $\|\theta\| > 2K + \sqrt{R}$ and any coupling (X, Y) of (P, ν_θ), $\{\|X - Y\| \le \sqrt{R}\} \subset \{\|X\| > K\} \cup \{\|Y - \theta\| > K\}$. Since ρ is non-decreasing, $\mathbb{E}h(X, Y) \ge \rho(R)(1 - 2\eta_0)$. Hence $\liminf_{\|\theta\| \to \infty} J(\theta) \ge \rho(R)(1 - 2\eta_0)$; letting $R \to \infty$ and $\eta_0 \to 0$ gives $\ge \lambda$. Also $J \le \lambda$.

If $\inf J < \lambda$, the sublevel set $\{J \le (\inf J + \lambda)/2\}$ is closed (by continuity) and bounded (by the limit), so the minimum is attained and the minimizers form a compact set. For $r > 1$, $h < \lambda$ everywhere, so $J(\theta) < \lambda$ for every θ. For $r = 1$ and $\nu_0 = \delta_0$, pick θ with $P(\|X - \theta\|^2 < \lambda) > 0$ (it exists for every P, since $\mathbb{R}^d$ is covered by countably many balls of radius $\sqrt{\lambda}/2$); then $J(\theta) < \lambda$.

(ii) The inequality is Lemma E.12. $P_n \Rightarrow P$ almost surely (Varadarajan's theorem), and T_λ metrizes weak convergence, so $T_\lambda(P_n, P) \to 0$ almost surely.

(iii) By (i) (uniqueness forces $\inf J < \lambda$, because $J \le \lambda$), J is continuous, $J(\theta^\star) < \lambda$, and $J \to \lambda$ at infinity. Hence $\inf_{\|\theta - \theta^\star\| \ge e} J(\theta) > J(\theta^\star)$ for every $e > 0$: on the compact annulus the infimum is attained and exceeds $J(\theta^\star)$ by uniqueness, and outside a large ball $J > (J(\theta^\star) + \lambda)/2$. This well-separated minimum, together with (ii), is exactly the setting of van der Vaart (1998, Thm. 5.7), which gives $\hat\theta_n \to_p \theta^\star$. For the a.s. version, apply the deterministic argument of that theorem pathwise. □

Theorem E.13 extends Ma et al. (2026, Thm. 4.4), who prove existence and consistency of minimum-ROBOT estimators ($r = 1$, metric cost) without moment conditions, and the minimum-Kantorovich consistency results of Bassetti et al. (2006) and Bernton et al. (2019). Under contamination the limit $\theta^\star$ is the population minimizer for the contaminated law, not the clean center.

E.6 The 95% flag rule for the spread target

The flag rule of the experiments compares a quantile of each point's costs to the fitted Gaussian target with λ; we show that it is a radial threshold, so its margin sets are balls.

Lemma E.14 (The 95% rule is a radial threshold)**.** *Let $Z \sim \mathcal{N}(0, I_d)$ and, for $t \geq 0$ and $\alpha \in (0, 1)$, let $H_\alpha(t)$ be the α-quantile of $\|\sqrt{t}\, e_1 + Z\|^2$, with e_1 the first basis vector (noncentral χ^2_d with noncentrality t). Then H_α is continuous and strictly increasing on $[0, \infty)$ with $H_\alpha(t) \to \infty$, and for all $x, \mu \in \mathbb{R}^d$ the α-quantile of $\|x - Y\|^2$, $Y \sim \mu + \mathcal{N}(0, I_d)$, equals $H_\alpha(\|x - \mu\|^2)$. Consequently the rule "flag x if the α-quantile of its costs to $\mu + \mathcal{N}(0, I_d)$ exceeds λ" flags every point if $\lambda < H_\alpha(0)$, and otherwise flags exactly the points with $\|x - \mu\|^2 > \tilde{\lambda} := H_\alpha^{-1}(\lambda)$ (with $\tilde{\lambda} = \sup\{t : H_\alpha(t) \leq \lambda\}$). The 95% rule of the experiments is the case $\alpha = 0.05$.*

Interpretation. The 95% rule is equivalent to comparing $\|x - \mu\|^2$ with an adjusted threshold $\tilde{\lambda}$, so for the spread target the flags, like the point-target flags, are defined by a ball around the fitted location.

Proof. $x - Y = (x - \mu) - Z'$ with $Z' \sim \mathcal{N}(0, I_d)$, and $Z' \overset{d}{=} -Z'$ (equality in law) is rotation invariant, so $\|x - Y\|^2 \overset{d}{=} \|\sqrt{t}\, e_1 + Z\|^2$ with $t = \|x - \mu\|^2$. Let $\Phi_{d,t}$ be its distribution function (as in Remark D.14). For each t, $\Phi_{d,t}$ is continuous and strictly increasing on $(0, \infty)$ (the law has a positive density there). By the strict stochastic ordering of the noncentral χ^2 family in the noncentrality, used in Lemma E.5, $\Phi_{d,t'}(u) < \Phi_{d,t}(u)$ for $t < t'$ and $u > 0$. Hence, with $u = H_\alpha(t) > 0$, $\Phi_{d,t'}(H_\alpha(t)) < \Phi_{d,t}(H_\alpha(t)) = \alpha$, so $H_\alpha(t') > H_\alpha(t)$. The map $(t, u) \mapsto \Phi_{d,t}(u)$ is continuous (dominated convergence in the Poisson-mixture representation), and $\Phi_{d,t}$ is strictly increasing in u, so H_α is continuous. Finally $\|\sqrt{t}\, e_1 + Z\| \geq \sqrt{t} - \|Z\|$ gives $H_\alpha(t) \to \infty$. The flag statement follows because H_α is a continuous strictly increasing bijection from $[0, \infty)$ onto $[H_\alpha(0), \infty)$. □

With a finite sample of the fitted target, the empirical 5th percentile replaces $H_{0.05}(\|x - \mu\|^2)$; this adds a Monte Carlo error that we do not analyze.

F Full Mean-Estimation Results

This appendix reports the complete mean-estimation results behind Section 5: the symmetric-contamination figure, every cell of the grid, a detection comparison with F_1, and the paired tests.

Figures 5 and 6 extend Figure 2 with the masking and swamping rows for every method that produces flags. Tables 2 and 3 list every cell of the mean-estimation study, including k=1, which the main-text figures omit. Each cell gives the error $\|\hat{\mu} - \mu\|_2$ as mean (standard deviation) over 100 trials, then masking / swamping for methods that produce flags. At $\epsilon = 0$ there are no outliers, so masking is reported as 0. Generated by `scripts/make_full_tables_tex.py`.

