A geometric framework for fairness

Alessandro Maggio[1,*], Luca Giuliani[1], Roberta Calegari[1], Michele Lombardi[1] and Michela Milano[1]

[1]*Alma Mater Studiorum–Università di Bologna, Italy*

Abstract

Fairness has emerged as a critical concern in the field of machine learning impacting its application in various domains. While there have been successful attempts to tackle fairness, many existing analyses rely on sophisticated mathematical methods that may lack intuitive understanding. Drawing inspiration from successful applications in other areas of machine learning, in this study, we propose a GEOmetric Framework for Fairness – GEOFFair – that represents distributions, ML models, fairness constraints, and hypothesis spaces as vectors and sets. The geometric framework aims to provide a more intuitive and rigorous understanding of fairness in Artificial Intelligence (AI). It enables visualizing mitigation techniques as movements in the vector space and aids in constructing proofs-by-witness by quickly identifying examples or counter-examples. Furthermore, the geometric framework offers a platform for studying various fairness properties, including geometrical distances between fairness vectors, relative fairness comparisons, and the exploration of symmetries, invariances, and trade-offs between fairness metrics.

Keywords

AI fairness, geometric framework, GEOFFair

1. Introduction

Fairness issues in Machine Learning (ML) have been raised to the spotlight in recent years, and are regarded as a major roadblock for the application of data-driven AI in fields such as healthcare, economics, welfare, and policy-making [1].

From a mathematical point of view, studying fairness properties (or the lack thereof) in ML models is a difficult endeavour, since it requires reasoning on statistical distributions and (potentially) non-linear models [2]. While successful attempts in this direction exist [3], many of the existing analyses rely on sophisticated methods that may not be easily intuitive for a comprehensive understanding of all the objects involved in addressing fairness.

We propose that the field could gain advantages from a simplified framework that, while not as nuanced or powerful as other statistical methods, could provide an intuitive and rigorous grasp of some key concepts and mechanisms related to fairness in AI. Specifically, we propose adopting a *GEOmetric Framework for Fairness* – GEOFFair – to represent distributions, functions (e.g. ML models), fairness constraints, and hypothesis spaces as vectors and sets. The key

benefit of geometric frameworks is that they enable visualization, allowing us to gain insights into the data or the model operation.

In terms of motivation, our approach draws from successful attempts in other areas of ML [4, 5], where mapping models to points in vector spaces (e.g. by concatenating their parameters) have led to simplifications in their representation and analysis. Through this lens, distance metrics, projections, similarities, and algorithms can be applied to gain insights into the models. Papers like [6] demonstrate the effectiveness of vector representations in natural language processing tasks, where embracing the vector space model allows for a deeper understanding and comparison of machine learning models. Similarly, we believe our framework can enable a better understanding and visualization of fairness issues and facilitate the study of their properties. For example, using this approach it is possible to map mitigation techniques for addressing fairness concerns as movements in the vector space or assist the construction of proofs-by-witness by quickly finding examples or counter-examples.

Furthermore, the geometric framework provides a platform for studying various properties related to fairness. We can investigate the geometrical distances between fairness vectors, which may provide insights into the relative fairness of different models or interventions. Additionally, the geometric space allows for exploring symmetries, invariances, or trade-offs between different fairness metrics, contributing to a deeper understanding of the fairness landscape.

Accordingly, the paper is organized as follows. Section 2 introduces the formal framework of GEOFFair, discussing its mathematical foundation and exploring the potential relationships among its elements. It establishes the groundwork for understanding the subsequent sections. Then in Section 3 the paper interprets fairness mitigation techniques within the context of the proposed GEOFFair framework. It examines how these techniques can be applied and understood through the lens of GEOFFair, providing insights and analysis. Conclusion and future works are discussed in Section 4.

2. GEOFFair: a GEOmetric Framework for Fairness

The following section aims to introduce a formal framework. To achieve this goal, we will focus on two key points. Firstly, we will explore the vector representation of the main mathematical objects of the framework defining the core elements and the main existing properties (Subsection 2.1, Subsection 2.2). This vector representation allows us to formalize fairness concepts and measures in a clear and precise mathematical language. Secondly, we will discuss how these vector representations exist within the same space, providing a common basis for comparing and contrasting different fairness measures (Subsection 2.3).

2.1. From Distribution to Vectors in the Space

Classical formulations for both ML models and fairness metrics typically rely on probability theory and statistics: for example, the ground truth is viewed as a probability distribution, the ML model as a parameterized function, the training loss as a likelihood measure, and fairness metrics as functions over conditional expectations. The first challenge in the definition of our framework is therefore mapping such concepts into a vector representation, with no significant loss of generality.

