LEARNING GRADED SPARSE MECHANISMS: A REPRESENTATION THEOREM AND CERTIFIED GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In a variety of multi-agent and physical environments, an element's update depends strongly on a few other elements and weakly on many, and both the identity and the number of the elements that matter change with the state of the system. Such graded sparse mechanisms are common, for instance, in flocks, crowds, learned particle simulators and the neighborhood aggregation of graph networks, yet they have received little systematic study as a class of functions with the grading given. In this paper we study the supervised learning of the mechanism when the grading is provided by the environment, as a weight on every element. First, we define the class by axioms, where each element's influence is capped by its weight, the dependence on the weights is regular, and no element is distinguished, and prove a representation theorem: the class is exactly the Lipschitz ball of a transport metric on weighted multisets, and the metric is forced by the axioms. Next, the geometry of the representation gives generalization bounds that depend on the data in two ways. One is via the sparsity profile of the data—the law of how many elements matter and by how much—which bounds the rate over all designs with that profile, with an exponent from the profile and no dependence on the number of elements. In a number of natural cases we also show matching lower bounds, up to constants in the exponent. The other is empirical: due to the representation result, classical empirical discrepancy bounds apply to the class as they stand, and the resulting bound is the transport distance between the two halves of the sample in the class's own metric, computed exactly by a matching, with an additional concentration term. Finally, we study a simulated Vicsek flock under four interaction rules at two noise levels, and show that the certificate is non-vacuous, a small fraction of the trivial value on every rule.

1 INTRODUCTION

In models of complex systems, a property of an element depends on its own state and on the states of some other elements. The dependence is often *sparse* and *graded*: a few elements have a strong influence, and the rest a weaker and decreasing influence, or none at all. It is often *changing*: as the system evolves the grading changes, so the sparse set varies, and varies in size. And it is often *anonymous*: the elements form a set, with no enumeration that the mechanism reads. One mechanism is applied at every state of the system to the elements the state selects, and we call the environment's rule that assigns the weights from the state, which elements count and by how much, the *selector*.

The examples of this paper come from multi-agent and physical systems, and in each of them the selector is the environment's, a rule with at most a few tuned parameters that reads distance, speed, visibility or rank. The running example is the Vicsek flock (Vicsek et al., 1995). Its particles move at a common speed on a periodic square, and at every step each particle turns toward the mean heading of the particles within a fixed distance of itself, up to noise. The selector is the ball: the particles inside it count with equal weight and the rest not at all, and how many are inside varies with the local crowding. Graded variants of the rule weight the neighbors by a kernel of the distance instead, so that the influence fades rather than stops. The selectors of the other examples are of the same kind: distance and speed thresholds in drone flocks (Vásárhelyi et al., 2018), the visible neighbors

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.

in crowds (Wirth et al., 2023), the six or seven nearest in starling flocks (Ballerini et al., 2008) and fish schools (Gautrais et al., 2012), a view cone or a sensing radius in driving simulators (Vinitsky et al., 2022; Kazemkhani et al., 2025) and navigation policies (Hüttenrauch et al., 2019; Nayak et al., 2023; Zhang et al., 2025), and a cutoff in learned particle simulators (Sanchez-Gonzalez et al., 2020; Musaelian et al., 2023).

The learning problem is this. A configuration is a pair (x, μ), the values x of the d elements and the weights μ the selector gives them, a probability vector: the selector is the environment's and is given, not learned, so the weights are part of the input, and the unknown is the mechanism $f(x, \mu)$, a member of the class $\mathcal{G}_L$ of Section 3. A sample is n configurations drawn from the *design*, the law of (x, μ), with the labels $f(x, \mu)$, under the absolute loss, and what we bound is the uniform deviation of the class on the design, the generalization gap of every empirical risk minimizer.

While such systems are classical, the graded sparse mechanism as a class of functions, and the complexity of learning one with the selector given, have received little study beyond hard sparsity at a fixed count (Section 2), and this paper develops first results toward both. Taking the selector as given is the decision the graph-network bounds make in taking the graph as given. It is also given in simulators, and in the animal groups some aspects of it are inferred in some cases (Section 2).

Axioms and representation. We propose an axiomatic definition of the class. The alternative, a function of a fixed number k_0 of arguments fed a different selection of inputs at each configuration, is the form of neighborhood aggregation in the generalization bounds for graph networks, stated under a maximum degree k_0 (Garg et al., 2020; Liao et al., 2021), a worst-case parameter that sees neither the average sparsity across configurations nor a grading inside the k_0. We define the class instead by three axioms on a mechanism, a function of the elements' values and of a weight on each element, so that the number of arguments changes with the configuration, and changes gradually. An element's influence is capped by its weight, none at weight zero. The dependence on the weights is regular: shifting weight between elements changes the output by no more than the same shift can change a weighted mean of the values, the simplest mechanism, which the class must contain. And no element is distinguished, so elements of equal weight are interchangeable. The axioms concern the mechanism alone: the selector enters only through the data distribution, and it may be discontinuous, as a hard radius is.

The examples come in two semantics of influence, a mean, fixed in total and diluted by more participants, and a sum, growing with their number, and the framework has a class for each, *conservative* and *accumulative*. The main text works with the conservative class. The accumulative class has the same representation and rates of the same exponents, at the price of a factor of the count (Section 3.1, Appendix B.7).

The axioms are a list of inequalities, and the structure of the class is not visible in them. We show that the class is exactly the set of Lipschitz functions on a metric space of weighted configurations (Theorem 3.2): all of the locality structure is in the metric, above it the class is a generic Lipschitz class, and this is what makes the rest of the analysis possible. The metric is the class's own, each of its terms from one axiom, a transport distance between weighted configurations read as multisets of points on a cone over the value space. We note that the permutation-invariant representation results, Deep Sets (Weyl, 1939; Macdonald, 1995; Zaheer et al., 2017; Wagstaff et al., 2022), have a similar structure: there the sum embedding encodes the invariance, with a generic class above it. There is also an exact connection: permutation-invariant sum networks as above, with a Lipschitz, bounded per-element map and our weights on the sum, $\sum_i \mu_i \varphi(x_i)$, are readouts of finitely many linear functionals of our embedding, thus constituting a strict subset of our class.

Complexity. Sparsity is a property of the data, and we prove generalization bounds for the class that are data dependent in two ways, both resting on the representation. The first is through the *sparsity profile* of the design, the law of the weight that a configuration leaves beyond its k heaviest elements: Figure 1(a) is the profile of the flock, the mean weight beyond the k heaviest neighbors under the ball and under three graded kernels at its scale. The bounds have an exponent set by the profile, the worst case over the designs with that profile, and no dependence on the number of elements, and lower bounds in the profile show that the dependence is the right one (Section 4). The second is through the sample itself. It gives no rate, but a bound at the actual sample size, and often a better one. The bound is the classical maximum discrepancy of Bartlett et al. (2002), the largest

difference, over the class, between the averages of a function on the two halves of the sample. For a Lipschitz class this is a transport distance (von Luxburg & Bousquet, 2004), and here it is computed exactly by a matching. The result is an *empirical generalization certificate*, the value of the bound on the data with one concentration term, which on the flock is 0.034κ on the hard ball, with κ the value at which the bound says nothing (Section 5.1).

To summarize, the contributions in this paper are as follows. **(i)** *Axioms and representation.* We define the class of graded sparse mechanisms by axioms and prove that it is the Lipschitz ball of a specific metric space, whose metric encodes the locality (Theorem 3.2). **(ii)** *Rates from the profile.* We define the sparsity profile of the design and prove generalization bounds with exponent determined by the profile, with no dependence on the number of elements (Theorems 4.3 to 4.5), and lower bounds matching up to constants in the exponent in several cases (Theorems 4.4, 4.5 and D.5). **(iii)** *The certificate.* We show that the generalization gap of the class can be certified from the sample alone, by a matching between its two halves (Theorem 4.6). **(iv)** *The flock.* We study the Vicsek flock under four interaction rules at two noise levels, measure their sparsity profiles, compute the certificate for every rule, a thirtieth to a sixth of the trivial value, and read it against a covering certificate and against the gap of a member with zero training risk (Section 5).

2 LITERATURE

Graph networks. A message-passing layer is our object with a hard selector, the neighborhood of each node, and it is of the Deep Sets form, a per-neighbor map summed and read out, one readout of finitely many linear sums of our embedding (Section 1), at a fixed width, a bounded degree and no weighting by the environment. Graph networks gain their expressivity from depth, where their sparsity lives: the receptive field is set by the graph and the number of layers, and the bounds of Garg et al. (2020) and Liao et al. (2021) pay for it with the maximum degree k_0 per layer, compounded over the depth, Ju et al. (2023) with its spectral relaxation. Here the sparsity is explicit, the weights of the selector, and the bound reads it as the profile: the number of neighbors enters through its law, with rates where there is no nominal sparsity at all and no dependence on the maximum count where there is one (Section 4.1, Remark B.19). Whether any finer statistic of the data than the maximum degree can be used was posed as a question in that literature (Liao et al., 2021; Morris et al., 2023; 2024; Vasileiou et al., 2025b). The two kinds of bound also bound different things: theirs hold over a parametric family at a norm budget and, like our rates, are statements about scaling (Appendix E), and the class adds a bound of a different kind, the certificate, which holds for the whole class at once and is computed from the sample by a matching (Section 4.2). A second line puts a metric on the graphs under which message-passing networks are Lipschitz (Chuang & Jegelka, 2022; Levie, 2023; Böker et al., 2023; Rauchwerger et al., 2025; Amran et al., 2026), and covers the graph space at one scale, or with the sample's cell proportions at one radius (Carrasco et al., 2025), Covered Forest (Vasileiou et al., 2025a) being its latest form. We differ in the object, a general weighted class whose metric the class itself implies (Theorem 3.2) rather than one metric among several under which a model is Lipschitz, and in the data dependence: rates through the profile (Section 4.1) and, on the data, the two-sample transport in place of a cover (Section 4.2).

The matching distance. The metric of Theorem 3.2 is produced by the axioms, which determine it in closed form, a matching on a cone: the identification is new, and the closed form is what makes the rest computable. The metric space itself, a cone with a transport that may leave a point unmatched at a price, is known: unbalanced transport (Chizat et al., 2018; Liero et al., 2018), the matching distances of persistence diagrams (Cohen-Steiner et al., 2010; Turner et al., 2014), of multisets (Chuang & Jegelka, 2022; Davidson & Dym, 2025; Nikolentzos & Skianis, 2026) and of point patterns (Schuhmacher & Xia, 2008; Schuhmacher et al., 2008; Huang et al., 2026), none of them with the correspondence the theorem establishes, between an element's weight and the mechanism's continuity in it.

Rates from a dimension. The dimensions of a law that bound the transport distance between an empirical measure and its law, the upper Wasserstein dimension of Weed & Bach (2019) and the entropic dimension of Chakraborty & Bartlett (2025), rest on one bulk covered at fixed dimension, the rest of the mass charged at the diameter. Our bounds stratify the design by its weight tail, a device with earlier variants (Fournier & Guillin, 2015; Lei, 2020; Huang et al., 2026), made to work

on configurations that are never exactly sparse. Quasi-polynomial rates of the shape of Theorem 4.5 appear before in two other settings, empirical measures in a Hilbert space under an exponential-ellipsoid moment condition (Lei, 2020) and point processes with a geometric count law (Huang et al., 2026). Ours is for the mean weight tail of a configuration, with free weights and values, where the decay is inside each sample and there is no count to stratify by, and on fixed, geometrically decaying weights the class's metric resembles Lei's ellipsoid (Appendix D.4).

Interaction kernels. The closest setting to the flock is the learning of interaction laws from trajectories (Lu et al., 2019; 2021), where the mechanism is a mean-field sum with an unknown scalar kernel, a linear sum in the terms of Section 3.1, learned at the one-dimensional rate with no profile in play (Appendix E). On real and simulated flocks, attention networks learn the update with the weights learned and read off the neighbors that matter (Heras et al., 2019; Ha & Jeong, 2021), and relational inference learns the interaction graph (Kipf et al., 2018): the selector is the object learned there, and it is given here.

The nearest results in learning theory are on transformers: a target on a few tokens at given positions, a hard selection at a fixed count in a parametric class (Mousavi-Hosseini et al., 2025; Edelman et al., 2022; Sanford et al., 2023; Trauger & Tewari, 2024), an input-dependent importance of the coordinates (Takakura & Suzuki, 2023) or a per-sample sensitivity cap (Nagler, 2023), and further remarks on the literature are in Appendix E.

3 Locality Axioms and Representation

This section defines the class by its axioms and proves its representation.

3.1 Axioms

Mechanism and selector. Throughout, $(\mathcal{V}, \rho)$ is a compact metric space with at least two points, the *value space*, and

$$R \;:=\; \tfrac{1}{2}\,\mathrm{diam}\mathcal{V}$$

is the *value scale*, so that $\rho \le 2R$ with equality attained. A *configuration* is a pair $(x, \mu) \in \mathcal{C} := \mathcal{V}^d \times \Delta_d$: a value $x_i \in \mathcal{V}$ on each of d coordinates and a weight vector μ in the simplex. A *mechanism* is a function $f : \mathcal{C} \to \mathbb{R}$. A *selector* produces the weights from the state of the system: each element carries a state $\omega_i \in \Omega$, of which its value $x_i = \pi(\omega_i)$ is the part the mechanism reads, for a fixed map $\pi : \Omega \to \mathcal{V}$. The selector is a map $c : \Omega^d \to \Delta_d$. The object of the examples is the composite

$$g(\omega) \;=\; f\big(\pi(\omega), c(\omega)\big), \qquad \pi(\omega) := (\pi(\omega_1), \dots, \pi(\omega_d)), \tag{1}$$

one mechanism applied at every state to the coordinates the selector weights. In the flock the state of a particle is its position and heading, the values are the headings, the selector is uniform on the particles within the interaction radius, and the mechanism is the mean heading. The axioms concern f alone. The selector enters only through the *design*, the law Q of $(\pi(\omega), c(\omega))$ under the law of the state ω, and nothing is assumed about it: it may be discontinuous, as a hard radius is, and in the examples it usually is. Throughout, the selector is given: the learner observes the configuration (x, μ), weights included, and learns f. Which coordinates of the state are the values, and in which semantics the mechanism is read, are part of the specification (Appendix C.1, Corollary A.15). Permutations σ of $[d]$ act on both arguments, $(\sigma x)_i = x_{\sigma^{-1}(i)}$ and likewise for μ. Write $\mathrm{supp}\mu = \{i : \mu_i > 0\}$ and $\mu_{\max} = \max_i \mu_i$.

Cap weights. The dependence allowed on a coordinate is a function of its weight, and the axioms are stated for a generic such function. A *cap weight* is a continuous map $w : \Delta_d \to [0,1]^d$ that commutes with permutations, is injective, and has $w_i(\mu) = 0$ exactly when $\mu_i = 0$. The two instances of this paper are

$$w^{\mathrm{c}}(\mu) \;=\; \mu \qquad \text{and} \qquad w^{\mathrm{a}}(\mu) \;=\; \mu/\mu_{\max},$$

the weights themselves, the *conservative* cap of the mean semantics, and the weights relative to the largest one, the *shares*, the *accumulative* cap of the sum semantics. On weights uniform over k coordinates the first allots influence L/k to each selected coordinate and the second L, a fixed

budget that more participants dilute against a total that accumulates with the count. The *conservative* class $\mathcal{G}_L := \mathcal{G}_L^{w^{\mathrm{c}}}$ is the class of the main text. The *accumulative* class $\widetilde{\mathcal{G}}_L := \mathcal{G}_L^{w^{\mathrm{a}}}$ has the same representation, Theorem 3.2 below holding for every cap weight, and rates of the same exponents at the price of a factor of the count (Appendix A.4, Appendices B.7 and B.8). Table 1 in Appendix A.4 sorts the common aggregations between the two classes, with their constants.

Definition 3.1 (The class $\mathcal{G}_L^w$)**.** *For a cap weight w and $L \geq 0$, $\mathcal{G}_L^w$ is the set of mechanisms $f : \mathcal{C} \to \mathbb{R}$ such that*

- *(V)* (value cap) *if $x, y \in \mathcal{V}^d$ differ in coordinate i only, then for every μ, $|f(x,\mu) - f(y,\mu)| \leq L\, w_i(\mu)\, \rho(x_i, y_i)$.*
- *(W)* (weight regularity) *for every $x \in \mathcal{V}^d$ and all $\mu, \nu \in \Delta_d$, $|f(x,\mu) - f(x,\nu)| \leq L\, R\, \|w(\mu) - w(\nu)\|_1$.*
- *(A)* (anonymity) *$f(\sigma x, \sigma\mu) = f(x,\mu)$ for every permutation σ.*

(V) is the locality and (A) the anonymity of the introduction, and together they make f a function of the multiset of pairs $\{(x_i, \mu_i) : \mu_i > 0\}$ (Lemma A.3). On an interval and for differentiable f the axioms are bounds on partial derivatives (Remark A.1). The weight axiom (W) is the one that needs a justification.

Where the weight axiom comes from. The class is meant to contain the weighted sums of a score, $\sum_i w_i(\mu)\, \varphi(x_i)$ with φ 1-Lipschitz and bounded by R, the *linear sums* of Appendix A.1. (W) is exactly their sensitivity to the weights: a move of weight from μ to ν changes a linear sum by at most $R\, \|w(\mu) - w(\nu)\|_1$, and some score attains this (Proposition A.4), so (W) is the tightest weight axiom under which every linear sum is still a member at constant 1. The axioms are chosen so that the simplest mechanisms have constant 1, and what Theorem 3.2 adds is that the metric they induce is the matching distance at the same constant (the coordinatewise form of the axiom and the readouts it excludes are in Appendix A.1). The modulus in (W) is stated in the weights alone, and this is a choice: pricing a weight move by the values it moves between gives the flat metric on the weighted value measure $\sum_i \mu_i \delta_{x_i}$, a smaller class in which the weights are mass rather than individuals, that of distribution regression and the measure-input networks (Appendix E), which excludes the selected maximum and the top weight, members here (Remark A.6). A mechanism may also read a distinguished input of its own, the state of the agent evaluating it, as in Section 5, and everything below then holds fiberwise in it (Remark A.2).

3.2 Representation

The cone and the lift. The *cone* over $\mathcal{V}$ is $\mathcal{K} := (\mathcal{V} \times [0,1])/(\mathcal{V} \times \{0\})$, the bottom edge collapsed to one point, the *apex* o. A point other than the apex is a pair (x, u) of a value and a *height* $u \in (0, 1]$, and the *cone cost* is

$$\varrho\big((x,u),(y,v)\big) \;:=\; \min(u,v)\, \rho(x,y) + R\,|u - v|, \qquad \varrho\big((x,u),o\big) \;:=\; R\,u, \tag{2}$$

a metric on $\mathcal{K}$ (Lemma A.7). The *lift* of a configuration is the counting measure on the cone

$$\Phi(x,\mu) \;:=\; \sum_{i \in \mathrm{supp}\mu} \delta_{(x_i,\, w_i(\mu))}, \tag{3}$$

one atom per coordinate of positive weight, at its value and at the height of its cap weight. Write $\mathcal{M}_d := \Phi(\mathcal{C})$, and n_P for the number of atoms of $P \in \mathcal{M}_d$. Two configurations have the same lift exactly when their multisets of (value, weight) pairs coincide, so Φ identifies $\mathcal{C}$ modulo permutations and modulo the values at zero-weight coordinates. Every member of $\mathcal{G}_L^w$ factors through it (Lemma A.8).

The matching distance. For $P, Q \in \mathcal{M}_d$, a *partial matching* is a bijection σ between a sub-multiset of the atoms of P and a sub-multiset of the atoms of Q. Its cost is the cone cost summed over the matched pairs plus R times the height of every unmatched atom on either side, and the *matching distance* is the least cost,

$$M_w(P,Q) \;:=\; \min_{\sigma} \Big[\sum_{p \text{ matched}} \varrho\big(p, \sigma(p)\big) + \sum_{p \in P \text{ unmatched}} \varrho(p,o) + \sum_{q \in Q \text{ unmatched}} \varrho(q,o) \Big]. \tag{4}$$

It is a metric on $\mathcal{M}_d$, and it is a transport distance: padding each side with the other's number of atoms at the apex,

$$M_w(P,Q) \;=\; W_1^{\varrho}\big(P + n_Q\,\delta_o,\; Q + n_P\,\delta_o\big), \tag{5}$$

the Kantorovich distance with ground cost ϱ between two measures of equal mass, whose optimal plans are matchings because the masses are integers (Lemma A.9). We write $D := M_{w^{\mathrm{c}}}$ for the conservative instance, heights equal to the weights, and $\widetilde{D} := M_{w^{\mathrm{a}}}$ for the accumulative one, heights equal to the shares.

Theorem 3.2 (Representation)**.** *Let w be a cap weight and $L \geq 0$.*

(i) A mechanism f belongs to $\mathcal{G}_L^w$ if and only if $f = F \circ \Phi$ for a function $F : \mathcal{M}_d \to \mathbb{R}$ with $|F(P) - F(Q)| \leq L\, M_w(P,Q)$ for all $P, Q \in \mathcal{M}_d$. The constant is 1 in both directions, for all configurations.

(ii) For all $P, Q \in \mathcal{M}_d$,

$$M_w(P,Q) \;=\; \sup\big\{\, f(z) - f(z') \;:\; f \in \mathcal{G}_1^w,\; \Phi(z) = P,\; \Phi(z') = Q \,\big\}.$$

Consequently, as for any metric and its Lipschitz ball, every metric on $\mathcal{M}_d$ for which the unit ball $\mathcal{G}_1^w$ is 1-Lipschitz dominates M_w, and M_w is the unique metric whose Lipschitz balls, composed with Φ, are the classes $\mathcal{G}_L^w$.

Part (i) is the representation: the class is the Lipschitz ball of $(\mathcal{M}_d, M_w)$ in the axioms' own constant, with nothing lost. Its proof is a routing: the axioms bound f along three moves, one value at fixed weights, one weight at fixed values, a permutation, so f is Lipschitz for the path metric the moves generate, and the theorem computes that metric in closed form as (4) by moving each matched value at the smaller of its two heights, before the weight move where the height rises and after it where it falls, at exactly the matching's cost (Appendix A.3; compare Balan et al., 2022; Cahill et al., 2024). Part (ii) is a corollary of (i), a metric being the supremum of the differences of its own 1-Lipschitz functions, and it is the sense in which the metric is the class's own, each of its ingredients one of the axioms (the price R of a unit of height is the modulus of Proposition A.4). The class is convex, closed under Lipschitz readouts of finitely many members, and contains the vector-valued linear sums, so attention pooling and the flocking resultant are members (Appendix A.3, Lemma A.5). On a fixed hard selection of k coordinates the class is the symmetric Lipschitz functions of the selected values with the constant L/k, the factor k attained by the selected maximum (Corollary A.15).

4 Complexity

A *design* is a Borel probability measure Q on $\mathcal{C}$, the law of the configurations, in the composite (1) the law of $(\pi(\omega), c(\omega))$, and its *lifted design* is $\Gamma := \Phi_{\#}Q$, a probability measure on $(\mathcal{M}_d, D)$. The data are $Z_1, \ldots, Z_n$ i.i.d. $\sim Q$, with lifts $P_i = \Phi(Z_i)$ and empirical lifted design $\hat{\Gamma}_n := n^{-1}\sum_{i \leq n} \delta_{P_i}$, and $\kappa := 2R$ bounds the diameter of $(\mathcal{M}_d, D)$ (Proposition A.20). The unit ball $\mathcal{G}_1$ is the set of $F \circ \Phi$ with F 1-Lipschitz on $(\mathcal{M}_d, D)$ (Theorem 3.2), so the generalization gap of the class on a sample is a transport,

$$\Delta_S \;:=\; \sup_{f \in \mathcal{G}_1} \Big(\mathbb{E}_Q f - \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \Big) \;=\; W_1(\hat{\Gamma}_n, \Gamma),$$

by Kantorovich–Rubinstein duality (Sriperumbudur et al., 2012). Every bound of this section bounds the mean $\Delta_n(Q) := \mathbb{E}\,\Delta_S$ or its value at the sample at hand, a statement about the design's geometry: the class itself is never covered.

4.1 Rates from the sparsity profile

Definition 4.1 (Tail mass and profile)**.** *For $(x, \mu) \in \mathcal{C}$ let $\mu_{(1)} \geq \mu_{(2)} \geq \cdots$ be the weights in decreasing order and, for $0 \leq k \leq d$, $T_k(x,\mu) := \sum_{i>k} \mu_{(i)}$ the weight beyond the k heaviest coordinates, a continuous function of the lift alone (Lemma B.9). The* profile *of the design is $G_k(t) := Q(T_k > t)$ for $0 \leq k \leq d$ and $t \in [0,1)$, nonincreasing in k and in t.*

Definition 4.2 (Sparsity profiles). *A design is* exactly k_0-sparse *if* $Q(T_{k_0} = 0) = 1$*, a selector that never reads more than* k_0 *coordinates. It has an* exponential profile *at rate* $\beta > 0$ *if* $\mathbb{E}_Q[T_k] \le Ae^{-\beta k}$ *for some* $A \ge 1$ *and all* $k \ge 0$*, a statement about the mean tail (Figure 1(a)).*

The profile decays in t at fixed k, the weights of one configuration, and in k at fixed t, the law of how many coordinates matter. The rates are stated through the *stratum exponent* $e(k)$, the covering exponent of the lifts carried by at most k atoms, for a value space of metric dimension $\dim \mathcal{V}$, $N(\varepsilon, \mathcal{V}, \rho) \le (C_{\mathcal{V}} R/\varepsilon)^{\dim \mathcal{V}}$. On interval values $e(k) = 2k - 1$, one value and one free weight per atom. On a *selector design*, one whose cropped weights are, at every crop level, an L_c-Lipschitz function of the selected values or are uniform on the support, the ball of Section 5.1 being one, $e(k) = k \dim \mathcal{V}$ (Appendix A.5). No upper bound below depends on d. The stratum of k atoms has covering count $N_k(\varepsilon) \le (C_1 R/\varepsilon)^{e(k)}$ at resolution $\varepsilon \le \kappa$, by balls of radius $\varepsilon/2$, with C_1 depending on the covering constant of $\mathcal{V}$, and on L_c on a selector design (Appendix A.5).

Theorem 4.3 (Generalization over the class). *Fix* $K \le d$ *and resolutions* $0 < \varepsilon_1 \le \cdots \le \varepsilon_K$ *with thresholds* $\theta_k := \varepsilon_k/(2\kappa) < 1$*, and let* $w_k := Q(T_k \le \theta_k,\ T_{k-1} > \theta_{k-1})$*, with* $\theta_0 := \theta_1$*, be the mass of the configurations that* k *atoms resolve at* ε_k *and* $k-1$ *do not at* ε_{k-1}*. Then for every design* Q*,*

$$\Delta_n(Q) \ \le\ 2\sum_{k=1}^{K} w_k\, \varepsilon_k \ +\ \frac{\kappa}{2}\sum_{k=1}^{K}\sqrt{\frac{w_k\, N_k(\varepsilon_k)}{n}} \ +\ \kappa\, G_K(\theta_K). \tag{6}$$

The three terms are each shell's resolution at its mass, its cell-frequency fluctuation at its count, and the trim beyond the last shell at the diameter, or twice its own tail if cropped (Theorem B.14). The two theorems below follow from it and a multiscale refinement of its coupling (Proposition B.18), each with its converse.

Theorem 4.4 (A single sparsity level). *For every design and every* $K \le d$ *with* $e(K) \ge 3$*,*

$$\Delta_n(Q) \ \le\ 2\kappa\, \mathbb{E}_Q T_K \ +\ CR\, n^{-\frac{1}{e(K)}},$$

with C *depending on the covering constant of* $\mathcal{V}$ *only, and on* L_c *on a selector design, uniformly in* K *and in* d *(the case* $e(K) \le 2$ *is in Appendix B.3): an exactly* k_0*-sparse design learns at* $n^{-1/e(k_0)}$ *once* $e(k_0) \ge 3$*, which is* $n^{-1/(2k_0-1)}$ *on interval values and* $n^{-1/(k_0 \dim \mathcal{V})}$ *on a selector design. Conversely, on interval values there is an exactly* k_0*-sparse design* Q_{k_0}*, with weights bounded below, such that every* n*-atom measure* ν_n *has* $W_1(\nu_n, \Gamma_{k_0}) \ge c_{k_0}\, n^{-1/(2k_0-1)}$*, for a constant* c_{k_0}*: on interval values the worst-case exponent of a sparsity level* $k_0 \ge 2$ *is* $1/(2k_0 - 1)$ *from both sides.*

The first term is the mean tail paid as a bias, the cost of cropping every configuration to K atoms, and the second is the rate of the exactly K-sparse design so obtained, by a multiscale coupling on the cropped stratum (Proposition B.18), whose exponent on interval values is the converse's (Appendix D.4), and on the converse's design no learner, from any number of noiseless labels, has risk below $c\, n^{-1/(2k_0-1)}$ uniformly over the class (Theorem D.5). The simplest members escape the rate: the conservative linear sums (7), the shape of a sum network, learn at $n^{-1/(\dim \mathcal{V}+2)}$ whatever the count and the profile (Proposition B.20), and the class pays the profile for everything beyond Lipschitz readouts of finitely many of them.

Theorem 4.5 (Exponential profiles). *Let* $\mathcal{V} = [-R, R]$ *or a circle of diameter* $2R$*. Under an exponential profile at rate* β*, for every* $n \ge 1$*,*

$$\Delta_n(Q) \ \le\ C\, e^{-\sqrt{\beta \ln n/2}},$$

and $\Delta_n(Q) \le C\, e^{-\sqrt{\beta \ln n}}$ *on a selector design, with* C *depending on* (β, A, R)*, and on* L_c *on a selector design. Conversely there is a design* Q^β *with* $T_k \le e^{-\beta k}$ *at every configuration such that, for every* n *with* $d \ge \sqrt{2\ln(2n)/\beta} + 1$*, every* n*-atom measure* ν_n *has* $W_1(\nu_n, \Gamma^\beta) \ge c\, e^{-\sqrt{2\beta \ln(2n)}}$*, with* c *depending on* (β, R)*.*

The rate is slower than every polynomial in n and faster than every power of $\ln n$, and it is the problem's own, the constant in the exponent lying between $1/\sqrt{2}$ and $\sqrt{2}$. The optimizing cutoff in Theorem 4.4 is $K_n \asymp \sqrt{\ln n/\beta}$, so the dimension the rate sees grows without bound, until it reaches d, and the profile costs nothing for being a statement about the mean tail, the crop paying each configuration its own tail (Remark B.25). The circle is the value space of the flock (Appendix C.1).

4.2 THE CERTIFICATE

Theorem 4.6 (The two-sample certificate)**.** *Let n be even and let S be split, by a rule fixed in advance, into halves A and B of $n/2$ points each, with empirical lifted designs $\hat{\Gamma}_A$ and $\hat{\Gamma}_B$. With probability at least $1-\delta$,*

$$W_1(\hat{\Gamma}_A, \Gamma) \;\le\; W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) + \kappa\sqrt{\frac{\ln(1/\delta)}{n}},$$

and $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) = \min_\pi \frac{2}{n}\sum_{i\le n/2} D(P_i, P'_{\pi(i)})$, the minimum over the bijections π between the lifted points of the two halves: an assignment problem on $n/2$ points, each of whose costs is itself an assignment between the atoms of two lifts (Lemma B.3).

What is ours is the object: on the class the maximum discrepancy (Bartlett et al., 2002) is a transport by duality (von Luxburg & Bousquet, 2004, eq. (11)), a matching computes it, and the symmetrization of Vapnik & Chervonenkis (1971) carries it to the population with one term (Theorem B.1, Remarks B.2, B.7 and B.8). The matching is the least of the certificates on the halves, every covering-type bound on them being one feasible transport (Proposition B.5): on the flock the cell certificate is twenty times the matching on the ball and trivial on every graded kernel (Section 5.1).

5 EXPERIMENTS

5.1 FLOCKING

The Vicsek flock of the introduction (Vicsek et al., 1995) is the reference model of collective motion (Vicsek & Zafeiris, 2012; Ginelli, 2016), and its update is a mechanism of the class on the ordered region (Appendix C.1): the configuration is the set of headings, the selector the ball, uniform on the $m+1$ particles within distance r of the focal one, itself included, and the *update* the direction of the mean heading vector, up to an angular noise of amplitude η. We simulate the standard setting, 3200 particles at density 2 with $r=1$, at noise $\eta = 1.2$ in the ordered phase (Chaté et al., 2008). Alongside the ball rule we simulate three flocks whose rules are graded, a bump, a Gaussian and a Cauchy kernel, each at the range at which its mean tail matches the ball's, at a finite cutoff, and each read through the selector of its own family with its lift cropped at a priced tail (Appendix C.1). A *sample* is one focal particle of one replica at one instant, one sample per replica, so the samples are independent: $64{,}000$ replicas of each flock, each started at random and read once at step 3000, after the flock has ordered, the view's headings rotated by an independent uniform angle: the design is the law of that view. The regime $\eta = 2.4$ and the kernels' parameters are in Appendix C.1.

The profile. Figure 1(a) is the mean tail of each selector, and Figure 2 in Appendix C.1 the count law of the ball flock. On the replicas the ball holds 11 particles on average and 67 at most, other than the self, from 2 to 23 between the 5% and 95% samples. Every sample is exactly sparse with uniform weights on its $m+1$ atoms, so the mean tail is a functional of the count law. It is geometric beyond the knee at the mean support because the count law is, at a decay rate of 0.23 per atom. The graded kernels are the other case, a decay inside each sample at a rate set by the local crowding, the Gaussian's geometric, the average-exponential profile of Theorem 4.5 on a design whose weights fade rather than stop, and the Cauchy's polynomial. The profile bounds give the rate and the certificate gives the number: the ball design's support reaches 68 atoms and a kernel's is every particle within its cutoff, so a bound in the dimension of the support says nothing at these n, where the certificate does. The rate the sample shows is the design's own: over the last three doublings the ball's transport decays with the exponent of a three-dimensional sample and the graded kernels' with that of thirteen to fifteen, rising by a dimension per doubling as the sample resolves more atoms of the weight vector, both far below the profile's worst-case exponent, eleven at the ball's mean count (Appendix C.1): the ordered flock's headings are nearly collinear, and the profile bound is over every design with that profile. That number bounds the gap of every Lipschitz function of the local configuration, on a design of strata of dimension one to sixty-eight, twelve on average (Figure 2), the 3200 headings of the flock's state nowhere in it.