Paired comparisons. The paired t-tests quoted in Section 5 compare two methods on the same 100 simulated data sets per cell (two-sided, $p < 0.05$, unadjusted for multiplicity); they are computed by `scripts/paired_power_tests.py` from the same log. Table standard deviations use the population convention (ddof = 0), the figures the sample convention; the two differ by at most 0.001.

Detection F_1 and constructed detectors. Table 4 compares the flag sets of the family's r=2 member with those of ROBOT and of detectors that we constructed from the trimming and χ^2 baselines (their papers do not propose outlier flags) at $\epsilon = 0.2$ (one-sided), adding F_1, the harmonic mean of precision and recall for the outlier class. On balanced F_1 the trimming and χ^2 baselines are best at small and moderate shifts and ROBOT at k=4; r=2 matches the lowest masking at $k \geq 2$ but swamps more.

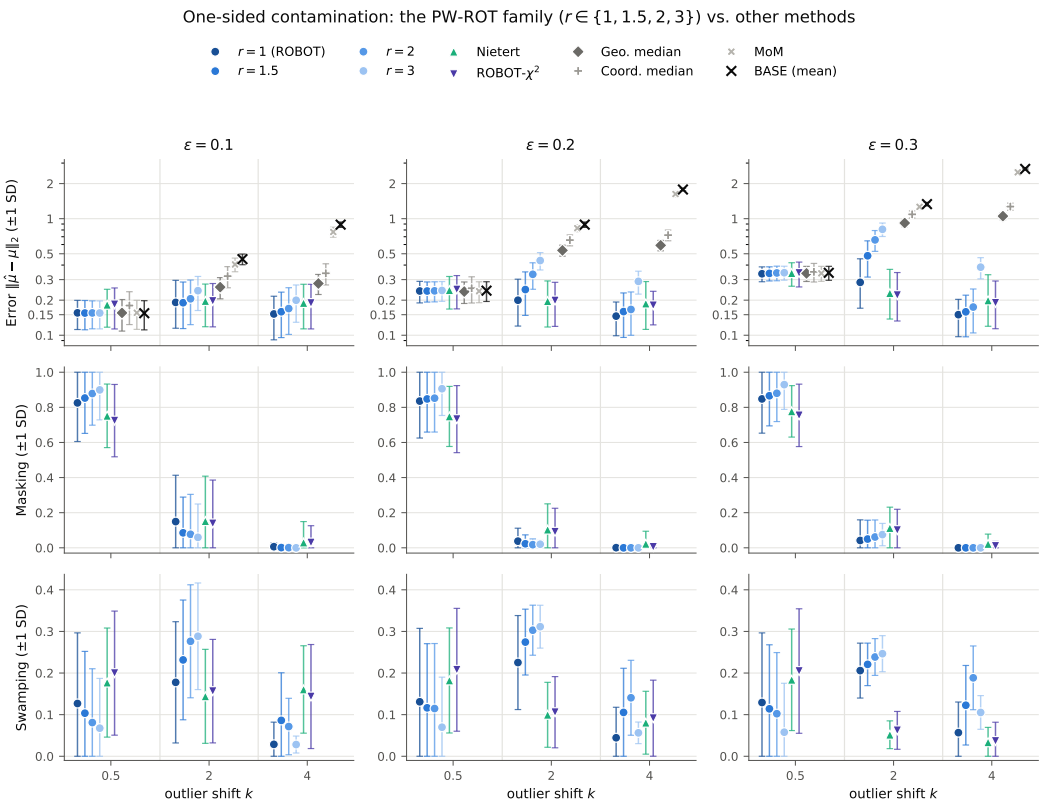

Figure 5: One-sided contamination, all metrics. Rows show location error $\|\hat{\mu} - \mu\|_2$ (log scale), masking, and swamping, each as mean $\pm$ one standard deviation across 100 trials (bars clipped to the axis range); columns show contamination fraction ϵ; the horizontal axis is outlier shift $k \in \{0.5, 2, 4\}$. The blue ramp denotes the PW-ROT family, from $r = 1$ (dark) to $r = 3$ (light); within each k, methods are offset horizontally (no lines join different k). Detection rows include only methods that produce flags. Trials are paired across methods, so differences can be significant despite overlapping bars.

Soft outlier scores. For $r > 1$ the flagged points carry a continuous weight $w^\star(C) = (\lambda/C)^{1/(r-1)} \in (0, 1)$, which ranks candidate outliers by how far past the threshold they lie; the hard clip's weight is binary.

G MNIST: Per-Seed Results and Generated Samples

This appendix gives the per-seed MNIST detection results, relates detection failures to the threshold selector's fallback, and shows generated samples.

Table 6 lists every MNIST run: the selected threshold, whether the selector found a plateau (and its length) or fell back, detection under the 95% rule used in Figure 7 and under the stricter min-distance rule (flag x_i iff $w^\star(c(x_i, y)) < 1$ for every generated sample y), and a robust FID (Fréchet Inception Distance); generated samples are shown in Figure 10. The two flag rules differ most for ROBOT on white images, seed 0, which detects every outlier under the 95% rule but none under the min-distance rule; the largest other differences occur for ROBOT (FashionMNIST, seed 1) and r=1.5 (FashionMNIST, seed 0). The robust FID is *not* a standard FID: it uses the penultimate features of a small classifier trained on public MNIST and a reference set of the contaminated training images trimmed by 35%, so it is in-sample and not comparable to published values; we report it for completeness only. With three seeds it does not order the powers (e.g. FashionMNIST: r=1.5 10.1 ± 3.8, r=2 20.7 ± 15.1, mean $\pm$ sample sd), and the collapsed r=3 white seed reaches 1464.

