# Clone-Robust Weights in Metric Spaces: Handling Redundancy Bias in Benchmark Aggregation

Damien Berriaud
ETHZ
Zürich, Switzerland
dberriaud@ethz.ch

Roger Wattenhofer
ETHZ
Zürich, Switzerland
wattenhofer@ethz.ch

## ABSTRACT

We are given a set of elements in a metric space. The distribution of the elements is arbitrary, possibly adversarial. Can we weigh the elements in a way that is resistant to such (adversarial) manipulations? This problem arises in various contexts. For instance, the elements could represent data points, requiring robust domain adaptation. Alternatively, they might represent tasks to be aggregated into a benchmark; or questions about personal political opinions in voting advice applications. This article introduces a theoretical framework for dealing with such problems. We propose *clone-proof weighting functions* as a solution concept. These functions distribute importance across elements of a set such that similar objects ("clones") share (some of) their weights, thus avoiding a potential bias introduced by their multiplicity. Our framework extends the maximum uncertainty principle to accommodate general metric spaces and includes a set of axioms —symmetry, continuity, and clone-proofness— that guide the construction of weighting functions. Finally, we address the existence of weighting functions satisfying our axioms in the significant case of Euclidean spaces and propose a general method for their construction.

## KEYWORDS

Redundancy Bias; Near-Clones; Weight Sharing; Local Voting; Metric Space

**ACM Reference Format:**
Damien Berriaud and Roger Wattenhofer. 2026. Clone-Robust Weights in Metric Spaces: Handling Redundancy Bias in Benchmark Aggregation. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/MJWJ6521

## 1 INTRODUCTION

Morpheus: "You take the **blue** pill and the story ends. You wake up in your bed and believe whatever you want to believe. You take the **red** pill, you stay in Wonderland and I show you how deep the rabbit hole goes." Before Neo can make his choice, Morpheus continues: "Or you can take this **indigo** pill, and wake up in Wonderland with \$100 in your pocket. Or this **navy** pill –with a different hair color." Why would Morpheus present these insignificant shades of blue? Neo already feels manipulated, but Morpheus continues excitedly and adds more pills colored in **bordeaux**, **cyan**, and **green**.



*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.* © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/MJWJ6521

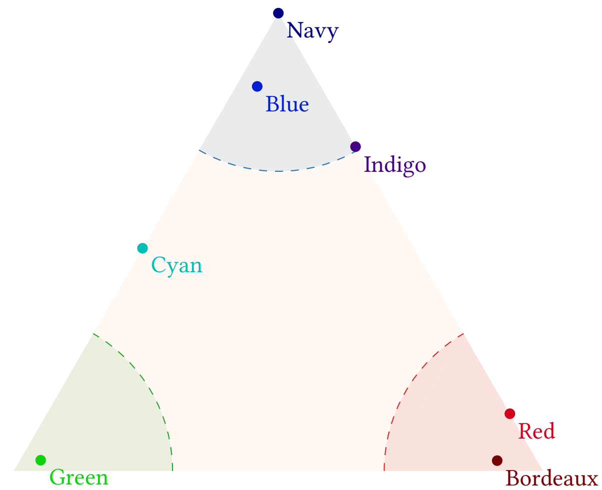



**Figure 1: Which weight should we give to each individual point? By symmetry, one would expect the areas in blue, red and green to sum up to the same value, even though they contain different numbers of points. How to deal with the addition of cyan though?**

Is there an objective way to make a choice without being bamboozled (see Figure 1)? This problem is at the heart of machine learning, but it can also be applied to various other areas, e.g., distributed systems or social choice. Let us set aside these related use cases and concentrate on our primary application: multi-task benchmark aggregation.

Consider a multi-task benchmark $\mathcal{B}$, e.g., GLUE [29]. The evaluation process of such a benchmark typically unfolds in two stages: First, a set of benchmark tasks $T = \{T_i\}_{i\in[n]}$ is defined, where each task $T_i$ maps a model $m \in \mathcal{M}$ to a score $T_i(m) \in \mathbb{R}$; Then, an aggregation rule $A$ combines the task-wise scores $(T_i(m))_{i\in[n]}$ into a single score $A(T(m)) \in \mathbb{R}$.

As discussed in prior works [9, 23, 31], the choice of an aggregation rule $A$ is perhaps best understood through the lens of social choice theory, where each task acts as a voter. However, multi-task benchmarking diverges from traditional voting scenarios in a key way: *anonymity*, or the equal treatment of votes, is not inherently required. In fact, it is well recognized that some tasks may exhibit significant similarity, such as CoLA [30] and SST-2 [24] in the case of GLUE. Perhaps tasks, similarly to colors in Figure 1, should share some of their weight? Indeed, the benchmark's outcome ought to remain unaffected by the inclusion of numerous highly similar tasks, as this could unfairly favor models that perform well on the original task over those excelling in other areas.

Although weighted aggregation rules have been proposed in the literature [23], the weights are typically chosen arbitrarily by the benchmark designer. In contrast, we propose a principled method for determining tasks' weights. Specifically, we suggest that the designer: (i) settles on a relevant measure of similarity between tasks, and embeds them in a metric space $(E, d)$; and (ii) calculates the tasks' weights using a weighting function with desirable axiomatic properties. In this article, we propose a set of axioms and weighting functions designed to be robust to noise and approximate clones, ensuring practical applicability in real-world settings. Using these weighting functions to assign task weights enables automatic scaling of the benchmark, i.e., the evaluation always benefits from adding new tasks, even if they are somewhat similar to previous tasks.

*Contribution & Outline.* This work proposes a mathematical framework for handling redundancy in a metric space. Specifically, we tackle the problem of determining the relative importance of elements in a finite set such that close-by elements, or "clones," share some of their weight. To extend the well-understood case of discrete metrics, where elements are either similar or equally dissimilar, we introduce in Section 3 the concept of weighting function and propose a set of axioms that such functions should satisfy in general metric spaces. These properties can be broadly categorized into three key principles: symmetry, continuity, and clone-proofness. Building on these foundations, we address in Section 4 the challenges of constructing functions that adhere to these axioms. For the specific case of Euclidean spaces, we resolve the question of existence and construct, in Theorems 1 and 2, a family of desirable weighting functions. We discuss in Section 5 the computational hurdles associated with their exact evaluation, and introduce Monte Carlo algorithms for their practical approximation. We finally explore in Section 6 possible extensions of our construction to more general spaces.

*Appendix.* Proofs and supporting material omitted from the main text are included in the full version of the paper [4]. In particular, the appendices contain further discussion of related work; a summary of notations and analytical tools; complete proofs of the results stated in the main text; a discussion of the axioms introduced in Section 3; and an extension of the framework to perfect clones.

## 2 RELATED WORK

In a recent line of work [9, 14, 23, 25, 31], multi-tasks benchmarking practices have been scrutinized through the lenses of social choice theory. In particular, these works question the usage of the arithmetical mean to aggregate scores of different tasks in popular benchmarks [15, 28] and investigate different aggregates such as the Pythagorean means [25], the Bradley-Terry model [21], or classical voting rules [9, 14, 23].

Contrary to usual voting scenarios where the equal treatment of voters is of utmost importance, there is apriori no requirement to treat each task equally in benchmark aggregation scenarios, and we may want to consider voting schemes with different weights, e.g., chosen arbitrarily by the benchmark creator. These weights may however carry more information than arbitrary preference. In [3], researchers proposed to model the evaluation of agents on different tasks through a zero-sum meta-game played between an "agent" and a "task" player, each choosing a probability distribution over the corresponding set. Scores on different tasks are then aggregated with a weighted average, where the weights correspond to the probability of playing each task in the entropy-maximizing Nash Equilibrium. One of the desirable properties of this technique is that it is invariant under the addition of *exact* copies of agents, a property which has been studied under the appellation *false-name-proofness* in social choice theory [10, 19, 27].