The certificate. Figure 1(b) is Theorem 4.6 computed on the split by replicas, half A the first half of the replicas and half B the second, fixed in advance. The transport is one assignment on the

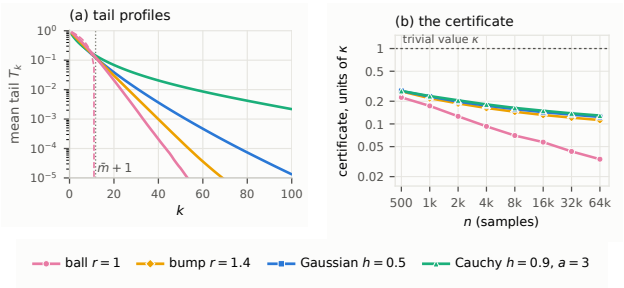

Figure 1: The flock at $\eta = 1.2$, 64,000 independent replicas per flock, one sample each. *(a)* The profile: the mean tail T_k, the weight beyond the k heaviest atoms, under the four selectors, the dashed line the tail of one sample at the mean count $\bar{m} + 1$. *(b)* The certificate of Theorem 4.6 for half A at $\delta = 0.05$, in units of κ, on samples after the burn-in of 3000 steps, by which the flock has ordered.

halves (Lemma B.3), 32,000 samples per half at the largest n. At $\eta = 1.2$ the certificate is 0.034κ on the ball and 0.11 to 0.13κ on the graded kernels, where the mean distance between two random samples is 0.61 to 0.68κ and $\kappa = \pi$: a component of the local mean velocity, of range 2, deviates by at most 0.11. Every step here is a theorem on the data: the samples are independent replicas, so the i.i.d. hypothesis of Theorem 4.6 holds as stated, and the concentration term is its 0.007κ at $n = 64{,}000$.

What the certificate bounds. The certificate is the exact uniform deviation of the class on the two halves, $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) = \max_{f \in \mathcal{G}_1}(\mathbb{E}_A f - \mathbb{E}_B f)$, so no bound uniform over the class on these halves is smaller, and by Remark B.22 it bounds the generalization gap of every empirical risk minimizer of $\mathcal{G}_L$ on every target in $\mathcal{G}_L$ by $2L$ times its value. On a member it is nearly attained: for the target $L\,D(\cdot, A)$, zero on half A, the lower Lipschitz extension $-L\,D(\cdot, A)$ is a member of $\mathcal{G}_L$ with zero training risk whose gap on half B is $2L$ times the mean nearest-neighbor distance from B to A, 0.038κ against the certified 0.068κ at $L = 1$ on the ball, and within a factor 1.13 of the certificate on every graded kernel (Appendix C.1). For one chosen learner on one target a holdout estimate on half B is tighter by definition, and the certificate does not compete with it: what it gives is uniformity at no labels, over every target on the design and every predictor in the class at once, in hand before a learner is chosen or a label seen, so that a choice made after seeing the data costs nothing (Theorem B.1(ii)). Against the covering certificates on the same halves it is the smaller by far: the cell certificate of Proposition B.5, optimized on the data, is 0.68κ on the ball, twenty times the matching, and at or above κ on every graded kernel at every n (Appendix C.1).

6 Conclusions and Future Work

We defined graded sparse mechanisms by axioms, showed that they are the Lipschitz ball of a transport metric on weighted configurations, and used that geometry for rates from the design's sparsity profile and for a certificate computed from the sample by a matching, with non-vacuous bounds on the flock. So although the class, the full Lipschitz ball of the lift, is prohibitively large for any data-independent bound, it is learnable, and the design alone can bound its rate.

We pose two questions. The first is the learning of the selector. Although in some situations it is dictated fully by the environment, and this is the setting we assume here, in other situations some aspects of it can be learned (Heras et al., 2019; Ha & Jeong, 2021; Kipf et al., 2018). What makes the problem interesting is its separation: the selector is the environment's, one rule common to all agents, while the mechanisms of the agents may differ. Second, empirically in this paper we concentrated on various rules on the flock, with the conservative cap. We believe the next natural step would be to evaluate additional scenarios, in terms of the caps, the observed sparsity profiles, and the implied empirical complexity.

AI USE STATEMENT

In this work the authors set the direction, the concepts and the outlines of the proof strategies of the new results, and generative AI tools executed under them. We used them to formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks or to provide critical ingredients for proving mathematical claims, and the generation of synthetic data sets and assistance with translation are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, suggest experimental parameters, create or edit software code, create artifacts, draft parts of the paper, summarize or analyse existing literature, edit the paper to improve readability, identify relevant literature, format references, brainstorm, and source or search for information. We have not used them to discover research topics or identify gaps, and the formulation of questions for surveys or interviews and the transcription of recordings are not applicable. We have reviewed all AI-assisted work: every mathematical statement and proof produced with AI assistance was checked line by line by the authors, every reference identified with AI assistance was read by the authors at its source before it was cited, and the code was reviewed by the authors and its outputs cross-checked against independent computations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every theorem, proposition and corollary of the main text is proved in the appendices, with its hypotheses stated in the statement itself: the representation theorem, the coverings of the strata and the crop cost in Appendix A, the generalization bounds in Appendix B, the lower bounds in Appendix D, and the membership of the flocking update, with its constants, in Appendix C.1. The measurements of Section 5.1 are on simulated replicas of the Vicsek model under four interaction rules, whose parameters, protocol, burn-in, split and seeds are recorded in Appendix C.1. The supplementary material contains the simulation and measurement code as three scripts run in order: the first regenerates every replica from its seed, $64{,}000$ per flock and regime, viewed after the burn-in, with every seed listed. The second computes from the replicas every table behind Figures 1 and 2 and the figures and the tables of Appendix C.1: the two-sample transports and certificates per flock, regime and sample size, the covering certificates on the same halves, the mean tail profiles and count laws, and the local polarization. The third draws the figures from the tables. The tables and the profiles are included, so that every figure is regenerated from them in seconds, and the paper's figures are included beside them for comparison.

REFERENCES

Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4(4): 259–264, 1984.

Tal Amir and Nadav Dym. Fourier sliced-Wasserstein embedding for multisets and measures. In *International Conference on Learning Representations (ICLR)*, 2025.

Tal Amir, Steven J. Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ofek Amran, Tom Gilat, and Ron Levie. A graphop analysis of graph neural networks on sparse graphs: Generalization and universal approximation. *arXiv preprint arXiv:2602.08785*, 2026.

Yair Ashlagi, Lee-Ad Gottlieb, and Aryeh Kontorovich. Functions with average smoothness: structure, algorithms, and learning. In *Conference on Learning Theory (COLT)*, 2021.

Radu Balan, Naveed Haghani, and Maneesh Singh. Permutation invariant representations with applications to graph deep learning, 2022. arXiv:2203.07546.

M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002. doi 10.1023/A:1013999503812.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

Peter J. Bickel and Bo Li. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, volume 54 of *IMS Lecture Notes–Monograph Series*, pp. 177–186. 2007.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under invariance and geometric stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.07148.

Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation using Wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37, 2022. arXiv:2106.04018.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539–563, 2014.

Jan Böker, Ron Levie, Ningyuan Huang, Soledad Villar, and Christopher Morris. Fine-grained expressivity of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03698.

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

Peter Bubenik and Alex Elchesen. Universality of persistence diagrams and the bottleneck and Wasserstein distances. *Computational Geometry*, 105–106:101882, 2022a. arXiv:1912.02563.

Peter Bubenik and Alex Elchesen. Virtual persistence diagrams, signed measures, Wasserstein distances, and Banach spaces. *Journal of Applied and Computational Topology*, 6:429–474, 2022b.

Peter Bubenik and Alexander Wagner. Embeddings of persistence diagrams into Hilbert spaces. *Journal of Applied and Computational Topology*, 4:339–351, 2020.

Jameson Cahill, Joseph W. Iverson, Dustin G. Mixon, and Daniel Packer. Group-invariant max filtering. *Foundations of Computational Mathematics*, 2024. arXiv:2205.14039.

Martin Carrasco, Caio F. Deberaldini Netto, Vahan A. Martirosyan, Ehimare Okoyomon, and Caterina Graziani. On the Rademacher complexity of graph neural networks: Unifying expressivity and geometry, 2025. arXiv:2510.10101, preprint.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2312.14820.

Saptarshi Chakraborty and Peter L. Bartlett. A statistical analysis for supervised deep learning with exponential families for intrinsically low-dimensional data, 2024. arXiv:2412.09779.

Saptarshi Chakraborty and Peter L. Bartlett. On the statistical properties of generative adversarial models for low intrinsic data dimension. *Journal of Machine Learning Research*, 26:1–57, 2025.

Hugues Chaté, Francesco Ginelli, Guillaume Grégoire, and Franck Raynaud. Collective motion of self-propelled particles interacting without cohesion. *Physical Review E*, 77:046113, 2008.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Ching-Yao Chuang and Stefanie Jegelka. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2210.01906.

Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. arXiv:2106.03314.

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have L_p-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007.

Yair Davidson and Nadav Dym. On the Hölder stability of multiset and graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2025.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *International Conference on Machine Learning (ICML)*, 2019.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: approximation by empirical measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49 (4):1183–1203, 2013.

Vincent Divol and Théo Lacombe. Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. *Journal of Applied and Computational Topology*, 5:1–53, 2021.

Richard M. Dudley. The speed of mean Glivenko–Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. arXiv:1703.11008.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, 2022.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139*, 2021. arXiv:2102.10333.

Pascal Mattia Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de Mathématiques Pures et Appliquées*, 94(2):107–130, 2010.

Massimo Fornasier, Pascal Heid, and Giacomo Enrico Sodini. Approximation theory, computing, and deep learning on the Wasserstein space, 2023. arXiv:2310.19548.

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. arXiv:1312.2128.

Vikas K. Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning (ICML)*, 2020.

Jacques Gautrais, Francesco Ginelli, Richard Fournier, Stéphane Blanco, Marc Soria, Hugues Chaté, and Guy Theraulaz. Deciphering interactions in moving animal groups. *PLoS Computational Biology*, 8(9):e1002678, 2012.

Francesco Ginelli. The physics of the Vicsek model. *The European Physical Journal Special Topics*, 225:2099–2117, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction. In *Algorithmic Learning Theory (ALT)*, 2013. arXiv:1302.2752; journal version in Theoretical Computer Science, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

Seungwoong Ha and Hawoong Jeong. Unraveling hidden interactions in complex systems with deep learning. *Scientific Reports*, 11:12804, 2021.

Leonid G. Hanin. Kantorovich–Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

Francisco J. H. Heras, Francisco Romero-Ferrero, Robert C. Hinz, and Gonzalo G. de Polavieja. Deep attention networks reveal the rules of collective motion in zebrafish. *PLoS Computational Biology*, 15(9):e1007354, 2019.

Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Andreas Krause, and Jonas Rothfuss. Instance-dependent generalization bounds via optimal transport. *Journal of Machine Learning Research*, 24:1–51, 2023. arXiv:2211.01258.

Dongzhou Huang, Tianyi Jiang, and Haonan Wang. Wasserstein convergence rates for empirical measures of point processes, 2026. arXiv:2604.23928.

Maximilian Hüttenrauch, Adrian Šošić, and Gerhard Neumann. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54):1–31, 2019.

Haotian Ju, Dongyue Li, Aneesh Sharma, and Hongyang R. Zhang. Generalization in graph neural networks: Improved PAC-Bayesian bounds on graph diffusion. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Leonid V. Kantorovich and Gennadii Sh. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad University*, 13(7):52–59, 1958.

Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness implies generalization via data-dependent generalization bounds. In *Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162*, 2022. arXiv:2206.13497.

Saman Kazemkhani, Aarav Pandya, Daphne Cornelisse, Brennan Shacklett, and Eugene Vinitsky. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In *International Conference on Learning Representations (ICLR)*, 2025.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning (ICML)*, 2018.

Andrey N. Kolmogorov and Vladimir M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. English translation: American Mathematical Society Translations (2) 17 (1961), 277–364.

Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. In Evarist Giné, David Mason, and Jon Wellner (eds.), *High Dimensional Probability II*, pp. 443–459. Birkhäuser, 2000.

Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011. arXiv:1110.4300.

Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020. arXiv:1804.10556.

Ron Levie. A graphon-signal analysis of graph neural networks. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. arXiv:2305.15987.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: learning, generalization, and sample complexity. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2302.06015.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A PAC-Bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Fei Lu, Mauro Maggioni, Sui Tang, and Ming Zhong. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116 (29):14424–14433, 2019. arXiv:1812.06003.

Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *Journal of Machine Learning Research*, 22(32):1–67, 2021. arXiv:1910.04832.

Shaogao Lv. Generalization bounds for graph convolutional neural networks via Rademacher complexity. *arXiv preprint arXiv:2102.10234*, 2021.

Ian G. Macdonald. *Symmetric Functions and Hall Polynomials*. Oxford University Press, 2nd edition, 1995.

Yishay Mansour and David McAllester. Generalization bounds for decision trees. In *Proceedings of the 13th Annual Conference on Computational Learning Theory (COLT)*, 2000.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sohir Maskey, Gitta Kutyniok, and Ron Levie. Generalization bounds for message passing networks on mixture of graphons, 2024. arXiv:2404.03473.

Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. WL meet VC. In *International Conference on Machine Learning (ICML)*, 2023.

Christopher Morris, Fabrizio Frasca, Nadav Dym, Haggai Maron, İsmail İlkan Ceylan, Ron Levie, Derek Lim, Michael Bronstein, Martin Grohe, and Stefanie Jegelka. Position: Future directions in the theory of graph machine learning. In *International Conference on Machine Learning (ICML)*, 2024.

Alireza Mousavi-Hosseini, Clayton Sanford, Denny Wu, and Murat A. Erdogdu. When do transformers outperform feedforward and recurrent networks? A statistical perspective, 2025. arXiv:2503.11272.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, 2023.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1902.04742.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning (ICML)*, 2023.

Siddharth Nayak, Kenneth Choi, Wenqi Ding, Sydney Dolan, Karthik Gopalakrishnan, and Hamsa Balakrishnan. Scalable multi-agent reinforcement learning through intelligent information aggregation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Giannis Nikolentzos and Konstantinos Skianis. On the Lipschitz continuity of set aggregation functions and neural networks for sets. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.24403.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021.

Tomáš Pevný and Vojtěch Kovařík. Approximation capability of neural networks on spaces of probability measures and tree-structured domains. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1906.00764.

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

Huyên Pham and Xavier Warin. Mean-field neural networks: learning mappings on Wasserstein space. *Neural Networks*, 2023. arXiv:2210.15179.

Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

Benedetto Piccoli and Francesco Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

Barnabás Póczos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013. arXiv:1302.0082.

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

Levi Rauchwerger, Stefanie Jegelka, and Ron Levie. Generalization, expressivity, and universality of graph neural networks on attributed graphs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.05464.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 10535 of *Lecture Notes in Computer Science*, pp. 737–753, 2017.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 8459–8468, 2020.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. Universal approximations of permutation invariant/equivariant functions by deep neural networks. arXiv:1903.01939, 2019.

Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *PMLR*, pp. 771–780, 2021.

Dominic Schuhmacher and Aihua Xia. A new metric between distributions of point processes. *Advances in Applied Probability*, 40(3):651–672, 2008. arXiv:0708.2777.

Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE Transactions on Signal Processing*, 56(8):3447–3457, 2008.

Nimrod Segol and Yaron Lipman. On universal equivariant set networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Justin Solomon, Kristjan Greenewald, and Haikady N. Nagaraja. k-variance: A clustered notion of variance. *SIAM Journal on Mathematics of Data Science*, 4(3):957–978, 2022. arXiv:2012.06958.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. DOI 10.1214/12-EJS722.

Marina Struleva, Shayan Hundrieser, Dominic Schuhmacher, and Axel Munk. Sharp convergence rates of empirical unbalanced optimal transport for spatio-temporal point processes. *arXiv preprint arXiv:2509.04225*, 2025.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17:1–40, 2016. arXiv:1411.2066.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.14269.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning (ICML)*, 2023.

Ekaterina Tolstaya, Fernando Gama, James Paulos, George Pappas, Vijay Kumar, and Alejandro Ribeiro. Learning decentralized controllers for robot swarms with graph neural networks. In *Conference on Robot Learning (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, 2020. arXiv:1903.10527.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Gábor Vásárhelyi, Csaba Virágh, Gergő Somorjai, Tamás Nepusz, Agoston E. Eiben, and Tamás Vicsek. Optimized flocking of autonomous drones in confined environments. *Science Robotics*, 3(20):eaat3536, 2018.

Antonis Vasileiou, Ben Finkelshtein, Floris Geerts, Ron Levie, and Christopher Morris. Covered forest: Fine-grained generalization analysis of graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML), PMLR 267*, 2025a. arXiv:2412.07106.

Antonis Vasileiou, Stefanie Jegelka, Ron Levie, and Christopher Morris. Survey on generalization theory for graph neural networks. *arXiv preprint arXiv:2503.15650*, 2025b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tamás Vicsek and Anna Zafeiris. Collective motion. *Physics Reports*, 517(3–4):71–140, 2012.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, 1995.

Eugene Vinitsky, Nathan Lichtlé, Xiaomeng Yang, Brandon Amos, and Jakob Foerster. Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*, 2022.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention, 2020. arXiv:2007.02876.

Edward Wagstaff, Fabian B. Fuchs, Martin Engelcke, Michael A. Osborne, and Ingmar Posner. Universal approximation of functions on sets. *Journal of Machine Learning Research*, 23(151): 1–56, 2022.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2406.06893.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. arXiv:1707.00087.

Hermann Weyl. *The Classical Groups: Their Invariants and Representations*. Princeton University Press, 1939.

Trenton D. Wirth, Gregory C. Dachner, Kevin W. Rio, and William H. Warren. Is the neighborhood of interaction in human crowds metric, topological, or visual? *PNAS Nexus*, 2(5):pgad118, 2023.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Rentian Yao, Xiaohui Chen, and Yun Yang. Mean-field nonparametric estimation of interacting particle systems. In *Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, 2022. arXiv:2205.07937.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Songyuan Zhang, Oswin So, Kunal Garg, and Chuchu Fan. GCBF+: A neural graph control barrier function framework for distributed safe multiagent control. *IEEE Transactions on Robotics*, 41: 1533–1552, 2025.

Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. arXiv:2111.05529.

Aaron Zweig and Joan Bruna. Exponential separations in symmetric neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A Proofs for Section 3

A.1 Basic lemmas, and the weight axiom

Two remarks on Definition 3.1 come first.

Remark A.1 (Differential form)**.** *On an interval $\mathcal{V} = [-R, R]$ and for f differentiable, (V) is the bound $|\partial f/\partial x_i| \le L\, w_i(\mu)$ at every configuration, and for the weight $w = \mu$, (W) is the bound $\max_i \partial f/\partial \mu_i - \min_i \partial f/\partial \mu_i \le 2LR$ along the simplex, since a tangent direction h of the simplex has $\sum_i h_i = 0$ and $\sup_{\|h\|_1 \le 1} \langle \nabla_\mu f, h\rangle$ is half the spread of the partial derivatives.*

Remark A.2 (Distinguished inputs)**.** *A mechanism may read, besides the anonymous coordinates, inputs of its own that no permutation moves: the state of the agent that evaluates it, or of its environment, as in Section 5. Such an input is one distinguished variable q in a metric space, and the class becomes the set of $f(q;\, x, \mu)$ that lie in $\mathcal{G}_L^w$ for every q and are Lipschitz in q. Everything in the paper holds fiberwise in q, the metric of Section 3.2 becomes a product metric, and the covering numbers of Appendix A.5 multiply by that of the space of q: with the metric $d_q(q, q') + M_w(P, P')$ the class is its L-Lipschitz ball, the covering number at ε is at most the product of the two covering numbers at $\varepsilon/2$, and the diameter of Theorem 4.6 is the sum of the two. We keep the notation without the slot.*

Lemma A.3 (Blindness, factorization, telescope)**.** *Let $f \in \mathcal{G}_L^w$.*

(i) If $\mu_i = 0$ then $f(x, \mu)$ does not depend on x_i.

(ii) If the multisets $\{(x_i, \mu_i) : \mu_i > 0\}$ and $\{(y_j, \nu_j) : \nu_j > 0\}$ coincide, then $f(x, \mu) = f(y, \nu)$.

(iii) For all $x, y \in \mathcal{V}^d$ and μ: $|f(x, \mu) - f(y, \mu)| \le L \sum_i w_i(\mu)\, \rho(x_i, y_i)$.

Proof. (i) is (V) with $w_i(\mu) = 0$. (ii) Choose a permutation σ matching the positive-weight coordinates of (y, ν) to those of (x, μ) pairwise. Then $(\sigma y, \sigma\nu)$ and (x, μ) agree on every positive-weight coordinate and on the weight vector, and differ at most in values at zero-weight coordinates, which are free by (i). Conclude by (A). (iii) Change the coordinates one at a time at fixed μ. Each step is a (V) move and every intermediate point is a configuration. □

Where the weight axiom comes from. The principle, one mechanism applied to every weighted coordinate, has a formula: a *score* $\varphi : \mathcal{V} \to \mathbb{R}$ applied to every value and summed with the cap weights. Let $\mathcal{S} := \{\varphi : \mathcal{V} \to \mathbb{R} \ : \ \mathrm{Lip}(\varphi) \le 1,\ |\varphi| \le R\}$ be the score ball, which contains a translate of every 1-Lipschitz score because $R = \frac{1}{2}\mathrm{diam}\mathcal{V}$, and let

$$\mathcal{L}^w \ := \ \Big\{ f_\varphi(x,\mu) = \sum_{i=1}^{d} w_i(\mu)\,\varphi(x_i) \ : \ \varphi \in \mathcal{S} \Big\} \tag{7}$$

be the *linear sums* of the cap weight. They contain the weighted means and value aggregations $\sum_i \mu_i\varphi(x_i)$ at $w = \mu$, the unnormalized sums $\sum_i(\mu_i/\mu_{\max})\varphi(x_i)$ at $w = \mu/\mu_{\max}$ and, with $\varphi \equiv R$, the count $R/\mu_{\max}$. Every linear sum satisfies (V) with $L = 1$ and (A) whatever the weight axiom (Lemma A.5), and the class is meant to contain them at constant 1. The weight axiom is the tightest one that does.

Proposition A.4 (The modulus of the linear sums)**.** *For all $\mu, \nu \in \Delta_d$,*

$$\sup_{x\in\mathcal{V}^d,\ \varphi\in\mathcal{S}} |\, f_\varphi(x,\mu) - f_\varphi(x,\nu)\,| \ = \ R\,\|w(\mu) - w(\nu)\|_1\,.$$

So (W) says exactly that f is no more sensitive to a change of weights than the linear sums are, uniformly over the values. Among weight axioms $|f(x,\mu) - f(x,\nu)| \le L\,c(\mu,\nu)$ with a modulus c that does not look at x, it is the tightest one under which every linear sum keeps constant 1, because any such c dominates the supremum above. In this sense $\mathcal{G}^w_L$ is the smallest class cut out by (V), (A) and a value-free weight axiom that contains the linear sums in its unit ball. A modulus that does look at x is a transport distance between the weighted value measures, and the class it cuts out is smaller and excludes the selection readouts of the next subsection: two coordinates at the same value R with equal shares, and a share δ shifted from one to the other, leave the weighted value measure unchanged while the selected maximum moves by δR, so no constant serves, where (W) prices the move at $2R\delta$ (Remark A.6). Weights are individuals and not mass, and this is why the metric of Section 3.2 is a matching distance on multisets and not a distance between measures. What (W) excludes is the rank readouts: the weighted median and quantiles, trimmed means and top-k selections jump when a weight crosses a threshold, two atoms at $\pm R$ with weights $\frac{1}{2} \pm \delta$ moving the weighted median by $2R$ as δ changes sign, so no constant serves; the selected maximum, which is weighted and not ranked, stays in.

Lemma A.5 (Linear sums are members)**.** *Every $f_\varphi \in \mathcal{L}^w$ satisfies (V) with $L = 1$ and (A). More generally, let $\varphi : \mathcal{V} \to \mathbb{R}^m$ satisfy $|\varphi(s) - \varphi(t)|_2 \le \rho(s,t)$ and $|\varphi|_2 \le R$, and put $f_\varphi(x,\mu) := \sum_i w_i(\mu)\,\varphi(x_i)$. Then for all configurations*

$$|\, f_\varphi(x,\mu) - f_\varphi(y,\nu)\,|_2 \ \le \ M_w\big(\Phi(x,\mu), \Phi(y,\nu)\big),$$

so every coordinate of f_φ, and every 1-Lipschitz function of f_φ, lies in $\mathcal{G}^w_1$.

Proof. A change of x_i alone changes f_φ by $w_i(\mu)|\varphi(x_i) - \varphi(y_i)| \le w_i(\mu)\rho(x_i,y_i)$, and anonymity is the equivariance of w. For the joint bound write $P = \Phi(x,\mu)$, $Q = \Phi(y,\nu)$, so that $f_\varphi(x,\mu) = \sum_{(s,u)\in P} u\,\varphi(s)$, and let σ be any partial matching. A matched pair $(s,u) \mapsto (t,v)$ contributes $|u\varphi(s) - v\varphi(t)|_2 \le \min(u,v)|\varphi(s) - \varphi(t)|_2 + |u - v|\sup|\varphi|_2 \le \varrho((s,u),(t,v))$, and an unmatched atom contributes $|u\varphi(s)|_2 \le Ru = \varrho((s,u),o)$. Summing gives the cost of σ, and minimizing over σ gives $M_w(P,Q)$. The last claim is Theorem 3.2(i), or directly: a coordinate of f_φ satisfies (V) and (A) as above and (W) because $|\sum_i(w_i(\mu) - w_i(\nu))\varphi(x_i)|_2 \le R\,\|w(\mu) - w(\nu)\|_1$. □

Proof of Proposition A.4. Upper bound: $|\sum_i(w_i(\mu) - w_i(\nu))\varphi(x_i)| \le \|w(\mu) - w(\nu)\|_1 \sup|\varphi| \le R\,\|w(\mu) - w(\nu)\|_1$. Lower bound: by compactness pick $a, b \in \mathcal{V}$ with $\rho(a,b) = 2R$, and let $\varphi(z) := R - \rho(z,a)$, which is 1-Lipschitz with $\varphi(a) = R$, $\varphi(b) = -R$ and $|\varphi| \le R$, so $\varphi \in \mathcal{S}$. Put $x_i = a$ where $w_i(\mu) > w_i(\nu)$ and $x_i = b$ where $w_i(\mu) < w_i(\nu)$, anywhere else otherwise. Then $\sum_i(w_i(\mu) - w_i(\nu))\varphi(x_i) = R\sum_i|w_i(\mu) - w_i(\nu)|$. □

Remark A.6 (The pointwise alternative)**.** *One could instead take the modulus at the given x, $c_x(\mu,\nu) := \sup_{\varphi\in\mathcal{S}}|f_\varphi(x,\mu) - f_\varphi(x,\nu)|$, which by Kantorovich–Rubinstein duality on $\mathcal{V}$ is a transport distance between the weighted value measures $\sum_i w_i(\mu)\delta_{x_i}$ and $\sum_i w_i(\nu)\delta_{x_i}$: the balanced Kantorovich distance for the conservative weight and the flat distance for the accumulative one.*

The resulting axiom is stronger and its class smaller, and under it a reweighting among coordinates at equal values is free. That is the right axiom when the members are functionals of the weighted value measure, averages and attention pooling, and the wrong one for the selection readouts. Take $\mathcal{V} = [-R, R]$ *and* $f = \max_i s_i x_i$*, with a leader at value* 0 *and share* 1 *and two coordinates at value* R *with equal shares* s*, and shift a share* δ *from one of the two to the other. The weighted value measure does not change at all, while* f *changes by* δR*, so no constant serves. Under (W) the move is priced at* $2R\delta$ *and the maximum has constant* 1 *(Proposition A.14).*

A.2 THE CONE AND THE MATCHING DISTANCE

Lemma A.7 (The cone is a compact metric space)**.** *ϱ is a metric on $\mathcal{K}$, $(\mathcal{K}, \varrho)$ is compact, and the distance to the apex is R times the height.*

Proof. ϱ is well defined on the quotient, since $\varrho((x,u),(y,0)) = Ru$ does not depend on y, it is symmetric, and it vanishes only on the diagonal: for $u, v > 0$, $\varrho = 0$ forces $u = v$ and then $\rho(x, y) = 0$. Triangle inequality: let $p = (x, u)$, $q = (z, t)$, $r = (y, v)$ with positive heights. We show $\varrho(p, r) \le \varrho(p, q) + \varrho(q, r)$ by the position of t. If $t \ge u \vee v$, the right side is $u\rho(x, z) + v\rho(z, y) + R(2t - u - v) \ge (u \wedge v)\rho(x, y) + R|u - v|$, since $2t - u - v \ge |u - v|$. If $t \le u \wedge v$, the right side is $t(\rho(x, z) + \rho(z, y)) + R(u + v - 2t) \ge t\rho(x, y) + R(u + v - 2t)$, and subtracting $\varrho(p, r) = (u \wedge v)\rho(x, y) + R(u \vee v - u \wedge v)$ leaves $(u \wedge v - t)(2R - \rho(x, y)) \ge 0$ because $\rho \le 2R$. If $u \le t \le v$, the other order being symmetric, the right side is $u\rho(x, z) + t\rho(z, y) + R(v - u) \ge u(\rho(x, z) + \rho(z, y)) + R(v - u) \ge \varrho(p, r)$. With the apex in the middle, $\varrho(p, o) + \varrho(o, r) = R(u + v) \ge (u \wedge v)\rho(x, y) + R|u - v|$ because $(u \wedge v)\rho(x, y) \le 2R(u \wedge v)$, and with the apex as an endpoint $\varrho(o, r) = Rv \le Rt + R|t - v| \le \varrho(o, q) + \varrho(q, r)$. Compactness: $\varrho((x, u), (y, v)) \le \rho(x, y) + R|u - v|$ and $\varrho((x, u), o) = Ru$, so the quotient map from the compact $\mathcal{V} \times [0, 1]$ is continuous into $(\mathcal{K}, \varrho)$ and onto. □

Lemma A.8 (Factorization)**.** *$\Phi(x, \mu) = \Phi(y, \nu)$ if and only if the multisets $\{(x_i, \mu_i) : \mu_i > 0\}$ and $\{(y_j, \nu_j) : \nu_j > 0\}$ coincide. Consequently every $f \in \mathcal{G}_L^w$ factors as $f = F \circ \Phi$ for a unique $F : \mathcal{M}_d \to \mathbb{R}$, and Φ is a bijection from the quotient of $\mathcal{C}$ by permutations and by values at zero-weight coordinates onto $\mathcal{M}_d$.*

Proof. If the multisets of (x_i, μ_i) coincide, so do those of $(x_i, w_i(\mu))$, by equivariance. Conversely, if the multisets of pairs (value, cap weight) coincide, there is a permutation π of $[d]$ carrying $\mathrm{supp}\nu$ onto $\mathrm{supp}\mu$ with $(y_{\pi^{-1}(i)}, w_{\pi^{-1}(i)}(\nu)) = (x_i, w_i(\mu))$ for $i \in \mathrm{supp}\mu$, so πy agrees with x on $\mathrm{supp}\mu$ and $w(\pi\nu) = w(\mu)$ by equivariance. Then $\pi\nu = \mu$ by injectivity. The factorization is Lemma A.3(ii). □

Lemma A.9 (Metric, measure form, duality)**.** *(i) M_w is a metric on $\mathcal{M}_d$.*

(ii) $M_w(P, Q) = W_1^{\varrho}\big(P + n_Q\,\delta_o,\ Q + n_P\,\delta_o\big)$, the Kantorovich distance with ground cost ϱ between the two apex-padded measures, both of total mass $n_P + n_Q$.

(iii) With $\mathrm{Lip}_0(\mathcal{K}) := \{\psi : \mathcal{K} \to \mathbb{R} \ :\ \mathrm{Lip}_{\varrho}(\psi) \le 1,\ \psi(o) = 0\}$,

$$M_w(P, Q) \ = \ \sup_{\psi \in \mathrm{Lip}_0(\mathcal{K})} \Big[\sum_{(x,u) \in P} \psi(x, u) - \sum_{(y,v) \in Q} \psi(y, v) \Big],$$

and every $\psi \in \mathrm{Lip}_0(\mathcal{K})$ satisfies $|\psi(x, u)| \le Ru$.

Proof. (i) Symmetry is clear. If $M_w(P, Q) = 0$ the optimal σ has zero cost, so no atom is unmatched, an unmatched atom costing $Ru > 0$, and every matched pair coincides, hence $P = Q$. Triangle inequality: given partial matchings σ_1 from P to Q and σ_2 from Q to S, define σ from P to S by $\sigma(p) = \sigma_2(\sigma_1(p))$ whenever both are defined, and leave every other atom unmatched. By the triangle inequality of ϱ with the apex included (Lemma A.7): a p matched through q to s costs $\varrho(p, s) \le \varrho(p, q) + \varrho(q, s)$, a p matched to a q that σ_2 leaves unmatched costs $\varrho(p, o) \le \varrho(p, q) + \varrho(q, o)$, an s matched by σ_2 to a q that σ_1 leaves unmatched costs $\varrho(s, o) \le \varrho(q, s) + \varrho(q, o)$, and atoms unmatched in σ_1 or σ_2 keep their apex cost. Each term of $\mathrm{cost}(\sigma_1) + \mathrm{cost}(\sigma_2)$ is used at most once, so $\mathrm{cost}(\sigma) \le \mathrm{cost}(\sigma_1) + \mathrm{cost}(\sigma_2)$. (ii) The Kantorovich problem between the padded measures is a transportation linear program between two

counting measures of the same mass, $n_P + n_Q$ unit atoms on each side, so by Birkhoff's theorem its polytope is the convex hull of the permutation matrices (Peyré & Cuturi, 2019, Prop. 2.1) and an optimal integer plan exists, a bijection between the atoms of the two sides. A bijection matches some atoms of P to atoms of Q, the remaining atoms of each side to apex copies, and apex copies to each other at cost 0, and its cost is the cost of the partial matching it induces. Conversely every partial matching extends to such a bijection, the counts of unmatched atoms and spare apex copies agreeing on both sides. Hence the two minima coincide. (iii) Kantorovich–Rubinstein duality on the compact metric space $(\mathcal{K}, \varrho)$, for two finite measures of equal mass: $W_1(\alpha, \beta) = \sup\{\int \psi \, d(\alpha - \beta) : \mathrm{Lip}_\varrho \psi \le 1\}$. Here $\alpha - \beta = P - Q + (n_Q - n_P)\delta_o$ has total mass 0, so $\psi \mapsto \psi - \psi(o)$ does not change the integral and one may restrict to $\psi(o) = 0$, which kills the apex term. The bound is $|\psi(x,u)| = |\psi(x,u) - \psi(o)| \le \varrho((x,u),o) = Ru$. □

Remark A.10 (The dual ball). *For $\varphi \in \mathcal{S}$ the function $\psi_\varphi(x,u) := u\,\varphi(x)$ lies in $\mathrm{Lip}_0(\mathcal{K})$ and $\int \psi_\varphi \, d\Phi(x,\mu) = f_\varphi(x,\mu)$, so the linear sums are the sub-ball of scores that read the height linearly. The dual ball is larger: it contains scores that read the height nonlinearly, such as $\psi(x,u) = \frac{R}{2}u^2$, whose functional is $\frac{R}{2}\sum_i w_i(\mu)^2$. This is where the members beyond the linear sums live: the supremum over x in Proposition A.4 enlarged the score ball from $\{u\varphi(x)\}$ to all Lipschitz functions on the cone vanishing at the apex, and M_w is the dual of the enlarged ball.*

Lemma A.11 (Elementary moves). *Let $(x,\mu) \in \mathcal{C}$.*

(i) (value move, exact) If y differs from x in coordinate i only, then $M_w\big(\Phi(x,\mu), \Phi(y,\mu)\big) = w_i(\mu)\,\rho(x_i, y_i)$.