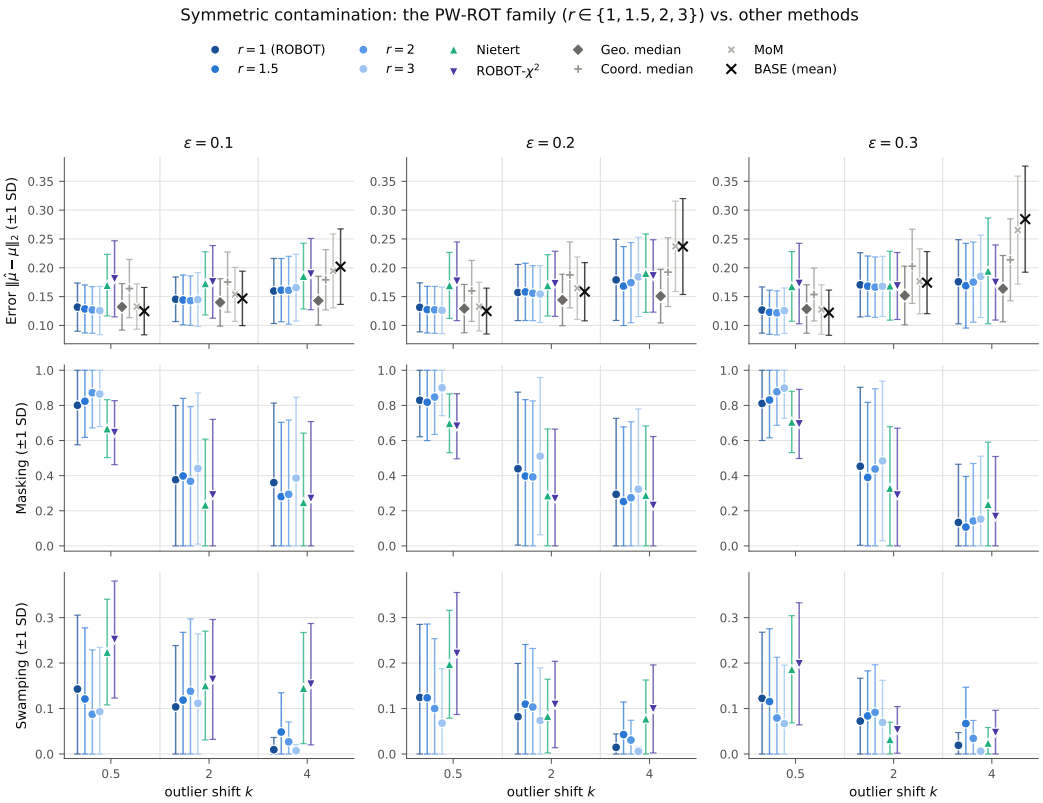

Figure 6: Symmetric contamination, with the same layout as Figure 5; the error axis is linear.

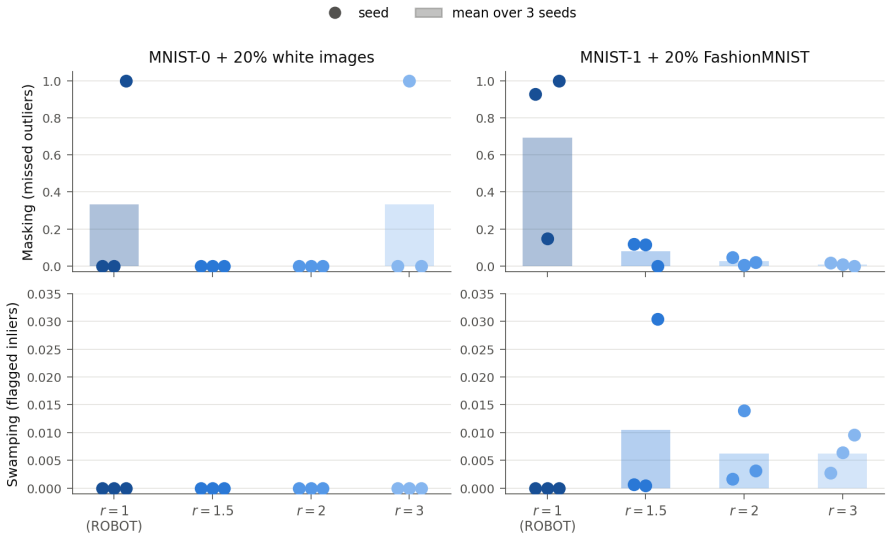

Figure 7: Outlier detection on contaminated MNIST at the selected λ (95% flag rule; bars: mean over three seeds; dots: seeds). Note the different y-scale of the swamping row. The same figure with each run marked by whether the selector found a plateau is Figure 9.

Plateau versus fallback. The selector falls back to the criterion alone when no stable-λ plateau (a run of neighboring thresholds with near-identical flag sets; Appendix H) exists. This happened in 4 of 6 ROBOT runs (white seed 2 and all three FashionMNIST seeds), 1 of 6 for r=1.5 (FashionMNIST seed 0) and for r=3 (white seed 1), and 0 of 6 for r=2. Every run that missed almost

Table 2: One-sided contamination, all cells.