Similarly, independence of clones and its stronger form, composition consistency [5, 16, 26], have been proposed as desirable principles for handling the cloning of alternatives. Our approach differs in that, without an internal order to distinguish between clones, we choose to treat them all symmetrically. Furthermore, all the above properties are very brittle and offer no guarantees whenever minimal noise is added to a clone.

Shortly before our preprint was posted, Procaccia et al. [22] proposed robustness to approximate clones as a desirable property for preference aggregation in reinforcement learning with human feedback. They propose a clone-robust weighing of the alternatives based on Voronoi diagrams. However, we show in the full version of the paper [4] that this weighting function has undesirable properties; in particular, it is discontinuous in each perfect clones, and close-by points may receive entirely dissimilar weights.

Importantly, our framework assumes that practitioners provide a suitable distance metric, as its selection lies beyond the scope of this work. Identifying an appropriate distance metric between tasks or datasets is a critical prerequisite for applying our approach effectively. Fortunately, this challenge has been extensively explored [2, 12, 17], particularly within the transfer learning literature [1, 20]. This existing body of research complements our work and provides valuable guidance for practitioners seeking to use our framework for benchmark aggregation.

## 3 WEIGHTING FUNCTIONS AND DESIRABLE AXIOMS

In this section, we formally introduce weighting functions and propose a set of axioms that we consider essential for generalizing the well-understood case of discrete metrics.

Consider a metric space $(E, d)$, that is a set $E$ equipped with a notion of distance in the form of an operator $d : E \times E \mapsto \mathbb{R}_{\geq 0}$ satisfying *separability*, *symmetry* and *triangular inequality*. We now formally define the object of interest of this work, called *weighting functions of* $(E, d)$.

Definition 1 (Weighting functions of $(E, d)$). *A weighting function of $(E, d)$ is a function $f$ that maps finite sets of $E$ to probability distributions over their elements, i.e.,*

$$f : S \in \mathcal{P}(E) \mapsto p_S \in \Delta(S),$$

*where $\mathcal{P}(E)$ denotes the set containing all finite subsets of $E$ (outside the empty set), and $\Delta(S) = \left\{ p_S : S \mapsto [0, 1] \mid \sum_{x \in S} p_S(x) = 1 \right\}$ denotes the simplex over the elements of $S$. We moreover refer to the probability distribution $f(S) : S \mapsto [0, 1]$ as the weighting of $S$.*

Note that this definition encompasses the *uniform distribution* as a particular case of weighting function. Indeed, consider the discrete metric space $(E, \rho)$, where $\rho(x, y)$ is equal to one if $x \neq y$ and zero otherwise. Then the maximum entropy principle compels us to use the *uniform weighting function* $\mathcal{U} : S \in \mathcal{P}(E) \mapsto \mathbb{1}_S(\cdot)/|S| \in \Delta(S)$.

Drawing inspiration from the properties of this particular weighting function, we next introduce a few axioms that we argue are desirable for a general metric space $(E, d)$ and weighting function $f$ thereof. The first desirable property that the uniform weighting $\mathcal{U}$ verifies is rather simple: it ensures that all elements of a finite set are represented with positive probability. This means that it never hurts to add new elements to a set as the support of the probability distribution given by the weighting function only increases.

Axiom 1 (Positivity). *Every element of a finite set is represented with positive probability, i.e., for all finite subset $S \in \mathcal{P}(E)$ and element $x$ in $S$, we have $f(S)(x) > 0$.*

The second property of $\mathcal{U}$ that we would want to extend to a generic $f$ is that of symmetry: when the distance is uninformative and some elements are isomorphic with respect to a distance preserving permutation $\sigma_S : S \mapsto S$, they receive similar weights. In particular, if all elements of a finite subset $S$ are equidistant, then $f(S)$ should be uniform over $S$.

Axiom 2 (Symmetry). *Elements of a set that are symmetric with respect to the metric are equally represented, i.e., for all finite subset $S \in \mathcal{P}(E)$ and self-isometry $\sigma_S : S \mapsto S$, it holds for all $x \in S$ that $f(S)(x) = f(S)(\sigma_S(x))$.*

Importantly, $\sigma_S$ need not be extendable to a full isometry on $E$. Moreover, determining the automorphism group of a set $S$ is an instance of the *graph automorphism problem*, which is known to be solvable in quasi-polynomial time [13], but is neither known to be in P nor to be NP-complete. Luckily, two symmetric elements $x$ and $\sigma_S(x)$ possess the same multi-set of distances $\{\{d(x, y)\}\}_{y \in S}$ and we only need to make sure that similar multi-sets lead to similar weightings.

Since perfect clones at distance precisely zero are always isomorphic with one another, Axiom 2 requires in particular that they get equal weights, and can hence be thought of as requiring *fairness among perfect clones.* We may want to extend this fairness requirement beyond perfect clones to include approximate ones as well. Unequal treatment between the two could undermine robustness – particularly in adversarial settings like data poisoning, where strategically crafted approximate clones could divert all the weight away from the original elements.

Axiom 3 (Uniform Clone Fairness). *Weighting is fair among approximate clones, i.e., for all $\varepsilon > 0$, there exists $\delta > 0$ such that, for all finite subset $S \in \mathcal{P}(E)$ and $x, y$ in $S$ satisfying $d(x, y) \leq \delta$, it holds that $|f(S)(x) - f(S)(y)| \leq \varepsilon$.*

Finally, the third property that $\mathcal{U}$ trivially satisfies is that of continuity, since the topology induced by the metric $\rho$ is the discrete one. Intuitively, we would want to ensure that slightly perturbing each element of a subset $X$ does not result in large variations in weighting. Formally, we define, for two finite subsets $X$ and $Y$ in $\mathcal{P}(E)$ of cardinality $k \in \mathbb{N}$, the *transport distance* $d_\Pi(X, Y) = d_\Pi(Y, X) = \min_{\pi \in \mathrm{Bij}(Y,X)} \max_{y \in Y} d(y, \pi(y))$, where $\mathrm{Bij}(Y, X)$ denotes the set of bijections from $Y$ to $X$. We similarly define the set of *minimal transport maps* from $Y$ to $X$ as $\Pi(Y, X) = \mathrm{argmin}_{\pi \in \mathrm{Bij}(Y,X)} \max_{y \in Y} d(y, \pi(y))$. Using $\pi \in \Pi(Y, X)$ to identify element $x$ in $X$ with a $\pi^{-1}(x)$ in $Y$, we then require that both get similar weights in their respective sets.

Axiom 4 (Uniform Individual Continuity). *Weighting is element-wise continuous, i.e., for all $\varepsilon > 0$ and $k \in \mathbb{N}$, there exists $\delta > 0$ such that, for all finite subsets $X, Y \in \mathcal{P}(E)$ of cardinality $|X| = |Y| = k$ such that $d_\Pi(X, Y) \leq \delta$, we have $\max_{x \in X} |f(X)(x) - f(Y)(\pi^{-1}(x))| \leq \varepsilon$, where $\pi \in \Pi(Y, X)$.*

We show in the full version of the paper [4] that Axiom 4 implies continuity with the Wasserstein metric on the codomain of $f$. Though similar in formulation, note that Axiom 3 cannot be derived from Axiom 4 by simply plugging in $Y = X$, since the identity is always the unique minimal transport map (in the absence of perfect clones).