(ii) (weight move) For every ν, $M_w\big(\Phi(x,\mu), \Phi(x,\nu)\big) \le R\,\|w(\mu) - w(\nu)\|_1$, with strict inequality possible: two atoms at the same value exchanging heights are at distance 0 while $\|w(\mu) - w(\nu)\|_1 > 0$.

(iii) (re-representation) $\Phi(\sigma x, \sigma\mu) = \Phi(x,\mu)$, and changing x_i with $\mu_i = 0$ does not change Φ.

Proof. (i) The identity matching gives $\le \varrho((x_i,w_i),(y_i,w_i)) = w_i\rho(x_i,y_i)$. For $\ge$, take $\psi(z,u) := u\,(R - \rho(z,x_i))$, which is ψ_φ for $\varphi = R - \rho(\cdot, x_i) \in \mathcal{S}$, hence in $\mathrm{Lip}_0(\mathcal{K})$ by Remark A.10. The two lifts differ in one atom and $\psi(x_i,w_i) - \psi(y_i,w_i) = w_i\rho(x_i,y_i)$, so Lemma A.9(iii) gives the lower bound. (ii) Match coordinate i to itself where both weights are positive, at cost $R|w_i(\mu) - w_i(\nu)|$ with no value motion, and send an atom with a zero partner to the apex, at cost Rw_i, again $R|w_i(\mu) - w_i(\nu)|$. The example: $\{(t,u),(t,v)\}$ and $\{(t,v),(t,u)\}$ are the same multiset. (iii) is the definition of the lift. □

Lemma A.12 (Continuity, compactness, diameter). *(i) For all $(x,\mu),(y,\nu) \in \mathcal{C}$,*

$$M_w\big(\Phi(x,\mu),\Phi(y,\nu)\big) \;\le\; |w(\mu)| \max_i \rho(x_i,y_i) + R\,\|w(\mu) - w(\nu)\|_1\,,$$

where $|w(\mu)| := \sum_i w_i(\mu)$. In particular $\Phi : \mathcal{C} \to (\mathcal{M}_d, M_w)$ is continuous.

(ii) $(\mathcal{M}_d, M_w)$ is compact and Φ is a quotient map: a function on $\mathcal{M}_d$ is continuous if and only if its composition with Φ is. In particular the tail masses $T_k(\mu) = \sum_{i>k}\mu_{(i)}$, with the weights sorted decreasingly, and $\mu_{\max}$ are continuous on $(\mathcal{M}_d, M_w)$.

(iii) $M_w(P,Q) \le R\,(|P| + |Q|)$ with $|P| := \int u\,dP(x,u)$ the total height, hence $\mathrm{diam}(\mathcal{M}_d, M_w) \le 2R\sup_\mu |w(\mu)|$.

Proof. (i) Go through $\Phi(y,\mu)$: the first leg costs at most $\sum_i w_i(\mu)\rho(x_i,y_i)$ by the identity matching, the second at most $R\,\|w(\mu) - w(\nu)\|_1$ by Lemma A.11(ii). Continuity follows with the continuity of w. (ii) $\mathcal{C}$ is compact and Φ continuous and onto, so $(\mathcal{M}_d, M_w)$ is compact. A continuous map from a compact space onto a Hausdorff space is closed, hence a quotient map. T_k and $\mu_{\max}$ are continuous on $\mathcal{C}$ and constant on the fibers of Φ. (iii) The empty matching sends every atom to the apex. □

A.3 Proof of Theorem 3.2

Proof of Theorem 3.2(i), the easy direction. If $|F(P)-F(Q)| \le L\, M_w(P,Q)$ for all P, Q and $f = F \circ \Phi$, then (V), (W) and (A) are Lemma A.11(i), (ii), (iii) respectively, read through the Lipschitz bound. □

Proof of Theorem 3.2(i), the routing direction. Let $f \in \mathcal{G}_L^w$. By Lemma A.8, $f = F \circ \Phi$ for a unique F, and it remains to show that for all $(x,\mu), (y,\nu) \in \mathcal{C}$,

$$|\, f(x,\mu) - f(y,\nu)\,| \;\le\; L \cdot M_w\big(\Phi(x,\mu),\, \Phi(y,\nu)\big).$$

Write $P = \Phi(x,\mu)$, $Q = \Phi(y,\nu)$, $S = \mathrm{supp}\mu$, $T = \mathrm{supp}\nu$, and let σ be an optimal partial matching, viewed at the level of coordinates as a bijection from $S' \subseteq S$ onto $T' \subseteq T$. We may take σ *maximal*: if $i \in S$ and $j \in T$ were both unmatched, adding the pair (i,j) would change the cost by $\varrho((x_i, w_i(\mu)), (y_j, w_j(\nu))) - \varrho((x_i, w_i(\mu)), o) - \varrho((y_j, w_j(\nu)), o) \le 0$, by the triangle inequality through the apex. So either $S' = S$ or $T' = T$.

Relabeling. Define an injection t of T into $[d]$ by $t(\sigma(i)) := i$ for $i \in S'$, and, if $T \setminus T' \neq \emptyset$ (so $S' = S$), by sending $T \setminus T'$ injectively into $[d] \setminus S$, which has $d - |S| \ge |T| - |T'|$ elements. Let π be a permutation of $[d]$ with $\pi(j) = t(j)$ for $j \in T$, and put $\nu' := \pi\nu$, $y' := \pi y$. Then $\nu'_{t(j)} = \nu_j$, $y'_{t(j)} = y_j$, ν' vanishes off $t(T)$, and by equivariance $w_{t(j)}(\nu') = w_j(\nu)$. By (A), $f(y',\nu') = f(y,\nu)$, and $\Phi(y',\nu') = Q$. Every coordinate $i \in [d]$ is now of one of four kinds: *matched* ($i \in S'$, hosting $(x_i, w_i(\mu))$ and $(y'_i, w_i(\nu')) = (y_{\sigma(i)}, w_{\sigma(i)}(\nu))$), *leaving* ($i \in S \setminus S'$, weight $w_i(\mu) > 0 = w_i(\nu')$), *arriving* ($i \in t(T \setminus T')$, weight $0 = w_i(\mu) < w_i(\nu')$), or *empty*. By maximality, leaving and arriving coordinates do not both occur.

The chain. Split the matched coordinates into $I^{\uparrow} := \{i \in S' : w_i(\mu) \le w_i(\nu')\}$ and $I^{\downarrow} := S' \setminus I^{\uparrow}$. Define $z \in \mathcal{V}^d$ by $z_i := y'_i$ for $i \in I^{\uparrow}$ and for arriving i, and $z_i := x_i$ otherwise. Then:

- Step 1. $(x,\mu) \to (z,\mu)$: value moves at weights μ on $I^{\uparrow}$ and on the arriving coordinates. By Lemma A.3(iii), $|f(x,\mu) - f(z,\mu)| \le L \sum_{i \in I^{\uparrow}} w_i(\mu)\, \rho(x_i, y'_i)$, the arriving coordinates having $w_i(\mu) = 0$ and contributing nothing.
- Step 2. $(z,\mu) \to (z,\nu')$: one weight move. By (W), $|f(z,\mu) - f(z,\nu')| \le LR\, \|w(\mu) - w(\nu')\|_1$.
- Step 3. $(z,\nu') \to (y',\nu')$: value moves at weights ν' on $I^{\downarrow}$ (on the leaving coordinates y' may be taken equal to x, since their ν'-weight is zero and the lift does not see them). By Lemma A.3(iii), $|f(z,\nu') - f(y',\nu')| \le L \sum_{i \in I^{\downarrow}} w_i(\nu')\, \rho(x_i, y'_i)$.

On $I^{\uparrow}$ the weight in Step 1 is $w_i(\mu) = \min(w_i(\mu), w_i(\nu'))$, and on $I^{\downarrow}$ the weight in Step 3 is $w_i(\nu') = \min(w_i(\mu), w_i(\nu'))$, so the value costs of Steps 1 and 3 add up to $\sum_{i \in S'} \min\big(w_i(\mu), w_{\sigma(i)}(\nu)\big)\, \rho(x_i, y_{\sigma(i)})$, while the ℓ^1 sum in Step 2 is $\sum_{i \in S'} |w_i(\mu) - w_{\sigma(i)}(\nu)|$ plus the weights $w_i(\mu)$ of the leaving coordinates plus the weights $w_{t^{-1}(i)}(\nu)$ of the arriving ones. Adding the three steps,

$$|f(x,\mu) - f(y,\nu)| \;\le\; L\Big[\sum_{i \in S'} \varrho\big((x_i, w_i(\mu)), (y_{\sigma(i)}, w_{\sigma(i)}(\nu))\big) + \sum_{\text{unmatched}} R\, w\Big] \;=\; L\, \mathrm{cost}(\sigma),$$

which is $L\, M_w(P,Q)$ by the optimality of σ. □

Proof of Theorem 3.2(ii). The inequality $\le$ is the routing direction. For $\ge$ take $F := M_w(\cdot, Q)$, which is 1-Lipschitz for M_w, so $F \circ \Phi \in \mathcal{G}_1^w$ by the easy direction, and $F(P) - F(Q) = M_w(P,Q)$. If a metric d on $\mathcal{M}_d$ makes every F with $F \circ \Phi \in \mathcal{G}_1^w$ 1-Lipschitz, then $M_w(P,Q) = \sup_f [f(z) - f(z')] \le d(P,Q)$. And if $\mathrm{Lip}_L(d) \circ \Phi = \mathcal{G}_L^w$ for all L, the same argument with d in place of M_w gives $d(P,Q) = \sup_f [f(z) - f(z')] = M_w(P,Q)$. □

In the language of persistence diagrams, part (ii) is the universality of the Wasserstein distance on the metric pair $(\mathcal{K}, \varrho, \{o\})$: Bubenik & Elchesen (2022a, Cor. 1.4 and §5.3) prove that on any metric space with a distinguished subset the p-Wasserstein distance on finite multisets is the largest

p-subadditive metric compatible with the ground metric, and that at $p = 1$ it satisfies Kantorovich–Rubinstein duality, which is Lemma A.9(iii), and Bubenik & Elchesen (2022b) carry the construction to signed measures and their Banach spaces. What is specific here is the pair, the cone with its apex, and that the class comes first.

The closure properties named after Theorem 3.2 are those of a Lipschitz ball: the class is convex, and if $f_1, \dots, f_m$ are members with constants $L_1, \dots, L_m$ and $h : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz then $h(f_1, \dots, f_m)$ is a member with constant $\sup_y \sum_a |\partial_a h(y)|\, L_a$, since $h(F_1, \dots, F_m)$ is Lipschitz for M_w with that constant. The linear sums extend to the vector-valued case, $\sum_i w_i(\mu)\varphi(x_i)$ with $\varphi : \mathcal{V} \to \mathbb{R}^m$ 1-Lipschitz and bounded by R in norm, which is 1-Lipschitz from $(\mathcal{M}_d, M_w)$ into $\mathbb{R}^m$ jointly (Lemma A.5), so that attention pooling and the flocking resultant are members by post-composition.

A.4 The two instances

Three levels, and the sum embeddings. Three nested sets order the class: the linear sums (7), of which a weighted Deep Sets network is a Lipschitz readout of finitely many (Remark B.19), the functionals of the lift, the dual ball of the matching distance (Lemma A.9, Remark A.10), and the class of Theorem 3.2, every Lipschitz function of the lift. The conservative linear sums learn at the rate of the value space alone, with no count and no profile (Proposition B.20), while the class sees the law of the lift and pays the profile (Section 4).

The two classes of Section 3.1 are the following.

Definition A.13 (The two classes)**.** *The* conservative *class $\mathcal{G}_L := \mathcal{G}_L^{w^{\mathrm{c}}}$ is the class of Definition 3.1 for the cap weight $w^{\mathrm{c}}(\mu) = \mu$, the weights themselves, and the* accumulative *class $\widetilde{\mathcal{G}}_L := \mathcal{G}_L^{w^{\mathrm{a}}}$ is the class for the cap weight $w^{\mathrm{a}}(\mu) = \mu/\mu_{\max}$, the vector of* shares *$s(\mu)$, the leader at share* 1.

Table 1 sorts the common aggregations by what they do to influence. The same aggregations are debated in graph networks on expressivity grounds, the sum separating multisets that the mean and the maximum identify (Xu et al., 2019) and a degree-dependent scaler interpolating between the two (Corso et al., 2020). In both classes the weight bound sums over all d coordinates, and $\|s(\mu) - s(\nu)\|_1$ can reach $d - 1$: that is the count of Table 1, a member at constant 1, and the size is intended.

mechanism	form	class	L
weighted mean	$h\big(\sum_i \mu_i x_i\big)$	conservative	$\|h'\|_\infty$
value aggregation	$\sum_i \mu_i\, \varphi(x_i)$, $\varphi \in \mathcal{S}$	conservative	1
attention pooling	$h\big(\sum_i \mu_i\, v(x_i)\big)$, $\mathrm{Lip} v \le 1$, $\|v\| \le R$	conservative	$\mathrm{Lip}(h)$
top weight	$R\,\mu_{\max}$	conservative	1
flocking update	direction of $\sum_i \mu_i\, e^{\mathrm{i}\theta_i}$	conservative	$1/\varpi_0$ at polarization $\varpi \ge \varpi_0$
count	$R/\mu_{\max} = R\sum_i s_i$	accumulative	1
unnormalized sum	$\sum_i s_i\, \varphi(x_i)$, $\varphi \in \mathcal{S}$	accumulative	1
force sum	$\sum_i s_i\, F(x_i)$, $\mathrm{Lip} F \le 1$, $\|F\| \le R$	accumulative	1
selected maximum	$\max_i s_i\, x_i$	accumulative	1
weighted mean*	$\sum_i \mu_i x_i$	accumulative	2

Table 1: Common aggregations, the class they belong to and the constant. Values in $[-R, R]$ where a value bound is used. The starred row is the mechanism of the first row read against the share cap. The flocking row is Appendix C.1, the rest is Proposition A.14. No constant depends on d or on the number of selected coordinates.

Proposition A.14 (Membership)**.** *Each mechanism of Table 1 belongs to the class shown with the constant shown. Moreover $\mathcal{G}_L \subseteq \widetilde{\mathcal{G}}_{2L}$.*

Proof of Proposition A.14. Each (V) line is the differential form of Remark A.1 or a direct two-point estimate, and each (W) line is the displayed one-line estimate. *Conservative rows*, on $\mathcal{V} = [-R, R]$. Weighted means $h(\langle \mu, x\rangle)$: (V) $|h'|\mu_i| x_i - y_i|$, (W) $|h'|\,|\langle \mu - \nu, x\rangle| \le \|h'\|_\infty R\, \|\mu - \nu\|_1$. Value aggregations: Lemma A.5. Attention pooling: the vector-valued linear sum of Lemma A.5 followed by the Lipschitz map h. The top weight $R\,\mu_{\max}$: (V) 0, (W) $R|\mu_{\max} - \nu_{\max}| \le R\, \|\mu - \nu\|_\infty$.

The flocking update is Appendix C.1. *Accumulative rows*, on $\mathcal{V} = [-R, R]$, with $s = s(\mu)$, $s' = s(\nu)$ and $|s| := \|s\|_1$. The count $R\sum_i s_i$: (V) 0, (W) $R\,||s| - |s'|| \le R\,\|s - s'\|_1$. The sums and the force sums: Lemma A.5 with the share weight. The selected maximum $\max_i s_i x_i$, on values in $[0, R]$ so that the coordinates off the support, at $s_i x_i = 0$, never attain it: (V) $|s_i x_i - s_i y_i| \le s_i |x_i - y_i|$, (W) $|\max_i s_i x_i - \max_i s'_i x_i| \le \max_i |s_i - s'_i|\,|x_i| \le R\,\|s - s'\|_\infty$. The weighted mean $\langle \mu, x\rangle = \langle s, x\rangle/|s|$: (V) $\mu_i \le s_i$, and for (W), $\langle \mu, x\rangle - \langle \nu, x\rangle = \langle s - s', x\rangle/|s| + \langle s', x\rangle\big(1/|s| - 1/|s'|\big)$ with $|\langle s', x\rangle| \le R|s'|$ and $|1/|s| - 1/|s'|| \le \|s - s'\|_1/(|s||s'|)$, so the difference is at most $2R\,\|s - s'\|_1/|s| \le 2R\,\|s - s'\|_1$ because $|s| \ge 1$. *Containment.* $\mu_i \le s_i$ gives the accumulative (V) from the conservative one, and for the weight axioms $\|\mu - \nu\|_1 \le 2\,\mu_{\max}\,\|s - s'\|_1 \le 2\,\|s - s'\|_1$: indeed $\mu = s/|s|$, so $\mu - \nu = (s - s')/|s| + s'(1/|s| - 1/|s'|)$, of ℓ^1 norm at most $(\|s - s'\|_1 + ||s'| - |s||)/|s| \le 2\,\|s - s'\|_1/|s|$, and $1/|s| = \mu_{\max}$. Hence $\mathcal{G}_L \subseteq \widetilde{\mathcal{G}}_{2L}$, and by Theorem 3.2(ii) also $D \le 2\,\widetilde{D}$ on $\mathcal{M}_d$. □

The corollary read in Section 3.2 is the following. Reading a sum as conservative inflates its constant by the count, the price of the semantics.

Corollary A.15 (Fibers: the two semantics made exact)**.** *Let μ_S be uniform on $S \subseteq [d]$ with $|S| = k$. On the fiber $\{\mu = \mu_S\}$ the accumulative distance is the matching distance between the two k-point multisets of values,*

$$\widetilde{D}\big(\Phi(x, \mu_S), \Phi(y, \mu_S)\big) \;=\; \min_{\tau} \sum_{i \in S} \rho\big(x_i, y_{\tau(i)}\big)$$

over the bijections τ of S, and the restriction of $\widetilde{\mathcal{G}}_L$ to the fiber is exactly the set of symmetric functions g of x_S with per-coordinate constant L, that is $|g(x_S) - g(x_S^{i \to y_i})| \le L\,\rho(x_i, y_i)$ with $x^{i \to y_i}$ the vector x with coordinate i replaced by y_i, every such g extending to a member of $\widetilde{\mathcal{G}}_L$. The restriction of $\mathcal{G}_L$ has per-coordinate constant L/k, and the factor k is attained by the selected maximum.

Proof of Corollary A.15. On the fiber all shares equal 1, so a matched pair costs $\varrho((x_i, 1), (y_j, 1)) = \rho(x_i, y_j) \le 2R$ while an unmatched atom costs R. Leaving a pair unmatched on both sides costs $2R \ge \rho$, so a full matching is optimal and the formula follows. A symmetric g with per-coordinate constant L satisfies $|g(x_S) - g(y_S)| \le L\sum_i \rho(x_i, y_{\tau(i)})$ for every τ, by symmetry and then the telescope of Lemma A.3(iii) with all weights 1, hence is L-Lipschitz for $\widetilde{D}$ on the fiber. Extend it to $\mathcal{M}_d$ with the same constant by McShane's extension and apply the easy direction of Theorem 3.2(i). Conversely the accumulative (V) at $s_i = 1$ gives the per-coordinate constant L, and the conservative (V) at $\mu_i = 1/k$ gives L/k. The selected maximum has accumulative constant 1 (Proposition A.14) and, on the fiber, changes by $|x_i - y_i|$ under a move of the leading coordinate, so its conservative constant is at least k. □

A.5 Strata, coverings, and the crop cost

These are the inputs of Section 4: for each number of atoms, a covering of the lifts carried by that many atoms, the cost of cropping a configuration to its heaviest atoms, and the diameter of the space.

Definition A.16 (Strata and cells)**.** *For $k \le d$, $\mathcal{A}_k := \{P \in \mathcal{M}_d : n_P \le k\}$, the lifts of configurations with at most k participants. For the accumulative instance and $m \ge 0$, $\mathcal{A}_k^{[m]} := \{P \in \mathcal{A}_k : |P| \le 2^m\} = \mathcal{A}_k \cap \{\mu_{\max} \ge 2^{-m}\}$. Since $|P| \le n_P \le k$ on $\mathcal{A}_k$, only $2^m < k$ gives a proper subset. Covering numbers $N(\varepsilon, S, M_w)$ use the radius convention.*

Theorem A.17 (Covering of the strata, conservative)**.** *For every $k \le d$ and $\varepsilon > 0$,*

$$N\big(\varepsilon, \mathcal{A}_k, D\big) \;\le\; N\big(\tfrac{\varepsilon}{2}, \mathcal{V}, \rho\big)^k \cdot N\big(\tfrac{\varepsilon}{2R}, \Delta_k, \ell^1\big) \;\le\; N\big(\tfrac{\varepsilon}{2}, \mathcal{V}, \rho\big)^k \Big(1 + \frac{16R}{\varepsilon}\Big)^{k-1}.$$

On $\mathcal{V} = [-R, R]$ and $\varepsilon \le R$: $\log N(\varepsilon, \mathcal{A}_k, D) \le (2k - 1)\log(17R/\varepsilon)$.

Proof. Parametrize: for $v \in \mathcal{V}^k$ and $a \in \Delta_k$ let $\Psi(v, a) := \Phi(z)$ where z carries values v and weights a on the first k coordinates and weight 0 elsewhere. By (A) every $P \in \mathcal{A}_k$ is some $\Psi(v, a)$.

Matching coordinate j to coordinate j, with the apex where one side has $a_j = 0$,

$$D\big(\Psi(v,a),\Psi(v',a')\big) \;\le\; \sum_{j=1}^{k}\Big[\min(a_j,a'_j)\,\rho(v_j,v'_j) + R|a_j - a'_j|\Big] \\ \le\; \max_j \rho(v_j,v'_j) + R\,\|a-a'\|_1\,,$$

using $\sum_j \min(a_j,a'_j) \le 1$. Take an $\varepsilon/2$-net of $\mathcal{V}$ in each value slot and an $\varepsilon/(2R)$-net of (Δ_k,ℓ^1). Each term is then at most $\varepsilon/2$. For the simplex count: dropping the last coordinate maps Δ_k into the unit ℓ^1-ball of $\mathbb{R}^{k-1}$ with $\|a-a'\|_1 \le 2\,\|b-b'\|_1$. A $\delta/4$-net of that ball has at most $(1+8/\delta)^{k-1}$ points, a ball in its own norm, and re-centering the net at points of the image gives a $\delta/2$-net of the image and hence a δ-net of Δ_k. With $\delta = \varepsilon/(2R)$ this is $(1+16R/\varepsilon)^{k-1}$. On the interval, $N(\varepsilon/2,[-R,R]) \le 4R/\varepsilon + 1 \le 5R/\varepsilon$ for $\varepsilon \le R$, and $k\log(5R/\varepsilon) + (k-1)\log(17R/\varepsilon) \le (2k-1)\log(17R/\varepsilon)$. □

Theorem A.18 (Covering of the cells, accumulative). *For every $k \le d$, $m \ge 0$ and $\varepsilon > 0$,*

$$N\big(\varepsilon,\mathcal{A}_k^{[m]},\widetilde{D}\big) \;\le\; N\big(2^{-m-1}\varepsilon,\mathcal{V},\rho\big)^k \cdot k\Big(1+\frac{2^{m+3}R}{\varepsilon}\Big)^{k-1},$$

with net points in $\mathcal{A}_k$. Equivalently, at the cell's own scale, $N\big(2^{m-1}\varepsilon,\mathcal{A}_k^{[m]},\widetilde{D}\big) \le N(\varepsilon/4,\mathcal{V},\rho)^k \cdot k(1+16R/\varepsilon)^{k-1}$: the count of a cell resolved at $2^{m-1}\varepsilon$ does not depend on m. On $\mathcal{V} = [-R,R]$ and $\varepsilon \le R$: $\log N(\varepsilon,\mathcal{A}_k^{[m]},\widetilde{D}) \le (2k-1)\log(2^{m+5}R/\varepsilon)$, and $2^m \le 2k$ suffices on $\mathcal{A}_k$.

Proof. Parametrize by the share vector: let $\Sigma_k := \{s \in [0,1]^k : \max_j s_j = 1\}$ and, for $v \in \mathcal{V}^k$, $s \in \Sigma_k$, let $\widetilde{\Psi}(v,s) := \Phi(z)$ with z carrying values v and weights $s/\|s\|_1$ on the first k coordinates, so that z has share vector s there. Every $P \in \mathcal{A}_k^{[m]}$ is $\widetilde{\Psi}(v,s)$ with $\|s\|_1 = |P| \le 2^m$. Matching coordinatewise,

$$\widetilde{D}\big(\widetilde{\Psi}(v,s),\widetilde{\Psi}(v',s')\big) \;\le\; \sum_j \min(s_j,s'_j)\,\rho(v_j,v'_j) + R\,\|s-s'\|_1 \\ \le\; 2^m \max_j \rho(v_j,v'_j) + R\,\|s-s'\|_1\,.$$

Take a $2^{-m-1}\varepsilon$-net of $\mathcal{V}$ in each value slot and an $\varepsilon/(2R)$-net of the share vectors of the cell in ℓ^1. Each term is then at most $\varepsilon/2$. For the share count: the share vectors of the cell lie in the union over $j \le k$ of the faces $\{s_j = 1\}$, and on the face $\{s_j = 1\}$ the remaining coordinates form a nonnegative vector of ℓ^1-norm at most $2^m - 1 < 2^m$, a point of the ℓ^1-ball of radius 2^m in $\mathbb{R}^{k-1}$. A $\delta/2$-net of that ball has at most $(1+2^{m+2}/\delta)^{k-1}$ points, and re-centering at points of the face gives a δ-net of the face inside the face. With $\delta = \varepsilon/(2R)$ this is $k(1+2^{m+3}R/\varepsilon)^{k-1}$. The net points $\widetilde{\Psi}(v',s')$ are lifts in $\mathcal{A}_k$. The second display is the first with $2^{m-1}\varepsilon$ in place of ε. On the interval, $N(2^{-m-1}\varepsilon,[-R,R]) \le 2^{m+2}R/\varepsilon + 1 \le 2^{m+3}R/\varepsilon$, and $k\log(2^{m+3}R/\varepsilon) + \log k + (k-1)\log(2^{m+4}R/\varepsilon) \le (2k-1)\log(2^{m+5}R/\varepsilon)$, using $\log k \le (2k-1)\log 2$. Finally $|P| \le k$ on $\mathcal{A}_k$, so $\mathcal{A}_k^{[m]} = \mathcal{A}_k$ once $2^m \ge k$. □

Both bounds have exponent $k\dim_\varepsilon(\mathcal{V}) + (k-1)$, which is $2k-1$ on intervals, with the ambient d absent and the cell factor 2^m inside the logarithms only, and not at all once the cell is resolved at its own scale. On a *selector design*, the law of $(\pi(\omega),c(\omega))$ for a selector c, the weights are functions of the values and the simplex directions are not there to cover. If the cropped weights depend on the k selected values x_S through a map that is L_c-Lipschitz from $(\mathcal{V}^S,\max_i\rho)$ to (Δ_k,ℓ^1), or are uniform on the support, as for a hard selector, then by Lemma A.12(i) the map $x_S \mapsto \Phi(x_S,\hat{\mu}(x_S))$ is $(1+RL_c)$-Lipschitz into $(\mathcal{M}_d,D)$, and the cropped support of the stratum is covered by $N(\varepsilon/(2(1+RL_c)),\mathcal{V},\rho)^k$ points: exponent $k\dim_\varepsilon(\mathcal{V})$. Both counts are used in Section 4.

Proposition A.19 (Crop cost). *Let $\hat{\mu}$ be a k-crop of μ for $k \ge 1$, k of its largest weights kept and renormalized, by a fixed measurable rule where weights tie (the cost below is the same for every choice), and $T_k = T_k(\mu)$ the weight beyond them. Then*

$$D\big(\Phi(x,\mu),\Phi(x,\hat{\mu})\big) \;\le\; 2RT_k, \qquad \widetilde{D}\big(\Phi(x,\mu),\Phi(x,\hat{\mu})\big) \;\le\; \frac{RT_k}{\mu_{\max}(\mu)}.$$

Proof. Match each kept atom to itself and send the tail to the apex. Conservative: a kept atom moves in height from μ_i to $\mu_i/(1-T_k)$, in total $R\sum_{\text{kept}}\mu_i T_k/(1-T_k) = R\,T_k$, and the tail costs $R\,T_k$. Accumulative: $\hat{\mu}_{\max} = \mu_{\max}/(1-T_k)$, so the shares of kept atoms are unchanged and they cost nothing, and the tail costs $R\sum_{\text{tail}} s_i = R\,T_k/\mu_{\max}$. □

Proposition A.20 (Qualitative inputs)**.** *For both instances $(\mathcal{M}_d, M_w)$ is compact, T_k and $\mu_{\max}$ are continuous on it, the shells $\{T_k \le \theta < T_{k-1}\}$ and the cells $\{\mu_{\max} \ge 2^{-m}\}$ are Borel, and Kantorovich–Rubinstein duality holds for $W_1(M_w)$ on probability measures on $\mathcal{M}_d$. Diameters:* $\operatorname{diam}(\mathcal{M}_d, D) \le 2R$, *on $\{|P| \le 2^m\}$ one has $\widetilde{D}(P,Q) \le 2^{m+1}R$, and globally* $\operatorname{diam}(\mathcal{M}_d, \widetilde{D}) \le 2Rd$.

Proof. Lemma A.12, and duality on a compact metric space is standard (Peyré & Cuturi, 2019). □

B Proofs for Section 4

Inputs: strata, crop and counts. The stratum $\mathcal{A}_k \subseteq \mathcal{M}_d$ of lifts carried by at most k atoms, the crop of a configuration to its k heaviest coordinates and the covering numbers of the strata are the inputs, all from Appendix A.5. A configuration with tail T_k lies within κT_k of its crop, a point of $\mathcal{A}_k$ (Proposition A.19), and, when the value space has metric dimension $\dim\mathcal{V}$, $N(\varepsilon,\mathcal{V},\rho) \le (C_{\mathcal{V}}R/\varepsilon)^{\dim\mathcal{V}}$, the strata are covered at

$$N_k(\varepsilon) := N\big(\tfrac{\varepsilon}{2}, \mathcal{A}_k, D\big) \;\le\; \Big(\frac{C_1 R}{\varepsilon}\Big)^{e(k)} \qquad (\varepsilon \le \kappa),$$

with C_1 depending on $C_{\mathcal{V}}$ only and $e(k)$ the *stratum exponent*, one copy of the value space per atom and one simplex direction per free weight: $e(k) = k\dim\mathcal{V} + k - 1$ in general, which is $2k-1$ on interval values with $C_1 = 34$, and $e(k) = k\dim\mathcal{V}$ on a selector design whose cropped weights are Lipschitz in the selected values or constant given the support, since its weights are not there to cover (Theorem A.17, and the paragraph after Theorem A.18 for selector designs). The ball of Section 5.1 is a selector design, its weights being uniform on the support, and the graded kernels read there are not, since their weights are set by positions that the mechanism does not read (Definition C.4). Nothing in these counts depends on d.

B.1 The two-sample certificate

Theorem B.1 (The two-sample certificate, in full)**.** *Let n be even and let S be split, by a rule fixed in advance, into halves A and B of $n/2$ points each, with empirical lifted designs $\hat{\Gamma}_A$ and $\hat{\Gamma}_B$.*

(i) $\mathbb{E}\,W_1(\hat{\Gamma}_n,\Gamma) \le \mathbb{E}\,W_1(\hat{\Gamma}_A,\Gamma) \le \mathbb{E}\,W_1(\hat{\Gamma}_A,\hat{\Gamma}_B) \le 2\,\mathbb{E}\,W_1(\hat{\Gamma}_A,\Gamma)$.