ϵ	method	k=0.5	k=1	k=2	k=4
0	BASE		0.121 (0.039) / 0.00 / 0.00		
0	r=1 (ROBOT)		0.126 (0.039) / 0.00 / 0.12		
0	r=1.5		0.123 (0.039) / 0.00 / 0.14		
0	r=2		0.122 (0.039) / 0.00 / 0.12		
0	r=3		0.121 (0.038) / 0.00 / 0.08		
0	Nietert		0.161 (0.059) / 0.00 / 0.20		
0	ROBOT-χ^2		0.166 (0.065) / 0.00 / 0.21		
0	Geo. median		0.127 (0.039)		
0	Coord. median		0.157 (0.050)		
0	MoM		0.128 (0.040)		
0.1	BASE	0.155 (0.043) / 1.00 / 0.00	0.241 (0.046) / 1.00 / 0.00	0.450 (0.047) / 1.00 / 0.00	0.889 (0.047) / 1.00 / 0.00
0.1	r=1 (ROBOT)	0.156 (0.043) / 0.83 / 0.13	0.208 (0.054) / 0.62 / 0.17	0.191 (0.104) / 0.15 / 0.18	0.152 (0.064) / 0.01 / 0.03
0.1	r=1.5	0.155 (0.043) / 0.85 / 0.10	0.221 (0.051) / 0.62 / 0.17	0.190 (0.095) / 0.09 / 0.23	0.159 (0.075) / 0.00 / 0.09
0.1	r=2	0.155 (0.042) / 0.88 / 0.08	0.230 (0.049) / 0.67 / 0.14	0.205 (0.092) / 0.08 / 0.28	0.169 (0.086) / 0.00 / 0.07
0.1	r=3	0.155 (0.042) / 0.90 / 0.07	0.234 (0.048) / 0.77 / 0.09	0.241 (0.078) / 0.06 / 0.29	0.200 (0.070) / 0.00 / 0.03
0.1	Nietert	0.183 (0.065) / 0.75 / 0.18	0.202 (0.068) / 0.52 / 0.18	0.197 (0.078) / 0.15 / 0.14	0.189 (0.084) / 0.03 / 0.16
0.1	ROBOT-χ^2	0.184 (0.070) / 0.72 / 0.20	0.211 (0.074) / 0.52 / 0.19	0.197 (0.081) / 0.14 / 0.16	0.189 (0.084) / 0.03 / 0.14
0.1	Geo. median	0.155 (0.047)	0.209 (0.050)	0.259 (0.053)	0.278 (0.054)
0.1	Coord. median	0.180 (0.056)	0.239 (0.064)	0.322 (0.066)	0.340 (0.070)
0.1	MoM	0.156 (0.043)	0.232 (0.050)	0.405 (0.054)	0.769 (0.075)
0.2	BASE	0.241 (0.046) / 1.00 / 0.00	0.450 (0.047) / 1.00 / 0.00	0.889 (0.047) / 1.00 / 0.00	1.780 (0.047) / 1.00 / 0.00
0.2	r=1 (ROBOT)	0.240 (0.050) / 0.84 / 0.13	0.379 (0.076) / 0.63 / 0.16	0.200 (0.102) / 0.04 / 0.23	0.146 (0.047) / 0.00 / 0.04
0.2	r=1.5	0.240 (0.047) / 0.85 / 0.12	0.401 (0.063) / 0.61 / 0.20	0.247 (0.102) / 0.02 / 0.27	0.160 (0.071) / 0.00 / 0.11
0.2	r=2	0.240 (0.047) / 0.85 / 0.11	0.425 (0.055) / 0.69 / 0.15	0.333 (0.085) / 0.02 / 0.30	0.166 (0.068) / 0.00 / 0.14
0.2	r=3	0.242 (0.046) / 0.91 / 0.07	0.433 (0.051) / 0.74 / 0.12	0.437 (0.074) / 0.02 / 0.31	0.290 (0.064) / 0.00 / 0.06
0.2	Nietert	0.243 (0.075) / 0.75 / 0.18	0.316 (0.104) / 0.57 / 0.16	0.196 (0.096) / 0.10 / 0.10	0.187 (0.101) / 0.02 / 0.08
0.2	ROBOT-χ^2	0.247 (0.078) / 0.73 / 0.21	0.313 (0.114) / 0.57 / 0.16	0.198 (0.086) / 0.09 / 0.11	0.180 (0.057) / 0.01 / 0.09
0.2	Geo. median	0.237 (0.049)	0.388 (0.052)	0.535 (0.057)	0.591 (0.060)
0.2	Coord. median	0.253 (0.064)	0.423 (0.064)	0.655 (0.070)	0.723 (0.076)
0.2	MoM	0.239 (0.048)	0.434 (0.053)	0.829 (0.056)	1.628 (0.089)
0.3	BASE	0.343 (0.047) / 1.00 / 0.00	0.668 (0.047) / 1.00 / 0.00	1.334 (0.047) / 1.00 / 0.00	2.673 (0.047) / 1.00 / 0.00
0.3	r=1 (ROBOT)	0.337 (0.049) / 0.85 / 0.13	0.596 (0.084) / 0.70 / 0.16	0.284 (0.168) / 0.04 / 0.21	0.150 (0.053) / 0.00 / 0.06
0.3	r=1.5	0.339 (0.047) / 0.87 / 0.11	0.621 (0.065) / 0.67 / 0.19	0.481 (0.164) / 0.05 / 0.22	0.159 (0.062) / 0.00 / 0.12
0.3	r=2	0.342 (0.049) / 0.88 / 0.10	0.641 (0.055) / 0.73 / 0.16	0.658 (0.133) / 0.06 / 0.24	0.174 (0.074) / 0.00 / 0.19
0.3	r=3	0.344 (0.047) / 0.93 / 0.06	0.655 (0.051) / 0.81 / 0.11	0.810 (0.107) / 0.08 / 0.25	0.384 (0.079) / 0.00 / 0.11
0.3	Nietert	0.341 (0.077) / 0.78 / 0.18	0.519 (0.131) / 0.65 / 0.16	0.230 (0.137) / 0.11 / 0.05	0.199 (0.131) / 0.02 / 0.03
0.3	ROBOT-χ^2	0.342 (0.082) / 0.75 / 0.20	0.496 (0.145) / 0.63 / 0.16	0.221 (0.124) / 0.10 / 0.06	0.190 (0.102) / 0.01 / 0.04
0.3	Geo. median	0.339 (0.050)	0.608 (0.053)	0.918 (0.064)	1.052 (0.072)
0.3	Coord. median	0.350 (0.063)	0.642 (0.068)	1.090 (0.077)	1.270 (0.090)
0.3	MoM	0.340 (0.050)	0.655 (0.055)	1.264 (0.065)	2.500 (0.109)

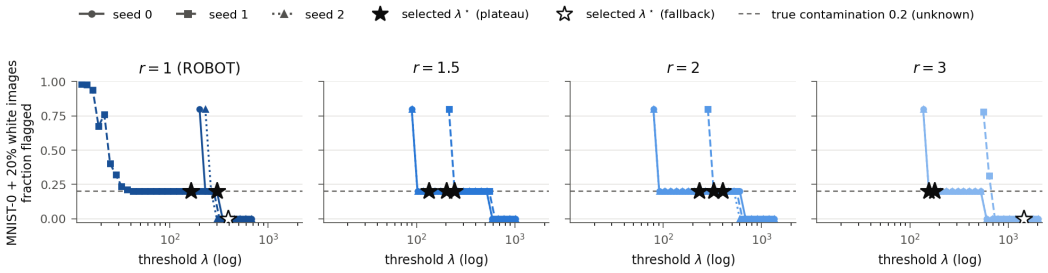

Figure 8: Threshold paths on MNIST digit 0 with 20% all-white outliers, with the layout of Figure 4. All powers have a flat criterion; the far outliers are flagged over a wide range of λ.

all outliers (masking ≥ 0.9: ROBOT 3 of 6 runs, r=3 1 of 6) is a fallback run, in which the criterion picked a threshold too large to flag most outliers; two fallback runs still flagged most outliers (ROBOT FashionMNIST seed 1, masking 0.15; r=1.5 FashionMNIST seed 0, masking 0.12). Figure 9 marks each run accordingly, and Table 5 gives the λ-path statistics quoted in Section 5.2.

H Implementation Details and Reproducibility

This appendix specifies the estimator and the ϵ-free selector in enough detail to reproduce every number and figure in Section 5. All code, the data-generating scripts, and the raw per-trial logs are released with the paper; the mean-estimation results derive from a single log file

Table 3: Symmetric contamination, all cells.