Probability Distributions and Functions   We focus on a supervised learning setting, and we start by defining a representation for (joint) probability distributions, which we approximate to arbitrary precision via an *infinite sample.* Formally:

Notation 1 (Probability Distributions). *Let X, Y be random variables with support in $\mathcal{X}$ and $\mathcal{Y}$ and joint distribution $P(X, Y)$. Then we encode the distribution as a vector $(x, y) = \{x_i, y_i\}_{i=1}^{n}$, with $x_i, y_i \sim P(X, Y)$ and $n \to \infty$.*

Intuitively, X represents an observable that may serve as the input for an ML model, while Y represents the quantity (or class) to be estimated. We make no assumption about the support of the random variables, i.e. the range of their possible values. The same representation can be applied for the individual distributions of X and Y, which are therefore denoted as x and y. Our approach makes it particularly easy to represent functions over random variables (e.g. Machine Learning models evaluated over their input). Formally:

Notation 2 (Functions). *A deterministic function f over X and Y can then be naturally viewed as a vector $f(x, y) = \{f(x_i, y_i)\}_{i=1}^{n}$ with $n \to \infty$, i.e. just the vector with the function evaluation over all the samples.*

Functions that depend only on X or only on Y are sub-cases of the above definitions and are respectively denoted as $f(x)$ and $f(y)$.

There are a few observations worth making. First, while we use the term "vector" for simplicity, our definitions are closer to functions that map an index i to an object such as x_i or y_i. In other words, x, y, $f(x)$, etc. can be thought of as points in a Hilbert space. Second, our representations are not exact, but they will be sufficient to approximate key statistical properties with arbitrarily high probability. Exact representations for distributions exist and are well known, e.g. the Probability Mass Function or Probability Density Function; however, they do not enable constructing a simple 1-1 mapping between components in the vector (e.g. x_i) and function evaluations (e.g. $f(x_i)$), which is instead trivial with our approach.

Equivalence of Expectation Predicates   Many of the existing fairness metrics are expressed in terms of (conditional) expectations, i.e. averages, or can be reduced in such a form. For example, assuming X is a binary protected attribute, the DIDI metric from [7] is defined in terms of the discrepancy between the global average outcome and the average outcome for each protected group, i.e. $|\mathbb{E}[Y \mid X = 0] - \mathbb{E}[Y]| + |\mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y]|$. Statistical parity in classification, which advocates for similar probabilities of a positive outcome across all groups, can be defined as $|\mathbb{E}[Y \mid X = 0] - \mathbb{E}[Y \mid X = 1]|$, and so on. Intuitively, this means that many fairness constraints can be viewed as predicates over (conditional) expectations.

The sample expectation function, represented by $\mu(\cdot)$, tends to converge towards the true expectation $\mathbb{E}[\cdot]$ as the sample size grows: we use this result to establish a form of equivalence between predicates expressed over a distribution and those expressed over a sample.

Theorem 1. *Let $\Pi(X, Y)$ be a predicate over (conditional) expectations for X and Y and let $\pi\left(\{x_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n}\right)$ be its sample counterpart. Then we have that:*

$$P\left(\Pi(X, Y) \Leftrightarrow \lim_{n \to \infty} \pi\left(\{x_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n}\right)\right) = 1 \tag{1}$$

i.e. the two predicates are equivalent almost surely as the sample size grows if the involved expectations are finite.

Proof. The two predicates are identical except for the use of the true and sample expectations. For the sake of simplicity and without loss of generality, let us assume the involved expectations are respectively $\mathbb{E}[Y]$ and $\mu\left(\{y_i\}_{i=1}^n\right)$. Since the samples are drawn independently from the same distribution, due to the strong law of large numbers we have that:

$$P\left(\lim_{n\to\infty} \mu\left(\{y_i\}_{i=1}^n\right) = \mathbb{E}[Y]\right) = 1 \tag{2}$$

Equivalence of the sample and true expectations then implies equivalence of Π and π. $\square$

Notation 1 and Notation 2 give us the ability to transition from the conventional distribution paradigm of ML to the realm of vector spaces. Theorem 1 enables reasoning over the vector representation and translates almost certainly any result to the original distribution, at least as far as fairness metrics are concerned. Together, these tools allow us to leverage the power and interpretability of vector space representations in the context of fairness metrics, expanding the scope of analysis and decision-making.

2.2. The Formal Model

As mentioned in Section 2.1, we focus on a supervised learning setting where the goal is to learn a model that maps inputs (always observable) to outputs (observable at training time and to be estimated at inference time). In this context, we introduce four key mathematical objects that play a major role in the analysis of fairness issues in AI.