(ii) With probability at least $1-\delta$,

$$W_1(\hat{\Gamma}_A,\Gamma) \;\le\; W_1(\hat{\Gamma}_A,\hat{\Gamma}_B) + \kappa\sqrt{\frac{\ln(1/\delta)}{n}},$$

and with probability at least $1-\delta$,

$$W_1(\hat{\Gamma}_n,\Gamma) \;\le\; W_1(\hat{\Gamma}_A,\hat{\Gamma}_B) + \kappa\sqrt{\frac{\ln(2/\delta)}{n}}.$$

(iii) $W_1(\hat{\Gamma}_A,\hat{\Gamma}_B) = \min_\pi \frac{2}{n}\sum_{i\le n/2} D(P_i, P'_{\pi(i)})$, *the minimum over the bijections π between the lifted points P_i of A and P'_j of B: an assignment problem on $n/2$ points, each of whose costs is itself an assignment between the atoms of two lifts (Lemma B.3).*

Proof of Theorem B.1, which contains Theorem 4.6. (i). $\hat{\Gamma}_n = \frac{1}{2}(\hat{\Gamma}_A + \hat{\Gamma}_B)$ and $W_1(\cdot,\Gamma)$ is convex, so $W_1(\hat{\Gamma}_n,\Gamma) \le \frac{1}{2}W_1(\hat{\Gamma}_A,\Gamma) + \frac{1}{2}W_1(\hat{\Gamma}_B,\Gamma)$, whose expectation is $\mathbb{E}\,W_1(\hat{\Gamma}_A,\Gamma)$, the halves being exchangeable. Next, by duality $W_1(\mu,\nu) = \sup_{F\in\mathrm{Lip}_1}(\int F\,d\mu - \int F\,d\nu)$ is a supremum of functionals affine in ν, and $\int F\,d\Gamma = \mathbb{E}_B\int F\,d\hat{\Gamma}_B$ for every F, so $W_1(\hat{\Gamma}_A,\Gamma) =$

$\sup_F \mathbb{E}_B[\int F\, d\hat{\Gamma}_A - \int F\, d\hat{\Gamma}_B] \le \mathbb{E}_B[W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) \mid A]$; take expectations over A. The last inequality is the triangle inequality through Γ and exchangeability. *(ii).* If ν and ν' are empirical measures on m points differing in one point then $W_1(\nu, \nu') \le \kappa/m$, one atom of mass $1/m$ moved by at most the diameter κ (Proposition A.20). So $B \mapsto W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ has bounded differences $2\kappa/n$ over the $n/2$ points of B, and McDiarmid's inequality (Boucheron et al., 2013, Thm. 6.2) bounds its conditional mean, which dominates $W_1(\hat{\Gamma}_A, \Gamma)$ by the argument of (i), by its value plus $\kappa\sqrt{\ln(1/\delta)/n}$ with probability $1 - \delta$: the first display. For the second, apply the first display at $\delta/2$ to A and, with the roles of the halves exchanged, to B: on the intersection, of probability at least $1 - \delta$, both $W_1(\hat{\Gamma}_A, \Gamma)$ and $W_1(\hat{\Gamma}_B, \Gamma)$ are at most $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) + \kappa\sqrt{\ln(2/\delta)/n}$, and $W_1(\hat{\Gamma}_n, \Gamma) \le \frac{1}{2}W_1(\hat{\Gamma}_A, \Gamma) + \frac{1}{2}W_1(\hat{\Gamma}_B, \Gamma)$ by the convexity used in (i). *(iii).* Transport between two uniform measures on $n/2$ atoms each is a linear program over doubly stochastic matrices, whose extreme points are the permutation matrices (Birkhoff), so the minimum is attained at a bijection. □

Remark B.2 (One split suffices, and averaging over splits)**.** *(a) The split must be a rule fixed in advance or drawn independently of the data, and the second display of Theorem B.1(ii) certifies the learner on the whole sample from the one observed split at the price of $\delta/2$ in place of δ, a factor $\sqrt{\ln(2/\delta)/\ln(1/\delta)}$, 1.11 at $\delta = 0.05$. (b) For a fixed sample, the average of $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ over splits drawn independently of the data has, by exchangeability, the same expectation, and McDiarmid over all n points with differences $2\kappa/n$ puts the expected deviation of a random half, $\mathbb{E}\, W_1(\hat{\Gamma}_{n/2}, \Gamma)$, within $\kappa\sqrt{2\ln(1/\delta)/n}$ above it. It does not certify the realized half A: conditionally on A the expectation of a mixed split need not dominate $W_1(\hat{\Gamma}_A, \Gamma)$, so the argument of (ii) does not apply to it, and Section 5.1 uses the one observed split. (c) Section 5.1 reports the certificate of the learner on half A and the supplementary material that of the learner on the whole sample.*

Lemma B.3 (D between two lifts is a rectangular assignment)**.** *Let P carry atoms (x_i, u_i), $i \le s$, and P' atoms (y_j, v_j), $j \le s'$, with $s \le s'$. Then*

$$D(P, P') \;=\; R\sum_{j \le s'} v_j \;+\; \min_{\pi} \sum_{i \le s} c_{i\pi(i)}, \qquad c_{ij} := \min(u_i, v_j)\, \rho(x_i, y_j) + R|u_i - v_j| - Rv_j,$$

the minimum over injections $\pi : [s] \to [s']$. With uniform weights $u_i = 1/s$, $v_j = 1/s'$, $D(P, P') = M/s' + \kappa(1 - s/s')$ with M the cost of a minimum-cost injection of the s values into the s' under ρ.

Proof. By (5), $D(P, P')$ is the Kantorovich distance with the cone cost between the two apex-padded counting measures, and a transport between counting measures of unit atoms is a bijection between the padded atom sets (Birkhoff). An atom matched to an apex point costs R times its height, and a matched real pair costs $\varrho = \min(u, v)\rho + R|u - v| \le R(u + v)$ since $\rho \le 2R$. So if an atom of P and an atom of P' are both matched to apex points, matching them to each other and the two apex points to each other, at cost 0, does not increase the cost, and iterating, some optimal bijection matches every atom of the smaller lift to a real atom. Its cost is $\sum_{\text{matched}} \varrho_{i\pi(i)} + R\sum_{\text{unmatched } j} v_j = \sum_i (\varrho_{i\pi(i)} - Rv_{\pi(i)}) + R\sum_j v_j$. With uniform weights $c_{ij} = \rho_{ij}/s' + R/s - 2R/s'$, and summing over the s matched atoms gives $M/s' + R - 2Rs/s'$, which with $R\sum_j v_j = R$ is $M/s' + 2R(1 - s/s')$. □

The bound is stated for a *resolution profile*, a nondecreasing sequence $0 < \varepsilon_1 \le \cdots \le \varepsilon_K$, $K \le d$, one resolution per number of resolved atoms, with thresholds $\theta_k := \varepsilon_k/(2\kappa) < 1$. It is nondecreasing so that the sets of Definition B.4 nest: a configuration that needs more atoms is resolved more coarsely, never more finely.

Definition B.4 (Shells)**.** *For $\theta \in [0, 1)$ and $0 \le k \le d$ let $B_k(\theta) := \{P \in \mathcal{M}_d : T_k(P) \le \theta\}$, the lifts within $\kappa\theta$ of the stratum $\mathcal{A}_k$, closed sets that increase with k, with $B_0(\theta) = \emptyset$ and $B_d(\theta) = \mathcal{M}_d$. For a resolution profile the sets $B_k(\theta_k)$ are nested, since $T_k \le T_{k-1}$ and $\theta_{k-1} \le \theta_k$, and the* shells

$$\mathsf{Sh}_k := B_k(\theta_k) \setminus B_{k-1}(\theta_{k-1}), \qquad 1 \le k \le K, \quad \theta_0 := \theta_1,$$

partition $B_K(\theta_K)$. Their masses are increments of the profile,

$$w_k := \Gamma(\mathsf{Sh}_k) = G_{k-1}(\theta_{k-1}) - G_k(\theta_k) \;\ge\; 0 \qquad (G_0 \equiv 1),$$

and the mass beyond the last shell, $\Gamma(B_K(\theta_K)^c) = G_K(\theta_K)$, is the trim.

Shell k holds the configurations that k atoms resolve to within $\varepsilon_k/2$ and $k-1$ atoms do not resolve to within $\varepsilon_{k-1}/2$: the sparse configurations sit in the low shells and are resolved finely, the crowded ones in the high shells and coarsely, each shell at its own mass. A *celling* of the shells partitions each Sh_k into cells, each contained in a closed D-ball of radius ε_k (Definition B.10). One celling is the theorem's own: the crop of a configuration of shell k to its k heaviest coordinates is a point of the stratum $\mathcal{A}_k$ within $\varepsilon_k/2$ of it (Proposition A.19), so the pullback of an $\varepsilon_k/2$-net of $\mathcal{A}_k$ cells the shell into at most $N_k(\varepsilon_k)$ cells (Lemma B.12). Another, used in the computations, quantizes the crop on a grid and reads the cell off the point. Both are determined by nothing random, and the shell and the cell of a sample point are computable since T_k is.

Proposition B.5 (The shells bound the transport)**.** *For every resolution profile and every celling of its shells, both of which may be chosen after seeing the sample, write $\hat{w}_k^A$, $\hat{w}_k^B$ for the fractions of the two halves in shell k and $\hat{\mu}_j^A$, $\hat{\mu}_j^B$ for their fractions in cell j, each point beyond the last shell a cell of its own. Then*

$$W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) \;\le\; \sum_{k=1}^{K} \big(\hat{w}_k^A + \hat{w}_k^B\big)\varepsilon_k \;+\; \frac{\kappa}{2} \sum_j \big|\hat{\mu}_j^A - \hat{\mu}_j^B\big|.$$

The inequality is deterministic, one feasible transport between the two halves: move each point to its cell's center, couple the common mass within each cell for free, and the residuals at the diameter. The second term is the observed discrepancy of the two halves' cell frequencies, the points beyond the last shell included as cells of their own at $2/n$ each, a number of the sample, and its expectation under i.i.d. halves is what the fluctuation term of the next theorem bounds. With Theorem 4.6 the right-hand side is a certificate, with the profile and the celling free to optimize on the data.

Proof of Proposition B.5. Move each point of A to the center of its cell, at cost at most $\sum_k \hat{w}_k^A \varepsilon_k$ in W_1, and likewise B; a point beyond the last shell is its own center and moves at no cost. Between the two measures on the centers, couple mass $\min(\hat{\mu}_j^A, \hat{\mu}_j^B)$ within each center at cost 0 and the residuals arbitrarily at cost at most κ, whose total mass is $\frac{1}{2}\sum_j |\hat{\mu}_j^A - \hat{\mu}_j^B|$. The triangle inequality adds the three. □

Proposition B.6 (The expectation of the sample bound)**.** *For every resolution profile, with the theorem's own celling of the shells, at most $N_k(\varepsilon_k)$ cells of radius ε_k on shell k, fixed before sampling,*

$$\mathbb{E}\, W_1(\hat{\Gamma}_A, \hat{\Gamma}_B) \;\le\; 2\sum_k w_k \varepsilon_k + \kappa \sum_k \sqrt{\frac{w_k N_k(\varepsilon_k)}{n}} + \kappa\, G_K(\theta_K),$$

which is (8) *with κ in place of $\kappa/2$. Hence, by Theorem B.1(i), $\Delta_n(Q)$ obeys the same bound, and every corollary of Theorem B.14 holds along the two-sample route with its constants at most doubled.*

Proof. Take expectations in Proposition B.5. *Floor:* $\mathbb{E}(\hat{w}_k^A + \hat{w}_k^B) = 2w_k$. *Discrepancy:* for a cell of mass μ_j, $\hat{\mu}_j^A - \hat{\mu}_j^B$ has mean 0 and variance $2\mu_j(1-\mu_j)/(n/2) \le 4\mu_j/n$, so $\mathbb{E}|\hat{\mu}_j^A - \hat{\mu}_j^B| \le 2\sqrt{\mu_j/n}$, and over the at most N_k cells of shell k, whose masses sum to w_k, Cauchy–Schwarz gives $\sum_j \mathbb{E}|\hat{\mu}_j^A - \hat{\mu}_j^B| \le 2\sqrt{w_k N_k/n}$; times $\kappa/2$ this is $\kappa\sqrt{w_k N_k/n}$. *Trim:* a trimmed point is a singleton cell of discrepancy $2/n$, so the trimmed points contribute $(\kappa/2)\,\mathbb{E}(\hat{w}_{\mathrm{trim}}^A + \hat{w}_{\mathrm{trim}}^B) = \kappa\, G_K(\theta_K)$. The last claim: $a+2b+c \le 2(a+b+c)$ for the three nonnegative terms, so the bound is at most twice (8) at every profile, hence at most twice its infimum, which is what the corollaries optimize. □

Remark B.7 (The Rademacher form)**.** *The empirical Rademacher complexity of the class on the sample, $\widehat{\mathcal{R}}_S := \mathbb{E}_\sigma \sup_F \frac{1}{n}\sum_i \sigma_i F(P_i)$ over the 1-Lipschitz F bounded by $b = \kappa/2$ and independent uniform signs σ_i, gives the classical certificate $\Delta_S \le 2\widehat{\mathcal{R}}_S + 3\kappa\sqrt{\ln(2/\delta)/(2n)}$ with probability $1-\delta$ (Koltchinskii & Panchenko, 2000; Bartlett & Mendelson, 2002): McDiarmid on Δ_S, the symmetrization of Vapnik & Chervonenkis (1971), McDiarmid on $\widehat{\mathcal{R}}_S$. For the Lipschitz ball, $\widehat{\mathcal{R}}_S$ is at most the expected cost of an optimal partial matching of the plus points to the minus points in D with every unmatched point sent to a reservoir at cost b, by pairing the terms of $\sum_i \sigma_i F(P_i)$ along the matching, and conditionally on the sizes of the sign classes the plus and minus points*

are a random split of the sample, so the balanced part of $\widehat{\mathcal{R}}_S$ is half the two-sample distance of a random unbalanced split and the rest is the reservoir term, at most $b/\sqrt{n}$. Against Theorem 4.6 the form pays the factor 2 and three concentration terms, and gains the whole sample in one statement without a split. Its use for a Lipschitz class over the sample's own geometry is the Rademacher theorem of Gottlieb et al. (2013), in the entropy form through the doubling dimension of the sample, and the cell bound of Proposition B.5 in the Rademacher form is Carrasco et al. (2025, Prop. 4).

Remark B.8 (The certificate, read in full)**.** *Of Theorem B.1, part (ii) is the certificate. The observed distance between the halves, plus one concentration term, bounds the generalization gap of the class for the learner trained on half A and for the learner trained on the whole sample, and by part (iii) it is computable: on the flock it is evaluated at up to $32{,}000$ samples per half in Section 5.1. Three remarks on the form. The split must be fixed in advance or drawn independently of the data, and one split suffices: the observed distance is within $\kappa\sqrt{2\ln(1/\delta)/n}$ of its mean by McDiarmid on the n points, and averaging over independent splits certifies the expected deviation of a half, not the realized one (Remark B.2). Any quantity computed from the sample that dominates $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ is a certificate too, whatever was looked at to compute it, a partition chosen on the sample or a profile optimized on it, since the event of (ii) does not depend on that choice: this is what lets the shells of the profile enter, in Appendix B.2, as a bound on the transport rather than as a hypothesis on the design. And the object is classical in its parts: the two-sample distance is the maximum discrepancy of Bartlett et al. (2002) on a Lipschitz ball, where it is a transport by duality (von Luxburg & Bousquet, 2004, eq. (11)) and a two-sample test statistic in its own right (Ramdas et al., 2017), and the certificate is the symmetrization of Vapnik & Chervonenkis (1971) in that form, while the empirical Rademacher complexity of the class (Koltchinskii & Panchenko, 2000; Bartlett & Mendelson, 2002) is the same object with a random sign pattern in place of a balanced one, and gives a certificate with the constant 2 and three concentration terms (Remark B.7). Across two domains the same transport, between the empirical source and target samples, bounds the target risk of a Lipschitz class in domain adaptation (Redko et al., 2017, Thm. 2), (Shen et al., 2018, Thm. 1), with the joint error of the two domains and a concentration term per sample; here the halves share a law, and one term suffices by the convexity step of (ii). What is not classical is that the class is exactly the Lipschitz ball of a metric the axioms determine, so that the transport is the class's own complexity and not an envelope of a model (Section 2).*

B.2 The shells and the coupling

Lemma B.9 (The tails descend to the lift)**.** *For every $0 \le k \le d$ there is a unique function $T_k : \mathcal{M}_d \to [0,1]$ with $T_k \circ \Phi = T_k(\cdot,\cdot)$, and it is continuous on $(\mathcal{M}_d, D)$ and on $(\mathcal{M}_d, \widetilde{D})$. In particular T_k is computable from the lifted point alone.*

Proof. Well defined: if $\Phi(x,\mu) = \Phi(y,\nu)$ then by Lemma A.8 the multisets of positive (value, weight) pairs coincide, hence so do the multisets of positive weights; padding by zeros does not change T_k for $k \ge 1$, and $T_0 \equiv 1$. *Continuous:* T_k is continuous on $\mathcal{C}$ (Definition 4.1), $\mathcal{C}$ is compact, and Φ is a continuous surjection onto the metric, hence Hausdorff, space $\mathcal{M}_d$ (Lemma A.12), so Φ is a closed map and therefore a quotient map; a continuous function constant on fibers descends to a continuous function through a quotient map. □

The sets $B_k(\theta) = \{T_k \le \theta\}$ of Definition B.4 are therefore closed, the shells are Borel, and their masses are the profile increments stated there.

Definition B.10 (Celling)**.** *Let $(\mathcal{X}, \varrho)$ be a metric space and $S \subseteq \mathcal{X}$ Borel. An (r, N)-*celling *of S is a partition of S into at most N Borel sets, each contained in a closed ϱ-ball of radius r about a point of $\mathcal{X}$, its* center*.*

Proposition B.11 (Single-partition coupling, grouped)**.** *Let $(\mathcal{X}, \varrho)$ be a metric space with $\operatorname{diam}\mathcal{X} \le \kappa_{\mathcal{X}}$, Γ a Borel probability measure on $\mathcal{X}$, and $\hat{\Gamma}_n$ the empirical measure of n i.i.d. draws from Γ. Let $B \subseteq \mathcal{X}$ be Borel, let $B = \bigsqcup_g S_g$ be a finite Borel partition, and suppose each group S_g admits an (r_g, N_g)-celling. Then, with $w_g := \Gamma(S_g)$ and W_1 the Kantorovich distance with ground metric ϱ,*

$$\mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\sum_g r_g\, w_g \;+\; \frac{\kappa_{\mathcal{X}}}{2}\sum_g \sqrt{\frac{N_g\, w_g}{n}} \;+\; \kappa_{\mathcal{X}}\, \Gamma(B^c).$$

Proof. Let $\{V_c\}_{c\in g}$ be the celling of S_g; together the cells V_c, $c \in \bigcup_g g$, form a Borel partition of B, the cells of group g having ϱ-diameter at most $2r_g$. Write $a_c = \Gamma(V_c)$, $b_c = \hat{\Gamma}_n(V_c)$, $m_c = a_c \wedge b_c$, and $M = \sum_c m_c$.

One coupling. Inside each cell pair mass m_c of Γ with mass m_c of $\hat{\Gamma}_n$, for instance the product of the two normalized restrictions, at cost at most $2r_g$ per unit; pair the two residual measures, each of total mass $1 - M$, arbitrarily, at cost at most $\kappa_{\mathcal{X}}$ per unit. From $a \wedge b = \frac{a+b}{2} - \frac{|a-b|}{2}$ and $\sum_c a_c = \Gamma(B)$, $\sum_c b_c = \hat{\Gamma}_n(B)$,

$$1 - M \;=\; \tfrac{1}{2}\sum_c |a_c - b_c| + \tfrac{1}{2}\big(\Gamma(B^c) + \hat{\Gamma}_n(B^c)\big),$$

hence, using $m_c \le a_c$ and $\sum_{c\in g} a_c = w_g$,

$$W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\sum_g r_g\, w_g + \kappa_{\mathcal{X}}\Big[\tfrac{1}{2}\sum_c |a_c - b_c| + \tfrac{1}{2}\big(\Gamma(B^c) + \hat{\Gamma}_n(B^c)\big)\Big].$$

Binomial fluctuations, Cauchy–Schwarz per group. The cells are fixed before sampling, so $n\hat{\Gamma}_n(V_c) \sim \mathrm{Bin}(n, a_c)$ and $\mathbb{E}|a_c - b_c| \le \sqrt{\mathbb{E}(a_c - b_c)^2} \le \sqrt{a_c/n}$. Summing over the at most N_g cells of group g, $\sum_{c\in g} \mathbb{E}|a_c - b_c| \le \sqrt{N_g \sum_{c\in g} a_c/n} = \sqrt{N_g w_g/n}$, and $\mathbb{E}\hat{\Gamma}_n(B^c) = \Gamma(B^c)$. Take expectations in the display. □

Lemma B.12 (Celling from a tube)**.** *Let $S \subseteq \mathcal{X}$ be Borel, $A \subseteq \mathcal{X}$, and $0 \le s < r$. If every point of S lies within ϱ-distance s of A, and A has a net of N points at radius $r - s$, then S admits an (r, N)-celling centered at the net points.*

Proof. A point p of S is within $s + \eta$ of some point of A for every $\eta > 0$, which is within $r - s$ of a net point, so p is within $r + \eta$ of one of the N net points, and one net point z_j serves for a sequence $\eta \downarrow 0$: $\varrho(p, z_j) \le r$, so the closed balls $\mathsf{B}(z_j, r)$, $j \le N$, cover S. Disjointify: $V_j := S \cap \mathsf{B}(z_j, r) \setminus \bigcup_{j'<j} \mathsf{B}(z_{j'}, r)$, Borel, a partition of S, and $V_j \subseteq \mathsf{B}(z_j, r)$. □

Proposition B.13 (Single-partition coupling, reference-point form)**.** *Let $(\mathcal{X}, \varrho)$ be a metric space, Γ a Borel probability measure on $\mathcal{X}$ and $\hat{\Gamma}_n$ its empirical measure, and fix a reference point $z_0 \in \mathcal{X}$ with $\varrho_0(\cdot) := \varrho(\cdot, z_0)$. Let $B = \bigsqcup_g S_g$ be a finite Borel partition of a Borel $B \subseteq \mathcal{X}$, each S_g with an (r_g, N_g)-celling and with $\varrho_0 \le \varrho_g$ on S_g. Then, with $w_g := \Gamma(S_g)$,*

$$\mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\sum_g r_g\, w_g \;+\; \sum_g \varrho_g \sqrt{\frac{N_g\, w_g}{n}} \;+\; 2\int_{B^c} \varrho_0 \, d\Gamma.$$

The point of the form is that it needs no bound on the diameter: residual mass is priced at its own distance to the reference point.

Proof. Couple within cells as in the proof of Proposition B.11, at cost at most $2r_g$ per unit. The two residual measures, Γ's, consisting of the mass $a_c - m_c$ in each cell c and of $\Gamma|_{B^c}$, and $\hat{\Gamma}_n$'s, consisting of $b_c - m_c$ in each cell and of $\hat{\Gamma}_n|_{B^c}$, have equal total mass, and any coupling π of them costs $\int \varrho \, d\pi \le \int (\varrho_0(P) + \varrho_0(Q))\, d\pi(P, Q)$ by the triangle inequality through z_0, that is at most $\int \varrho_0$ against each residual. A residual unit in a cell of group g costs at most ϱ_g, and the two residuals of a cell together have mass $(a_c - m_c) + (b_c - m_c) = |a_c - b_c|$; the residual mass on B^c costs $\int_{B^c} \varrho_0 \, d\Gamma + \int_{B^c} \varrho_0 \, d\hat{\Gamma}_n$. Hence

$$W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\sum_g r_g w_g + \sum_g \varrho_g \sum_{c\in g} |a_c - b_c| + \int_{B^c} \varrho_0 \, d(\Gamma + \hat{\Gamma}_n),$$

and taking expectations with $\sum_{c\in g} \mathbb{E}|a_c - b_c| \le \sqrt{N_g w_g/n}$ as in that proof and $\mathbb{E}\int_{B^c} \varrho_0 \, d\hat{\Gamma}_n = \int_{B^c} \varrho_0 \, d\Gamma$ gives the display. □

Theorem B.14 (Generalization over the class)**.** *For every design Q and every resolution profile,*

$$\Delta_n(Q) \;\le\; 2\sum_{k=1}^{K} w_k\,\varepsilon_k \;+\; \frac{\kappa}{2}\sum_{k=1}^{K}\sqrt{\frac{w_k\,N_k(\varepsilon_k)}{n}} \;+\; \kappa\,G_K(\theta_K). \tag{8}$$

The count $N_k(\varepsilon_k)$ may be replaced by the covering number at radius $\varepsilon_k/2$ of any set within $\varepsilon_k/2$ of Sh_k, in particular of the crop of the shell's support, whose exponent on a selector design is $k\dim\mathcal{V}$. The same bound with κ in place of $\kappa/2$ is the expectation of Proposition B.5 with the theorem's own celling (Proposition B.6). The trim may be traded for a crop: cropping every configuration beyond the last shell to its K heaviest atoms, which lands it in some shell, (8) *holds with $\kappa\,G_K(\theta_K)$ replaced by $2\kappa\,\mathbb{E}_Q[T_K;\ T_K>\theta_K]$ and each w_k by w_k+g_k, where g_k is the mass of the cropped configurations whose crop lies in Sh_k, $\sum_k g_k = G_K(\theta_K)$.*

Remark B.15 (Shells by the first threshold met)**.** *The monotonicity of the profile serves only to make the shells a partition. Defining instead $\mathsf{Sh}'_k := \{T_k\le\theta_k,\ T_j>\theta_j \text{ for } j<k\}$, the configurations whose first threshold met is the kth, gives a partition of the configurations that meet some threshold, for every sequence $\varepsilon_1,\dots,\varepsilon_K$, with $\mathsf{Sh}'_k\subseteq B_k(\theta_k)$ still, and Theorem B.14 and Proposition B.5 hold verbatim with $w_k := \Gamma(\mathsf{Sh}'_k)$ and the trim the mass that meets no threshold (Appendix B.2). For a nondecreasing profile the two definitions agree. The nested form is the one computed in Section 5.1.*

Proof of Theorem B.14. By Kantorovich–Rubinstein duality on the compact space $(\mathcal{M}_d, D)$ (Proposition A.20), $W_1(\hat\Gamma_n,\Gamma) = \sup\{\int F\,d(\hat\Gamma_n-\Gamma) : \mathrm{Lip}_D(F)\le 1\}$, and by Theorem 3.2 the functions $F\circ\Phi$ with $\mathrm{Lip}_D(F)\le 1$ are exactly the members of $\mathcal{G}_1$, so that $\int F\,d\hat\Gamma_n = n^{-1}\sum_i f(Z_i)$ and $\int F\,d\Gamma = \mathbb{E}_Q f$ for $f = F\circ\Phi$. The class is symmetric under $f\mapsto -f$, so the supremum of the signed deviation is the supremum of its absolute value, and $\Delta_n(Q) = \mathbb{E}\,W_1(\hat\Gamma_n,\Gamma)$. Apply Proposition B.11 on $(\mathcal{X},\varrho) = (\mathcal{M}_d, D)$ with $\kappa_{\mathcal{X}} = \kappa$ (Proposition A.20), $B = B_K(\theta_K)$, and the groups $S_k = \mathsf{Sh}_k$, $k\le K$, Borel by Lemma B.9, of masses w_k, with $\Gamma(B^c) = G_K(\theta_K)$. Each group admits an $(\varepsilon_k, N_k(\varepsilon_k))$-celling by Lemma B.12: a point of $\mathsf{Sh}_k\subseteq B_k(\theta_k)$ has $T_k\le\theta_k$, hence lies within $\kappa\theta_k = \varepsilon_k/2$ of its crop, a point of $\mathcal{A}_k$ (Proposition A.19), and $\mathcal{A}_k$ has a net of $N_k(\varepsilon_k) = N(\varepsilon_k/2,\mathcal{A}_k,D)$ points at radius $\varepsilon_k/2$, finite since $\mathcal{A}_k$ is compact (Lemma A.12). So $r_k=\varepsilon_k$, and the proposition's three terms are the display. For the replacement of the count: Lemma B.12 takes any set A within $\varepsilon_k/2$ of the shell, in particular the image of $\mathsf{Sh}_k\cap\mathrm{supp}\Gamma$ under the k-crop, which lies in $\mathcal{A}_k$ and within $\varepsilon_k/2$ of the shell by the same crop cost; on a selector design its covering number has exponent $k\dim\mathcal{V}$ by the paragraph after Theorem A.18. For Remark B.15: the sets $\mathsf{Sh}'_k = \{T_k\le\theta_k\}\cap\bigcap_{j<k}\{T_j>\theta_j\}$ are Borel by Lemma B.9, they partition $\bigcup_{k\le K}B_k(\theta_k)$ for every sequence of thresholds, and $\mathsf{Sh}'_k\subseteq B_k(\theta_k)$, so the same cellings apply and Proposition B.11 gives the display with $w_k=\Gamma(\mathsf{Sh}'_k)$ and the trim $\Gamma(\bigcap_{k\le K}\{T_k>\theta_k\})$; Proposition B.5 needs only a partition into cells of the stated radii.

The crop clause. Let φ be the identity on $B_K(\theta_K)$ and the K-crop cr_K on its complement, a measurable map into $B_K(\theta_K)$ since $T_K(\mathrm{cr}_K P)=0$, with $D(P,\varphi(P))\le\kappa\,T_K(P)\,\mathbf{1}\{T_K(P)>\theta_K\}$ by Proposition A.19. Then $\varphi_\#\hat\Gamma_n$ is the empirical measure of n i.i.d. draws from $\Gamma' := \varphi_\#\Gamma$, and coupling along φ on both sides,

$$W_1(\hat\Gamma_n,\Gamma) \;\le\; \frac{\kappa}{n}\sum_{i\le n} T_K(P_i)\,\mathbf{1}\{T_K(P_i)>\theta_K\} + W_1(\varphi_\#\hat\Gamma_n,\Gamma') + \kappa\,\mathbb{E}_Q[T_K;\ T_K>\theta_K],$$

whose expectation is $2\kappa\,\mathbb{E}_Q[T_K;T_K>\theta_K] + \mathbb{E}\,W_1(\varphi_\#\hat\Gamma_n,\Gamma')$. Apply Proposition B.11 to Γ' with the same B and the same groups: the cellings are properties of the sets, $\Gamma'(B^c)=0$, and $\Gamma'(\mathsf{Sh}_k) = w_k+g_k$ with $g_k = \Gamma(\{T_K>\theta_K\}\cap\mathrm{cr}_K^{-1}(\mathsf{Sh}_k))$. □

Remark B.16 (The three terms of Theorem B.14, read in full)**.** *The three terms are a resolution floor, a fluctuation and a trim, and the profile enters each. The floor is mass weighted: shell k pays its own resolution at its own mass, so a coarsely resolved high shell does not tax the finely resolved bulk. The fluctuation is the design's mixture of local coverings: shell k pays the count of the k-atom stratum, of exponent $e(k)$ and not $e(d)$, weighted by the mass the design puts there, and the term is the binomial fluctuation of the sample's cell counts, which is where the measure enters, with no function class in sight. The trim is the mass beyond the last shell, paid at the diameter. The crop distance $\kappa\theta_k$ is absorbed into the radius of the cells and paid once, in the floor, never as a bias and*

never in the counts. Under the crop clause a crowded configuration pays its own tail in place of the diameter, and Theorem 4.4 below is the constant-profile case, with the multiscale coupling of Proposition B.18 on the cropped stratum in place of the single scale. The constant profile $\varepsilon_k \equiv \varepsilon$ is the single-scale form, one resolution and one trim, the shape of the trimmed empirical-measure bounds of Weed & Bach (2019); the stratified form, one resolution per shell, is what a profile needs. The stratified form is not the single-scale bound with more indices: the scales index disjoint parts of the design, not refinements of one set, and the mass per shell is what lets each part pay its own count, the per-shell step of Proposition B.11. Every profile gives a valid bound and the left side does not depend on the profile, so the bound is the infimum over K and over profiles. The shells are observable: $T_k(P_i)$ is computable from the sample, so an estimator can bin by shell (Remark B.23). The profile is an assumption on one part of the design, its sparsity across and within configurations, and on nothing else: within a shell the distribution is arbitrary and the bound charges the stratum at its worst, which is where the certificate, which sees the sample's own configurations, improves on it (Section 5.1).

B.3 Proof of Theorem 4.4

The upper bound is proved here, from a multiscale coupling on the cropped stratum (Proposition B.18), with $C = 51C_1$ for the constant C_1 of the stratum count, 1734 on interval values. The lower bound of Theorem 4.4 is Lemma D.4 in Appendix D: its bound is the smaller of $c_{k_0} n^{-1/(2k_0-1)}$ and $\varepsilon_{\max}(k_0)/2$, and since $n^{-1/(2k_0-1)} \le 1$ the statement holds with the smaller of the two constants c_{k_0} and $\varepsilon_{\max}(k_0)/2$ in place of c_{k_0}.

Lemma B.17 (Nested cellings from covers)**.** *Let $(\mathcal{X}, \varrho)$ be a metric space and $S \subseteq \mathcal{X}$ Borel with $\mathrm{diam} S \le \varepsilon_0$ and finite covering numbers, and put $\varepsilon_j := 3^{-j}\varepsilon_0$. For every $J \ge 0$ there are Borel partitions $\mathcal{Q}_0 = \{S\}, \mathcal{Q}_1, \dots, \mathcal{Q}_J$ of S, each cell of $\mathcal{Q}_j$ of diameter at most ε_j and contained in a cell of $\mathcal{Q}_{j-1}$, with $|\mathcal{Q}_j| \le N(\varepsilon_{j+1}/2, S, \varrho)$.*

Proof. This is the construction of Boissard & Le Gouic (2014, Lemma 2.1) and Weed & Bach (2019, Prop. 3). For $1 \le j \le J$ let $B_1^j, \dots, B_{N_j}^j$ be closed balls of radius $\varepsilon_{j+1}/2$ covering S, $N_j := N(\varepsilon_{j+1}/2, S, \varrho)$. Disjointify the finest cover: $Q_\ell^J := S \cap B_\ell^J \setminus \bigcup_{\ell'<\ell} B_{\ell'}^J$ gives a Borel partition $\mathcal{Q}_J$ of S into at most N_J cells of diameter at most $\varepsilon_{J+1} \le \varepsilon_J$. Given $\mathcal{Q}_{j+1}$, let Q_ℓ^j be the union of the cells of $\mathcal{Q}_{j+1}$ that meet B_ℓ^j and were not assigned to $Q_{\ell'}^j$ for some $\ell' < \ell$. Every cell of $\mathcal{Q}_{j+1}$ meets some B_ℓ^j, so the Q_ℓ^j form a Borel partition of S into at most N_j cells, each a union of cells of $\mathcal{Q}_{j+1}$; and two points of Q_ℓ^j lie in cells of $\mathcal{Q}_{j+1}$, of diameter at most ε_{j+1}, that both meet the ball B_ℓ^j, of diameter ε_{j+1}, so their distance is at most $3\varepsilon_{j+1} = \varepsilon_j$. Finally $\mathcal{Q}_0 := \{S\}$. □

Proposition B.18 (Multiscale coupling)**.** *Let $(\mathcal{X}, \varrho)$, S and ε_j be as in Lemma B.17, Γ a Borel probability measure on S and $\hat{\Gamma}_n$ the empirical measure of n i.i.d. draws from Γ. If $N(\varepsilon_{j+1}/2, S, \varrho) \le (A/\varepsilon_j)^e$ for every $j \ge 0$, with $A \ge \varepsilon_0$ and $e > 0$, then*

$$\mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 17\, A\, n^{-1/e} \quad \text{if } e \ge 3, \qquad \mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 4A\,(1 + \ln n)\, n^{-1/2} \quad \text{if } e \le 2.$$

For $2 < e < 3$ the first bound holds with a constant depending on e in place of 17, by the same proof. The exponents used below are integers.

Proof. The coupling. Fix $J \ge 0$ and the partitions of Lemma B.17, and for two probability measures μ, ν on S write $a_C := \mu(C)$, $b_C := \nu(C)$ for a cell C. We claim

$$W_1(\mu, \nu) \;\le\; 2\varepsilon_J + \sum_{j=1}^{J} \varepsilon_{j-1} \cdot \tfrac{1}{2} \sum_{C \in \mathcal{Q}_j} |a_C - b_C|. \tag{9}$$

Pick a point x_C in each cell $C \in \mathcal{Q}_J$ and let $\mu_J := \sum_C a_C\, \delta_{x_C}$, $\nu_J := \sum_C b_C\, \delta_{x_C}$: moving every point to its cell's representative costs at most ε_J, so $W_1(\mu, \nu) \le 2\varepsilon_J + W_1(\mu_J, \nu_J)$. Couple μ_J with ν_J from the finest level up. First pair, at each x_C, the mass $a_C \wedge b_C$ of the two, at cost 0; what remains in a cell $C \in \mathcal{Q}_J$ is $(a_C - b_C)_+$ of μ_J and $(b_C - a_C)_+$ of ν_J. Then, for $j = J, J-1, \dots, 1$ in turn, inside each cell $P \in \mathcal{Q}_{j-1}$ pair as much as possible of the remaining μ_J-mass of the cells $C \in \mathcal{Q}_j$, $C \subseteq P$, with the remaining ν_J-mass of those cells, at cost at most

$\mathrm{diam}P \le \varepsilon_{j-1}$ per unit. By induction on the level, before the step at level j the mass remaining in a cell $C \in \mathcal{Q}_j$ is $(a_C - b_C)_+$ of μ_J and $(b_C - a_C)_+$ of ν_J: the two sums $\sum_{C\subseteq P}(a_C - b_C)_+$ and $\sum_{C\subseteq P}(b_C - a_C)_+$ differ by $a_P - b_P$ and add to $\sum_{C\subseteq P}|a_C - b_C|$, so the step at level j inside P pairs $\frac{1}{2}\sum_{C\subseteq P}|a_C - b_C| - \frac{1}{2}|a_P - b_P|$ units and leaves $(a_P - b_P)_+$ of μ_J and $(b_P - a_P)_+$ of ν_J in P. At the top, $P = S$ has $a_S = b_S = 1$ and nothing remains, so the pairing is a coupling of μ_J and ν_J, of cost at most $\sum_{j=1}^{J} \varepsilon_{j-1} \sum_{P\in\mathcal{Q}_{j-1}} \frac{1}{2}\sum_{C\subseteq P}|a_C - b_C|$, which is (9).