ϵ	method	k=0.5	k=1	k=2	k=4
0	BASE		0.121 (0.039) / 0.00 / 0.00		
0	r=1 (ROBOT)		0.126 (0.039) / 0.00 / 0.12		
0	r=1.5		0.123 (0.039) / 0.00 / 0.14		
0	r=2		0.122 (0.039) / 0.00 / 0.12		
0	r=3		0.121 (0.038) / 0.00 / 0.08		
0	Nietert		0.161 (0.059) / 0.00 / 0.20		
0	ROBOT-χ^2		0.166 (0.065) / 0.00 / 0.21		
0	Geo. median		0.127 (0.039)		
0	Coord. median		0.157 (0.050)		
0	MoM		0.128 (0.040)		
0.1	BASE	0.125 (0.041) / 1.00 / 0.00	0.129 (0.042) / 1.00 / 0.00	0.147 (0.047) / 1.00 / 0.00	0.202 (0.065) / 1.00 / 0.00
0.1	r=1 (ROBOT)	0.132 (0.042) / 0.80 / 0.14	0.137 (0.042) / 0.68 / 0.12	0.145 (0.038) / 0.38 / 0.10	0.160 (0.056) / 0.36 / 0.01
0.1	r=1.5	0.129 (0.041) / 0.82 / 0.12	0.132 (0.042) / 0.68 / 0.11	0.144 (0.043) / 0.40 / 0.12	0.161 (0.055) / 0.28 / 0.05
0.1	r=2	0.127 (0.040) / 0.87 / 0.09	0.131 (0.044) / 0.75 / 0.08	0.143 (0.043) / 0.37 / 0.14	0.161 (0.059) / 0.29 / 0.03
0.1	r=3	0.126 (0.042) / 0.86 / 0.09	0.132 (0.043) / 0.78 / 0.08	0.145 (0.047) / 0.44 / 0.11	0.166 (0.058) / 0.39 / 0.01
0.1	Nietert	0.170 (0.053) / 0.67 / 0.22	0.176 (0.058) / 0.42 / 0.20	0.173 (0.055) / 0.23 / 0.15	0.186 (0.057) / 0.25 / 0.15
0.1	ROBOT-χ^2	0.181 (0.066) / 0.64 / 0.25	0.178 (0.066) / 0.45 / 0.21	0.175 (0.063) / 0.29 / 0.16	0.189 (0.062) / 0.27 / 0.15
0.1	Geo. median	0.132 (0.040)	0.135 (0.040)	0.140 (0.041)	0.143 (0.042)
0.1	Coord. median	0.164 (0.050)	0.166 (0.052)	0.175 (0.052)	0.179 (0.052)
0.1	MoM	0.133 (0.039)	0.138 (0.042)	0.154 (0.047)	0.194 (0.064)
0.2	BASE	0.125 (0.040) / 1.00 / 0.00	0.132 (0.041) / 1.00 / 0.00	0.158 (0.050) / 1.00 / 0.00	0.237 (0.083) / 1.00 / 0.00
0.2	r=1 (ROBOT)	0.131 (0.042) / 0.83 / 0.12	0.139 (0.046) / 0.64 / 0.13	0.157 (0.048) / 0.44 / 0.08	0.179 (0.070) / 0.29 / 0.01
0.2	r=1.5	0.127 (0.040) / 0.82 / 0.12	0.138 (0.043) / 0.65 / 0.13	0.158 (0.049) / 0.40 / 0.11	0.168 (0.068) / 0.25 / 0.04
0.2	r=2	0.127 (0.040) / 0.85 / 0.10	0.135 (0.042) / 0.78 / 0.07	0.156 (0.047) / 0.39 / 0.10	0.174 (0.069) / 0.27 / 0.03
0.2	r=3	0.126 (0.040) / 0.90 / 0.07	0.134 (0.041) / 0.79 / 0.07	0.154 (0.049) / 0.51 / 0.07	0.184 (0.068) / 0.32 / 0.01
0.2	Nietert	0.169 (0.057) / 0.70 / 0.20	0.179 (0.058) / 0.49 / 0.16	0.169 (0.053) / 0.29 / 0.08	0.191 (0.068) / 0.29 / 0.08
0.2	ROBOT-χ^2	0.177 (0.068) / 0.68 / 0.22	0.177 (0.067) / 0.50 / 0.17	0.172 (0.056) / 0.27 / 0.11	0.186 (0.062) / 0.23 / 0.10
0.2	Geo. median	0.129 (0.042)	0.135 (0.042)	0.144 (0.044)	0.151 (0.046)
0.2	Coord. median	0.160 (0.053)	0.170 (0.052)	0.188 (0.057)	0.192 (0.059)
0.2	MoM	0.133 (0.042)	0.139 (0.044)	0.165 (0.054)	0.237 (0.078)
0.3	BASE	0.122 (0.039) / 1.00 / 0.00	0.134 (0.041) / 1.00 / 0.00	0.174 (0.054) / 1.00 / 0.00	0.284 (0.092) / 1.00 / 0.00
0.3	r=1 (ROBOT)	0.127 (0.040) / 0.81 / 0.12	0.137 (0.042) / 0.73 / 0.09	0.170 (0.055) / 0.45 / 0.07	0.176 (0.073) / 0.13 / 0.02
0.3	r=1.5	0.123 (0.038) / 0.83 / 0.12	0.135 (0.041) / 0.67 / 0.11	0.168 (0.052) / 0.39 / 0.08	0.169 (0.073) / 0.11 / 0.07
0.3	r=2	0.122 (0.038) / 0.88 / 0.08	0.135 (0.041) / 0.74 / 0.08	0.166 (0.052) / 0.44 / 0.09	0.175 (0.069) / 0.14 / 0.03
0.3	r=3	0.126 (0.039) / 0.90 / 0.07	0.135 (0.042) / 0.84 / 0.04	0.168 (0.052) / 0.48 / 0.07	0.185 (0.071) / 0.15 / 0.01
0.3	Nietert	0.168 (0.060) / 0.71 / 0.19	0.170 (0.065) / 0.56 / 0.11	0.169 (0.059) / 0.33 / 0.03	0.195 (0.091) / 0.24 / 0.02
0.3	ROBOT-χ^2	0.173 (0.069) / 0.69 / 0.20	0.176 (0.070) / 0.55 / 0.13	0.168 (0.058) / 0.29 / 0.05	0.174 (0.065) / 0.17 / 0.05
0.3	Geo. median	0.128 (0.042)	0.137 (0.044)	0.152 (0.051)	0.164 (0.057)
0.3	Coord. median	0.154 (0.046)	0.172 (0.050)	0.203 (0.064)	0.214 (0.071)
0.3	MoM	0.128 (0.043)	0.140 (0.045)	0.176 (0.056)	0.265 (0.093)

Table 4: Outlier detection at $\epsilon = 0.2$ (one-sided; 100 trials): masking / swamping / F_1. Bold: best per column and metric (lowest masking and swamping, highest F_1). Generated by `scripts/make_detbench_table.py`. ROBOT-χ^2 flags use a post-hoc cutoff on its per-point weight w_i ($w_i < 0.5/n$), since its weights are not pinned at 1 for inliers.

method	k=0.5	k=1	k=2	k=4
ROBOT	0.84/0.13/0.13	0.63/0.16/0.29	0.04/0.23/0.69	**0.00**/**0.04**/**0.93**
ROBOT-χ^2	**0.73**/0.21/0.21	**0.57**/0.16/0.35	0.09/0.11/**0.79**	0.01/0.09/0.86
Nietert	0.75/0.18/**0.22**	**0.57**/0.16/**0.38**	0.10/**0.10**/**0.79**	0.02/0.08/0.86
PW-ROT (r=2)	0.85/**0.11**/0.12	0.69/**0.15**/0.23	**0.02**/0.30/0.62	**0.00**/0.14/0.80

(`logs/table1_final.jsonl`, 2.86×10^4 records) and the figures are produced from it by one plotting script.