One might wonder why we restrict our continuity requirement to sets of the same cardinality. Indeed, for two finite subsets $X$ and $Y$ with cardinality $|Y| \geq |X|$, it is possible to extend the definition of $d_\Pi(Y, X) = d_\Pi(X, Y) = \min_{\pi \in \mathrm{Surj}(Y,X)} \max_{y \in Y} d(y, \pi(y))$ by requiring that $\pi \in \mathrm{Surj}(Y, X)$ is only a surjection (we similarly extend the definition of $\Pi(Y, X)$). We then show in the extended version that $d_\Pi$ constitutes a metric on the whole domain $\mathcal{P}(E)$. Importantly, note that only sets $Y$ of cardinality greater or equal to that of $X$ satisfy $d_\Pi(X, Y) \leq \delta$ for small enough $\delta$; indeed, for $\delta$ smaller than $\underline{d}(X) = \min_{x \neq x' \in X} d(x, x')/2$, no element $y$ in $E$ can be simultaneously $\delta$-close to distinct $x$ and $x'$ in $X$ (c.f. Figure 2a). For such $Y$, a surjection $\pi : Y \mapsto X \in \Pi(X, Y)$ still offers a natural way to identify elements of $Y$ with those of $X$, and we could think of each $\pi^{-1}(x) = \{y \in Y \mid \pi(y) = x\}$ as a *class of clones* since all $y, y'$ in $\pi^{-1}(x)$ are at distance at most $2\delta$ by the triangle inequality.

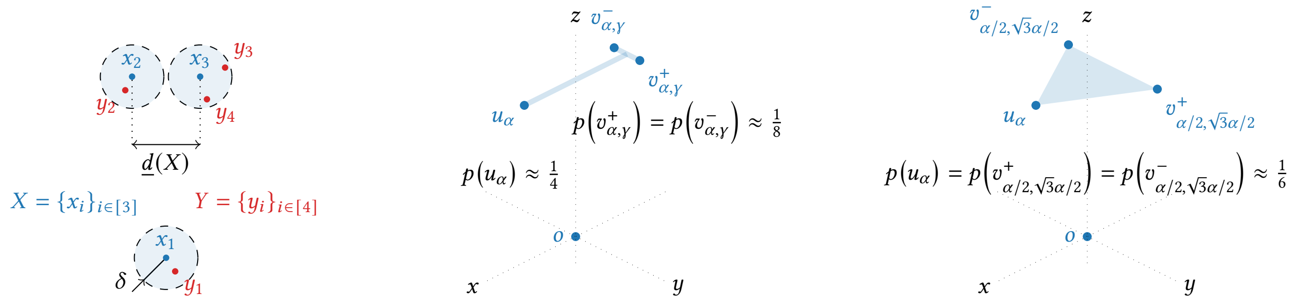
Summing weights locally over each class of clones, we could then collapse some of the dimensions of the codomain $\Delta(Y)$ and identify it with $\Delta(X)$. With this intuition, we then define *class continuity* as follows.

Axiom 5 (Class Continuity). *Weights are class-wise continuous, i.e., for a finite subset $X \in \mathcal{P}(E)$ and $\varepsilon > 0$, there exists $\delta$ verifying $\min_{x \neq x' \in X} d(x, x')/2 > \delta > 0$ such that, for each finite subset $Y \in \mathcal{P}(E)$ satisfying $d_\Pi(X, Y) \leq \delta$, we have $\max_{x \in X} \left| f(X)(x) - \sum_{y \in \pi^{-1}(x)} f(Y)(y) \right| \leq \varepsilon$, where $\pi \in \Pi(X, Y)$.*

Note that Axiom 5 ensures a form of *cloneproofness*, i.e., robustness of weighting under the addition of clones. Intuitively, a $\delta$-neighboring set $Y$ of $X$ with greater cardinality contains many $\delta$-clones, and Axiom 5 ensures that both sets get "similar weightings" when summing probabilities locally over the redundancies in $Y$.

However, we argue that the local summation in Axiom 5, although intuitive, is too strong of a requirement. Indeed, we show hereafter that a weighting function $f$ verifying Axioms 2 and Axiom 5 gives very different individual weights to points in nearby sets, and breaks Axiom 3. Note that similar arguments explain the choice of $d_\Pi$ over a perhaps more standard *Hausdorff distance.*

Example (Diverging Individual Weights.). *Let $f$ be a weighting function on the three-dimensional Euclidean space $(\mathbb{R}^3, d_2)$ satisfying both Axioms 2 and 5, and define the parametric family $S_{\alpha,\beta,\gamma} =$*




**(a) Sets $X$ and $Y$ at distance $d_\Pi(X,Y) \leq \delta < \underline{d}(X)/2$.** **(b) $S_{\alpha,\alpha,\gamma}$ for $1 \gg \alpha \gg \gamma > 0$.** **(c) $S_{\alpha,\alpha/2,\sqrt{3}\alpha/2}$ for $1 \gg \alpha > 0$**

**Figure 2: Visualization of the neighborhoods of $d_\Pi$, and of the divergence of individual weightings under Axioms 2 and 5. The edges in 2b highlight the symmetries of $S_{\alpha,\alpha,\gamma}$ in the limit $\gamma \to 0$; the equilateral triangle in 2c displays the symmetry of $S_{\alpha,\alpha/2,\sqrt{3}\alpha/2}$.**

*$\{o, u_\alpha, v^+_{\beta,\gamma}, v^-_{\beta,\gamma}\}$, where $o = (0,0,0)$, $u_\alpha = (1,\alpha,0)$, $v^+_{\beta,\gamma} = (1,-\beta,\gamma)$ and $v^-_{\beta,\gamma} = v^+_{\beta,-\gamma}$. Figure 2 summarizes our construction.*

*On one hand, consider the set $S_{\alpha,\alpha,\gamma}$, where $\alpha > 0$ is fixed and $\gamma > 0$ is much smaller than $\alpha$. Since $S_{\alpha,\alpha,\gamma}$ converges to the set $S_\alpha = \{o, u_\alpha, u_{-\alpha}\}$ in $(\mathcal{P}(\mathbb{R}^3), d_\Pi)$ when $\gamma$ goes to zero, Axiom 5 implies $\lim_{\gamma\to 0} f(S_{\alpha,\alpha,\gamma})(u_\alpha) = f(S_\alpha)(u_\alpha)$. Moreover, $S_\alpha$ converges in turn to the symmetric set $S_0 = \{o, u_0\}$, and Axioms 2 and 5 together imply that $\lim_{\alpha\to 0} f(S_\alpha)(u_\alpha) = f(S_0)(u_0)/2 = 1/4$. Combining these two results, we get*

$$\lim_{\alpha\to 0}\lim_{\gamma\to 0} f(S_{\alpha,\alpha,\gamma})(u_\alpha) = 1/4.$$

*On the other hand, consider the set $S_{\alpha,\alpha/2,\sqrt{3}\alpha/2}$. Note that the points $u_\alpha$, $v^+_{\alpha/2,\sqrt{3}\alpha/2}$ and $v^-_{\alpha/2,\sqrt{3}\alpha/2}$ form an equilateral triangle centered in $u_0$ and orthogonal to the origin, hence by Axiom 2 they must receive similar weights. As $S_{\alpha,\alpha/2,\sqrt{3}\alpha/2}$ also converges to the symmetric $S_0$ when $\alpha$ goes to zero, Axioms 2 and 5 finally imply*

$$\lim_{\alpha\to 0} f\big(S_{\alpha,\alpha/2,\sqrt{3}\alpha/2}\big)(u_\alpha) = 1/6.$$

*We hence constructed two arbitrarily close sets of the same cardinality whose individual weightings differ. Moreover, $u_\alpha$ and $v^+_{\alpha,\gamma}$ receive vastly different weights in the limit $\alpha \to 0, \gamma \to 0$, although their distance tends to zero, i.e.,*

$$\frac{1}{4} = \lim_{\alpha\to 0}\lim_{\gamma\to 0} f(S_{\alpha,\alpha,\gamma})(u_\alpha) \neq \lim_{\alpha\to 0}\lim_{\gamma\to 0} f(S_{\alpha,\alpha,\gamma})(v^+_{\alpha,\gamma}) = \frac{1}{8},$$

*and Axiom 3 breaks.*

Given this incompatibility, we preserve Axiom 3 and opt for the restriction to sets of similar cardinality in Axiom 4. However, note that Axiom 4, unlike Axiom 5, does not directly address the addition of clones. To account for this, we introduce another weakening of Axiom 5, essentially requiring continuity of weighting everywhere except in the vicinity of the newly added clone. This relaxation allows for greater flexibility in how the mass is redistributed locally.