The fluctuations. Take $\mu = \hat{\Gamma}_n$ and $\nu = \Gamma$. The cells are fixed before sampling, so as in the proof of Proposition B.11 $\mathbb{E}\sum_{C\in\mathcal{Q}_j}|a_C - b_C| \le \sqrt{|\mathcal{Q}_j|/n} \le (A/\varepsilon_j)^{e/2} n^{-1/2}$, and with $\varepsilon_{j-1} = 3\varepsilon_j$,

$$\mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\varepsilon_J + \tfrac{3}{2}\, A^{e/2}\, n^{-1/2} \sum_{j=1}^{J} \varepsilon_j^{1-e/2}.$$

The rate. Let $e \ge 3$, so that $\varepsilon_j^{1-e/2} = \varepsilon_J^{1-e/2}\, 3^{-(J-j)(e/2-1)} \le \varepsilon_J^{1-e/2}\, 3^{-(J-j)/2}$ and the sum is at most $\varepsilon_J^{1-e/2}/(1 - 3^{-1/2}) \le 2.37\, \varepsilon_J^{1-e/2}$: the bound reads $2\varepsilon_J + 3.56\, \varepsilon_J\, (A/\varepsilon_J)^{e/2} n^{-1/2}$. Put $\varepsilon^* := A\, n^{-1/e}$. If $\varepsilon^* \ge \varepsilon_0$ the claim holds since $W_1 \le \mathrm{diam}S \le \varepsilon_0$. Otherwise take $J := \lfloor \log_3(\varepsilon_0/\varepsilon^*) \rfloor \ge 0$, so that $\varepsilon^* \le \varepsilon_J < 3\varepsilon^*$; then $(A/\varepsilon_J)^{e/2} n^{-1/2} \le (A/\varepsilon^*)^{e/2} n^{-1/2} = 1$ and the bound is at most $5.56\, \varepsilon_J < 16.7\, \varepsilon^* \le 17 A\, n^{-1/e}$. For $e \le 2$ take $J := \lceil \frac{1}{2}\log_3 n \rceil$, so that $\varepsilon_J \le \varepsilon_0 n^{-1/2}$; each summand is at most $\varepsilon_0^{1-e/2} \le A^{1-e/2}$, so the bound is at most $2\varepsilon_0 n^{-1/2} + \frac{3}{2} A\, J\, n^{-1/2} \le A\, (\frac{7}{2} + \frac{3}{4}\log_3 n)\, n^{-1/2} \le 4A\, (1 + \ln n)\, n^{-1/2}$. □

Proof of Theorem 4.4. A single level. Let $\mathrm{cr}_K : \mathcal{M}_d \to \mathcal{A}_K$ be the K-crop, which keeps the K heaviest atoms of a lift and renormalizes their weights, a measurable map with $D(P, \mathrm{cr}_K P) \le \kappa T_K(P)$ (Proposition A.19). The lifted design $\Gamma^K := (\mathrm{cr}_K)_{\#}\Gamma$ is that of an exactly K-sparse design, the empirical lifted design of the cropped sample is $(\mathrm{cr}_K)_{\#}\hat{\Gamma}_n$, and by the triangle inequality

$$W_1(\hat{\Gamma}_n, \Gamma) \;\le\; W_1\big(\hat{\Gamma}_n, (\mathrm{cr}_K)_{\#}\hat{\Gamma}_n\big) + W_1\big((\mathrm{cr}_K)_{\#}\hat{\Gamma}_n, \Gamma^K\big) + W_1(\Gamma^K, \Gamma),$$

where the first term is at most $\kappa n^{-1}\sum_i T_K(P_i)$ and the third at most $\kappa\, \mathbb{E}_Q T_K$, by the coupling along the crop; take expectations and apply the exactly sparse case, proved next, to the middle term.

Exact sparsity. Γ^K is a probability measure on $\mathcal{A}_K$, of diameter at most κ (Proposition A.20), and $N(\varepsilon/2, \mathcal{A}_K, D) \le (C_1 R/\varepsilon)^e$ for $\varepsilon \le \kappa$ with $e := e(K)$ (Theorem A.17; on a selector design the cropped support of the stratum replaces $\mathcal{A}_K$, with the count of the paragraph after Theorem A.18). Apply Proposition B.18 on $S = \mathcal{A}_K$ with $\varepsilon_0 := \kappa$: at the scales $\varepsilon_j = 3^{-j}\kappa$, $N(\varepsilon_{j+1}/2, \mathcal{A}_K, D) \le (C_1 R/\varepsilon_{j+1})^e = (3C_1 R/\varepsilon_j)^e$, so $A := 3C_1 R \ge \kappa$ serves, and the middle term is at most $51 C_1 R\, n^{-1/e}$ for $e \ge 3$ and $12 C_1 R\, (1 + \ln n)\, n^{-1/2}$ for $e \le 2$. So $C := 51 C_1$ serves in both cases, $C = 1734$ on interval values. □

Remark B.19 (Graph-network bounds in this currency)**.** *The generalization bounds for graph networks bound a different quantity: the gap between the training and the test loss of one trained network, of fixed width and norm budget, over m i.i.d. graphs. The rate in m is $m^{-1/2}$ whatever the graphs, and the graph enters the constant: the maximum degree d as d^{l-1} for an l-step message-passing network, through the Lipschitz constant of the sum over neighbors, and the width h and the stable ranks of the weights as, in effect, the square root of the parameter count (Liao et al., 2021, Theorems 3.2 and 3.4), with the spectral norm of the diffusion matrix in place of d in Ju et al. (2023, Theorem 3.1). The class behind these bounds is a per-neighbor map shared across the neighbors, summed and read out, the weighted Deep Sets form of Section 1: parameter sharing makes its entropy independent of how many neighbors there are, and the sum makes the count a scale, so nothing in these bounds is a rate in the count, and at one layer with a mean the degree is absent altogether. To learn a member of $\mathcal{G}$ with such a network the width cannot stay fixed: a sum decomposition that represents every symmetric function of k elements needs latent dimension k (Wagstaff et al., 2022, Theorem 20), and the readout is then a function of that many coordinates. What no parametrization changes is the lower bound: on the stratum design of Appendix D at k atoms the generalization gap of the class is at least $n^{-1/(2k-1)}$ (Theorem 4.4), and so is the risk of any learner (Theorem D.5), so the count is in the exponent. In shape, asked to learn the whole class, the graph-network route, one architecture serving every configuration with a readout of $k_{\max}$ coordinates, sees the design*

through its maximum count, Theorem 4.4 at $k_0 = k_{\max}$, where for its own parametric class the degree is a constant in front of the parametric rate. On the flock, where $k_{\max} = 68$, that bound is trivial (Appendix B.3). Where the count is a scale on both sides, in the accumulative class, ours is the mean effective count $\bar{\zeta}$ of Corollary B.28, 12 *on the flock, against the maximum degree of the sample,* 68*. The conservative linear sums, the members a sum network with a linear readout learns without the count, are count-free at the rate of Proposition B.20, and the profile is the price of the rest.*

The graph-network bounds, and Theorem 4.4 at the maximum count. The PAC-Bayes bound of Liao et al. (2021, Theorem 3.2) for an l-layer graph convolutional network of width h trained on m graphs has the form

$$L_{\mathcal{D}}(f_w) \;\le\; L_{S,\gamma}(f_w) + O\left(\sqrt{\frac{B^2 d^{\,l-1} l^2 h \log(lh)\, \prod_i \|W_i\|_2^2 \, \sum_i \|W_i\|_F^2 \,/\, \|W_i\|_2^2 + \log(ml/\delta)}{\gamma^2 m}}\right),$$

with $d-1$ the maximum degree of the training graphs, and their Theorem 3.4 for the message-passing network carries the geometric series $((dC)^{l-1}-1)/(dC-1)$ through a different weight term. Ju et al. (2023, Theorem 3.1) replace the degree by the spectral norm of the diffusion matrix, $\|P_G\|^{2(l-1)}$, at most $d^{2(l-1)}$ for a sum and at most one for a normalized mean, under a root that carries the width d_i of each layer and the stable ranks $r_i^2 \prod_j s_j^2$ explicitly, summed over the layers. Three features are common to the line. The rate is $m^{-1/2}$. The count enters as a scale, a power of the degree per layer that comes from the norm of a sum of d bounded vectors, and it is absent at one layer with a mean. And the width enters as roughly the square root of the parameter count, h times the stable rank, so that at a fixed architecture the bound is a constant over $\sqrt{m}$. These are bounds on the gap of one network, not on the risk of learning a class, and the class they are bounds for reads one neighbor at a time through a shared map and sums. Its entropy does not see the count because the parameters do not, which is the mechanism that makes the node count drop out (Section 2), and the same mechanism keeps the count out of the exponent.

To learn a member of $\mathcal{G}$ with such a network on a design of maximum count k, the width has to grow with the accuracy. By Wagstaff et al. (2022, Theorem 20) a continuous sum decomposition $\rho(\sum_i \varphi(x_i))$ that approximates every continuous symmetric function of k scalar elements needs latent dimension at least k, and below it there is a Lipschitz symmetric function, affine in the sorted vector, that no decomposition approximates better than a constant does. So the readout ρ is a function of $p \ge k$ coordinates, $k \dim \mathcal{V}$ for vector elements. What the count costs does not depend on the parametrization: on the stratum design of Appendix D at k atoms the generalization gap of the class is at least $n^{-1/(2k-1)}$ (Theorem 4.4, Lemma D.4), and the risk of learning likewise (Theorem D.5), so for the class the count is in the exponent whatever the architecture, while the conservative linear sums are count-free, at the rate of Proposition B.20. One architecture serves every configuration, so the count in the exponent is the maximum over the design, where Theorem 4.4 charges the count of the shell and Theorem B.14 the profile. On the flock the maximum support is 68, and the bound at $k_0 = 68$ is the trivial value κ on the flock, under either exponent convention.

B.4 Proof of Proposition B.20

Proposition B.20 (The linear sums are count-free)**.** *Let $w = \mu$, and let $\Delta_n^{\mathrm{lin}}(Q) := \mathbb{E}\sup_{\varphi\in\mathcal{S}}\big(\mathbb{E}_Q f_\varphi - n^{-1}\sum_{i\le n} f_\varphi(Z_i)\big)$ be the generalization process of the linear sums* (7)*. Write $\nu_z := \sum_i \mu_i\, \delta_{x_i}$ for the weighted value measure of a configuration $z = (x, \mu)$, a probability measure on $\mathcal{V}$. Then*

$$\Delta_n^{\mathrm{lin}}(Q) \;=\; \mathbb{E}\, W_1^{\rho}\Big(\frac{1}{n}\sum_{i\le n}\nu_{Z_i},\; \mathbb{E}_Q\,\nu_Z\Big) \;\le\; \inf_{\varepsilon>0}\Big[\,2\varepsilon + R\sqrt{\frac{N(\varepsilon,\mathcal{V},\rho)}{n}}\,\Big] \;\le\; C_{\mathcal{V}}\, R\, n^{-\frac{1}{\dim\mathcal{V}+2}},$$

with $\dim\mathcal{V}$ the covering dimension of the value space and $C_{\mathcal{V}}$ a constant of its covering numbers alone. Neither d, nor the count, nor the profile enters. In one dimension, $\mathcal{V}$ an interval or the circle, $\Delta_n^{\mathrm{lin}}(Q) \le 2R/\sqrt{n}$.

Remark B.21 (The two regimes: the count-free members and the rest)**.** *The gap of the linear sums is a transport on the value space, between the mean of n weighted value measures and its expectation,*

one level below the transport on the lifted space of Section 4, and the count is absent because a weighted mean of k atoms is one point of $\mathcal{P}(\mathcal{V})$ whatever k: crowding lowers the variance of the cell masses and helps. On the flock $\dim \mathcal{V} = 1$ and the rate is $n^{-1/2}$ at every crowding, the one-dimensional case being exact (Appendix B.4). This is the mechanism by which the parameter sharing of a sum network keeps the count out of its entropy (Remark B.19); a Lipschitz readout of p linear sums sees the law of a p-dimensional feature of the mean lift and stays count-free at the price of p, the bottleneck. The class pays the profile for everything beyond, the members that are nonlinear across atoms: the selected maximum, the count, and every functional whose score reads the height nonlinearly (Remark A.10). The proof follows.

Proof of Proposition B.20. For $z = (x, \mu)$ and $\varphi \in \mathcal{S}$, $f_\varphi(z) = \int \varphi \, d\nu_z$, so $\mathbb{E}_Q f_\varphi - n^{-1} \sum_i f_\varphi(Z_i) = \int \varphi \, d(\bar\nu - \bar\nu_n)$ with $\bar\nu := \mathbb{E}_Q \, \nu_Z$ and $\bar\nu_n := n^{-1} \sum_i \nu_{Z_i}$, both probability measures on $\mathcal{V}$. The score ball $\mathcal{S}$ contains a translate of every 1-Lipschitz function on $\mathcal{V}$ and $\bar\nu - \bar\nu_n$ has mass zero, so the supremum over $\mathcal{S}$ is the supremum over the 1-Lipschitz functions, which is $W_1^\rho(\bar\nu_n, \bar\nu)$ by Kantorovich–Rubinstein duality on the compact space $(\mathcal{V}, \rho)$; this is the identity. For the bound let $c_1, \dots, c_N$ be an ε-net of $\mathcal{V}$, $N = N(\varepsilon, \mathcal{V}, \rho)$, and $A_1, \dots, A_N$ its Voronoi cells, every point within ε of its center. Moving a measure to the centers of its cells costs at most ε in W_1, and between two measures on the centers $W_1 \le \operatorname{diam}\mathcal{V} \cdot \frac{1}{2} \sum_j |\bar\nu_n(A_j) - \bar\nu(A_j)|$, so by the triangle inequality

$$W_1^\rho(\bar\nu_n, \bar\nu) \;\le\; 2\varepsilon + R \sum_{j \le N} |\bar\nu_n(A_j) - \bar\nu(A_j)|.$$

The cell masses $\nu_{Z_i}(A_j)$, $i \le n$, are i.i.d. in $[0,1]$ with mean $\bar\nu(A_j)$, so $\mathbb{E}|\bar\nu_n(A_j) - \bar\nu(A_j)| \le \sqrt{\operatorname{Var} \nu_Z(A_j)/n} \le \sqrt{\bar\nu(A_j)/n}$, using $\nu_Z(A_j)^2 \le \nu_Z(A_j)$, and by Cauchy–Schwarz $\sum_j \sqrt{\bar\nu(A_j)/n} \le \sqrt{N/n}$. Taking expectations gives the first bound. With $N(\varepsilon, \mathcal{V}, \rho) \le (CR/\varepsilon)^{\dim \mathcal{V}}$, the choice $\varepsilon = R\, C^{\dim \mathcal{V}/(\dim \mathcal{V}+2)}\, n^{-1/(\dim \mathcal{V}+2)}$ balances the two terms and gives the rate, with a constant depending on C and $\dim \mathcal{V}$ alone. Nothing in the argument looks at how many atoms carry ν_z, which is where the count and the profile would enter.

In one dimension the net is not needed. On an interval I of length ℓ with its length metric, $W_1(\bar\nu_n, \bar\nu) = \int_I |\bar F_n(t) - \bar F(t)| \, dt$ with $\bar F_n(t) = n^{-1} \sum_i \nu_{Z_i}((-\infty, t])$ an average of n i.i.d. variables in $[0,1]$ of mean $\bar F(t)$, so $\mathbb{E}|\bar F_n(t) - \bar F(t)| \le 1/(2\sqrt{n})$ and $\mathbb{E}\, W_1 \le \ell/(2\sqrt{n}) = R/\sqrt{n}$. On the circle cut at a point, the geodesic distance is at most the distance on the cut interval, of length $2\pi = 4R$, so $W_1^{S^1} \le W_1^I$ and $\mathbb{E}\, W_1 \le 2R/\sqrt{n}$. On the flock this is the rate of the linear sums at every crowding. □

B.5 Regression

Remark B.22 (Regression with the absolute loss)**.** *The two-sample distance is the maximum discrepancy of the unit ball on the halves, $W_1(\hat\Gamma_A, \hat\Gamma_B) = \sup_{f \in \mathcal{G}_1} |\mathbb{E}_A f - \mathbb{E}_B f|$ (Section 4.2), which Theorem 4.6 carries to the population with one concentration term. For regression with the absolute loss on a target $f^* \in \mathcal{G}_1$, the loss $|f - f^*|$ of a predictor $f \in \mathcal{G}_1$ is itself a member, of $\mathcal{G}_2$, so with R_A and R_B the empirical risks on the halves, $\sup_{f \in \mathcal{G}_1} |R_A(f) - R_B(f)| \le 2\, W_1(\hat\Gamma_A, \hat\Gamma_B)$: the certificate bounds the generalization gap of every predictor at once, at twice its value, and of the empirical risk minimizer in particular. For a target with m components that is 1-Lipschitz into ℓ^2, under the Euclidean loss, a predictor whose components are each in $\mathcal{G}_1$ is $\sqrt{m}$-Lipschitz into ℓ^2, so the loss $|f - f^*|_2$ has constant $\sqrt{m} + 1$ and the factor is $1 + \sqrt{m}$ in place of 2 (and $2\sqrt{m}$ for a target that is only componentwise in $\mathcal{G}_1$), which is $1 + \sqrt{2}$ for the flock's update direction, jointly Lipschitz by Proposition C.7 (Appendix C.1). For labels in a range of length b the loss $|f(z) - y|$ is 1-Lipschitz in the pair (z, y) under $D + |y - y'|$, so the same holds with the transport between the labelled halves in that metric, the same matching with the label distance added to each pair's cost, and the concentration term at the diameter $\kappa + b$. A loss that is ℓ_1-Lipschitz in the prediction costs the factor ℓ_1 in place of 1, by Talagrand's contraction lemma (Mohri et al., 2018).*

Remark B.23 (The square loss)**.** *For the square loss the rate is a different one: least squares has excess risk at the square of a rate, and what governs it is the entropy of the class under the design rather than the transport. The same shells give that entropy and the same fibers give the matching lower bounds, so we sketch the route and omit the statements, the regression that the paper's number*

concerns being the absolute-loss one (Remark B.22). Entropy. *Cover each shell of Definition B.4 by the celling of Theorem B.14 at its own resolution ε_k, assign to every cell a value from a grid of step ε_k on $[-b, b]$, and the value 0 beyond the cutoff shell: this is a cover of the members of $\mathcal{G}_1$ bounded by b in $L_2(\Gamma)$, of log-cardinality $\sum_{k\le K} N_k(\varepsilon_k)\ln(3b/\varepsilon_k)$ at a radius η with $\eta^2 \le \frac{9}{4}\sum_{k\le K} w_k\varepsilon_k^2 + b^2 G_K(\theta_K)$, on interval values the entropy of a Lipschitz class (Kolmogorov & Tikhomirov, 1959) in the dimension $2K-1$ of the cutoff shell, with the mass beyond it charged at b and no dependence on d.* Upper bound. *Least squares over the cover is the shell-binning regressogram, on each cell the grid value nearest the cell's sample mean of Y, the cells being observable since T_k of a sample is computable from the sample. Bernstein's inequality with a union bound over the cover gives an excess risk of order $\eta^2 + b^2 \ln N/n$ over the best member, and the resolution that balances the two terms gives $(\ln n/n)^{2/(2k_0+1)}$ at exact sparsity, the nonparametric rate of a Lipschitz class in the dimension $2k_0 - 1$ of the cutoff shell, and $e^{-2\sqrt{\beta \ln n/7}}$ on an exponential profile, the shape of Theorem 4.5's rate with another constant in the exponent.* Lower bound. *On a fiber of fixed weights the metric is bi-Lipschitz to a weighted supremum distance with the weights as semi-axes (Lemma D.1), so bumps of height h with disjoint supports in the cells of the stratum design of Definition D.2 or of the fiber design of Theorem D.6 are 1-Lipschitz for D, and Assouad's hypercube (Tsybakov, 2009, §2.7) on Gaussian regression over them gives a minimax risk of $n^{-2/(2k+1)}$ on the stratum design, $n^{-2/(k+2)}$ on a selector design at count k and $e^{-\sqrt{8\beta \ln n}}$ on the fiber design, with constants polynomial in k; on a mixture of separated populations the hypercubes combine, and the bound is the mass-weighted sum of the populations' risks at their effective sample sizes np_j. By this sketch the square loss learns at the nonparametric rate of the cutoff shell's dimension, matched from below in the exponent at exact sparsity and in shape on an exponential profile.*

B.6 Exponential profiles: proof of Theorem 4.5

The upper bound is proved here, with its constant $2e^{3\beta/2}\Lambda$, $\Lambda := CR \vee 2\kappa A$ for the constant C of Theorem 4.4 (1734 on interval values), and $2e^{3\beta}\Lambda$ with the selector count's constant in place of C_1 on a selector design. The lower bound of Theorem 4.5 is Theorem D.6 in Appendix D, with the constant $\gamma^2 R/16$, $\gamma := 1 - e^{-\beta}$.

Proof. Let $n \ge 2$ and $d \ge 2$: the bound at $n = 1$ exceeds the diameter κ, and at $d = 1$ the lift is the interval, where $\Delta_n(Q) \le 2R/\sqrt{n} \le 2Re^{\beta/4}e^{-\sqrt{\beta \ln n/2}}$ by the one-dimensional bound of Proposition B.20. Put $x := \sqrt{2 \ln n/\beta}$ and $K := d \wedge (\lceil x/2 \rceil \vee 2)$, so that $e(K) = 2K - 1 \ge 3$, that $\beta K \ge \sqrt{\beta \ln n/2}$ unless $K = d$, and that $2K - 1 \le x + 3$ in every case. Theorem 4.4 at this K, with its constant C, gives

$$\Delta_n(Q) \;\le\; 2\kappa\, \mathbb{E}_Q T_K + CR\, n^{-1/(2K-1)} \;\le\; 2\kappa A\, e^{-\beta K} + CR\, e^{-\ln n/(x+3)},$$

where the first term vanishes at $K = d$ and is at most $2\kappa A\, e^{-\sqrt{\beta \ln n/2}}$ otherwise, and $\ln n/(x + 3) \ge \ln n/x - 3 \ln n/x^2 = \sqrt{\beta \ln n/2} - 3\beta/2$. So $\Delta_n(Q) \le (2\kappa A + CR\, e^{3\beta/2})\, e^{-\sqrt{\beta \ln n/2}} \le 2e^{3\beta/2}\Lambda\, e^{-\sqrt{\beta \ln n/2}}$. On a selector design, $e(K) = K$: take $x := \sqrt{\ln n/\beta}$ and $K := d \wedge (\lceil x \rceil \vee 3)$, so that $e(K) \ge 3$ once $d \ge 3$, $K \le x + 3$ and $\ln n/(x + 3) \ge \sqrt{\beta \ln n} - 3\beta$; for $d \le 2$ the design is exactly d-sparse and the clause $e \le 2$ of Theorem 4.4 at $K = d$, $CR(1 + \ln n)n^{-1/2}$, is below the claimed rate at a constant depending on β. □

The shells route. The same rate with the constant $\frac{1}{3}$, at $n \ge n_0$, follows from Theorem B.14 at a constant profile with the trim at the diameter, and the computation serves the accumulative class (Corollary B.30), whose theorem has no crop clause. Write $L := \ln n$, let $n_0 := \exp(48\beta \vee 9 \ln^2 \Lambda'/\beta)$, and let $C_1 := 136$, the constant of the accumulative count in Appendix B.7, which dominates the 34 of Theorem A.17 and lets one Λ' serve both classes, so that $N_k(\varepsilon) \le (C_1 R/\varepsilon)^{2k-1}$ for $\varepsilon \le \kappa$ (Theorem A.17) and $\Lambda' := C_1 R \vee 2\kappa A \vee e$, a different constant from the crop route's Λ. Apply Theorem B.14 with the constant profile at

$$\varepsilon := \Lambda' e^{-u}, \qquad \theta := \varepsilon/(2\kappa), \qquad u := \tfrac{1}{3}\sqrt{\beta L},$$

so that $\ln(\Lambda'/\varepsilon) = u$ and $N_k(\varepsilon) \le e^{(2k-1)u}$. Three consequences of $n \ge n_0$, that is $48\beta \le L$ and $9 \ln^2 \Lambda' \le \beta L$, are used repeatedly:

(a) $u \le L/20$, from $\beta \le L/48$;

(b) $\beta \le u$ and $1 \le \ln \Lambda' \le u$, from $u \ge \sqrt{\beta \cdot 48\beta}/3 \ge 2\beta$ and $u \ge \sqrt{9 \ln^2 \Lambda'}/3 = \ln \Lambda'$;

(c) $\kappa \le \Lambda'$, since $C_1 \ge 2$, and $L \ge \sqrt{48 \cdot 9}\, \ln \Lambda' \ge 20$.

In particular $\varepsilon = \Lambda' e^{-u} \le 1 \le \kappa$, using $R \ge \frac{1}{2}$ (so $\kappa = 2R \ge 1$), so the count bound applies and $\theta \le \frac{1}{2}$.

Markov's inequality at the working threshold converts the moment profile into the quantile the theorem consumes:

$$G_k(\theta) \;\le\; 1 \wedge \frac{A e^{-\beta k}}{\theta} \;=\; 1 \wedge e^{-\beta(k-h)}, \qquad h := \beta^{-1} \ln(2\kappa A/\varepsilon) \in [0,\, u/\beta].$$

Set $m := \lceil h \rceil + 1 \le u/\beta + 2$ and apply the theorem with $K := (\lceil 3u/\beta \rceil \vee 1) \wedge d$. *Trim:* either $K = d$, or $\beta K \ge 3u$ and $\kappa\, G_K(\theta) \le \kappa\, e^{-\beta(K-h)} \le \kappa\, e^{-2u} \le \Lambda' e^{-u}$. *Shells:* $w_k \le G_{k-1}(\theta) \le e^{-\beta(k-m)_+}$, as $k - 1 - h \ge k - m$, so the k-th shell term is at most $n^{-1/2} e^{\psi(k)}$ with $\psi(k) := (2k-1)\frac{u}{2} - \frac{\beta}{2}(k-m)_+$. Split the sum at m: below m it is geometric with ratio $e^u \ge e$; above m, geometric with ratio $e^{u-\beta/2} \ge e^{u/2} \ge e^{1/2}$ by (b), and we extend it upward to $\tilde{K} := \lceil 3u/\beta \rceil \ge K$. Hence

$$\sum_{k=1}^{K} \sqrt{\frac{w_k N_k}{n}} \;\le\; \frac{2\, e^{\psi(m)} + 3\, e^{\psi(\tilde{K})}}{\sqrt{n}},$$

$$\psi(m) \;\le\; \tfrac{u^2}{\beta} + \tfrac{3u}{2} \;=\; \tfrac{L}{9} + \tfrac{3u}{2}, \qquad \psi(\tilde{K}) \;\le\; \tfrac{3u^2}{\beta} + \tfrac{u}{2} - \big(u - \beta\big) \;=\; \tfrac{L}{3} - \tfrac{u}{2} + \beta,$$

using $m \le u/\beta + 2$, $\tilde{K} \le 3u/\beta + 1$, and $\tilde{K} - m \ge 2u/\beta - 2 \ge 0$ by (b). The middle term of the theorem is thus at most $\Lambda' \big(e^{\psi(m)} + \frac{3}{2} e^{\psi(\tilde{K})}\big)/\sqrt{n}$, and each piece is at most e^{-u}:

$$\psi(m) - \tfrac{L}{2} \;\le\; \tfrac{L}{9} + \tfrac{3u}{2} - \tfrac{L}{2} \;\le\; -\tfrac{L}{4} \;\le\; -u,$$

$$\ln \tfrac{3}{2} + \psi(\tilde{K}) - \tfrac{L}{2} \;\le\; \tfrac{L}{48} + \tfrac{L}{3} + \tfrac{L}{48} - \tfrac{u}{2} - \tfrac{L}{2} \;=\; -\tfrac{L}{8} - \tfrac{u}{2} \;\le\; -u,$$

using $\frac{3u}{2} \le \frac{3L}{40}$, $\beta \le L/48$, $\ln \frac{3}{2} \le L/48$, and $u \le L/4$, all from (a)–(c). So the middle term is at most $2\Lambda' e^{-u}$, and the total is $2\Lambda' e^{-u} + 2\Lambda' e^{-u} + \Lambda' e^{-u} = 5\Lambda' e^{-u}$.

Remark B.24 (The exponential rate, read in full)**.** *The rate $e^{-c\sqrt{\beta \ln n}}$ is slower than every polynomial in n and faster than every power of $\ln n$. The optimizing cutoff is $K_n \asymp \sqrt{\ln n/\beta}$, so the dimension the rate sees at sample size n grows without bound, like $\sqrt{\ln n}$, and no fixed dimension explains the answer. In sample-complexity form, accuracy ε costs $n(\varepsilon) = \exp\big(c' \ln^2(1/\varepsilon)/\beta\big)$, quasi-polynomial in $1/\varepsilon$, between every finite-dimensional class and generic nonparametrics, and free of d. The profile is a statement about the mean tail and costs nothing for being one: the crop pays each configuration its own tail and the bound takes the mean (Remark B.25). Since every design is exactly d-sparse, Theorem 4.4 at $K = d$ holds unconditionally and the bound is the minimum of the two branches, which cross at $\ln n = \beta(2d-1)^2/2$; on the measured flock, with d in the thousands, the crossover is unreachable at any rate the profiles of Section 5.1 show. Two qualifiers. The content of these bounds is the exponent and the absence of d, and the number at hand is the certificate's. And the rate is a single level's, Theorem 4.4 at its best K with the mean tail as bias: what the shells add is the across-configuration axis, where each shell pays its own rate at its own mass.*

Remark B.25 (The constants in the exponent)**.** *The constant $1/\sqrt{2}$ is the crop's. A trim at the diameter, whether Theorem B.14 without its crop clause at a constant profile or the finite-sample dimension of Weed & Bach (2019, Prop. 10) on the trimmed coverings of Appendix A.5, needs the tail to fall below the working threshold on all but the trimmed mass, which a mean tail gives only through Markov's inequality, at twice the cutoff: that route returns $\frac{1}{2}$ at best, and $\frac{1}{3}$ as computed (the shells route of Appendix B.6, at $n \ge n_0$ and $R \ge \frac{1}{2}$), and the factor is the trim's price, not the profile's; a pointwise bound $T_k \le A e^{-\beta k}$ at every configuration would hand the trim its $1/\sqrt{2}$ back, and nothing here needs one. The lower bound of Theorem D.6 has $\sqrt{2}$.*

B.7 THE ACCUMULATIVE CLASS

The theorems above are instances of Proposition B.11, and everything specific to the metric entered through four inputs: the stratum covering, the crop cost, the diameter of $(\mathcal{M}_d, D)$, and the qualitative facts behind Lemma B.9. This subsection instantiates the argument for the accumulative class $\widetilde{\mathcal{G}}_L$, the Lipschitz ball of $\widetilde{D}$, the matching distance with the shares $s_i = \mu_i/\mu_{\max}$ as heights. The lifts of the two instances are bijections from the same quotient of $\mathcal{C}$, so we regard $\widetilde{D}$ as a second metric on the same set $\mathcal{M}_d$, with T_k, $\mu_{\max}$ and the shells the same functions and sets, and write $|P| := \sum_i s_i = 1/\mu_{\max}(P)$ for the total height of P, the effective count, with the leader at height 1. Three inputs are imported from Appendix A.5. The crop cost: for $\hat{\mu}$ the k-crop, $\widetilde{D}(\Phi(x,\mu), \Phi(x,\hat{\mu})) \le R\,T_k/\mu_{\max} = \kappa T_k/(2\mu_{\max})$, and $\mu_{\max}(\hat{\mu}) = \mu_{\max}/(1-T_k) \ge \mu_{\max}$, the crop does not decrease the top weight (Proposition A.19 and its proof). The qualitative facts: $(\mathcal{M}_d, \widetilde{D})$ is compact, T_k and $\mu_{\max}$ are continuous on it, and duality holds for $\widetilde{W}_1$, the Kantorovich distance on measures on $\mathcal{M}_d$ with ground metric $\widetilde{D}$ (Proposition A.20); moreover $\widetilde{D}(P,Q) \le R(|P| + |Q|)$, by sending every atom of both to the apex, so for a fixed one-atom lift P_0,

$$\varrho_0(P) := \widetilde{D}(P, P_0) \;\le\; R\Big(\frac{1}{\mu_{\max}(P)} + 1\Big) \;\le\; \frac{\kappa}{\mu_{\max}(P)},$$

while $\operatorname{diam}(\mathcal{M}_d, \widetilde{D}) \le 2Rd$ and no better in general, the count being 1-Lipschitz for $\widetilde{D}$ and ranging over $[R, Rd]$. The celled covering: for $m \ge 0$ let $\mathcal{A}_k^{[m]} := \{P \in \mathcal{A}_k : \mu_{\max}(P) \ge 2^{-m}\} = \{P \in \mathcal{A}_k : |P| \le 2^m\}$; then for every $k \le d$, $m \ge 0$ and $\varepsilon > 0$, by the second display of Theorem A.18,

$$N\big(2^{m-1}\varepsilon,\ \mathcal{A}_k^{[m]},\ \widetilde{D}\big) \;\le\; \widetilde{N}_k(\varepsilon) \;:=\; N\big(\tfrac{\varepsilon}{4}, \mathcal{V}, \rho\big)^k \cdot k\Big(1 + \frac{16R}{\varepsilon}\Big)^{k-1},$$

with net points in $\mathcal{A}_k$: resolving the cell $\{|P| \le 2^m\}$ at scale $2^{m-1}\varepsilon$ costs the conservative count up to the factor k and a constant, independently of m. On interval values and $\varepsilon \le \kappa$, $\widetilde{N}_k(\varepsilon) \le (68R/\varepsilon)^{2k-1}$.