The PW-ROT estimator. The location estimator solves the OT problem with the robust ground cost $\rho_{\lambda,r}$ (Theorem 3.1) by a stochastic semi-discrete scheme: the target is the Gaussian generator $\mu + \mathcal{N}(0, I_d)$, so its "atoms" are drawn fresh each step and the dual potential $v \in \mathbb{R}^n$ is maintained on the n data points. Algorithm 1 gives the procedure. The inner loop is an entropic (Sinkhorn) dual ascent on v; the outer loop takes an Adam step on μ along the robust transport gradient, in which, because of the linear penalty $\lambda(1-w)$, only points with squared distance $s_i = \|x_i - y\|^2 \le \lambda$ to the current target draw y (weight $w_i^\star = w^\star(s_i) = 1$) enter at full strength, while flagged points are damped by $W_r(s_i) = (w_i^\star)^r = (\lambda/s_i)^{r/(r-1)}$. Here $\mathrm{softmax}(u)_i = e^{u_i}/\sum_j e^{u_j}$, and $\mathrm{Adam}(\cdot)$ is the update direction of Adam in its AMSGrad variant ($\beta_1 = 0.9$, $\beta_2 = 0.95$).

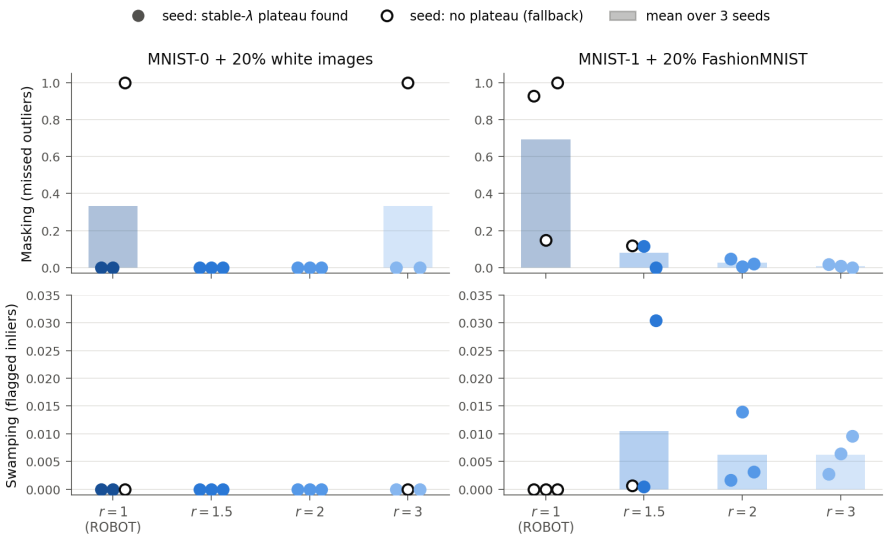

Figure 9: Figure 7 with each run marked by the selector's outcome: filled, a stable-λ plateau was found; hollow, no plateau (fallback). Every near-total detection failure is a fallback run.

Table 5: λ-path statistics on MNIST (`scripts/plot_mnist_story.py`). In-band: number of grid values of λ at which the flagged fraction lies in $(0.02, 0.5)$. Median adjacent Jaccard: median Jaccard index between the flag sets at neighboring grid values (increasing λ), over pairs whose smaller-λ member is in band. $\mathcal{W}_2$ step factor: exp of the median over the path of $|\log(\mathcal{W}_{2,j+1}/\mathcal{W}_{2,j})|$ for the trimmed Sinkhorn-$\mathcal{W}_2$ criterion, with j indexing the λ grid.

task	method	in-band λ (s0/s1/s2)	median adj. Jaccard	median W_2 step factor
white	r=1 (ROBOT)	3/19/1	1.00/1.00/0.00	1.03/1.09/1.06
white	r=1.5	13/7/13	1.00/1.00/1.00	1.05/1.03/1.03
white	r=2	15/5/14	1.00/1.00/1.00	1.04/1.06/1.02
white	r=3	10/1/10	1.00/0.00/1.00	1.03/1.00/1.01
FMNIST	r=1 (ROBOT)	3/3/3	0.00/0.00/0.00	6.85/7.48/5.27
FMNIST	r=1.5	5/20/18	0.49/0.89/0.89	2.01/1.14/1.21
FMNIST	r=2	9/18/20	0.74/0.93/0.87	1.25/1.17/1.09
FMNIST	r=3	14/20/11	0.93/0.92/0.88	1.15/1.12/1.11

Algorithm 1 PW-ROT robust mean estimator (stochastic semi-discrete solver). Code: `pwrot/mean_estimation.py::fit_mean`.

Require: data $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$; power $r \geq 1$; threshold $\lambda > 0$; step η_A; entropic scale ε_{ent}; inner steps N_{in}; iterations N_{out}; initial value $\mu^{(0)}$

initialize $\mu \leftarrow \mu^{(0)}$; dual $\tilde{v} \leftarrow 0 \in \mathbb{R}^n$, $v \leftarrow 0$; empirical weights $a \leftarrow \frac{1}{n}\mathbf{1}$
for $k = 1, \ldots, N_{\text{out}}$ **do**
  for $j = 1, \ldots, N_{\text{in}}$ **do** ▷ entropic dual ascent, fresh target atom
    draw $z \sim \mathcal{N}(0, I_d)$; $y \leftarrow \mu + z$; $s_i \leftarrow \|x_i - y\|^2$, $C_i \leftarrow \rho_{\lambda,r}(s_i)$ ($\min(s_i, 2\lambda)$ for ROBOT)
    $\tilde{v} \leftarrow \tilde{v} + \big(a - \text{softmax}\big((\tilde{v} - C)/\varepsilon_{\text{ent}}\big)\big)$; $v \leftarrow (\tilde{v} + \lfloor (k-1)/10^4 \rfloor v)/(\lfloor (k-1)/10^4 \rfloor + 1)$
  end for
  $\pi \leftarrow \text{softmax}\big((v - C)/\varepsilon_{\text{ent}}\big) \in \Delta^n$ ▷ semi-discrete plan at the last atom
  $\hat{g} \leftarrow \nabla_\mu \sum_i \pi_i \rho_{\lambda,r}(\|x_i - y\|^2)$ ▷ only $s_i \leq \lambda$ contribute fully; flagged points damped by $W_r(s_i) = (w_i^\star)^r$
  $\mu \leftarrow \mu - \eta_A \cdot \text{Adam}(\hat{g})$
end for
return μ ▷ flags: 95% rule, `scripts/hybrid_trimval.py::detect_95`