Axiom 6 (Uniform $\alpha$-Locality under Addition of Clones). *The addition of a clone only changes the weights of points in the $\alpha$-neighborhood of the clone, i.e., for all $\varepsilon > 0$, there exists $\delta > 0$ such that, for each finite subset $S \in \mathcal{P}(E)$ and elements $x \in S$ and $x' \in E \setminus S$ satisfying $d(x,x') \leq \delta$, we have for all $z \in S$ such that $d(x,z) \geq \alpha$ that $|f(S)(z) - f(S \cup \{x'\})(z)| \leq \varepsilon$.*

Note that Axioms 3 and 6 provide orthogonal restrictions in the presence of clones: the former dictates how to shift weights around the recently introduced clone, while the latter ensures weights do not change away from it. We further discuss the relationship between these axioms in the full version of the paper [4].

We finally denote by $\mathcal{R}_\alpha(E,d)$ the set of weighting functions on $(E,d)$ satisfying Axioms 1, 2, 3, 4 and 6 with parameter $\alpha > 0$. The burning question is now this: does the set $\mathcal{R}_\alpha(E,d)$ contain any elements at all?

*Back to the Motivating Example.* Let's now examine the guarantees provided by our axioms in the context of multi-task benchmark aggregation. First, Axiom 1 ensures that each task has a contribution to the final aggregated score (in fact, we show a stronger guarantee for $g_r$ in the proof of Theorem 1). Axioms 2 and 3 both reflect principles of anonymity in social choice. Axiom 2 ensures some level of *isotropy* in the embedding space, meaning that tasks which are indistinguishable under the symmetry of the embedding space receive equal weights. Axiom 3 guarantees that tasks that are deemed similar based on their embedding will be assigned similar weights. Axiom 4 addresses the *continuity of the weighting*. It ensures that small changes in a task –such as adding a few elements to a test set– should not lead to large changes in the assigned weight. Finally, Axiom 6 enforces *weight sharing for similar tasks*. It ensures that tasks with sufficiently different characteristics remain unaffected by the introduction of new yet partially redundant tasks to the benchmark.

Philosophically, our framework offers insight into the assumptions about the space of tasks that are baked into a given choice of weights. For instance, giving equal weight to each task could reflect various possible scenarios: it could suggest that tasks are either all equally dissimilar, all similar, or simply symmetric within the space. The continuity in Axiom 4 is also of philosophical significance. While the construction of a benchmark is inherently a discrete process, one may hope that small changes around the benchmark (such as adding noise) do not result in completely different outcomes.

## 4 LOCAL VOTING APPROACH

To construct weighting functions that satisfy our axioms, the first step is to identify invariant objects under the addition of clones: we argue that the open balls around an element $x \in E$, that is $B_r(x) := \{y \in E \mid d(x,y) < r\}$ for some radius $r > 0$, are natural invariants for our problem. Indeed, they are stable under the addition of clones, in the sense that for some $\delta$-*clones* $x, y$ in $E$ satisfying $d(x,y) \leq \delta$, the triangle inequality ensures that $B_r(x) \subseteq B_r(x) \cup B_r(y) \subseteq B_{r+\delta}(x)$. If we then equip our space with a measure $\mu$ defined on the open balls of the space[1] and associate with each finite subset $X \subseteq E$ its neighborhood $B_r(X) := \bigcup_{x \in X} B_r(x)$, we obtain a map invariant under clone addition. Indeed, for each neighboring finite set $Y \subseteq B_\delta(X)$ with $r > \delta > 0$, we have $\mu\big(B_{r-\delta}(X)\big) \leq \mu\big(B_r(Y)\big) \leq \mu\big(B_{r+\delta}(X)\big)$ and the map $X \in \mathcal{P}(E) \mapsto \mu\big(B_r(X)\big)$ is continuous with respect to the distance $d_\Pi$.[2]

Note however that further requirements are needed to satisfy the symmetry in Axiom 2, essentially regarding the homogeneity and the isotropy of the underlying measure space. For this reason, we focus on Euclidean spaces $(\mathbb{R}^n, d_2)$ for the remainder of the section, where $d_2^2(x,y) = \sum_{i=1}^n (x_i - y_i)^2$ for all $x = (x_i)_{1\leq i\leq n}$ and $y = (y_i)_{1\leq i\leq n}$ in $\mathbb{R}^n$. We will discuss in Section 6 how to adapt our approach to more general metric spaces.

Based on the above invariant, we construct a weighting function as a local voting scheme. For a fixed $r > 0$ and finite subset $S \subseteq \mathbb{R}^n$, we consider each element of $B_r(S)$ as a voter that approves only of the candidates in $S$ close to him, and as such spreads his voting power equally among them. Formally, we define the grade that each voter $z$ in $B_r(S)$ attributes to a candidate $x$ in $S$ as follows

$$g_{r,S,x}(z) = \frac{\mathbb{1}_{B_r(x)}(z)}{\sum_{y \in S} \mathbb{1}_{B_r(y)}(z)}.$$

We then aggregate the ballots with the Lebesgue measure $\mu$ and finally define the weighting function $g_r$, for each finite subset $S \subseteq \mathbb{R}^n$, i.e.,

$$g_r(S) : x \in S \mapsto \int_{B_r(S)} \frac{g_{r,S,x}(z)}{\mu\big(B_r(S)\big)}\, d\mu(z).$$

As illustrated in Figure 3, the weighting function $g_r$ computes a weighted average of the inverse depth of each cell, with the depth defined as the number of intersecting balls forming the cell and the weights based on the cell's size.

We next show that this class of weighting functions satisfies the desirable axioms introduced above.

Theorem 1. *For $r > 0$, the weighting function $g_r$ is well-defined and belongs in $\mathcal{R}_{2r}(\mathbb{R}^n, d_2)$.*

The detailed proof of Theorem 1 is included in the full version of the paper [4]; we provide here a sketch of the proof. As depicted in Figure 3, the weight $g_r(S)(x)$ is in fact a weighted average of positive elements, hence it is positive and Axiom 1 trivially holds. Showing that $g_r$ is symmetric (Axiom 2) is also relatively straightforward after observing the following two properties of Euclidean spaces: first, the fact that one can uplift any self-isometry $\sigma_S$ on a finite subset $S$ to an isometry on the entire space $\mathbb{R}^n$ ; second, the fact that the Lebesgue measure is invariant under *translations*, *rotations* and *reflections*, which generate the group of Euclidean isometries [11]. The most challenging aspect of the proof is verifying that $g_r$ satisfies Axioms 3, 4 and 6. While these proofs are technically intricate, they fortunately follow a similar structure. We illustrate our approach by focusing on the simpler case of Axiom 3 below. In order to bound the difference of weightings $|g_r(S)(x) - g_r(S)(y)|$ between two approximate clones $x$ and $y$ in a given set $S$, we first show that the grading functions $g_{r,S,x}$ and $g_{r,S,y}$ are equal outside of a thin spherical shell parametrized by $\delta$, the distance between $x$ and $y$ (c.f. Figure 4a). This allows us to obtain a difference of Lebesgue measure $\mu(B_r(x)) - \mu(B_{r-\delta}(x))$, which we then bound in terms of $\delta$ by taking the limit of this difference as $\delta$ approaches zero (c.f. Figure 4b). The formalization of this argument relies on tools from *geometric measure theory*, particularly the $n-1$-dimensional Minkowski content. These arguments are illustrated in Figure 4.