Definition B.26 (Concentration index, cells, concentration weights)**.** *For $P \in \mathcal{M}_d$ let $m(P) := \lceil \log_2(1/\mu_{\max}(P)) \rceil \in \{0, 1, \dots\}$, so that $1/\mu_{\max}(P) \le 2^{m(P)} < 2/\mu_{\max}(P)$; $m(P) = 0$ exactly on point masses. For the stratified shells Sh_k of Definition B.4 set*

$$\mathsf{Sh}_{k,m} := \mathsf{Sh}_k \cap \{m(P) = m\}, \qquad w_{k,m} := \Gamma(\mathsf{Sh}_{k,m}),$$

the first and second concentration weights *of shell k,*

$$\zeta_k := \sum_{m \ge 0} 2^m\, w_{k,m} = \int_{\mathsf{Sh}_k} 2^{m(P)}\, d\Gamma, \qquad \xi_k := \sum_{m \ge 0} 4^m\, w_{k,m},$$

and the count-weighted trim

$$H_K := \int_{B_K(\theta_K)^c} \frac{d\Gamma}{\mu_{\max}} \;\in\; \big[\, G_K(\theta_K),\ d\,G_K(\theta_K) \,\big].$$

Since $\mu_{\max} \ge (1-\theta_k)/k$ on Sh_k, only indices $m < \log_2(2k/(1-\theta_k))$ occur there, at most $M_k := \lceil \log_2(2k/(1-\theta_k)) \rceil$ cells per shell, and, from $1/\mu_{\max} \le 2^{m(P)} < 2/\mu_{\max}$,

$$w_k \;\le\; \zeta_k \;\le\; 2\int_{\mathsf{Sh}_k} \frac{d\Gamma}{\mu_{\max}} \;\le\; \frac{2k}{1-\theta_k}\, w_k, \qquad \xi_k \;\le\; 4\int_{\mathsf{Sh}_k} \frac{d\Gamma}{\mu_{\max}^2} \;\le\; \Big(\frac{2k}{1-\theta_k}\Big)^2 w_k.$$

The design's concentration moments *are $\bar{\zeta} := \mathbb{E}_Q[\mu_{\max}^{-1}]$ and $\bar{\xi} := \mathbb{E}_Q[\mu_{\max}^{-2}]$.*

Theorem B.27 (Generalization over the accumulative class)**.** *For every design Q and every resolution profile, with the shells and masses of Definition B.4 and the cells, weights and trim of Definition B.26,*

$$\tilde{\Delta}_n(Q) \;=\; \mathbb{E}\,\widetilde{W}_1(\hat{\Gamma}_n, \Gamma) \;\le\; 2\sum_{k=1}^{K} \zeta_k\, \varepsilon_k \;+\; \kappa \sum_{k=1}^{K} \sum_{m < M_k} 2^m \sqrt{\frac{w_{k,m}\, \widetilde{N}_k(\varepsilon_k)}{n}} \;+\; 2\kappa\, H_K.$$

Supplementary bounds: the floor is at most $4\sum_k \varepsilon_k \int_{\mathsf{Sh}_k} \mu_{\max}^{-1}\, d\Gamma$, hence at most $4\varepsilon\bar{\zeta}$ for the constant profile; the inner sum is at most $\sqrt{M_k\, \xi_k}$; and $H_K \le \sqrt{\bar{\xi}\, G_K(\theta_K)}$.

Proof. The identity is duality for $\widetilde{W}_1$ and Theorem 3.2 for the accumulative instance, as in the proof of Theorem B.14. Apply Proposition B.13 on $(\mathcal{X}, \varrho) = (\mathcal{M}_d, \widetilde{D})$ with reference point P_0 a point mass, $B = B_K(\theta_K)$, and the groups $S_{k,m} = \mathsf{Sh}_{k,m}$, $k \le K$, $m < M_k$, Borel, of masses $w_{k,m}$. *Reference costs:* on $S_{k,m}$, $1/\mu_{\max} \le 2^m$, so $\varrho_0 \le \kappa/\mu_{\max} \le 2^m\kappa =: \varrho_{k,m}$; on B^c, $\varrho_0 \le \kappa/\mu_{\max}$, so $\int_{B^c} \varrho_0 \, d\Gamma \le \kappa \, H_K$. *Cellings:* a point $P \in \mathsf{Sh}_{k,m}$ has $T_k(P) \le \theta_k$, so its k-crop $\hat{P}$ satisfies $\widetilde{D}(P, \hat{P}) \le \kappa\theta_k/(2\mu_{\max}(P)) \le 2^{m-1}\kappa\theta_k = 2^{m-2}\varepsilon_k$, with $\hat{P} \in \mathcal{A}_k$ and $\mu_{\max}(\hat{P}) \ge \mu_{\max}(P) \ge 2^{-m}$, that is $\hat{P} \in \mathcal{A}_k^{[m]}$; and $\mathcal{A}_k^{[m]}$ has a net of $\widetilde{N}_k(\varepsilon_k)$ points at radius $2^{m-1}\varepsilon_k \le 2^m\varepsilon_k - 2^{m-2}\varepsilon_k$, so Lemma B.12 gives a $(2^m\varepsilon_k, \widetilde{N}_k(\varepsilon_k))$-celling of $S_{k,m}$. With $r_{k,m} = 2^m\varepsilon_k$, $N_{k,m} = \widetilde{N}_k(\varepsilon_k)$ and $\varrho_{k,m} = 2^m\kappa$ the proposition's three terms are the display: $2\sum_{k,m} 2^m\varepsilon_k \, w_{k,m} = 2\sum_k \zeta_k\varepsilon_k$, then $\sum_{k,m} 2^m\kappa\sqrt{w_{k,m}\widetilde{N}_k/n}$, then $2\kappa H_K$. The supplementary bounds: $2^{m(P)} < 2/\mu_{\max}(P)$ for the floor; Cauchy–Schwarz over m for the inner sum, $\sum_m 2^m\sqrt{w_{k,m}} \le \sqrt{M_k \sum_m 4^m w_{k,m}}$; and Cauchy–Schwarz for $H_K \le \sqrt{\int \mu_{\max}^{-2} \, d\Gamma} \, \sqrt{\Gamma(B_K^c)}$. □

Against Theorem B.14 on the same design and the same shells, the profile, the shells and the masses are identical: the accumulative class does not induce a different partition, it is the same partition with cell (k, m) resolved at scale $2^m\varepsilon_k$ instead of ε_k. The counts are identical in shape. The price of the semantics is a count-weighting of all three terms, the floor by the mean inverse top weight on the bulk, the fluctuation by the second concentration moment, the trim by the count of the trimmed configurations. This is forced, not slack: the count $R/\mu_{\max}$ is a member of $\widetilde{\mathcal{G}}_1$ (Proposition A.14), so two configurations of counts c and c' are at $\widetilde{D}$-distance at least $R|c - c'|$, and mass misplaced across counts must pay for the difference. In the conservative class the total height is 1 everywhere and all three weights are 1.

Corollary B.28 (Exact sparsity, accumulative class, the single-partition route)**.** *If $Q(T_{k_0} = 0) = 1$ then $1/\mu_{\max} \le k_0$ Q-a.s., $\bar{\zeta} \le k_0$, and with $K = k_0$ and the constant profile, for every $\varepsilon > 0$,*

$$\tilde{\Delta}_n(Q) \;\le\; 4\varepsilon\,\bar{\zeta} + 2k_0\kappa\sqrt{\frac{k_0\, M\, \widetilde{N}_{k_0}(\varepsilon)}{n}}, \qquad M := \lceil \log_2(2k_0) \rceil.$$

On interval values, optimizing ε, for $n \ge 68^{2k_0+1}k_0^3$, so that the optimizing ε is at most κ,

$$\tilde{\Delta}_n(Q) \;\le\; C\,R\,\bar{\zeta}\, n^{-\frac{1}{2k_0+1}} \;\le\; C\,R\,k_0\, n^{-\frac{1}{2k_0+1}}$$

with C absolute.

Proof. $G_{k_0}(\theta) \le G_{k_0}(0) = 0$ kills the trim, $H_{k_0} = 0$. On the support of Q at most k_0 weights are positive, so $1/\mu_{\max} \le k_0$ there, $2^m < 2k_0$ on every occupied cell, at most M values of m occur, and the fluctuation term is at most $2k_0\kappa\sum_{k,m}\sqrt{w_{k,m}\widetilde{N}_k/n} \le 2k_0\kappa\sqrt{k_0 M\, \widetilde{N}_{k_0}(\varepsilon)/n}$, by Cauchy–Schwarz over the at most k_0M occupied cells, $\sum_{k,m} w_{k,m} = 1$, and monotonicity of the counts in k. For the rate insert $M\,\widetilde{N}_{k_0}(\varepsilon) \le (C_1R/\varepsilon)^{2k_0-1}$ with $C_1 = 136$, from $\widetilde{N}_{k_0}(\varepsilon) \le (68R/\varepsilon)^{2k_0-1}$ and $M \le 2k_0 \le 2^{2k_0-1}$, and choose

$$\varepsilon \;:=\; C_1R\left(\frac{k_0^3}{\bar{\zeta}^2 n}\right)^{\frac{1}{2k_0+1}} \;\le\; \kappa;$$

both terms are then at most $4(C_1+1)R\,(k_0^3)^{1/(2k_0+1)}\,\bar{\zeta}^{(2k_0-1)/(2k_0+1)}\,n^{-1/(2k_0+1)}$, using $\kappa = 2R$; conclude with $\bar{\zeta} \ge 1$ and $\sup_{k\ge1}(k^3)^{1/(2k+1)} \le 2$. □

Lemma B.29 (Concentration moments from the profile)**.** *Let $A \ge 1$, $\beta > 0$, and $k_1 := \lceil \beta^{-1}\ln(2A) \rceil \vee 1$.*

(i) If $\mathbb{E}_Q[T_k] \le Ae^{-\beta k}$ for all k, then $\bar{\zeta} \le 2k_1 + 2 + 2/\beta$, and more generally $\bar{\zeta} \le 2k + 4Ae^{-\beta k}/(1 - e^{-\beta})$ for every $k \ge 1$.

(ii) Under the same hypothesis, $\bar{\xi} \le 4k_1^2 + 4(2k_1+1)(1+1/\beta) + 8/\beta^2$, so $\sqrt{\bar{\xi}} \le 4k_1 + 2 + 4/\beta$; and for every $K \ge 1$ and θ,

$$\int_{B_K(\theta)^c} \frac{d\Gamma}{\mu_{\max}} \;\le\; 2K\,G_K(\theta) + \frac{4Ae^{-\beta K}}{1 - e^{-\beta}}.$$

The concentration moments are constants of order $\beta^{-1}(1+\ln 2A)$, the effective neighbor count, and its square, independent of d and of n.

Proof. For any $k \geq 1$, $\mu_{\max} \geq (1-T_k)/k$, so $\{1/\mu_{\max} > 2k\} \subseteq \{T_k > \frac{1}{2}\}$, and Markov gives $\Gamma(1/\mu_{\max} > 2k) \leq 2Ae^{-\beta k}$. (i) Layer cake: $\mathbb{E}[\mu_{\max}^{-1}] \leq 2k + \sum_{j\geq k} 2\Pr(\mu_{\max}^{-1} > 2j) \leq 2k + 4A\sum_{j\geq k} e^{-\beta j}$. At $k = k_1$, $4Ae^{-\beta k_1} \leq 2$ and $(1-e^{-\beta})^{-1} \leq 1 + 1/\beta$. (ii) For $t := 1/\mu_{\max}$ and any integer $j_0 \geq 1$, pointwise $t \leq 2j_0 + 2\sum_{j\geq j_0} \mathbf{1}\{t > 2j\}$ and $t^2 \leq 4j_0^2 + \sum_{j\geq j_0}(8j+4)\,\mathbf{1}\{t > 2j\}$. Integrating the second against Γ at $j_0 = k_1$: $\bar{\xi} \leq 4k_1^2 + 8A\sum_{j\geq k_1}(2j+1)e^{-\beta j} \leq 4k_1^2 + 8Ae^{-\beta k_1}\big[(2k_1+1)(1+1/\beta) + 2/\beta^2\big]$, using $\sum_{i\geq 0} e^{-\beta i} \leq 1 + 1/\beta$ and $\sum_{i\geq 0} ie^{-\beta i} = e^{-\beta}/(1-e^{-\beta})^2 \leq 1/\beta^2$, and $8Ae^{-\beta k_1} \leq 4$; the square root by $2\sqrt{ab} \leq a+b$. Integrating the first against Γ restricted to $B_K(\theta)^c$ at $j_0 = K$ gives the last display. □

Corollary B.30 (Exponential profiles, accumulative class, the single-partition route)**.** *Let $\mathcal{V} = [-R, R]$ with $R \geq \frac{1}{2}$, let A and β be as in Theorem 4.5, $\Lambda' := 136R \vee 2\kappa A \vee e$ as in the shells route of Appendix B.6, k_1 as in Lemma B.29, and let $n_0 := \exp(48\beta \vee 9\ln^2 \Lambda'/\beta)$. Then for $n \geq n_0$, under an exponential profile at rate β, $\tilde{\Delta}_n(Q) \leq 32\,(k_1 + 1 + 1/\beta)\,\Lambda'\, e^{-\frac{1}{3}\sqrt{\beta \ln n}}$.*

Proof. Run the shells route of Appendix B.6 on Theorem B.27 with the constant profile: $\varepsilon := \Lambda' e^{-u}$, $\theta := \varepsilon/(2\kappa) \leq \frac{1}{2}$ (as $\varepsilon \leq 1 \leq \kappa$), $u = \frac{1}{3}\sqrt{\beta L}$, $L = \ln n$; the consequences (a)–(c) hold verbatim, and $M_k\, \widetilde{N}_k(\varepsilon) \leq (C_1 R/\varepsilon)^{2k-1} \leq e^{(2k-1)u}$ with $C_1 = 136$, since $M_k \leq \lceil \log_2 4k \rceil \leq 2k$ for $\theta \leq \frac{1}{2}$.

Take $K := (\lceil 3u/\beta \rceil \vee 1) \wedge d$ as there, so that either $K = d$ or $\beta K \geq 3u$ and $G_K(\theta) \leq e^{-2u}$. *Floor:* at most $4\varepsilon\bar{\zeta} = 4\Lambda'\bar{\zeta}e^{-u}$. *Fluctuation:* by the supplementary bound of the theorem and $\xi_k \leq 4\bar{\xi}$ it is at most $2\kappa\sqrt{\bar{\xi}}\sum_{k\leq K}\sqrt{M_k \widetilde{N}_k(\varepsilon)/n} \leq 2\kappa\sqrt{\bar{\xi}}\, n^{-1/2}\sum_{k\leq K} e^{(k-\frac{1}{2})u} \leq 4\kappa\sqrt{\bar{\xi}}\, e^{(K-\frac{1}{2})u - L/2}$, a geometric sum with ratio $e^u \geq e$, and $(K - \frac{1}{2})u \leq 3u^2/\beta + u/2 = L/3 + u/2$, so the exponent is at most $-L/6 + u/2 \leq -u$ as $u \leq L/20$: the term is at most $4\Lambda'\sqrt{\bar{\xi}}\, e^{-u}$. *Trim:* if $K < d$, Lemma B.29(ii) gives $2\kappa H_K \leq 4\kappa K e^{-2u} + 8\kappa A(1+1/\beta)e^{-\beta K} \leq 4\Lambda'(3u/\beta + 1)e^{-2u} + 4\Lambda'(1 + 1/\beta)e^{-3u}$, using $\kappa \leq \Lambda'$, $2\kappa A \leq \Lambda'$, $K \leq 3u/\beta + 1$ and $\beta K \geq 3u$; since $ue^{-u} \leq e^{-1}$ and $e^{-u} \leq e^{-1}$, this is at most $4\Lambda' e^{-u}\big[(3e^{-1} + e^{-2})/\beta + 2e^{-1}\big] \leq 8\Lambda'(1+1/\beta)e^{-u}$. If $K = d$ the trim is zero. *Total:* at most $4\Lambda' e^{-u}\,[\bar{\zeta} + \sqrt{\bar{\xi}} + 2 + 2/\beta]$, and Lemma B.29(i)–(ii) give $\bar{\zeta} + \sqrt{\bar{\xi}} \leq 6k_1 + 4 + 6/\beta$, so the bracket is at most $8(k_1 + 1 + 1/\beta)$. □

Corollary B.31 (Exact sparsity, accumulative class, the multiscale route)**.** *Let $\mathcal{V} = [-R, R]$. If $Q(T_{k_0} = 0) = 1$ with $k_0 \geq 2$, then*

$$\tilde{\Delta}_n(Q) \;\leq\; C\, R\, k_0\, n^{-\frac{1}{2k_0 - 1}}, \qquad C := 17 \cdot 408 = 6936.$$

Conversely $\tilde{\Delta}_n(Q) \geq \frac{1}{2}\Delta_n(Q)$ on every design, since $\mathcal{G}_1 \subseteq \widetilde{\mathcal{G}}_2$ (Proposition A.14), so on the stratum design Q_{k_0} of Lemma D.4 the exponent $1/(2k_0 - 1)$ holds from both sides.

Proof. On the support of Q at most k_0 weights are positive, so $1/\mu_{\max} \leq k_0$ there and every lift lies in $S := \mathcal{A}_{k_0}^{[M]}$ with $M := \lceil \log_2 k_0 \rceil$, so that $k_0 \leq 2^M < 2k_0$. Its diameter: $\widetilde{D}(P, P') \leq R(|P| + |P'|) \leq 2Rk_0 =: \varepsilon_0$. Its covering: at a radius $\varrho \leq 2^{M-1}\kappa$ put $\varepsilon := \varrho/2^{M-1} \leq \kappa$, so that $N(\varrho, S, \widetilde{D}) \leq \widetilde{N}_{k_0}(\varepsilon) \leq (68R/\varepsilon)^{2k_0-1} = (68R\,2^{M-1}/\varrho)^{2k_0-1} \leq (68Rk_0/\varrho)^{2k_0-1}$. Apply Proposition B.18 on S with $\varepsilon_j := 3^{-j}\varepsilon_0$: the radii $\varepsilon_{j+1}/2 = \varepsilon_j/6 \leq \varepsilon_0/6 = Rk_0/3 \leq 2^{M-1}\kappa$ are admissible, and $N(\varepsilon_{j+1}/2, S, \widetilde{D}) \leq (6 \cdot 68Rk_0/\varepsilon_j)^{2k_0-1} = (A/\varepsilon_j)^e$ with $A := 408Rk_0 \geq \varepsilon_0$ and $e := 2k_0 - 1 \geq 3$. The proposition gives $\mathbb{E}\,\widetilde{W}_1(\hat{\Gamma}_n, \Gamma) \leq 17A\, n^{-1/e}$. The converse: a member of $\mathcal{G}_1$ is a member of $\widetilde{\mathcal{G}}_2$, so $\sup_{\widetilde{\mathcal{G}}_1}(\mathbb{E}_Q f - \mathbb{E}_n f) \geq \frac{1}{2}\sup_{\mathcal{G}_1}(\mathbb{E}_Q f - \mathbb{E}_n f)$ pointwise in the sample, and Lemma D.4 bounds the expectation of the right side from below on Q_{k_0}. □

Corollary B.32 (Exponential profiles, accumulative class, the crop route)**.** *Let $\mathcal{V} = [-R, R]$ and let Q have an exponential profile at rate β with constant $A \geq 1$. Then for every $n \geq 2$ and $d \geq 2$,*

$$\tilde{\Delta}_n(Q) \;\leq\; C\, R\,\big[A(1 + 1/\beta) + e^{3\beta/2}\big]\Big(\sqrt{\tfrac{\ln n}{2\beta}} + 3\Big)\, e^{-\sqrt{\beta \ln n/2}}, \qquad C = 6936:$$

the exponent of Theorem 4.5, with the effective count at the cutoff, $K_n \asymp \sqrt{\ln n/\beta}$, as a factor.

Proof. The crop. For $K \le d$ let cr_K be the K-crop, with $\widetilde{D}(P, \mathrm{cr}_K P) \le R\, T_K(P)/\mu_{\max}(P)$ (Proposition A.19), and $\Gamma^K := (\mathrm{cr}_K)_\# \Gamma$, the lift of an exactly K-sparse design. As in the proof of Theorem 4.4, coupling along the crop on both sides,

$$\tilde{\Delta}_n(Q) \;\le\; 2R\,\mathbb{E}_Q\Big[\frac{T_K}{\mu_{\max}}\Big] + \mathbb{E}\,\widetilde{W}_1\big((\mathrm{cr}_K)_\#\hat{\Gamma}_n, \Gamma^K\big),$$

and once $K \ge 2$ the second term is at most $6936\, RK\, n^{-1/(2K-1)}$ by Corollary B.31, the cropped lifts having K atoms of share at most 1. *The crop cost.* On $\{T_K \le \frac{1}{2}\}$, $\mu_{\max} \ge (1 - T_K)/K \ge 1/(2K)$, so pointwise $T_K/\mu_{\max} \le 2K\, T_K + \mu_{\max}^{-1}\mathbf{1}\{T_K > \frac{1}{2}\}$, and Lemma B.29(ii) at $\theta = \frac{1}{2}$, with $G_K(\frac{1}{2}) \le 2\,\mathbb{E}_Q T_K \le 2Ae^{-\beta K}$ by Markov, gives

$$\mathbb{E}_Q\Big[\frac{T_K}{\mu_{\max}}\Big] \;\le\; 2KAe^{-\beta K} + 4KAe^{-\beta K} + \frac{4Ae^{-\beta K}}{1 - e^{-\beta}} \;\le\; Ae^{-\beta K}\,(6K + 4 + 4/\beta),$$

and $6K + 4 + 4/\beta \le 6(1 + 1/\beta)(K + 1)$. *The cutoff.* As in the proof of Theorem 4.5, put $x := \sqrt{2 \ln n/\beta}$ and $K := d \wedge (\lceil x/2 \rceil \vee 2)$, so that $\beta K \ge \sqrt{\beta \ln n/2}$ unless $K = d$, where $T_d = 0$ and the crop costs nothing, that $2K - 1 \le x + 3$, whence $n^{-1/(2K-1)} \le e^{-\ln n/(x+3)} \le e^{3\beta/2} e^{-\sqrt{\beta \ln n/2}}$, and that $K + 1 \le x/2 + 3$. The two terms are then at most $12RA(1 + 1/\beta)(K + 1)\, e^{-\sqrt{\beta \ln n/2}}$ and $6936\, R\, e^{3\beta/2}(K + 1)\, e^{-\sqrt{\beta \ln n/2}}$, and $12 \le 6936$. □

B.8 The two caps compared, and the accumulative certificate

Against Theorem 4.4 and Theorem 4.5, the multiscale and crop routes give the accumulative class the same exponents (Corollaries B.31 and B.32), and the accumulative price is a factor of the count, paid once: the maximal count k_0 at exact sparsity, and the effective count at the cutoff, $\sqrt{\ln n/(2\beta)} + 3$, on exponential profiles. The single-partition route of Theorem B.27 gives the weaker exponents $1/(2k_0 + 1)$ and $1/3$ (Corollaries B.28 and B.30), with the mean count $\bar{\zeta}$ and a constant of order $\beta^{-1}(1 + \ln 2A)$ in place of the maximal count, the profile bounding the concentration moments (Lemma B.29). This is the two semantics of Section 3.1 as arithmetic. On a hard selection of k coordinates the accumulative class is the conservative class at scale kL (Corollary A.15), and the design pays that scale.

Remark B.33 (The accumulative certificate at a truncated diameter)**.** *Theorem 4.6 needs a diameter for its concentration step, and $(\mathcal{M}_d, \widetilde{D})$ has $2Rd$ and no better, the count being a member of the class. Let $h(P) := \sum_i s_i = 1/\mu_{\max}(P)$ be the total height, so that $\widetilde{D}(P, P') \le R(h(P) + h(P'))$ through the reservoir. Fix $\zeta > 0$, let $\tau := Q(h > \zeta)$ and $\Gamma_\zeta := \Gamma(\cdot \mid h \le \zeta)$. On the event that all n sample points have $h \le \zeta$, of probability at least $1 - n\tau$, the points are i.i.d. from Γ_ζ, whose support has $\widetilde{D}$-diameter at most $2R\zeta$, so Theorem 4.6 applies with $\kappa_\zeta := 2R\zeta$ in place of κ and Γ_ζ in place of Γ; and $\widetilde{W}_1(\Gamma_\zeta, \Gamma) \le R\tau\,(\zeta + \mathbb{E}[h \mid h > \zeta])$, by coupling the heavy mass to light points through the reservoir. Hence, with probability at least $1 - \delta - n\tau$,*

$$\widetilde{W}_1(\hat{\Gamma}_A, \Gamma) \;\le\; \widetilde{W}_1(\hat{\Gamma}_A, \hat{\Gamma}_B) + 2R\zeta\sqrt{\frac{\ln(1/\delta)}{n}} + R\tau\big(\zeta + \mathbb{E}[h \mid h > \zeta]\big).$$

The level comes from the profile, by the first line of the proof of Lemma B.29: $\mu_{\max} < 1/(2k)$ forces $T_k > \frac{1}{2}$, so $Q(h > 2k) \le Q(T_k > \frac{1}{2})$, and Markov gives $Q(h > 2k) \le 2Ae^{-\beta k}$, and $\zeta_n := 2\lceil \beta^{-1} \ln(2An/\delta) \rceil$, at most $(2/\beta)\ln(2An/\delta) + 2$, makes $n\tau \le \delta$, the display then holding with probability at least $1 - 2\delta$, with a tail term of order $\delta\zeta_n/n$; the price is the factor ζ_n on the concentration term, the effective count at the sample size, logarithmic in n, the quantity the bounds in expectation carry as $\bar{\zeta}$. The trim that works is on the heavy configurations, not on the light atoms: cropping a crowded configuration to K atoms costs R times the removed height, cheap in the conservative class where the height is the weight and expensive here where it is the share. The cell bound of Proposition B.5 and the exact matching need no diameter, a trimmed sample point being priced at its own observed height.

C EXAMPLES

C.1 FLOCKING

The protocol below stores a sample of Section 5.1, one focal particle at one instant, as a *view*, the term of the legend of Figure 2.

The dynamics. Particles $i = 1, \dots, d$ carry positions $p_i \in \mathbb{T}^2_\ell$, the torus of side ℓ, and headings $\theta_i \in S^1$. Writing $u(\theta) = (\cos\theta, \sin\theta)$, one step of the drift of the dynamics is

$$\theta_i' \;=\; \arg \sum_j c_{ij}\, e^{\mathrm{i}\theta_j}, \qquad p_i' \;=\; p_i + v_0\, u(\theta_i'),$$

where $c_{i\cdot} \in \Delta_d$ are the neighborhood weights of particle i, the selector, computed from the distances $|p_j - p_i|$. The classical model (Vicsek et al., 1995) takes $c_{i\cdot}$ uniform on $\{j : |p_j - p_i| \le r\}$, the metric interaction on a hard ball. Two variants are in the literature, the second in our roster: the *topological* interaction, uniform on the particle itself and its k nearest neighbors, the form supported by the starling observations of Ballerini et al. (2008) with $k \approx 7$, and *soft kernels*, $c_{ij} \propto K(|p_j - p_i|)$.

Convention C.1 (Noise)**.** *The stochastic term, angular noise added to θ_i', is exogenous: the mechanism is the drift map above, the noise is added afterwards, and all statements are about the drift.*

The four flocks. Every flock has $d = 3200$ particles on the torus of side $\ell = 40$, density 2, at speed $v_0 = 0.5$, and the flocks differ in the rule. The ball rule is the one above at $r = 1$. The three graded rules have the same drift with soft weights $c_{ij} \propto K(|p_j - p_i|)$: the bump $K(t) = (1 - t^2/r_b^2)_+^2$ at $r_b = 1.4$, the Gaussian $K(t) = e^{-t^2/2h^2}$ at $h = 0.5$ truncated at radius 3.5, and the Cauchy $K(t) = (1 + t^2/h^2)^{-3}$ at $h = 0.9$ truncated at radius 4, beyond which lies 0.2% of the untruncated kernel's weight. The range of each graded rule is the one at which its mean tail matches the ball's at ten atoms, T_{10} between 0.16 and 0.19 for every flock in both regimes, so the four rules have the same effective number of partners. Each flock is read through the selector of its own rule at the rule's parameters: the lift of a sample is the weights the flock's own update puts on it, and the mechanism whose class the certificate bounds is that update. The two regimes are $\eta = 1.2$, the main text's, and $\eta = 2.4$, both in the ordered phase (Chaté et al., 2008).

The protocol. The protocol is the same for every flock and regime: a flock, a list of seeds, one replica per seed, and every replica viewed at the same steps. A replica draws uniformly random positions and headings from its seed and runs the drift with its noise, $64{,}000$ replicas per flock and regime. A view is one focal particle, drawn uniformly at random at a fixed step of its replica, with its neighborhood within radius 3.5 (4 on the Cauchy flock) stored as positions relative to the focal particle and headings rotated by an independent uniformly random angle, the self atom included at position 0: the design certified is the law of the rotated view, a design of the same mechanism, which is equivariant under the rotation. Every replica is viewed at step 3000, after the burn-in, by which the flock has ordered (the polarization below), and the focal particle and the rotation of a view are drawn from the replica's generator after its last step. One view per replica: the views are independent samples of the flock's law at that age, the i.i.d. hypothesis of Theorem 4.6 holds as stated, and nothing is assumed about mixing within a run.

The split, the lift and the certificate. Half A is the replicas of the first $32{,}000$ seeds and half B those of the last $32{,}000$, a split fixed before any data were seen, and the nested sample sizes are the prefixes of the seed order, $n = 500$ to $64{,}000$ samples in doublings, both halves counted. The lift of a view is the selector's weights on its stored atoms, normalized to sum to one, and on the Gaussian and the Cauchy flock the lift is cropped to the heaviest atoms carrying all but 10^{-3} of the weight, at the cost of Proposition A.19 per sample, paid in the certificate by the triangle inequality: 0.0017κ and 0.0019κ at the largest n. Every distance between a lift of A and a lift of B is the assignment of Lemma B.3, and the transport is one exact assignment on the $32{,}000 \times 32{,}000$ matrix of them, by the Jonker–Volgenant algorithm, at every prefix. The certificate is that of Theorem 4.6 for half A at $\delta = 0.05$, the transport plus the crop cost plus $\kappa\sqrt{\ln(1/\delta)/n}$, and the whole-sample certificate of Theorem B.1(ii), with its one term at $\delta/2$, is larger by 0.0008κ at the largest n.

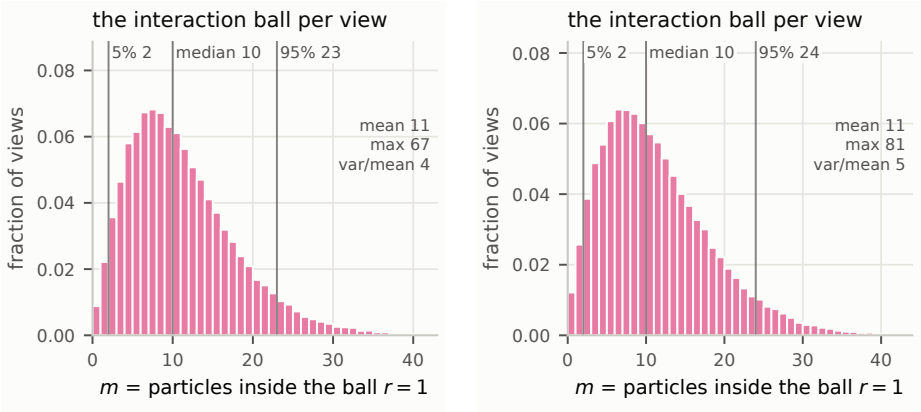

Figure 2: The count law of the ball flock after the burn-in: the number m of particles inside the ball $r = 1$ other than the self, per sample, on the 64,000 replicas at *(left)* $\eta = 1.2$, those of Figure 1, and *(right)* $\eta = 2.4$, those of Figure 3.

Table 2: The two-sample transport $W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ and the certificate of Theorem 4.6 for half A at $\delta = 0.05$, in units of κ, at $n = 64{,}000$ samples (32,000 per half), each flock through the selector of its own rule, after the burn-in of 3000 steps, in both regimes, with the cell certificate of Proposition B.5 on the same halves at the same δ, its shells, resolution and cutoff optimized on the data (the covering script of the supplementary material). The trivial value is κ, and the mean distance D between two random samples is 0.56 to 0.68κ.

	$\eta = 1.2$			$\eta = 2.4$		
flock	W_1	cert.	cell cert.	W_1	cert.	cell cert.
ball, $r = 1$	0.0272	0.0340	0.681	0.0435	0.0503	0.881
bump, $r_b = 1.4$	0.1057	0.1125	1.005	0.1386	0.1454	1.004
Gaussian, $h = 0.5$	0.1152	0.1238	1.006	0.1498	0.1583	1.004
Cauchy, $h = 0.9$, $a = 3$	0.1205	0.1292	1.006	0.1555	0.1643	1.005

The simulator and the seeds. The simulator is a compiled implementation of the drift above, exact against the reference implementation step by step, and a matching takes from an hour for the ball flock to five hours for the Cauchy flock on a many-core machine, the assignment on the full matrix half an hour of it. The seeds are consecutive integers, one range per flock and regime, stored with every replica and listed in the supplementary material, which regenerates every table and figure of this appendix and of Section 5.1 from them.

The count law and the profile. Figure 2 is the count law of the ball flock after the burn-in in both regimes, the number m of particles inside the ball other than the self. At $\eta = 1.2$ the ball holds 10.7 particles on average and 67 at most, from 2 to 23 between the 5% and 95% samples, with a variance 4.3 times the mean, and at $\eta = 2.4$ it holds 10.9 on average and 81 at most, from 2 to 24, with a variance 4.6 times the mean. The survival $P(m \geq k)$ is geometric beyond the knee, at 0.20 per atom on $15 \leq k \leq 50$ in both regimes. Every sample of the ball flock is exactly sparse with uniform weights on its $m+1$ atoms and tail $T_k = (1 - k/(m+1))_+$, so the ball's mean tail in Figure 1(a) is the functional $\mathbb{E}_Q T_k = \sum_m P(m)\,(1 - k/(m+1))_+$ of the count law, and it is geometric beyond the knee because the count law is, at a decay rate of 0.23 per atom. The graded rules are the other case, a decay inside each sample at a rate set by the local crowding. Their profiles are Figure 1(a) at $\eta = 1.2$ and Figure 3(a) at $\eta = 2.4$: matched to the ball's at ten atoms, the bump's tail is the ball's shape on a wider support with the boundary atoms discounted, the Gaussian's is geometric and the Cauchy's polynomial, above the Gaussian's from the seventh atom on. At the one-percent level the selectors resolve 24 (ball), 27 (bump), 32 (Gaussian) and 57 (Cauchy) atoms at $\eta = 1.2$, and 24, 29, 35 and 66 at $\eta = 2.4$. The profiles of the two regimes are close, the high-noise flock's tails a little heavier.