Table 6: Per-seed MNIST results (`scripts/plot_mnist_story.py`).

task	method	seed	$\lambda^\star$ (selection)	p95 rule mask	p95 rule swamp	min-distance rule mask	min-distance rule swamp	rFID
white	ROBOT (r=1)	0	300.5 (plateau, 3)	0.000	0.0000	1.000	0.0000	65.3
white	ROBOT (r=1)	1	162.9 (plateau, 16)	0.000	0.0000	0.000	0.0000	27.6
white	ROBOT (r=1)	2	393.6 (fallback)	1.000	0.0000	1.000	0.0000	61.6
white	r=1.5	0	200.6 (plateau, 13)	0.000	0.0000	0.000	0.0000	15.7
white	r=1.5	1	244.1 (plateau, 7)	0.000	0.0000	0.000	0.0000	38.6
white	r=1.5	2	133.9 (plateau, 13)	0.000	0.0000	0.000	0.0000	57.5
white	r=2	0	401.0 (plateau, 15)	0.000	0.0000	0.000	0.0000	34.5
white	r=2	1	325.8 (plateau, 5)	0.000	0.0000	0.000	0.0000	28.7
white	r=2	2	233.8 (plateau, 14)	0.000	0.0000	0.000	0.0000	36.9
white	r=3	0	156.1 (plateau, 10)	0.000	0.0000	0.000	0.0000	24.3
white	r=3	1	1438.2 (fallback)	1.000	0.0000	1.000	0.0000	1464.3
white	r=3	2	178.6 (plateau, 10)	0.000	0.0000	0.000	0.0000	45.6
FMNIST	ROBOT (r=1)	0	107.0 (fallback)	0.927	0.0000	0.958	0.0000	9.4
FMNIST	ROBOT (r=1)	1	48.4 (fallback)	0.149	0.0000	0.357	0.0000	19.3
FMNIST	ROBOT (r=1)	2	187.7 (fallback)	1.000	0.0000	1.000	0.0000	24.6
FMNIST	r=1.5	0	93.8 (fallback)	0.119	0.0006	0.183	0.0000	8.8
FMNIST	r=1.5	1	94.8 (plateau, 9)	0.117	0.0004	0.145	0.0000	14.5
FMNIST	r=1.5	2	37.5 (plateau, 10)	0.000	0.0304	0.002	0.0102	7.1
FMNIST	r=2	0	72.8 (plateau, 4)	0.048	0.0017	0.067	0.0008	15.1
FMNIST	r=2	1	48.6 (plateau, 11)	0.006	0.0140	0.016	0.0031	37.8
FMNIST	r=2	2	65.1 (plateau, 9)	0.021	0.0031	0.048	0.0015	9.2
FMNIST	r=3	0	63.8 (plateau, 8)	0.018	0.0027	0.034	0.0015	17.3
FMNIST	r=3	1	56.2 (plateau, 13)	0.008	0.0065	0.022	0.0013	9.1
FMNIST	r=3	2	49.9 (plateau, 5)	0.001	0.0096	0.010	0.0021	7.7

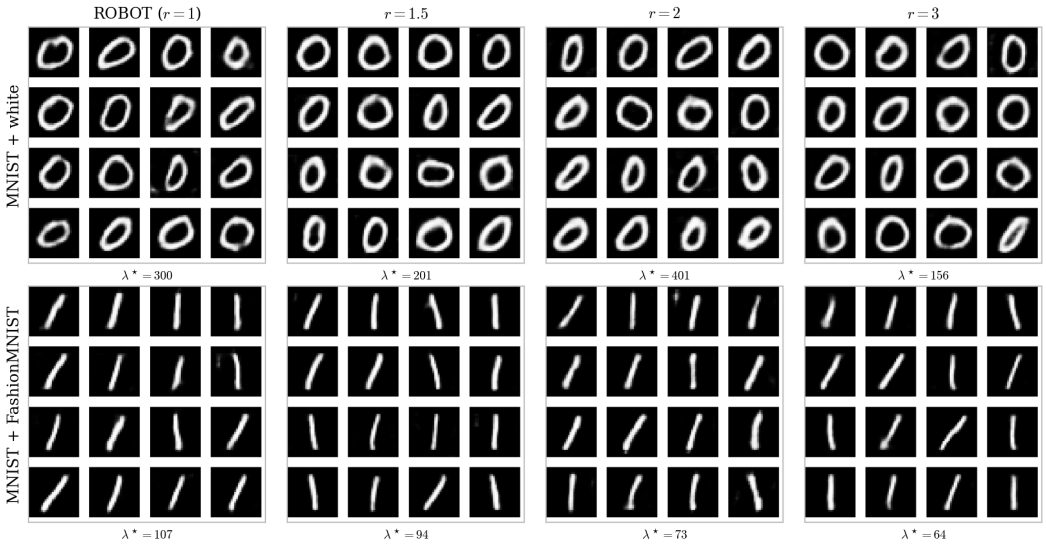

Figure 10: Generated samples at the selected λ (seed 0; top-left 4×4 of each 8×8 panel). Rows: contamination types; columns: powers. The samples contain no outlier images, but this is *not* evidence of robust generation: the stochastic semi-discrete solver drops far, disconnected modes even without robust weighting, and these runs include no non-robust baseline. A bounded cost alone cannot robustify an unrestricted generator, since $\mathrm{OT}_\rho(P_{\mathrm{data}}, Q)$ is minimized over all laws Q at $Q = P_{\mathrm{data}}$. r=3 collapsed onto the outlier mode on one seed.

The ϵ-free selector. The threshold λ is chosen *without* knowledge of ϵ. In the mean-estimation study (Figures 2, 5 and 6, Appendix F) the rule is *pure trimmed validation* (Algorithm 2): fit PW-ROT along a data-derived λ path on an 80% training split, compute the held-out error after dropping the worst $\zeta = 35\%$ of validation residuals—a level above the largest contamination studied ($\epsilon \leq 0.3$)—and take the minimizer among thresholds flagging fewer than 50% of training points. The flag set uses the 95% rule: a point is flagged iff the 5th percentile of its squared distances to samples of the fitted target exceeds the threshold.

For MNIST (Section 5.2) we use a plateau-gated variant (`pwrot/metrics.py::select_lambda_hybrid_gan`): among thresholds whose flag sets form a stable run (adjacent Jaccard ≥ 0.9, flag fraction in $(0.02, 0.5)$, run length ≥ 3), pick the one minimizing an in-sample, 35%-trimmed Sinkhorn-$\mathcal{W}_2$ between generated and real images in pixel space; if no plateau exists it falls back to the minimizer over all thresholds (no 50% cap). The fallback was used in 6 of 24 MNIST runs, including all three ROBOT runs on FashionMNIST.