Since $\mathcal{R}_{2\alpha}(E,d)$ is monotonically increasing in positive $\alpha$, Theorem 1 actually ensures that the whole collection $\{g_r\}_{\alpha \geq r > 0}$ belongs in $\mathcal{R}_{2\alpha}(\mathbb{R}^n, d_2)$. Moreover, it is relatively straightforward to see that $\mathcal{R}_{2\alpha}(\mathbb{R}^n, d_2)$ is a convex set and, as such, contains all finite convex combinations of $\{g_r\}_{\alpha \geq r > 0}$. Since the weighting functions $g_r$ are well-behaved, we generalize this result as follows.

Theorem 2. *Let $\nu$ be a probability density function over $[0, \alpha]$. Then the weighting function $f_\nu : S \in \mathcal{P}\big(\mathbb{R}^n\big) \mapsto \int_0^\alpha \nu(r) g_r(S)\, dr$ belongs in $\mathcal{R}_{2\alpha}(\mathbb{R}^n, d_2)$.*

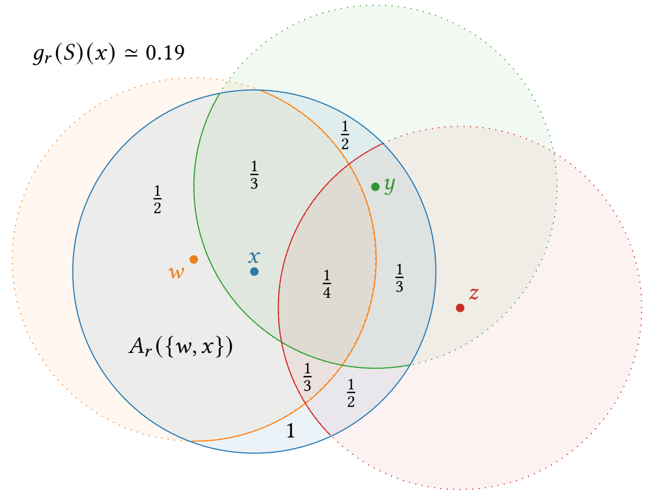
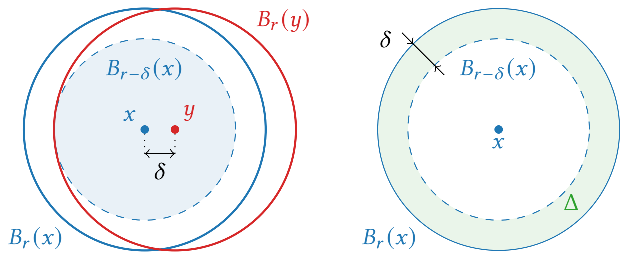



**Figure 3: Computation of $g_r(S)(x)$ in the two-dimensional Euclidean space $(\mathbb{R}^2, d_2)$, where the set $S = \{w, x, y, z\}$ contains four elements. A cell $A_r(U)$ is uniquely defined by the subset $U \subseteq S$ as the possibly empty intersection of the balls around each element in $U$ and the complement of the balls of each element of $S$ absent from $U$. For each subset $U$ containing $x$, the grading function $g_{r,S,x}$ is constant on the cell $A_r(U)$ and equal to the inverse depth of the cell, i.e., $g_{r,S,x}(z) = 1/U$ for all $z$ in $A_r(U)$. The weight of $x$ in $S$ is then equal to the weighted average of $g_{r,S,x}$ on the ball centered in $x$, where the weight of each cell corresponds to its area normalized by the total area of the balls' union. We estimated the value $g_r(S)(x) \simeq 0.19$ via Monte Carlo sampling, c.f. Algorithm 1.**

[1] I.e., on the Borel $\sigma$-algebra.
[2] At least when $\mu$ is locally finite.




**(a) The ball $B_{r-\delta}(x)$ belongs in the intersection of $B_r(x)$ and $B_r(y)$, hence $x$ and $y$ receive the same grade from every voter on $B_{r-\delta}(x)$.**

**(b) The $n^{\text{th}}$-dimensional volume of the green set $\mu(\Delta)$ can be approximated as $\delta$ times $S_r^{n-1}$, the $n-1^{\text{th}}$-dimensional surface of a ball of radius $r$.**

**Figure 4: Key steps in demonstrating that $g_r$ satisfies Axioms 3, 4 and 6.**

The detailed proof of Theorem 2, provided in the extended version of the paper [4], relies on inequalities derived for the proof of Theorem 1.

## 5 COMPUTATIONAL CONSIDERATIONS

Our focus thus far has been on identifying weighting functions with theoretically desirable properties. However, from a practical standpoint, such tools are of limited utility if they cannot be computed efficiently. This concern is encapsulated in the following principle.

Axiom 7 (Exact Computability). *The weighting of any given subset is efficiently computable, i.e., for any subset $S \in \mathcal{P}(E)$, the probability distribution $f(S)$ can be exactly computed in time polynomially bounded by the cardinality $|S|$ of the subset, and the dimension $n$ of the space when $E = \mathbb{R}^n$.*

It is worth noting that the weighting functions introduced in Section 4 are unlikely to meet this criterion. For example, computing $g_r(S)(x)$ would a priori involve averaging $g_{r,S,x}$ over as many as $O(2^{|S|})$ disjoint cells, making the approach computationally infeasible. Even the simpler task of evaluating the volume $\mu(B_r(S))$ of the union of Euclidean balls becomes increasingly challenging in higher dimensions (see [7] for the case $n = 3$). While hardness results for this exact problem are not readily available, related geometric problems –such as computing the volume of the union of axis-aligned boxes– are known to be #P-hard [6], suggesting that exact computation of $g_r(S)$ is unlikely to admit a polynomial-time solution in general.

In light of these challenges, we may want to relax Axiom 7 and settle for efficient *approximate* evaluations of $g_r(S)$ and $f_v(S)$. Monte Carlo sampling techniques [6, 18] offer a natural route in this direction: by sampling random points in metric neighborhoods of the elements of $S$, one can estimate the relevant integrals within a prescribed accuracy $\varepsilon > 0$ and confidence level $1 - \delta$.

We next formalize this idea and introduce a simple Monte Carlo sampling procedure –namely Algorithm 1– which provides a consistent and asymptotically unbiased estimator of $g_r(S)$. The following theorem establishes explicit $(\varepsilon, \delta)$-style guarantees on its accuracy.

**Algorithm 1** Naive Monte Carlo Estimation of $g_r(S)$

**Input:** Finite set $S \subset \mathbb{R}^n$, radius $r > 0$, sample size $k \in \mathbb{N}$.
**Precomputation:** For each $x \in S$, compute the closed neighborhood $B_{2r}(x) \cap S$.
**for** each $x \in S$ **do**
  Initialize $\widehat{N}_x \leftarrow 0$.
  **for** $i = 1, \dots, k$ **do**
    Sample $z_i \sim \mathrm{Unif}(B_r(x))$.
    Compute count $c_i := \big|\{y \in B_{2r}(x) \cap S \mid d_2(z_i, y) \le r\}\big|$.
    Update $\widehat{N}_x \leftarrow \widehat{N}_x + \frac{1}{c_i}$.
  **end for**
  Set $\widehat{N}_x \leftarrow \widehat{N}_x / k$.
**end for**
Compute normalization factor $\widehat{D} := \sum_{x \in S} \widehat{N}_x$.
**Output:** Estimates $\widehat{g}_r(S)(x) := \widehat{N}_x / \widehat{D}$ for each $x \in S$.