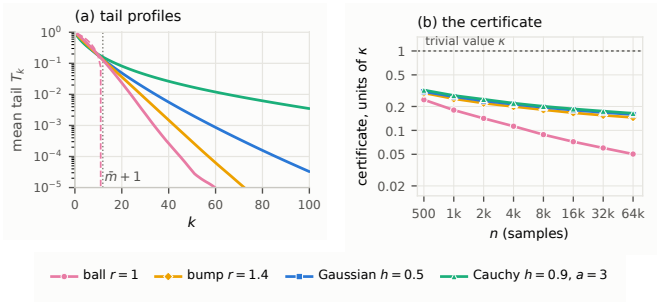

Figure 3: The flock at $\eta = 2.4$, the twin of Figure 1: 64,000 independent replicas per flock, one sample each. *(a)* The profile, the mean tail T_k under the four selectors. *(b)* The certificate of Theorem 4.6 for half A at $\delta = 0.05$, in units of κ, after the burn-in of 3000 steps.

The certificate. Table 2 is the certificate of Theorem 4.6 at the largest n on every flock, in both regimes, after the burn-in, with the transport it is built on, and Figure 1(b) and Figure 3(b) are the curves against n. Four readings. *The bound is non-vacuous by a wide margin* on every flock and regime: the largest value in the table is below 0.17κ, against the trivial value κ and a mean distance between two random samples of 0.56 to 0.68κ, and on the ball at $\eta = 1.2$ the certificate is 0.034κ, a thirtieth of the trivial value. *The graded flocks sit together*, at four times the ball's transport at $\eta = 1.2$ and three times at $\eta = 2.4$: the weight mismatch between two samples is a cost the ball's uniform weights never pay, and it depends little on how the grading is done, the three graded certificates lying within 0.02κ of one another in both regimes. *The matching is close to its floor.* The mean distance from a sample of A to its nearest sample of B is 0.95 to 0.96 of the matching cost on the graded flocks and 0.69 (0.79 at $\eta = 2.4$) on the ball, 0.0188κ at $\eta = 1.2$, so what remains is the sample size and not slack in the transport. *The covering certificates sit at the trivial value.* The cell certificate of Proposition B.5 on the cells of the quantized k-crops of the two halves (on the ball, where the weights tie, the crop takes the k headings nearest the view's circular mean), the resolution floor, the observed discrepancy of the halves' cell frequencies and the trim, with the concentration term of Theorem 4.6, at the shells, the resolution and the cutoff $K \le 24$ that are best on the data, is 0.68κ on the ball at $\eta = 1.2$ and 0.88κ at $\eta = 2.4$, twenty and eighteen times the matching certificate, and at or above κ on every graded kernel at every n, where the continuous weights make almost every occupied cell a singleton, and the stratified bound of Theorem B.14 at a fixed covering constant, from the sample's own tails with the shells optimized, is above 0.98κ on every flock and regime. The comparison is between forms of the same bound on the same class, the same halves and the same δ: every cell bound is one feasible transport and the matching is below every one of them (Proposition B.5), and a celling optimized on the sample is a certificate on the same terms (Remark B.8). Not computed is a chained form, the entropy integral over the scales, or the doubling-dimension covers of Gottlieb et al. (2017), which could sit between the two.

The implied dimension. The *implied dimension* of a measured transport is $-1/(\mathrm{d}\ln W_1/\mathrm{d}\ln n)$, the exponent it shows at the sample sizes attained, read here as the reciprocal of the least-squares slope of $\ln W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ against $\ln n$ over a window of three consecutive doublings, at the middle n. No constant enters: the transport is computed exactly and the slope is free of units. On interval values an exactly k-sparse design with $k \ge 2$ shows at most $2k - 1$ as n grows and the stratum design of Appendix D at least $2k - 1$ (Theorem 4.4), as the bipartite matching on the d-cube decays at $n^{-1/d}$ (von Luxburg & Bousquet, 2004, Thm. 15). In general the number is the dimension of the lifted design at the scales the sample resolves. Figure 4 draws it for the four selectors after the burn-in in both regimes. Over the last three doublings the ball's transport decays with the exponent of a 3-dimensional sample at $\eta = 1.2$ and of a 5-dimensional one at $\eta = 2.4$, the graded flocks' with 13 to 15: about ten dimensions of weights at the sample's scale, against the $k - 1 \approx 10$ that the stratum count charges at the mean support. The dimensions are local. On the graded flocks the exponent rises with n, from 5 to 8 at $n = 1000$ to 13 at $n = 32{,}000$ at $\eta = 1.2$, about one dimension per doubling and decelerating: at a finer scale the sample resolves more atoms of the weight vector, the cutoff $K_n \asymp \sqrt{\ln n/\beta}$ of Theorem 4.5 read off the data, and the number is the dimension of the lifted design at the scales the sample resolves, not a limit. On the ball it is between 2 and 3.5

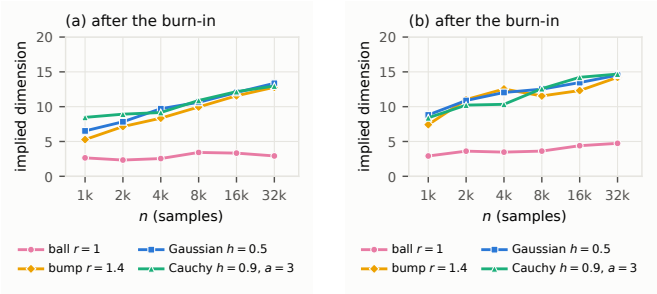

Figure 4: The implied dimension, the reciprocal of the local slope of $\ln W_1(\hat{\Gamma}_A, \hat{\Gamma}_B)$ against $\ln n$ over three consecutive doublings, at the middle n, after the burn-in: *(a)* on the replicas of Figure 1 at $\eta = 1.2$ and *(b)* on those of Figure 3 at $\eta = 2.4$. The curves are the transports behind panel (b) of the two figures.

at every n at $\eta = 1.2$, where the support has up to 68 atoms, and the more ordered flock is the lower-dimensional for every selector.

The two regimes. Figure 3 is Figure 1 at $\eta = 2.4$, on the replicas of that regime. The profile is nearly the same, the transport is larger, 0.044κ on the ball against 0.027κ: the high-noise flock is less ordered, with a local polarization of 0.78 against 0.94 (below), and the bound sees that too.

The certificate on a member with zero training risk. For a target $f^* \in \mathcal{G}_L$ observed on half A, the lower Lipschitz extension $f_-(z) := \max_{i \in A}(f^*(z_i) - L\, D(z, z_i))$ agrees with f^* on A and is L-Lipschitz (von Luxburg & Bousquet, 2004, Lemma 7), an empirical risk minimizer of $\mathcal{G}_L$ with zero training risk, and on the target $f^* = L\, D(\cdot, A)$, which is zero on A, it is $-L\, D(\cdot, A)$: its error at z is $2L\, D(z, A)$ and its gap on half B is exactly $2L\, \overline{\mathrm{NN}}$, twice the mean nearest-neighbor distance from B to A. By Remark B.22 the certificate bounds the gap of every empirical risk minimizer of $\mathcal{G}_L$ on every target in $\mathcal{G}_L$ by $2L$ times its value, so on this member the ratio of the bound to the gap is the certificate over $\overline{\mathrm{NN}}$: 1.8 on the ball at $\eta = 1.2$ (0.034 over 0.0188) and 1.5 at $\eta = 2.4$, and 1.10 to 1.13 on the graded kernels in both regimes, where the matching is at its nearest-neighbor floor (Table 2 and the floor above). The bound uniform over the class is within that factor of the truth on this member, and a member is what a uniform bound is judged by. For one chosen learner on one target the gap can be far smaller, the target's own smoothness at the neighbor scale, which only the labels see (Ashlagi et al., 2021), and a holdout estimate on half B reads it: the certificate competes with nothing there. What it gives is the bound before any learner is chosen or label seen, at once for every target and predictor in the class, so that a choice made after seeing the data costs nothing (Theorem B.1(ii)).

The bandwidth sweep. The graded kernels of Section 5.1 are at the ball's own scale: the Gaussian's $h = 0.5$ is the second-moment match to the disc of radius 1 (the disc has mean squared distance $\frac{1}{2}$, the kernel $2h^2$), and the bump's radius and the Cauchy's bandwidth match the ball's tail at ten atoms. Their certificates lie within 0.02κ of one another, four times the ball's: the price of a design of dimension thirteen against one of dimension three. What the bound reads is the profile, and the bandwidth sets the profile, so we sweep it rather than choose it: the Gaussian rule at $h \in \{0.125, 0.18, 0.25, 0.35, 0.5\}$, each flock simulated under its own rule as in the protocol, on 8,000 replicas per bandwidth read at 4,000 samples per half and on 32,000 replicas read at 16,000 per half, and the flock of Section 5.1 at the same prefixes (Table 3). The certificate follows the profile, slowly: halving the bandwidth sharpens the profile by a factor of nearly three, from 32 to 12 atoms carrying all but one percent of the weight, lowers the implied dimension from 10 to 7, and lowers the certificate by a tenth at this n, and a quarter of the bandwidth, at 9 atoms and dimension 6, lowers it by three tenths, monotone in the bandwidth over the five values. At 16,000 samples per half every certificate is lower by 15 to 26%, 0.133κ, 0.128κ, 0.113κ, 0.097κ and 0.081κ at $h = 0.5$, 0.35, 0.25, 0.18 and 0.125, against the ball's 0.043κ at that n. The weights the nearest atoms carry vary continuously with their positions whatever the bandwidth, and those are the coordinates the transport pays for; the gain is in the rate, $n^{-1/6}$ against $n^{-1/10}$, so it widens with n. The narrow

kernels also order the flock less, the polarization falling from 0.94 to 0.78, since each particle then follows fewer neighbors.

Table 3: The bandwidth sweep of the Gaussian rule at $\eta = 1.2$, each flock under its own rule, read at $4{,}000$ samples per half (the pilot sets of $8{,}000$ replicas) and at $16{,}000$ per half (the full sets of $32{,}000$ replicas), $h = 0.5$ the flock of Section 5.1 at the same prefixes: the atoms carrying all but 1% and 0.1% of the mean weight and the decay rate per atom between them, the mean local polarization within $r = 1$, the implied dimension of the transport over the last three doublings, and the certificate of Theorem 4.6 in units of κ.

h	$k(1\%)$	$k(0.1\%)$	decay	ϖ	dim	cert.
$4{,}000$ *samples per half*						
0.125	9	16	0.33	0.78	5.8	0.109
0.18	11	19	0.29	0.85	5.9	0.127
0.25	12	21	0.26	0.90	7.2	0.142
0.35	20	33	0.18	0.93	7.2	0.155
0.5	32	53	0.11	0.94	9.7	0.157
$16{,}000$ *samples per half*						
0.125	9	16	0.33	0.78	5.8	0.081
0.18	11	19	0.29	0.86	6.7	0.097
0.25	13	21	0.29	0.90	8.0	0.113
0.35	20	33	0.18	0.93	9.9	0.128
0.5	32	53	0.11	0.94	12.0	0.133

The noise sweep. Both regimes of the paper are in the ordered phase. To see where the certificate stops being small we simulated the ball rule above the transition, at $\eta = 3$, 4 and 5, on $8{,}000$ replicas each under the protocol of Section 5.1, read at $4{,}000$ samples per half, with the flocks of the paper at the same prefix (Table 4). The certificate grows slowly, from 0.070κ to 0.107κ at $\eta = 3$, and then holds, as the crowding falls with the order, from 10.7 to 6.5 neighbors, and the transport's implied dimension rises from 2.5 to 4.9: a disordered lift is a count and a spread of headings with no common direction, a few dimensions more than an ordered one. The polarization gate $\varpi_0 = 0.6$ of the $\eta = 2.4$ regime, below which the update leaves the class (Proposition C.7), is met on 69%, 27% and 13% of the samples at $\eta = 3$, 4 and 5: above the transition the update is a member on a minority of the samples, while the certificate is the design's and does not need the order.

Table 4: The noise sweep of the ball rule at $4{,}000$ samples per half: the paper's regimes at that prefix and the pilots above the transition ($8{,}000$ replicas each). The columns of Table 3, with the mean count of neighbors other than the self and the share of samples at the polarization gate ϖ_0 of Proposition C.7.

η	$k(1\%)$	$k(0.1\%)$	β	count	pol.	gate	dim.	cert$_A$	ϖ_0
1.2	24	34	0.23	10.7	0.94	1.00	2.5	0.070	0.8
2.4	24	34	0.23	10.9	0.78	0.94	3.5	0.088	0.6
3	26	39	0.18	10.8	0.66	0.69	4.0	0.107	0.6
4	15	20	0.46	7.5	0.47	0.27	4.4	0.101	0.6
5	12	15	0.77	6.5	0.37	0.13	4.9	0.106	0.6

The fully symmetric form. The drift is invariant under torus shifts, and depends on particle i only through the configuration relative to it, with no internal memory. Frame by the focal particle with the shift $T_i : p \mapsto p - p_i$:

$$\tilde{p}_j := p_j - p_i \ \in \ \mathbb{T}^2_\ell, \qquad x_j := \theta_j \ \in \ \mathcal{V} := S^1,$$

for all j, including $j = i$: the framed position $\tilde{p}_j$ is what the selector reads, the heading x_j is the value the mechanism reads, and the self atom sits at position $\tilde{p}_i = 0$ with its own heading. Framed positions keep the tilde, and $t_j := |\tilde{p}_j|$ is the distance from the focal particle. Headings are untouched (shift-only framing). Then the selector becomes an ordinary anonymous selector $c : (\mathbb{T}^2_\ell)^d \to \Delta_d$ of the framed positions with per-coordinate scores: with i fixed by the framing,

$c = c(\tilde{p}) \in \Delta_d$ is the weight vector, its j-th coordinate $c_j \propto K(t_j)$ depends on $\tilde{p}_j$ alone, and $j = i$ contributes $K(0)$, the maximal weight. The aggregation is fully anonymous, self included,

$$m \;=\; \sum_j c_j \, e^{\mathrm{i}\theta_j}, \qquad \theta_i' = \arg m, \qquad p_i' = p_i + v_0 \, m/|m|,$$

and the focal state appears only in the framing and un-framing shifts, which are isometries, never as an argument of the aggregation. The composite is therefore $\theta_i' = g(x, \tilde{p}) := f(x, c(\tilde{p}))$ with f anonymous, the mechanism reading the headings and the selector the positions, and the design the law of $(x, c(\tilde{p}))$: a move of particle $j \neq i$ moves the framed positions in coordinate j only, by exactly the world distance, and a move of the focal particle moves the whole frame, an ordinary telescope priced with no new axiom.

Convention C.2 (The self atom stays in)**.** *The self atom is always part of the configuration. Beyond physical fidelity (the Vicsek average includes the particle itself) it pins the normalization from below: the sum $Z = \sum_l K(t_l)$ contains the self term $K(0)$, so $Z \geq K(0)$ however small every neighbor term is, which keeps the selector well defined for an isolated particle and carries the additive locality of compact windows.*

Remark C.3 (What the framing removes, and what stays)**.** *Removed: the self as a distinguished argument. f has no query slot, is symmetric in all d atoms, and cannot tell which atom is the self; it sees a multiset of weighted headings, and the selector sees an atom at position 0 carrying the maximal weight, so a particle exactly on top of the focal one is rightly treated identically. Kept: the self data, as an ordinary atom, necessarily so, because shift-only framing pins the self position but not the self heading, and θ_i is a variable the alignment average reads. Deleting the self atom would force f to remember θ_i, a query slot again. Full $SE(2)$ framing, rotating by $-\theta_i$, removes the self entirely but is wanted only for heading-dependent interactions (vision cones, blind spots) or to make a learned mechanism equivariant by construction; rotations are not isometries of the torus, so that variant lives on a planar relative chart cropped to a ball of a few interaction ranges, at a crop cost that is exactly zero for compact windows and exponentially small for Gaussian kernels. Default: shift-only, self atom in. From here on (x, μ) denotes the framed configuration and neighbor moves are moves of coordinates $j \neq i$.*

The value space and its constants.

Definition C.4 (Value space)**.** *$\mathcal{V} = S^1$, the heading, with $\rho = d_{S^1}$, the angle in radians. The framed positions are not values: they are read by the selector alone, and the design is the law of the headings and the weights.*

$\mathcal{V}$ is compact and geodesic, $\mathrm{diam}\mathcal{V} = \pi$, so $R = \pi/2$ and $\kappa = \pi$, and $N(\varepsilon, \mathcal{V}, \rho) = \lceil \pi/\varepsilon \rceil$: metric dimension 1. A learner who took the positions as values would pay their dimension in the exponents of Section 4, which carry $\dim \mathcal{V}$. Every statement of Section 4 made on an interval $[-R, R]$ holds on S^1 with the same constants. With $R = \pi/2$, $N(\varepsilon/2, S^1) = \lceil 2\pi/\varepsilon \rceil \leq 5R/\varepsilon$ for $\varepsilon \leq R$, the bound the proof of Theorem A.17 uses for the interval, so $C_1 = 34$ and the constant of Theorem 4.5 are unchanged, and the lower bounds of Theorems 4.4 and 4.5 transfer by placing the design on a half-circle, which is isometric to an interval of length π. By Theorem A.17 the stratum of k-sparse lifted configurations has

$$\log N(\varepsilon, \mathcal{A}_k, D) \;\lesssim\; k \, \log(C/\varepsilon) + (k-1) \log(C'/\varepsilon) \;=\; (2k-1) \, \log(C''/\varepsilon),$$

exponent $2k - 1$ when the weights are free, as they are for the graded kernels, whose weights are functions of the positions, and k for the ball, whose weights are uniform on its support.

Remark C.5 (The domain size is in the selector)**.** *The torus enters only through the selector's scores. The class's constants are those of the circle: the reservoir price $R = \pi/2$ is the worst change of a heading average per unit of weight removed, the crop constant $\kappa = \pi$ is a half turn, and neither the side ℓ nor the interaction range r appears in them. The framing shift and the position update $p_i' = p_i + v_0 u'$ are an isometry and a Lipschitz image of the heading output, and never enter the aggregation.*

Membership: the Vicsek statistics are in the class. The update factors through two statistics of the configuration,

$$\mathbf{m}(x, \mu) \;=\; \sum_j \mu_j \, e^{\mathrm{i}\theta_j} \;\in\; \mathbb{R}^2, \qquad \varpi(x, \mu) \;=\; |\mathbf{m}(x, \mu)| \;\in\; [0, 1], \tag{10}$$

the local mean velocity, with headings read from the values and weights from the selector slot, and its resultant length, the local polarization: $\varpi = 1$ at perfect alignment, $\varpi \approx 0$ at cancellation, and with uniform weights over the whole flock ϖ is exactly Vicsek's global order parameter. Membership comes in two layers of different regularity, and the ordered-region hypothesis of the second is a lower bound $\varpi \geq \varpi_0$ on the first.

Proposition C.6 (Unconditional layer)**.** *For* $\mathbf{m}$ *and* ϖ *of* (10)*: (i) each component of* $\mathbf{m}$ *is in* $\mathcal{G}_1$*, and jointly* $|\mathbf{m}(P) - \mathbf{m}(Q)|_2 \leq D(P, Q)$ *for all lifted* P, Q*; (ii)* ϖ *is in* $\mathcal{G}_1$*, so the quantity gating the conditional layer is itself a member, unconditionally.*

Proof. (i) $\mathbf{m} = \sum_j \mu_j e^{\mathrm{i}\theta_j} = \int u\, \varphi(s)\, d\Phi$ with $\varphi(s) = e^{\mathrm{i}\theta(s)}$, the linear sum of Lemma A.5 with a value-only score. For values s, t with headings θ, θ', $|e^{\mathrm{i}\theta} - e^{\mathrm{i}\theta'}|_2 = 2\sin(d_{S^1}/2) \leq d_{S^1}(\theta, \theta') = \rho(s, t)$. Lemma A.5, in its vector-valued form, gives the joint bound. Anonymity is clear. (ii) Post-compose with the 1-Lipschitz map $|\cdot|_2$. □

$\mathbf{m}$ is position-blind: value moves in position cost nothing and change nothing, consistently with positions entering only through the selector. The update $\theta_i' = \arg \mathbf{m}$ factors as configuration $\mapsto \mathbf{m}$ (Proposition C.6) followed by the fixed normalization $\mathbf{m} \mapsto u' = \mathbf{m}/|\mathbf{m}|$ of $\mathbb{R}^2 \to S^1$, which discards the magnitude and keeps the direction. Both outputs pass through it: the position update $p_i' = p_i + v_0\, u'$ consumes the same direction at constant speed, and the magnitude ϖ never enters the dynamics.

Proposition C.7 (Direction layer)**.** *On the ordered region* $\{\varpi \geq \varpi_0\}$ *the normalization* $u' = \mathbf{m}/|\mathbf{m}|$ *satisfies* $|u'(\mathbf{m}) - u'(\mathbf{m}')|_2 \leq (1/\varpi_0)|\mathbf{m} - \mathbf{m}'|_2$ *and* $\|Du'(\mathbf{m})\| = 1/|\mathbf{m}| \leq 1/\varpi_0$*. Hence the heading-update components lie in the class at scale* $1/\varpi_0$ *on the ordered region, and the position displacement* $v_0\, \mathbf{m}/|\mathbf{m}|$ *at scale* v_0/ϖ_0*.*

Proof. For unit vectors u, v and $r, s \geq \varpi_0$, $|ru - sv|_2^2 = (r - s)^2 + rs|u - v|_2^2$, so $|u - v|_2 \leq |ru - sv|_2/\sqrt{rs} \leq |ru - sv|_2/\varpi_0$: take $u = u'(\mathbf{m})$, $r = |\mathbf{m}|$, $v = u'(\mathbf{m}')$, $s = |\mathbf{m}'|$. The differential form is the Jacobian $(I - u'u'^{\top})/|\mathbf{m}|$. Compose with Proposition C.6(i). □

Remark C.8 (The ϖ_0 caveat is the phase transition)**.** $\varpi \geq \varpi_0$ *fails where neighbors cancel, the rule in the disordered phase and the exception in the ordered one, half a percent of the samples at* $\eta = 1.2$*. There the true update is genuinely unstable, the direction of a near-zero resultant, so the divergence of the class constant as* $\varpi_0 \downarrow 0$ *is the constant pricing the physics rather than slack:* $L = 1/\varpi_0$ *diverges at the order–disorder transition, where the order-parameter fluctuations peak (Chaté et al., 2008). The resultant* ϖ *is a member unconditionally, and only the direction readout pays* $1/\varpi_0$*. The price is attached to the constant-speed rule, which forces the output onto the unit circle, not to alignment as such: a linear velocity-averaging dynamics of Cucker–Smale type (Cucker & Smale, 2007),* $v_i' = \sum_j c_{ij} v_j$ *with no renormalization, stops at the first layer and is a fully unconditional member.*

On the replicas after the burn-in the local polarization under the ball selector has mean 0.94 at $\eta = 1.2$, with $\varpi \geq 0.8$ on 99.5% of the samples and $\varpi \geq 0.6$ on 99.96%, and mean 0.78 at $\eta = 2.4$, with $\varpi \geq 0.6$ on 94% and $\varpi \geq 0.5$ on 97% ($64{,}000$ samples per regime, the resultant of the atoms within $r = 1$, self included).

The gate $\varpi \geq \varpi_0$ is a condition on (x, μ) jointly, hence with $\mu = c(\omega)$ apparently a restriction on the selector as well as on x. The next lemma decouples it: weighted averages of unit vectors respect cones, so a heading-coherence condition on x alone yields one ϖ_0 uniformly over all selectors with a given tail bound.

Lemma C.9 (Coherence decouples the gate from the selector)**.** *Fix a range* $r_c > 0$ *and* $\alpha < \pi/2$*, and call the state* (α, r_c)*-coherent at the focal particle if all headings of the particles within distance* r_c *of it, self included, lie in a closed arc of half-width* α*, a condition on the headings and the positions together. Then for every* $\mu \in \Delta_d$ *placing mass at least* $1 - t$ *on the* r_c*-ball,*

$$\varpi(x, \mu) \;\geq\; (1 - t)\cos\alpha - t \;\geq\; \cos\alpha - 2t.$$

Consequently, for any selector whose tail beyond r_c *is at most* t *on the design region, the ordered-region hypothesis holds with* $\varpi_0 = \cos\alpha - 2t$*, positive once* $\cos\alpha > 2t$*, uniformly over the selector family: the restriction is on the state, the standard locally ordered hypothesis of the flocking literature, and on the selector only through its standing tail bound.*

Proof. Let e be the arc axis. Each in-ball heading vector has $\langle e^{\mathrm{i}\theta_j}, e\rangle \geq \cos\alpha$, each out-of-ball vector at least -1, so $\langle \mathbf{m}, e\rangle \geq (1-t)\cos\alpha - t$, and $\varpi \geq \langle \mathbf{m}, e\rangle$. □

Remark C.10 (Region restrictions are benign)**.** *ϖ is a functional of the lift, $\varpi = |\int u\, e^{\mathrm{i}\theta(s)}\, dP|$, so $\{\varpi \geq \varpi_0\}$ is a lift-defined region of the same kind as a stratum hypothesis: in the learning setup both are conditions on the lifted design. Membership on the region is direct. $\mathbf{m}$ is 1-Lipschitz on all lifts (Proposition C.6) and u' is $(1/\varpi_0)$-Lipschitz in $\mathbf{m}$ on the region (Proposition C.7), so each component of $u' \circ \mathbf{m}$, restricted to the lifts of the region, is $(1/\varpi_0)$-Lipschitz there, and McShane's extension, clipped to $[-1,1]$, is a member of $\mathcal{G}_{1/\varpi_0}$ that agrees with the update on the region. The characterization itself is not the route: the proof of Theorem 3.2 passes through intermediate configurations that can leave the region. The region is stable under the crop up to a margin: a t-crop moves ϖ by at most $2t/(1-t)$, so after the crop the gate holds at $\varpi_0 - 2t/(1-t)$.*

What the number bounds. The certificate bounds the uniform deviation of the unit ball $\mathcal{G}_1$ in the class's own units, against the trivial κ, and it translates directly into the flock's own statistics. The components of the local mean velocity are members of $\mathcal{G}_1$ with range 2 (Proposition C.6), so on the ball at $\eta = 1.2$ their deviation between the population and the sample is at most $0.034\kappa = 0.11$ against 2. The heading update is in $\mathcal{G}_L$ at $L = 1/\varpi_0$ on the samples whose local polarization is at least ϖ_0 (Proposition C.7), and its extension from that region, clipped to $[-1,1]$, is a member of $\mathcal{G}_{1/\varpi_0}$ on every sample (Remark C.10) that differs from the update only on the region $B := \{\varpi < \varpi_0\}$, by at most the range 2. So for a component u of the update and its extension $\bar{u}$,

$$|\Gamma u - \hat{\Gamma}_A u| \leq |\Gamma \bar{u} - \hat{\Gamma}_A \bar{u}| + 2\,\Gamma(B) + 2\,\hat{\Gamma}_A(B) \leq \frac{1}{\varpi_0}\, W_1(\hat{\Gamma}_A, \Gamma) + 2\,\Gamma(B) + 2\,\hat{\Gamma}_A(B),$$

where $\hat{\Gamma}_A(B)$ is observed and $\Gamma(B)$ is at most the share of B in the 64,000 samples plus $\sqrt{\ln(1/\delta)/(2\cdot 64{,}000)} = 0.005$ at $\delta = 0.05$, by Hoeffding's inequality on the indicator of B. At $\eta = 1.2$ and $\varpi_0 = 0.8$ the region holds on 99.5% of the samples of either half, so the deviation of the update between the population and the sample is at most $0.13 + 0.03 = 0.16$ against a range of 2. At $\eta = 2.4$ and $\varpi_0 = 0.6$ it holds on 94%, and the deviation is at most $0.26 + 0.25 = 0.51$ against 2, the samples outside the gate costing as much as the certificate at that noise. For regression with the absolute loss on a target in the class the gap of every predictor is at most twice the number (Remark B.22), and other losses cost the factor of their Lipschitz constant.

D LOWER BOUNDS FROM THE PROFILE

The upper bounds of Section 4 are functionals of the profile. This appendix shows that the profile is also what bounds the problem from below: at a fixed stratum the uniform law is at the $(2k-1)$-dimensional rate, and so is the risk of learning from noiseless labels (Theorem D.5), and on a design with an exponential profile it is quasi-polynomial in n. The technique is classical, a bound on the mass that the lifted design gives a ball, so that n atoms cover at most half of the design at the scale where the bound is $1/(2n)$, and the content is in the object: the metric D restricted to a fiber of fixed weights is bi-Lipschitz to a weighted supremum distance whose weights are the design's (Lemma D.1), so the lifted design is locally an anisotropic box with semi-axes the weights, and on an exponential profile its mass at scale ε is $\exp(-\ln^2(1/\varepsilon)/(2\beta))$ up to constants. The square loss has lower bounds of the same shape from the same fibers, by Assouad's hypercube, sketched in Remark B.23.

D.1 THE DEVICE: FIBERS OF FIXED WEIGHTS

Lemma D.1 (Fibers of fixed weights)**.** *Let $\mu \in \Delta_d$ have support $S := \{i : \mu_i > 0\}$ with pairwise distinct entries on S, and set, with $\gamma(\mu) := 1$ when S is a single atom,*

$$\gamma(\mu) \;:=\; \min_{i \neq j \in S} \frac{|\mu_i - \mu_j|}{\max(\mu_i, \mu_j)} \;\in\; (0,1].$$

(i) Φ is injective on $\mathcal{V}^S \times \{\mu\}$, values off S being forgotten, and for all $x, y \in \mathcal{V}^d$,

$$\frac{\gamma(\mu)}{2}\, \max_{i \in S} \mu_i\, \rho(x_i, y_i) \;\leq\; D\big(\Phi(x,\mu), \Phi(y,\mu)\big) \;\leq\; \sum_{i \in S} \mu_i\, \rho(x_i, y_i).$$

(ii) Write $F_\mu := \Phi(\mathcal{V}^d \times \{\mu\})$ for the fiber over μ. If μ, ν differ as multisets of positive weights then $D(F_\mu, F_\nu) \geq R\, m(\mu, \nu) > 0$, where $m(\mu, \nu)$ is the minimum, over partial matchings of the positive entries of μ with those of ν, of the sum of $|\mu_i - \nu_j|$ over matched pairs plus the sum of the unmatched entries.

Proof. (i) Injectivity: $\Phi(x, \mu) = \Phi(y, \mu)$ forces equal multisets of positive pairs (Lemma A.8); the weights being distinct, the pairing is the identity on S. Upper bound: match atom i to atom i, at cone cost $\mu_i \rho(x_i, y_i)$ each. Lower bound: fix $i \in S$ attaining the maximum and follow the atom (x_i, μ_i) in an optimal partial matching. Matched to (y_i, μ_i) it costs $\mu_i \rho(x_i, y_i)$. Matched to (y_j, μ_j), $j \neq i$, it costs at least the height term $R|\mu_i - \mu_j| \geq R\gamma(\mu)\mu_i \geq \frac{1}{2}\gamma(\mu)\, \mu_i \rho(x_i, y_i)$, as $\rho \leq 2R$. Unmatched it costs $R\mu_i \geq \frac{1}{2}\mu_i \rho(x_i, y_i)$. All other terms are nonnegative. (ii) Value terms are nonnegative, so every partial matching between $\Phi(x, \mu)$ and $\Phi(y, \nu)$ costs at least R times its height cost, which is at least $Rm(\mu, \nu)$; the minimum is over finitely many matchings and vanishes only if one of them pairs every entry with an equal entry, that is if the multisets coincide. □

D.2 Fixed stratum: the calibration from below

Definition D.2 (A region of separated weights)**.** *Let $\mathcal{V} = [-R, R]$, $k \geq 2$, and put*

$$\mu_i^0 := \frac{2i}{k(k+1)}, \qquad \delta_0 := \frac{2}{k(k+1)}, \qquad r := \frac{\delta_0}{4k}, \qquad m_0 := \frac{3\delta_0}{4},$$

so δ_0 is both the gap between consecutive coordinates of $\mu^0 \in \Delta_k$ and its minimum, and let

$$U_k := \big\{a \in \Delta_k : |a_i - \mu_i^0| \leq r,\ i \leq k-1\big\},$$

a cube of side $2r$ in the chart $(a_1, \ldots, a_{k-1})$. On U_k every coordinate is within $\delta_0/4$ of μ^0, the last one because $(k-1)r \leq \delta_0/4$, hence $a_i \geq m_0$ for all i, the coordinates of a single $a \in U_k$ are pairwise at distance at least $\delta_0/2$, and $|a_i - a'_j| \geq \delta_0/2$ for all $a, a' \in U_k$, $i \neq j$. The stratum design *is $Q_k := \mathrm{Unif}([-R, R]^k) \otimes \mathrm{Unif}(U_k)$ on the first k coordinates, weight 0 on the others, normalized Lebesgue measure in the chart: an exactly k-sparse design, $T_k = 0$ a.s., with lift Γ_k supported in $S_k := \Phi([-R, R]^k \times U_k) \subseteq \mathcal{A}_k$.*

At $k = 1$ the stratum is the interval at height 1, and $W_1(\nu_n, \Gamma_1) \geq c/n$ for every n-atom ν_n is the quantization bound of the uniform law on an interval, so the converse of Theorem 4.4 holds at $k_0 = 1$ as well.

Lemma D.3 (Local bi-Lipschitz bound on the separated region)**.** *For $(x, a), (y, a') \in [-R, R]^k \times U_k$ with $D(\Phi(x, a), \Phi(y, a')) \leq \varepsilon < R\delta_0/2$,*

$$m_0\ \|x - y\|_1 + R\ \|a - a'\|_1\ \leq\ \varepsilon.$$

Proof. Both lifts have k atoms, one per coordinate. A partial matching that pairs coordinate i of the first with coordinate $j \neq i$ of the second pays $R|a_i - a'_j| \geq R\delta_0/2$ on that pair, and one that leaves an atom unmatched pays $Ra_i \geq Rm_0 \geq R\delta_0/2$; either exceeds ε. So the optimal matching pairs each coordinate with itself, and $D = \sum_i \big[\min(a_i, a'_i)\, |x_i - y_i| + R|a_i - a'_i|\big] \geq m_0 \|x - y\|_1 + R\, \|a - a'\|_1$. □

Lemma D.4 (Mass bound, and the uniform law at a stratum)**.** *Let $\varepsilon_{\max}(k) := R\delta_0/5$,*

$$C_k := \frac{2^{2k-1}}{m_0^k\, R^{2k-1}\, r^{k-1}\, k!\, (k-1)!}, \qquad c_k := \tfrac{1}{2}\, (2C_k)^{-1/(2k-1)}.$$

For every $P \in \mathcal{M}_d$ and $\varepsilon \leq \varepsilon_{\max}(k)$, $\Gamma_k(\mathsf{B}(P, \varepsilon)) \leq C_k\, \varepsilon^{2k-1}$, and for every n and every n-atom probability measure ν_n on $\mathcal{M}_d$,

$$W_1(\nu_n, \Gamma_k)\ \geq\ \min\big\{\, c_k\, n^{-1/(2k-1)},\ \varepsilon_{\max}(k)/2\,\big\};$$

in particular this bounds $\mathbb{E}\, W_1(\hat{\Gamma}_n, \Gamma_k)$ from below.