Algorithm 2 ϵ-free trimmed-validation selection of λ (mean study). Code: `scripts/run_table1_final.py::_select_puretrim`.

Require: data X; power r; trim $\zeta = 0.35$; flag cap 0.5

split X into training X_{tr} (80%) and validation X_{val} (20%)
build a data-derived geometric grid Λ_r (anchored on the cost scale of an anchor fit; see Settings)
for all $\lambda \in \Lambda_r$ (warm-started along the path) **do**
    $\hat{\mu}_\lambda \leftarrow$ Algorithm 1 on X_{tr} at (r, λ)
    $\hat{\mathcal{O}}_\lambda \leftarrow 95\%$-rule flag set on X_{tr}; $\mathrm{err}_\lambda \leftarrow$ trimmed-validation error on X_{val} (drop worst ζ)
end for
$\lambda^\star \leftarrow \arg\min_{\lambda:\, |\hat{\mathcal{O}}_\lambda|/|X_{\mathrm{tr}}| < 0.5} \mathrm{err}_\lambda$ ▷ if no λ flags $< 50\%$, minimize over all λ
refit Algorithm 1 on all of X at $(r, \lambda^\star)$, warm-started at $\hat{\mu}_{\lambda^\star}$
return $\hat{\mu}$, $\lambda^\star$

Settings. Unless noted, $d = 5$, $n = 300$, 100 trials per cell. The solver uses $N_{\mathrm{out}} = 10^5$ outer iterations for the final fit, $N_{\mathrm{in}} = 100$ inner dual-ascent steps per outer step, Adam step $\eta_{\mathrm{A}} = 10^{-3}$, and entropic scale $\varepsilon_{\mathrm{ent}} = 10^{-3}$; $\mu^{(0)}$ is the warm start of Algorithm 2. The grid starts at $\lambda_{\max}$, a multiple of the largest squared residual of a standard-OT (non-robust) fit, and decreases geometrically (25 points, ratio 0.85); the λ path starts with a 3×10^4-iteration robust fit at $\lambda_{\max}$ and then uses at most 2500 warm-started iterations per grid point (early stop once μ settles). Trial j uses data seed $12345 + j$ and solver seed j; all powers share these settings, and the $r = 1$ (ROBOT) branch uses the 2λ hard-clip threshold consistent with its original formulation (the near-threshold study uses the analytic endpoint $\min(C, \lambda)$). The MNIST solver (`pwrot/gan.py`) trains a DCGAN-style generator (latent dimension 64; three transposed-convolution blocks with batch normalization and ReLU, then a convolution and a sigmoid) with Adam (step 2×10^{-3}) for 10^4 iterations, entropic scale $\varepsilon_{\mathrm{ent}} = 0.2$, the first 10% of iterations with the standard OT cost as warm-up. Each λ is trained from scratch on a 30-point geometric grid from $\lambda_{\max}$ (r times the 95th percentile of warm-up costs) down to $0.02\lambda_{\max}$, traversed from large to small λ and stopped once the min-distance rule flags $\geq 50\%$ of points. The selection criterion is the squared Sinkhorn-$\mathcal{W}_2$ distance between generated and real images after dropping the worst 35% of real images, with entropic scale 0.1 times the median cost on at most 2000 points. Package versions are in the supplement's `requirements.txt`. The near-threshold study (Figure 3; `experiments/breakdown_validation.py`, numbers in `results/near_threshold_validation.json`) uses $n = 400$, 300 trials, with the ϵ fraction of outliers placed as a tight cluster at squared distance $\approx \lambda$ (where the score norm is maximal) and λ the 0.85 quantile of χ^2_5.

Data generation. Inliers are $\mathcal{N}(\mu, I_5)$. For *asymmetric* contamination an ϵ fraction of outliers is $\mathcal{N}(\mu - k\mathbf{1}_5, I_5)$ (displacement magnitude $k\sqrt{5}$); for *symmetric* contamination each coordinate of an outlier is shifted by $+k$ or $-k$ independently (a random hypercube corner, magnitude $k\sqrt{5}$). We sweep $\epsilon \in \{0, 0.1, 0.2, 0.3\}$ and $k \in \{0.5, 1, 2, 4\}$.

Baselines. The external estimators of Figures 2–6 are the coordinate-wise median, the geometric median (Weiszfeld), median-of-means, Nietert et al.'s robust-OT estimator, and ROBOT-χ^2. Nietert's estimator is, for a point target, the ϵ-trimmed mean, computed by least-trimmed-squares iterations from the coordinate-wise median (trimmed mass on the grid $\{0, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$); ROBOT-$\chi^2$ is the reweighted mean with weights proportional to $(\eta_\chi - \ell_i)_+$ for squared residuals ℓ_i, with η_χ set by bisection so that the weights exhaust the χ^2 budget (same grid); MoM is the geometric median of the means of n_{b} random blocks, $n_{\mathrm{b}} \in \{1, 3, 5, 9, 15, 21, 31\}$. Each baseline's parameter is selected by the same trimmed-validation criterion ($\zeta = 0.35$) over its grid, fitting on all n points and scoring on the last 20%. For the detection table, Nietert flags the trimmed points and ROBOT-χ^2 flags points whose per-point weight w_i

falls below half the uniform share ($w_i < 0.5/n$); the latter is a post-hoc cutoff, because its weights are not pinned at 1 for inliers. On clean data the selector picks zero trimming only in a minority of trials, which is why these baselines lose efficiency at $\epsilon = 0$.

Reproducing the figures. `scripts/run_table1_final.py` regenerates `logs/table1_final.jsonl` (all methods, both contamination types, the full $\epsilon \times k$ grid); `scripts/plot_story_figs.py` produces Figures 2, 5 and 6, `scripts/make_family_detect_table.py` Table 1, and `scripts/make_full_tables_tex.py` and `scripts/make_detbench_table.py` the appendix tables, from that log. `scripts/plot_mnist_story.py` produces Figures 4, 8 and 7 and Table 6 from `results_mnist_hybrid/`. That directory is produced by `scripts/mnist_hybrid_perseed_sweep.sh` (which calls `experiments/run_gan.py`); `experiments/breakdown_validation.py` produces Figure 3, `experiments/plot_cost_functions.py -scaled` Figure 1, and `scripts/plot_mnist_samples.py` the MNIST sample figure. The anonymized supplement contains exactly these scripts, the package `pwrot/`, the logs and the MNIST results; its `README.md` lists the commands, and it regenerates every table byte for byte and every figure pixel for pixel.