Theorem 3. *Algorithm 1 yields a consistent and asymptotically unbiased estimate of $g_r(S)$. Moreover, for any target accuracy $\varepsilon > 0$ and confidence level $\delta \in (0, 1)$, setting the number of samples to satisfy*

$$k \ge \frac{(|S|^2 - 1)^2}{2\varepsilon^2 |S|^2} \ln \frac{2|S|}{\delta}$$

*guarantees with probability at least $1 - \delta$ that*

$$\left| \widehat{g}_r(S)(x) - g_r(S)(x) \right| \le \varepsilon.$$

The proof of Theorem 3, included in the full version of the paper [4],proceeds by separately estimating the unnormalized numerators –the average grade of a random voter $z$ sampled uniformly from $B_r(x)$– and the common normalization factor in the denominator –the relative volume of the union of balls $\mu(B_r(S))/\mu(B_r(x))$. Standard concentration results –specifically Hoeffding's inequality– are then applied to control the deviation of these empirical averages with high probability, and the final error bound is obtained by carefully propagating this deviation through the ratio of the two estimators $\widehat{N}_x / \widehat{D}$.

The resulting accuracy guarantee follows the typical Monte Carlo rate $\varepsilon \propto 1/\sqrt{k}$, which is optimal up to constant factors for independent sampling schemes. The total runtime of Algorithm 1 is $O(nk|S|^2)$, since each $c_i$ can be computed in $O(n|S|)$ time, and the algorithm produces an estimate for each of the $|S|$ elements in $S$. Importantly, the required number of samples per estimate, $k$, is *independent of the dimension $n$* of the Euclidean space $\mathbb{R}^n$. Note however that $k$ scales (up to a logarithmic factor) as $k \propto |S|^2$. This quadratic dependency arises because the errors on the numerators accumulate when estimating the denominator, which must therefore be approximated to precision $\varepsilon/|S|$ to achieve an overall accuracy of $\varepsilon$. More generally, a $|S|^2$ scaling is essentially unavoidable whenever we seek a simultaneous high-probability accuracy guarantee for all $x \in S$. Indeed, controlling the joint deviation of all $|S|$ estimates via a union bound forces each individual estimate to be of order $O(\varepsilon/|S|)$ in order to achieve a global accuracy of $\varepsilon$.

If the $O(|S|^2)$ cost is prohibitive and only a single weight $g_r(S)(x)$ is needed, one can instead consider Algorithm 2, which leverages the ApproxUnion algorithm from [6] as a subroutine to directly approximate the volume of the union of balls $\mu(B_r(S))$.

**Algorithm 2** Estimation of $g_r(S)(x)$ with ApproxUnion

**Require:** Finite set $S \subset \mathbb{R}^n$, element $x \in S$, radius $r > 0$, accuracy parameter $\varepsilon > 0$, and failure probability $\delta \in (0, 1)$.

**Ensure:** Output $\widehat{g}_r(S)(x)$ verifies $|\widehat{g}_r(S)(x) - g_r(S)(x)| \le \varepsilon$ with probability at least $1 - \delta$.

1: Set sample size $k \leftarrow \left\lceil \frac{8(|S|-1)^2}{\varepsilon^2 |S|^2} \ln \frac{4}{\delta} \right\rceil$
2: Sample $z_1, \dots, z_k \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(B_r(x))$.
3: Compute empirical average

$$\widehat{N}_x \;\leftarrow\; \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\left|\{y \in S \mid d_2(z_i, y) \le r\}\right|}.$$

4: Set amplification parameter $t \leftarrow \left\lceil \ln \frac{2}{\delta} \right\rceil$.
5: **for** $j = 1, \dots, t$ **do**
6: $\quad \widehat{U}_j \leftarrow \mathrm{ApproxUnion}(\{B_r(x) : x \in S\}, \varepsilon/4)$.
7: **end for**
8: Set $\widehat{U} \leftarrow \mathrm{median}(\widehat{U}_1, \dots, \widehat{U}_t)$.
9: Compute $\widehat{D} \leftarrow \widehat{U}/\mathrm{vol}(B_r(x))$.
10: **Output:** Return $\widehat{g}_r(S)(x) := \widehat{N}_x / \widehat{D}$.

A proof that Algorithm 2 achieves the stated high-probability accuracy guarantee –quantified in terms of target precision $\varepsilon$ and confidence level $\delta$– can be found in the extended version of the paper [4].In our setting, all geometric oracles required by ApproxUnion –membership testing within a ball, volume computation, and uniform sampling – can be implemented efficiently, each in time $O(n)$ with full numerical precision. Consequently, a single execution of $\mathrm{ApproxUnion}\big(\{B_r(x) : x \in S\}, \varepsilon/4\big)$ requires at most $T = 128 \ln(8)\,(1 + \varepsilon/4)\,|S|/\varepsilon^2$ random samples. This yields an overall runtime for Algorithm 2 of $O\big(n\,|S|\,\varepsilon^{-2} \ln \delta^{-1}\big)$. Rather than using ApproxUnion merely as a subroutine to estimate $\mu(B_r(S))$, one could instead implement a refined backtracking procedure throughout its execution to simultaneously produce an estimate of $\widehat{N}_x\,\mu(B_r(x))$. Such an approach could, in principle, produce a multiplicative $\varepsilon$-approximation of $g_r(S)(x)$, improving upon the additive accuracy guarantee obtained in the current formulation.

We now turn to the estimation of $f_\nu(S)$ and introduce Algorithm 3, a simple two-stage Monte Carlo sampling method that relies on Algorithm 1 as a subroutine to estimate $g_{r_i}(S)$ for multiple radii $r_i$ drawn independently from $\nu$. The theoretical guarantees of this procedure, including explicit confidence and precision bounds, are established in Theorem 4.

Theorem 4 (Naive Monte-Carlo Estimation for $f_\nu$). *For any target accuracy $\varepsilon > 0$ and confidence level $\delta \in (0, 1)$, setting the outer and inner sample sizes in Algorithm 3 such that*

$$M \;\ge\; \frac{8}{\varepsilon^2} \ln \frac{4|S|}{\delta}, \qquad k \;\ge\; \frac{2(|S|^2 - 1)^2}{\varepsilon^2 |S|^2} \ln \frac{4|S|M}{\delta},$$

*ensures with probability at least $1 - \delta$ the following bound for every $x \in S$, i.e.,*

$$\left|\widehat{f}_\nu(S)(x) - f_\nu(S)(x)\right| \le \varepsilon,$$

*and the estimator is consistent. Moreover, if $k = k(M) = \Omega(\log M)$ (e.g. $k(M)$ chosen as above), then Algorithm 3 also yields an asymptotically unbiased estimator, i.e.,* $\lim_{M\to\infty} \mathbb{E}[\widehat{f}_\nu(S)(x)] = f_\nu(S)(x)$.