Proof. Mass bound. If $\mathsf{B}(P,\varepsilon)$ misses S_k there is nothing to prove. Otherwise pick $P' = \Phi(x',a') \in S_k \cap \mathsf{B}(P,\varepsilon)$, so that $\mathsf{B}(P,\varepsilon) \subseteq \mathsf{B}(P',2\varepsilon)$. Every $(y,a) \in [-R,R]^k \times U_k$ with $\Phi(y,a) \in \mathsf{B}(P',2\varepsilon)$ satisfies, by Lemma D.3 as $2\varepsilon < R\delta_0/2$, $\|x'-y\|_1 \le t := 2\varepsilon/m_0$ and $\|a'-a\|_1 \le 2\varepsilon/R$. Since Q_k is a product measure,

$$\Gamma_k\big(\mathsf{B}(P',2\varepsilon)\big) \;\le\; \frac{(2t)^k}{k!\,(2R)^k} \cdot \frac{(4\varepsilon/R)^{k-1}}{(k-1)!\,(2r)^{k-1}} \;=\; C_k\,\varepsilon^{2k-1},$$

using the volume $(2t)^m/m!$ of the ℓ^1-ball of radius t in $\mathbb{R}^m$, in the chart for the second factor, where the ℓ^1-norm of the first $k-1$ coordinates is at most the full one. *Uniform law.* Let $\nu_n = \sum_{i\le n} a_i\,\delta_{P_i}$ and set $\varepsilon := \min\{(2nC_k)^{-1/(2k-1)}, \varepsilon_{\max}(k)\}$, so that $nC_k\varepsilon^{2k-1} \le \frac12$. By the mass bound, $\Gamma_k\big(\bigcup_i \mathsf{B}(P_i,\varepsilon)\big) \le \frac12$, and any coupling of ν_n with Γ_k transports the Γ_k-mass outside this union, at least $\frac12$, across distance $> \varepsilon$. Hence $W_1(\nu_n,\Gamma_k) \ge \varepsilon/2$. □

Theorem D.5 (Learning from below at a stratum)**.** *Let $k \ge 2$ and let Q_k be the stratum design of Definition D.2. A* learner *is a measurable map from a sample $(Z_i, f^*(Z_i))_{i\le n}$, Z_i i.i.d. $\sim Q_k$, to a function $\hat f$ on $\mathcal{C}$, and its risk on the target f^* is $\mathbb{E}_{Q_k}|\hat f - f^*|$. There are $c'_k > 0$ and n_k, depending on k and R only, such that for every $n \ge n_k$ and every learner*

$$\sup_{f^*\in\mathcal{G}_1} \mathbb{E}\,\mathbb{E}_{Q_k}|\hat f - f^*| \;\ge\; c'_k\, n^{-\frac{1}{2k-1}} :$$

at a fixed stratum the labels, however many and noiseless, do not lift the exponent of Theorem 4.4.

Proof. Write $\Gamma := \Gamma_k$, $S := S_k$, and fix $\varepsilon \le \varepsilon'_{\max}(k) := R\delta_0 \min\{1/20, 2/k\}$, so that $2\varepsilon \le \varepsilon_{\max}(k)$ and $\varepsilon/(8R) \le r$. *A packing.* Let $P_1,\dots,P_N$ be a maximal 2ε-separated subset of S, finite by compactness. By maximality the balls $\mathsf{B}(P_j,2\varepsilon)$ cover S, so $1 \le \sum_j \Gamma(\mathsf{B}(P_j,2\varepsilon)) \le N\,C_k(2\varepsilon)^{2k-1}$ by the mass bound of Lemma D.4, that is $N \ge 1/(C_k(2\varepsilon)^{2k-1})$. *The mass near a center.* Write $P_j = \Phi(x^j,a^j)$ with $(x^j,a^j) \in [-R,R]^k \times U_k$. Matching coordinate i with coordinate i (Equation (4)), $D(\Phi(y,a),P_j) \le \big\|y-x^j\big\|_1 + R\,\big\|a-a^j\big\|_1$, and the last coordinate of $a-a^j$ is bounded by the sum of the others, so $\{\big\|y-x^j\big\|_1 \le \varepsilon/4\} \times \{\big\|(a-a^j)_{\le k-1}\big\|_1 \le \varepsilon/(8R)\}$ lifts into $\mathsf{B}(P_j,\varepsilon/2)$. The ℓ^1-ball of radius t in $\mathbb{R}^m$ has volume $(2t)^m/m!$, and its intersection with a cube containing the center contains one of its 2^m orthant pieces, wherever the center sits, once t is at most half the cube's side, which holds for both factors ($\varepsilon/4 \le R$ and $\varepsilon/(8R) \le r$). As Q_k is the product of the uniform laws on the cube $[-R,R]^k$ and on the cube U_k of side $2r$ in the chart,

$$\Gamma\big(\mathsf{B}(P_j,\varepsilon/2)\big) \;\ge\; \frac{(\varepsilon/2)^k}{2^k\,k!\,(2R)^k} \cdot \frac{(\varepsilon/(4R))^{k-1}}{2^{k-1}\,(k-1)!\,(2r)^{k-1}} \;=:\; c_2\,\varepsilon^{2k-1}.$$

The bumps. Let $\varphi_j(P) := (\varepsilon - D(P,P_j))_+$, a 1-Lipschitz function of the lift supported in $\mathsf{B}(P_j,\varepsilon)$ and at least $\varepsilon/2$ on $\mathsf{B}(P_j,\varepsilon/2)$, so that $\int \varphi_j\,d\Gamma \ge \frac12 c_2\varepsilon^{2k}$. For $\sigma \in \{-1,1\}^N$ put $f_\sigma := \sum_j \sigma_j\,\varphi_j \circ \Phi$. It is in $\mathcal{G}_1$: the supports are disjoint, the centers being 2ε apart, and for P in the j-th support and P' in the l-th, $l \ne j$, $|f_\sigma(P) - f_\sigma(P')| \le \varphi_j(P) + \varphi_l(P') = 2\varepsilon - D(P,P_j) - D(P',P_l) \le D(P,P')$ by the triangle inequality through $D(P_j,P_l) \ge 2\varepsilon$, while for P' in no support $|f_\sigma(P) - f_\sigma(P')| = \varphi_j(P) \le \varepsilon - D(P,P_j) \le D(P',P_j) - D(P,P_j) \le D(P,P')$. *The signs.* Let σ be uniform on $\{-1,1\}^N$, independent of the sample, and J the set of j whose ball $\mathsf{B}(P_j,\varepsilon)$ contains no sample point. The labels $f_\sigma(Z_i)$ are functions of the sample and of $(\sigma_j)_{j\notin J}$ alone, so conditionally on both, $(\sigma_j)_{j\in J}$ is still uniform, and for $Z' \in \mathsf{B}(P_j,\varepsilon)$ with $j \in J$, where $f_\sigma = \sigma_j\varphi_j$, $\mathbb{E}[|\hat f(Z') - \sigma_j\varphi_j(Z')| \mid \text{sample}, (\sigma_j)_{j\notin J}] \ge \varphi_j(Z')$, as $\frac12|a-b| + \frac12|a+b| \ge |b|$. Integrating over Z' and summing over $j \in J$, the risk averaged over σ, which the supremum over f^* dominates, is at least $\sum_j \Pr(j \in J)\int \varphi_j\,d\Gamma$, and $\Pr(j \in J) = (1-\Gamma(\mathsf{B}(P_j,\varepsilon)))^n \ge 1 - nC_k\varepsilon^{2k-1}$ by the mass bound. *The scale.* Take $\varepsilon := (2nC_k)^{-1/(2k-1)} = 2c_k n^{-1/(2k-1)}$, which is at most $\varepsilon'_{\max}(k)$ for $n \ge n_k := (2c_k/\varepsilon'_{\max}(k))^{2k-1}$. Then $\Pr(j \in J) \ge \frac12$ and

$$\sup_{f^*\in\mathcal{G}_1} \mathbb{E}\,\mathbb{E}_{Q_k}|\hat f - f^*| \;\ge\; N \cdot \tfrac12 \cdot \tfrac12 c_2\varepsilon^{2k} \;\ge\; \frac{c_2\,\varepsilon}{4\,C_k\,2^{2k-1}} \;=\; \frac{c_2\,c_k}{4^k\,C_k}\,n^{-\frac{1}{2k-1}} \;=:\; c'_k\,n^{-\frac{1}{2k-1}}.$$ □

D.3 EXPONENTIAL PROFILES: QUASI-POLYNOMIAL FROM BELOW

Theorem D.6 (Exponential profile: the uniform law from below)**.** *Let $\mathcal{V} = [-R, R]$, $\beta > 0$, $q := e^{-\beta}$, $\gamma := 1 - q$, and $d \geq 2$. Let $\mu^\beta \in \Delta_d$ have entries $\mu_j^\beta := \gamma q^{j-1}/(1-q^d)$, and let $Q^\beta := \mathrm{Unif}([-R, R]^d) \otimes \delta_{\mu^\beta}$, with lift Γ^β supported in the fiber $F^\beta := F_{\mu^\beta}$. Then $T_k \leq e^{-\beta k}$ for all k at every configuration: an exponential profile at rate β with $A = 1$, to which Theorem 4.5 applies. For every $n \geq 1$ with $d \geq \sqrt{2\ln(2n)/\beta} + 1$ and every n-atom probability measure ν_n on $\mathcal{M}_d$,*

$$W_1(\nu_n, \Gamma^\beta) \;\geq\; \frac{\gamma^2 R}{16}\, e^{-\sqrt{2\beta \ln(2n)}};$$

in particular this bounds $\mathbb{E}W_1(\hat{\Gamma}_n, \Gamma^\beta) = \Delta_n(Q^\beta)$ from below.

Proof. Profile and gaps. $T_k(\mu^\beta) = \sum_{j>k} \mu_j^\beta = (q^k - q^d)/(1-q^d) \leq q^k$, as $q^{k+d} \leq q^d$. For $i < j$, $\mu_i^\beta - \mu_j^\beta = \mu_i^\beta(1 - q^{j-i}) \geq \gamma \mu_i^\beta$, so $\gamma(\mu^\beta) \geq \gamma$ in Lemma D.1(i):

$$D\big(\Phi(x, \mu^\beta), \Phi(y, \mu^\beta)\big) \;\geq\; \frac{\gamma}{2} \max_j \mu_j^\beta |x_j - y_j| \qquad (x, y \in [-R, R]^d), \tag{11}$$

and Φ is injective on $[-R, R]^d \times \{\mu^\beta\}$. Also $\mu_j^\beta \geq \gamma q^{j-1}$.

The anisotropic box. For $t \in (0, \gamma^2 R/4]$ put $a(t) := \ln(\gamma^2 R/(4t)) \geq 0$, $K(t) := \lfloor a(t)/\beta \rfloor + 1 \geq 1$, and $y_j(t) := \gamma \mu_j^\beta R/(2t) \geq \gamma^2 q^{j-1} R/(2t) = 2e^{a(t) - \beta(j-1)}$, so that $y_j(t) \geq 2$ for $j \leq K(t)$, as $\beta(K(t) - 1) \leq a(t)$. Consequently, for $K = K(t) \leq d$ and $a = a(t)$,

$$\prod_{j \leq K} \frac{1}{y_j(t)} \;\leq\; 2^{-K} e^{-aK + \beta K(K-1)/2} \;\leq\; 2^{-K} e^{-aK/2} \;\leq\; e^{-a^2/(2\beta)}, \tag{12}$$

using $\beta(K-1) \leq a$ and then $K > a/\beta$.

The bound. Fix $P \in \mathcal{M}_d$ and ε with $2\varepsilon \leq \gamma^2 R/4$ and $K' := K(2\varepsilon) \leq d$, and suppose $\mathsf{B}(P, \varepsilon)$ meets F^β at $P' = \Phi(x', \mu^\beta)$. Then $\mathsf{B}(P, \varepsilon) \subseteq \mathsf{B}(P', 2\varepsilon)$, and by (11) every $\Phi(y, \mu^\beta) \in \mathsf{B}(P', 2\varepsilon)$ has $|y_j - x'_j| \leq 4\varepsilon/(\gamma \mu_j^\beta) = R/y_j(2\varepsilon)$ for every j. As Q^β is uniform on the cube, $\Gamma^\beta(\mathsf{B}(P, \varepsilon)) \leq \prod_{j \leq K'} \min\{1, 1/y_j(2\varepsilon)\} \leq e^{-a(2\varepsilon)^2/(2\beta)}$ by (12). For an n-atom $\nu_n = \sum_i a_i \delta_{P_i}$ take $\varepsilon := \frac{1}{8}\gamma^2 R\, e^{-\sqrt{2\beta \ln(2n)}}$, so that $a(2\varepsilon)^2/(2\beta) = \ln(2n)$ and $K' = \lfloor \sqrt{2\ln(2n)/\beta} \rfloor + 1 \leq d$: then $\Gamma^\beta(\bigcup_i \mathsf{B}(P_i, \varepsilon)) \leq n e^{-\ln(2n)} = \frac{1}{2}$, every coupling of ν_n with Γ^β moves Γ^β-mass at least $\frac{1}{2}$ across distance $\geq \varepsilon$, and $W_1(\nu_n, \Gamma^\beta) \geq \varepsilon/2$. □

D.4 READING

Fixed stratum. The lower exponent of Lemma D.4, $1/(2k-1)$, is the upper exponent of Theorem 4.4 on interval values for $k \geq 2$, by the multiscale coupling of Proposition B.18 on the cropped stratum: at a fixed stratum the uniform law is $n^{-1/(2k-1)}$ from both sides. The constants are polynomial in k on the lower side and the comparison of constants is not pursued. At $k = 1$ the stratum is an interval, where the empirical measure converges at $n^{-1/2}$ and the quantization bound c/n of every n-atom measure is the converse's form.

The exponential rate is the rate of the problem. Between Theorem 4.5 and Theorem D.6 the uniform law of the design Q^β is $e^{-c\sqrt{\beta \ln n}}$ with the constant c lying between $1/\sqrt{2}$ and $\sqrt{2}$, up to lower-order factors. The fiber resembles the exponential ellipsoid of Lei (2020, Thm. 4.2) at $\log \gamma = \beta$, whose entropy $\ln^2(1/\varepsilon)/(2\beta)$ is his Lemma 4.3 and whose rate $e^{-\sqrt{2 \log \gamma \, \ln n}}$ has the shape of Theorem D.6. Neither the partition nor the crop is the reason for the quasi-polynomial shape: for every fixed k the lower bound exceeds $n^{-1/(2k-1)}$ eventually (for $d \geq \sqrt{2\ln(2n)/\beta} + 1$), so Q^β is harder than every stratum, while the upper bound places it below every logarithmic rate. The design is d-free in the sense that matters: d enters only through that condition, the number of coordinates the construction uses, which is the cutoff $K_n \asymp \sqrt{\ln n/\beta}$ of Theorem 4.5 read from below. The whole lower bound is the mass bound $\Gamma^\beta(\mathsf{B}(P, \varepsilon)) \leq \exp(-\ln^2(\gamma^2 R/(8\varepsilon))/(2\beta))$: the local dimension

of the lifted design at scale ε is $\ln(1/\varepsilon)/\beta$, the profile's own cutoff, and upper and lower bounds read the same quantity, one as a covering count and the other as a packing mass. The stratum count of Theorem A.17 resolves every value at the scale ε and covers the weights over all of Δ_k, whereas the fiber itself has the metric entropy of the anisotropic box $\prod_j [\pm R\mu_j^{\beta}]$ and no weights to cover. Whether a count resolving the atom of rank j at its own scale $\varepsilon/(k\mu_j)$, which would recover the fiber's entropy up to a polylogarithmic factor, raises the constant of Theorem 4.5 is left open.

E THE LITERATURE IN FULL

Section 2 keeps the contrasts the results rest on. This appendix gives the lines in full.

Symmetric functions and Deep Sets. The representation of Section 3.2 belongs to a tradition in which an invariant embedding absorbs a symmetry and the functions above it are free. For permutation symmetry the embedding is a sum of per-element features: the fundamental theorem of symmetric polynomials (Weyl, 1939; Macdonald, 1995) gives every continuous symmetric function of n scalars as $\rho(\sum_i \varphi(x_i))$ with φ the power sums, Deep Sets (Zaheer et al., 2017) is this statement with the two maps learned, and the sum decomposition is the form of the set networks that followed (Sannai et al., 2019; Segol & Lipman, 2020; Amir et al., 2023; Zweig & Bruna, 2022). The width is the price: Wagstaff et al. (2022) show that a continuous sum decomposition of all symmetric functions of n elements needs latent dimension n, and that below it there is a Lipschitz symmetric function that no decomposition approximates better than a constant does. The embedding of these results carries no metric, so the class it represents is all continuous symmetric functions and the complexity sits in the width. Here the embedding carries a metric, the class is its Lipschitz ball, the sum networks are its linear functionals with a readout (Section 3.2), and the complexity sits in the geometry of the design. The count-free rate of the linear sums (Proposition B.20) is the phenomenon of distribution regression, the learning of a function of a distribution from the mean embedding of a sample of it, at rates in the number of distributions and the size of each sample (Póczos et al., 2013; Szabó et al., 2016): a linear sum is a function of the mean lift of the configuration and learns at the value space's rate whatever the count, and the class beyond the linear sums is what the profile prices. The sample-complexity gain of an invariance is quantified in another line: on an invariant target the least-squares gap of a linear model exceeds the invariant model's by the dimension of the non-invariant directions over n (Elesedy & Zaidi, 2021), an invariant kernel gains a factor of the group's order for a finite group (Bietti et al., 2021), the covering of a quotient feature space gains $\sqrt{d!}$ for a permutation-invariant network (Sannai et al., 2021), and on a manifold under a Lie group the minimax exponent carries the dimension of the quotient (Tahmasebi & Jegelka, 2023). For the symmetric group on d coordinates the quotient of the configuration space still has dimension of order d, so anonymity alone buys a constant, and it is the profile that removes d from the exponent (Section 4.1). The stability results on sum embeddings ask how well a fixed embedding preserves a Wasserstein distance chosen in advance, under which sum-based multiset networks are Hölder in expectation and not bi-Lipschitz (Davidson & Dym, 2025), and which aggregators are Lipschitz for a multiset distance fixed in advance, the sum, the mean, the maximum and attention against three such distances, is Nikolentzos & Skianis (2026), and a sliced-Wasserstein embedding of multisets that is bi-Lipschitz for that distance, with no bi-Lipschitz embedding of measures at all, is Amir & Dym (2025), while De Bie et al. (2019) build networks on weighted point clouds that are Lipschitz for a Wasserstein distance chosen in advance. Networks on spaces of probability measures with a universal approximation theorem are Pevný & Kovařík (2019), and the learning of Lipschitz and Sobolev functions on Wasserstein space from samples is Fornasier et al. (2023) and Pham & Warin (2023): functions of the weighted value measure, the pointwise subclass of Remark A.6, in which the weights are mass. Here the question is which distance the class itself determines. On a fixed hard selection the two views agree, since the restriction of our class is every symmetric Lipschitz function of the selected values (Corollary A.15).

Graph networks. A graph network aggregates over the neighborhood of each node, a set that differs from node to node, and its neighborhood aggregation is our object with a hard selector. Its generalization bounds, Rademacher and PAC-Bayes bounds for a fixed architecture (Garg et al., 2020; Liao et al., 2021), charge the maximum degree of the training graphs per layer, or its spectral relaxation (Ju et al., 2023) or norms of the propagation matrix that reduce to degree bounds (Esser

et al., 2021), at the rate $m^{-1/2}$ with the degree a constant in front. Whether any finer statistic of the data can be used was posed as a question in that line (Liao et al., 2021; Morris et al., 2024). The graphon line (Maskey et al., 2022), with its extension to sparse graphs and mixtures of graphons (Maskey et al., 2024), points the other way: with mean aggregation on dense graphs its bounds improve with the node count, with constants that grow as the kernel's minimal degree shrinks, where ours improve as the design gets sparser and never see the node count. These bounds cover the parameter ball of one network, whose parameters are shared across nodes so that the node count drops out, and with it the neighbor count. What they become when asked to learn our class is Remark B.19, a bound at the maximum count. Lower bounds exist in graph learning, in Weisfeiler–Leman classes and in margins (Vasileiou et al., 2025b), and on the Rademacher complexity of a graph convolutional network at a constant degree on one fixed graph (Lv, 2021), none minimax, and none for a neighbor count that varies with the configuration.

A second line puts a metric on the graphs under which message-passing networks are Lipschitz. The metrics are the tree mover's distance (Chuang & Jegelka, 2022), the cut distance of graphon signals (Levie, 2023), unbalanced transport distances with a mass-difference term (Böker et al., 2023; Rauchwerger et al., 2025), and lately one metric on graphs of every size, sparse and dense, through graphops (Amran et al., 2026). It covers the graph space and applies a generic statistical step, a Rademacher bound or the robustness bound of Xu & Mannor (2012) or its occupied-cell form (Kawaguchi et al., 2022). Covered Forest (Vasileiou et al., 2025a) is its latest form, with covering numbers through Weisfeiler–Leman classes. Ours has the same architecture, the lift with D, the strata with their covering numbers and the coupling, and differs in three places. The direction of the geometry: they start from a model and find a metric in which it is Lipschitz, so the Lipschitz ball is an envelope of the model and the metric one choice among several, where Rauchwerger et al. (2025) come closest with a converse and the model class dense in the continuous functions of their space. We start from the axioms, and the representation theorem is an equality. What is covered: they cover the space at one scale, or with the sample's cell proportions at one radius (Carrasco et al., 2025, Prop. 4) and per-cell radii (Vasileiou et al., 2025a, Thm. 35), and we cover the measure, by the profile across the shells with a multiscale coupling within each, and by the sample's own configurations. What is proved, and what is non-vacuous, below the trivial value of the risk in the sense of Dziugaite & Roy (2017) and of the risk certificates of Pérez-Ortiz et al. (2021): the covering number of a graph space is astronomical where it is known (Levie, 2023) and unbounded where it is not (Amran et al., 2026), so the theorems' values are not reported on a benchmark, and the non-vacuous values Covered Forest reports evaluate the Weisfeiler–Leman classes the dataset occupies, with a Lipschitz constant estimated from the data, a counting of the kind of the occupied-cell bound of Kawaguchi et al. (2022), while the non-vacuous value of Ju et al. (2023) is a PAC-Bayes bound evaluated at the trained network through its loss Hessian, a bound on one trained network rather than over a class. The certificate of Theorem 4.6 is a theorem on the data at every step, and on the flock it is a thirtieth of the trivial value (Section 5.1). That a class large enough to hold a trained network makes every uniform bound vacuous (Nagarajan & Kolter, 2019) is an objection to the size of the class: here the class is the Lipschitz ball of the lift, its uniform deviation on the design is measured rather than bounded from above by a covering, and on the flock it is small.

Empirical complexity. The bound behind the certificate is the classical one, the maximum discrepancy of Bartlett et al. (2002) through the symmetrization of Vapnik & Chervonenkis (1971), and what is ours is that on the class it is the transport itself, by the representation, and is computed at one observed split. A similar quantity has recently been used in a different context. Chuang et al. (2021, Thm. 2) take the expected transport between two independent samples of a feature distribution, the k-variance of Solomon et al. (2022), as the complexity term of a margin bound for the readouts of a trained encoder, with an empirical version at one concentration term (their Lemma 5 and Cor. 6), and compute it on data as a measure that tracks generalization rather than evaluating the bound. Hou et al. (2023) evaluate a transport bound with the counts of a fixed partition, and Zhu et al. (2021) compute sample covering numbers on image data with the bound itself not evaluated, in the shape of the sample-cover bound of Mansour & McAllester (2000). None of these is the bound of Theorem 4.6: theirs is a complexity term or a partition count entering a bound on a model, in the metric of its features, ours the observed gap of the class itself in the metric the class implies. The closest form is in domain adaptation, where the target risk of a Lipschitz class is bounded by the source risk plus the transport between the empirical source and target samples (Redko et al., 2017, Thm. 2), (Shen et al., 2018, Thm. 1), with the joint error of the two domains and one concentration

term per sample: the two halves of Theorem 4.6 share a law, so the joint error is absent and one term serves, and the transport is a matching in the class's own metric, computed exactly.

Attention. Attention (Vaswani et al., 2017) gives each output its own field of view: the softmax weights over the inputs are a weight vector of the kind this paper is about, they vary with the input, and in some heads and on some inputs a few tokens carry most of the weight. Fixed-pattern attention restricts each output to a window or a sparse pattern chosen in advance (Child et al., 2019; Beltagy et al., 2020), a hard selector given, with the softmax inside the pattern a graded one. The Lipschitz constant of a whole attention map, weights included, is Kim et al. (2021), a statement about the composite rather than about the mechanism with its weights held as an input. Attention as a map on measures, a self-interacting particle system that is Lipschitz in a transport metric (Vuckovic et al., 2020), with a constant bounded by $\sqrt{n}$ in the sequence length on a compact set and independent of it in the mean-field regime (Castin et al., 2024), is the same composite with uniform weights and the selector inside it. The learning-theory results nearest to ours are on transformers whose target depends on a few tokens of the context: a sample complexity almost independent of the context length when the positions of the q tokens are given in the prompt (Mousavi-Hosseini et al., 2025), logarithmic in it for sparse variable creation (Edelman et al., 2022), with the q-sparse average as the representational primitive (Sanford et al., 2023), length-independent norm bounds (Trauger & Tewari, 2024), an input-dependent importance of the coordinates (Takakura & Suzuki, 2023), and a per-sample sensitivity cap as the locality condition (Nagler, 2023), a one-layer transformer that learns the sparse token selection by gradient descent where a fully connected network cannot (Wang et al., 2024), and a sample complexity in the inverse fraction of the label-relevant tokens, a statistic of graded sparsity, with an attention map that training sparsifies (Li et al., 2023). Those have a selection of tokens, hard at a fixed count or graded by an importance the network infers, an ordered link and a parametric class, and ask which architecture learns to retrieve the tokens. Ours has graded weights given as input and a varying count, anonymity, the full Lipschitz class, a profile with lower bounds, and a certificate, and asks the complexity of the mechanism once the weights are given. Nikolentzos & Skianis (2026) show that attention pooling has instances that are Lipschitz for none of the three multiset distances they consider, a statement about the composite (1) with the selector inside it; with the weights an input and the cap an axiom, the mechanism f is Lipschitz in the metric the class determines, and the composite is not treated here. The mechanism is nevertheless outside the scope of this paper. Its weights are learned, together with everything else, and an attention block is not trained on its own, so the problem studied here, learning a mechanism whose selector is given, does not arise in that form.

Interaction kernels. The learning of interaction laws in particle systems from trajectories is the subject of Lu et al. (2019; 2021): the mechanism is a mean-field sum $N^{-1}\sum_j \varphi(|x_j - x_i|)(x_j - x_i)$ with the scalar kernel φ unknown, and the least-squares estimator converges at the one-dimensional nonparametric rate, in the L^2 norm of the law of the pairwise distances along the trajectories, a measure-level norm. In the terms of Section 3.1 their class is the linear sums with the weights uniform over all agents and the score a fixed function of the pairwise displacement, so the rate is that of the one-dimensional kernel and no profile enters; here the class is every Lipschitz function of the lift, whose rate the profile sets, with the selector given. Their treatment of dependence along a trajectory is what the independent replicas of Section 5 sidestep. The same mean-field object with the drift a nonparametric function of the state and of the empirical measure is estimated from one trajectory per particle at the minimax rate of its Hölder class (Yao et al., 2022), uniform weights again and no selector. In the animal groups the object inferred is the selector itself, the interaction range of a starling flock from field data by a maximum-entropy fit, a fixed number of topological neighbors (Bialek et al., 2012), and in robot swarms a graph network learns the flocking controller with the communication graph given (Tolstaya et al., 2020): the second is the setting of this paper, the first its complement.

The matching distance. The metric of Section 3.2 is produced by the axioms, and the derivation is new: none of the lines that built its ingredients starts from a class of functions. Each ingredient was met before from the other end, introduced to compare measures of unequal mass. Mass must be allowed to appear or vanish, the cheapest way to price that is a reservoir at a linear price per unit, and the height coordinate of the cone is the mass itself. The Kantorovich–Rubinstein norm extends from measures of mass zero to all finite measures by paying the variation of the untransported part,

which gives the flat metric, whose dual is the space of bounded Lipschitz functions (Kantorovich & Rubinstein, 1958; Hanin, 1992), reached again from transport equations with a source term as the generalized Wasserstein distance (Piccoli & Rossi, 2014; 2016). Figalli & Gigli (2010) let mass leak through the boundary of the domain at the cost of its squared distance to it. In unbalanced transport, the least cost of carrying mass m_0 at x_0 to mass m_1 at x_1, with transport at unit price and creation or destruction at price δ per unit, is $\min(|x_1 - x_0|, 2\delta)\min(m_0, m_1) + \delta|m_1 - m_0|$ (Chizat et al., 2018), the cone cost (2) at $\delta = R$. The same function is the marginal perspective of the total-variation entropy-transport problem, which Liero et al. (2018) rewrite as transport on the cone over the base space with mass-weighted marginals. A persistence diagram is a multiset of birth–death pairs, and its distance to another diagram lets a feature vanish at the cost of its distance to the diagonal, available with infinite multiplicity (Cohen-Steiner et al., 2010; Turner et al., 2014): the matching form with a reservoir on multisets of unit masses. On multisets of vectors the matching distances of Chuang & Jegelka (2022), Davidson & Dym (2025) and Nikolentzos & Skianis (2026) charge an unmatched element its norm, the origin as the reservoir. Divol & Lacombe (2021) identify it with boundary transport and compute it by padding each side with the other's mass at the reservoir, the identity (5) (also Lacombe et al., 2018). Bubenik & Elchesen (2022a) prove that on any metric space with a distinguished subset the Wasserstein distance on such multisets is universal, the largest subadditive metric compatible with the ground metric, with Kantorovich–Rubinstein duality at $p = 1$: Lemma A.9(iii) is that theorem on the cone with its apex. What Theorem 3.2 adds is not the space but the identification. The configuration is a multiset because anonymity says so, the value cost at the smaller height is the value cap, the height is the cap weight rather than a mass, and the price R of a unit of height is the modulus of the linear sums (Proposition A.4). The axioms' per-coordinate moduli integrate, with constant 1 and for every pair of configurations, to exactly the matching cost (Appendix A.3). Theorem 3.2(ii), that the metric is the smallest making the class Lipschitz, is the class-side statement and a corollary of (i). Two consequences separate this from the mass-bookkeeping reading. The marginals are counting marginals, one atom per participant whatever its weight: with mass-weighted marginals the same cost gives the flat metric, whose Lipschitz ball is a strictly smaller class in which the selected maximum has no constant at all (Remark A.6). And the two semantics differ only in the height coordinate. The statistics done on these spaces asks different questions: consistency and minimax rates for the diagram of a sample under a support assumption (Chazal & Michel, 2021), populations of diagrams and their means (Mileyko et al., 2011; Turner et al., 2014), and plug-in rates for the Kantorovich–Rubinstein distance of unbalanced transport, for point-process intensities, under a covering dimension of the space (Struleva et al., 2025). On the diagram space a hypothesis on the space alone is weak, since with the bottleneck distance every separable bounded metric space embeds isometrically into it (Bubenik & Wagner, 2020), and Section 4 supplies what is missing, the design's profile over the strata of the lift.

Partitions, transport, and one-number dimensions. Bounding the transport distance between two measures by the mass they assign to the cells of a partition is the classical device behind the rates of convergence of empirical measures in Wasserstein distance from Dudley (1969) onward (Ajtai et al., 1984; Dereich et al., 2013; Boissard & Le Gouic, 2014; Fournier & Guillin, 2015), stated by Weed & Bach (2019, Prop. 4) in the form we use, one partition at one scale with a trimmed remainder. The stratified argument behind our rates has been used several times: Fournier & Guillin (2015) stratify $\mathbb{R}^d$ into annuli, Lei (2020) an unbounded function space into shells, and Huang et al. (2026), recently, the space of configurations into cardinalities, for the empirical measures of point processes, a setting that is the unweighted case of ours. We add two things. The first is *continuity*, the grading and the tubes: Theorem B.14 stratifies by the weight tail instead of the count, so that each stratum is a tube around a sparse set (Lemma B.12) with the crop in place of the truncation (Proposition B.11), and this is what gives rates on kernels of full support, where there is no count to stratify by, including the exponential rate from a decay inside the configuration (Theorem 4.5). The tail T_k itself is the best k-term ℓ^1 approximation error of the weight vector (Cohen et al., 2009), so an exponential profile is a geometric compressibility of the weights on average, taken here as a learning parameter. The second is that rate itself, for the mean tail of a configuration with free weights and values: quasi-polynomial rates of the same shape are those of Lei (2020, Thm. 4.2), for empirical measures in a Hilbert space under an exponential-ellipsoid moment condition, and of Huang et al. (2026, Thms. 2, 3), for point processes with a geometric count law, the unweighted case of a hard-ball selector, and neither hypothesis is ours (Appendix D.4 for the fiber, where the class's metric is Lei's ellipsoid). The rates of Weed and Bach are in one number of the measure, its upper Wasserstein dimension, a limit in the scale, or at finite n the trimmed covering count optimized over one scale

and one trim (Weed & Bach, 2019, Prop. 10). The entropic dimension of Chakraborty & Bartlett (2025), restated in Chakraborty & Bartlett (2024), is another such number, a trimmed covering in the limit of small scale, with a minimax lower bound in the upper Wasserstein dimension (Theorem 29 of the former), and the regression rates in the intrinsic dimension of the support (Bickel & Li, 2007; Kpotufe, 2011) read the support, which for the lifted flock is every stratum up to the crowded one. A design is not a measure with a dimension but a mixture of strata of different dimension with a mass on each, and the bulk of the lifted flock is such a union (Figure 2, counts from 2 to 23 other than the self between the 5% and 95% views). The certificate resolves the strata at the sample's own configurations, and the implied dimension of Appendix C.1 is the estimator of Block et al. (2022), who read the intrinsic dimension of a measure from the decay in n of the transport between two independent samples, applied to the lifted design. We keep from these works the point of contact, that a measure-level dimension controls rates, and the shape of the minimax lower bound.