**Algorithm 3** Naive Monte Carlo Estimation of $f_\nu(S)$

1: **Input:** Finite set $S \subset \mathbb{R}^n$, distribution $\nu$ on $[0, \alpha]$, outer and inner-sample sizes $M, k \in \mathbb{N}$, respectively.
2: Sample $M$ radii $r_1, \dots, r_M \overset{\text{i.i.d.}}{\sim} \nu$.
3: **for** $j = 1, \dots, M$ **do**
4: $\quad$ Run Algorithm 1 with radius $r_j$ and $k$ samples.
5: $\quad$ Obtain estimates $\hat{g}_{r_j}(S)(x)$ for all $x \in S$.
6: **end for**
7: For each $x \in S$, compute the average

$$\hat{f}_\nu(S)(x) \;=\; \frac{1}{M} \sum_{j=1}^{M} \hat{g}_{r_j}(S)(x).$$

8: **Output:** Estimates $\hat{f}_\nu(S)(x)$ for all $x \in S$.

The proof of Theorem 4, given in the full version of the paper [4], is standard and proceeds by separately controlling the errors of the inner and outer Monte Carlo estimates. Critically, the runtime of Algorithm 3 is $M$ times that of Algorithm 1, that is, $O(Mnk|S|^2)$. However, $M$ must scale as $\propto \varepsilon^{-2}$ to control the outer error, which results in an overall runtime $\propto \varepsilon^{-4}$ to achieve an additive accuracy of $\varepsilon > 0$. More sophisticated statistical techniques can further improve on this naive approach: as shown in the extended version, separating the inner error into deviation and bias terms already reduces the total runtime to $\propto \varepsilon^{-3}$.

However, we identify an alternative direction to reduce the total runtime: reusing samples across multiple radii. This approach relies on two key observations. First, by sampling points from balls in decreasing order of radius, each point also serves as a uniform sample for all smaller balls it falls into. Second, once the radii are sorted, the depth of a point across all radii can be computed efficiently by performing a dichotomic search for the smallest radius such that the point lies within the corresponding ball, which then determines membership for all larger-radius balls. A concrete algorithm implementing these ideas is described in the full version of the paper.

In the worst case, such an algorithm may still need to draw $k$ fresh samples for every radius and every center, for a total of $kM|S|$ samples. Each sample then requires $O(n|S|)$ operations to compute distances to all centers in $S$, as well as an extra $O(|S| \log M)$ operations for the binary search across radii. Altogether, this brings the total worst-case computational cost to $O(kM|S|^2(n + \log M))$.

In expectation, however, a substantial reduction in computation may be achieved through sample reuse. Suppose the radii are ordered in increasing order, i.e., $r_1 \le \cdots \le r_M$. For any $j \le i \in [M]$, the probability that a sample $z$ drawn uniformly from $B_{r_i}(x)$ also lies within $B_{r_j}(x)$ is exactly $(r_j/r_i)^n$. Let $T_j$ denote the (random) number of new samples drawn at radius $r_j$, after having already drawn samples for all $i > j$. Conditioning the expectation on the realization $r_1, \dots, r_M$, we then have the recursive relation

$$\mathbb{E}[T_j] = k - \sum_{i>j} \mathbb{E}[T_i]\,(r_j/r_i)^n.$$

In general, the closed-form solution will depend on the joint distribution of the ordered radii. If we consider the case where $\nu$ is uniform on $[0, \alpha]$, however, then the expected value of the $j$-th order statistic becomes $\mathbb{E}[r_j] = \frac{j}{M+1}\alpha$. Approximating the realized order $r_1 \leq \cdots \leq r_M$ by their respective quantiles $r_j \approx \frac{j}{M+1}\alpha$ and substituting them in the above relation, we then show with a strong induction that $\mathbb{E}[T_j] \approx k\big(1-(\frac{j}{j+1})^n\big)$ for all $j < M$, with $\mathbb{E}[T_M] = k$. The total expected number of samples per element of $S$ is therefore

$$\sum_{j=1}^{M} \mathbb{E}[T_j] \approx k\left(M - \sum_{j=1}^{M-1} \left(\tfrac{j}{j+1}\right)^n\right),$$

which is equivalent to $k(n \log M + O(1))$ in the limit $M \longrightarrow \infty$. Only the computation of the $M|S|$ averages incurs a cost of $O(kM|S|)$, so the total expected runtime of this sample-reuse algorithm would scale as $O\big(k|S|(M + |S|n^2 \log M)\big)$. For large $|S|$, this constitutes a substantial improvement over the $O(k|S|^2Mn)$ runtime of Algorithm 3.

While the above discussion primarily provides intuition for the potential benefits of sample reuse, a rigorous analysis is required and left for future work. All in all, the algorithms introduced in this section demonstrate that efficient computation of the proposed weighting functions is already feasible in practice. Besides, substantial efficiency gains are still achievable by combining sample reuse with ideas from Algorithm 2, and incorporating more advanced variance-reduction techniques.

## 6 DISCUSSION

We gather in this section different remarks on our results as well as possible extensions of our work.

*Extension to Perfect Clones.* The framework we considered until now only allows for $\delta$-clones with $\delta > 0$, but not perfect clones, i.e., with $\delta = 0$. The appropriate analytical tool to handle this is to consider a *pseudo-metric space* $(E, d)$ instead of a metric one, where the pseudo-metric $d$ verifies *non-negativity*, *symmetry*, *triangle inequality* like a full-fledged metric, but only verifies *identity* instead of *separability*. This exactly means that two different elements $x \neq y$ in $E$ may be perfect clones, i.e., $d(x, y) = 0$.

The axioms used in the definition of $\mathcal{R}_\alpha(E, d)$ directly extend to a pseudo-metric space $(E, d)$; this is also the case for the representation functions $f_v$ in Theorem 2, when the space induced by the vanishing of the pseudo-metric is $\mathbb{R}^n$. We refer the interested reader to the full version of the paper [4] for more details.

*Extension beyond Euclidean Spaces.* While the solution proposed in Section 4 is restricted to Euclidean spaces, similar ideas could be applied in more general metric spaces. Using a Radon measure $\mu$, one could define the weighting functions $g_r$ in full generality and show similarly as in Theorem 1 that Axioms 1, 3, 4 and 6 hold. The real challenge however is to satisfy Axiom 2.

Indeed, our proof relies on two convenient properties of Euclidean spaces: first, the uplifting of self-isometry $\sigma_S$ to the entire space; second the invariance of the Lebesgue measure under translations, rotations and reflections. What can be done without these properties? The first problem could be entirely shunned by arguing that only full-fledged isometries should be considered in the definition of Axiom 2. The second issue is however tougher to ward off. To extend invariance by translation beyond vector spaces, one should consider *uniformly distributed measures*, i.e., measures that give the same weight to all balls of the same radius. However, such measures turn out to be very rigid objects and are uniquely defined up to a multiplicative constant in most metric spaces.

Lemma 1 (From [8]). *Let $(E, d)$ be a locally compact metric space. There exists a Radon measure $\mu$ defined on the Borel $\sigma$-algebra of $E$ that is uniformly distributed, i.e., it verifies $0 < \mu(B_r(x)) = \mu(B_r(y)) < \infty$ for all $r > 0$ and $x, y$ in $E$. Moreover, this measure is unique up to a multiplicative constant if $E$ is separable.*

As a particular example, this essentially implies that the Lebesgue measure is the only Borel measure invariant by translation on $\mathbb{R}^n$. This indicates that our approach is doomed even in the simple case of $\mathbb{R}^n$ endowed with the $L^1$ distance $d_1(x, y) = \sum_{i=1}^n |x_i - y_i|$, as illustrated in Figure 5.

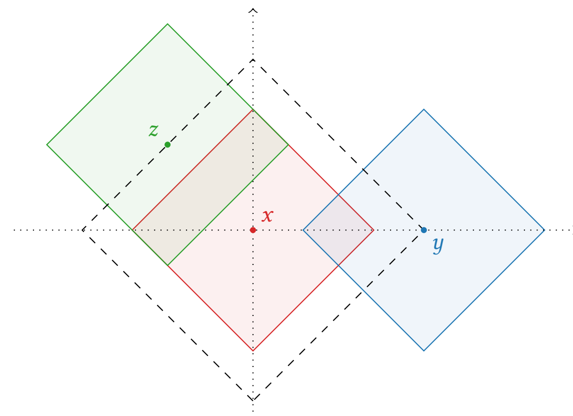



**Figure 5: The weighting function $g_r$ does not satisfy Axiom 2 in $(\mathbb{R}^2, d_1)$. As illustrated by the dashed $L^1$ ball centered in $x$, points $y$ and $z$ are indeed at the same distance of $x$, thus belong in a common isometry class in $S = \{x, y, z\}$ and should receive similar weights under Axiom 2. Note however that the Lebesgue measure, i.e., the area, of the intersection between the red and the green ball differs from that of the intersection between the red and the blue ball, hence $g_r(S)(y) \neq g_r(S)(z)$.**

Such metric spaces thus require developing techniques different from the ones introduced in this work. *Topologically independent* weighting functions would provide an elegant solution to this issue, i.e., functions that do not rely on the topological properties of $(E, d)$, but rather solely depend on the distance matrices associated with each finite set.

Axiom 8 (Topological Invariance). *Weighting only depends on the distance matrix associated with each finite set, i.e., there exists a family $(h_n)_{n\geq 1}$ with $h_n : \mathbb{R}^{n\times n} \mapsto \Delta(n)$ such that, for all $S \in \mathcal{P}(E)$ of cardinality $|S| = n$, we have $f(S) = h_n(M(S))$, where $M = (d(x, y))_{x,y\in S} \in \mathbb{R}^{n\times n}$ denotes the distance matrix associated to $S$ and $d$, unique up to permutations.*

Identifying weighting functions within $\mathcal{R}_\alpha(E, d)$ that adhere to Axiom 8 is a promising direction for future work.

## CREDIT AUTHOR STATEMENT

**Damien Berriaud**: Conceptualization, Formal Analysis, Investigation, Visualization, Writing – Original Draft. **Roger Wattenhofer**: Writing – Review & Editing, Supervision, Funding Acquisition.

# ACKNOWLEDGMENTS

A preliminary version of this paper was presented at GAIW'25; we thank the workshop attendees for their interesting questions and remarks. We also thank the anonymous reviewers for their constructive feedback and suggestions, which helped refine this final version.

# REFERENCES


[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, and al. 2019. Task2Vec: Task Embedding for Meta-Learning. https://doi.org/10.48550/arXiv.1902.03545

[2] David Alvarez-Melis and Nicolò Fusi. 2020. Geometric Dataset Distances via Optimal Transport. https://doi.org/10.48550/arXiv.2002.02923

[3] David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. 2018. Re-evaluating Evaluation. https://doi.org/10.48550/arXiv.1806.02643

[4] Damien Berriaud and Roger Wattenhofer. 2025. Clone-Robust Weights in Metric Spaces: Handling Redundancy Bias for Benchmark Aggregation. https://doi.org/10.48550/arXiv.2502.03576

[5] Florian Brandl, Felix Brandt, and Hans Georg Seedig. 2016. Consistent Probabilistic Social Choice. https://doi.org/10.48550/arXiv.1503.00694

[6] Karl Bringmann and Tobias Friedrich. 2010. Approximating the volume of unions and intersections of high-dimensional geometric objects. *Computational Geometry* 43, 6 (Aug. 2010). https://doi.org/10.1016/j.comgeo.2010.03.004

[7] Frederic Cazals, Harshad Kanhere, and Sébastien Loriot. 2011. Computing the volume of a union of balls: A certified algorithm. *ACM Trans. Math. Softw.* 38, 1 (Dec. 2011). https://doi.org/10.1145/2049662.2049665

[8] Jens Peter Reus Christensen. 1970. On Some Measures Analogous to Haar Measure. *Math. Scand.* 26 (June 1970). https://doi.org/10.7146/math.scand.a-10969

[9] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. 2022. What are the best systems? New perspectives on NLP Benchmarking. https://doi.org/10.48550/arXiv.2202.03799

[10] Vincent Conitzer and Makoto Yokoo. 2010. Using Mechanism Design to Prevent False-Name Manipulations. *AI Magazine* 31, 4 (Dec. 2010). https://doi.org/10.1609/aimag.v31i4.2315

[11] Joseph Gallian. 2020. *Contemporary Abstract Algebra* (10 ed.). CRC Press. https://doi.org/10.1201/9781003142331

[12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.* 13 (March 2012).

[13] Harald Andrés Helfgott, Jitendra Bajpai, and Daniele Dona. 2017. Graph isomorphisms in quasi-polynomial time. https://doi.org/10.48550/arXiv.1710.04574

[14] Anas Himmi, Ekhine Irurozki, Nathan Noiry, Stephan Clemencon, and Pierre Colombo. 2023. Towards More Robust NLP System Evaluation: Handling Missing Scores in Benchmarks. https://arxiv.org/abs/2305.10284v1

[15] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, and al. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. https://doi.org/10.48550/arXiv.2012.07421

[16] Gilbert Laffond, Jean Lainé, and Jean-François Laslier. 1996. Composition-consistent tournament solutions and social choice functions. *Social Choice and Welfare* 13, 1 (Jan. 1996), 75–93. https://doi.org/10.1007/BF00179100

[17] Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. 2022. Wasserstein Task Embedding for Measuring Task Similarities. https://doi.org/10.48550/arXiv.2208.11726

[18] Scott A. Mitchell, Muhammad A. Awad, Mohamed S. Ebeida, and Laura P. Swiler. 2018. *Fast Approximate Union Volume in High Dimensions with Line Samples.* Technical Report SAND–2018-8684, 1464880. Sandia National Laboratories. http://www.osti.gov/servlets/purl/1464880/

[19] Ilan Nehama, Taiki Todo, and Makoto Yokoo. 2022. Manipulation-resistant false-name-proof facility location mechanisms for complex graphs. *Autonomous Agents and Multi-Agent Systems* 36, 1 (Jan. 2022). https://doi.org/10.1007/s10458-021-09535-5

[20] Xingchao Peng, Yichen Li, and Kate Saenko. 2020. Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation. https://doi.org/10.48550/arXiv.2007.09257

[21] Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than Average: Paired Evaluation of NLP Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/2021.acl-long.179

[22] Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. 2025. Clone-Robust AI Alignment. https://doi.org/10.48550/arXiv.2501.09254

[23] Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, and al. 2023. Vote'n'Rank: Revision of Benchmarking with Social Choice Theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/2023.eacl-main.48

[24] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (Eds.). Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170/

[25] Shavrina Tatiana and Malykh Valentin. 2021. How not to Lie with a Benchmark: Rearranging NLP Leaderboards. https://doi.org/10.48550/arXiv.2112.01342

[26] T. N. Tideman. 1987. Independence of clones as a criterion for voting rules. *Social Choice and Welfare* 4, 3 (Sept. 1987), 185–206. https://doi.org/10.1007/BF00433944

[27] Taiki Todo, Atsushi Iwasaki, Makoto Yokoo, and Yuko Sakurai. 2009. Characterizing False-name-proof Allocation Rules in Combinatorial Auctions. (2009).

[28] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, and al. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. https://doi.org/10.48550/arXiv.1905.00537

[29] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, and al. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. https://doi.org/10.48550/arXiv.1804.07461

[30] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641. https://doi.org/10.1162/tacl_a_00290

[31] Guanhua Zhang and Moritz Hardt. 2024. Inherent Trade-Offs between Diversity and Stability in Multi-Task Benchmarks. https://arxiv.org/abs/2405.01719v2