We represent the *input distribution* by means of an *input vector* $x \in \mathcal{X}^n$, with $n \to \infty$, according to Notation 1. Concerning the output, we make a distinction between the distribution that can actually be observed and the one that we ideally wish to estimate. We start by introducing the following concept:

Definition 1 (Ground Vector). *The ground vector $y^+ \in \mathcal{Y}^n$ represents data that can be observed and used as ground truth to learn machine learning models. It is paired with the input vector x.*

As inspired by [3], we model the fact that the ground truth might be subject to systemic social biases, but with a key difference. That is, we directly define an "unbiased" output vector rather than an unbiased input matrix, as our framework allows us to reason in terms of vector components within the output space.

Definition 2 (Gold Vector). *The gold vector $y^* \in \mathcal{Y}^n$ represents the "unbiased" data that corresponds to the output distribution before it is corrupted by social biases; accordingly, we can derive the ground vector y^+ by considering the application of a biased mapping over the gold vector, i.e.:*

$$y^+ = b(y^*), \quad \text{where } b : \mathcal{Y}^n \to \mathcal{Y}^n \text{ is called the biased mapping}$$

Note that in practical applications, the gold vector is typically unobservable and therefore not accessible at training time. Still, explicitly modeling the unbiased distributions allows us to study in deeper detail the interplay between bias and fairness constraints.

In our framework, an ML model can be viewed as a function that maps input to output data. In supervised learning, the training process is typically viewed as that of selecting one model out of a pool of candidates, so as to minimize a loss metric. Formally, training amounts to solving in an exact or approximate fashion:

$$\underset{f \in \mathscr{F}}{\arg\min} \; \mathscr{L}(f(x), y^+) \tag{3}$$

where f is the ML model, $\mathscr{L}$ is the chosen loss metric and $\mathscr{F}$ represents the set of possible models, usually defined by specifying an architecture (e.g. a number and size of layers in a feed-forward neural network, number of estimators and maximum depth in a random forest).

In our framework, however, the input vector x is by construction fixed, thus making the model output the only relevant factor. In other words, two models are equivalent as long as they have the same output. This observation allows us to introduce a simplified representation of the classical notion of hypothesis space.

Definition 3 (Hypothesis Space). *The hypothesis space $\hat{\mathbb{Y}}$ is the set of possible outputs for the chosen class of ML models, i.e.*

$$\hat{\mathbb{Y}} = \{y \in \mathscr{Y}^n \mid \exists f \in \mathscr{F} \; : \; f(x) = y\}$$

Intuitively, the hypothesis space can be viewed as the set of possible model outputs for the considered sample. A linear regression model will have a limited hypothesis space due to its ability to represent linear relationships only, while more complex models such as random forests and neural networks will have a much larger hypothesis space.

Finally, as we are considering a fairness scenario, we need to model a final mathematical object in order to guarantee a proper analysis of the phenomenon, namely the region in the output space that is considered fair.

Definition 4 (Fair Space). *Let $\overline{\mathbb{Y}} \subseteq \mathscr{Y}^n$ be the set containing all the output vectors that are aligned with the fairness requirements.*

We make no assumption on the mathematical definition of the fair space. Nonetheless, it is worth noting that in many practical cases, this set is defined by means of a threshold t on a fairness metric K, i.e. $\overline{\mathbb{Y}} = \{y \in \mathscr{Y}^n \mid K(y) \leq t\}$.

Once all the elements are defined, we can examine how they interact with each other. In the most general setup, we can not make any assumption about the relationships between y^*, y^+, $\hat{\mathbb{Y}}$, and $\overline{\mathbb{Y}}$. Without specific contextual information on data, models, and constraints, the relationships between these entities can vary significantly.

It is worth mentioning that, in the defined framework, all vectors and sets we introduced exist in the same space, which facilitates easy visualization (see Figures in Section 3). This visual representation can assist with proof-by-witness, allowing us to analyze and demonstrate relationships between these vectors more effectively.

2.3. Relationships Between Elements

Let us now explore the four elements we discussed earlier. A summary of their potential relationships is presented in Table 1. Firstly, thanks to the problem's symmetry, we can acquire

Table 1
Possible one-to-one relationships. When comparing the two sets, we use $\not\cap$ and $\cap$ as aliases for $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} = \varnothing$ and $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} \neq \varnothing, \hat{\mathbb{Y}}, \overline{\mathbb{Y}}$.

	$\hat{\mathbb{Y}}$	$\overline{\mathbb{Y}}$	y^+	y^*
$\hat{\mathbb{Y}}$		$\not\cap, \subseteq, \supseteq, \cap$	$\ni, \not\ni$	$\ni, \not\ni$
$\overline{\mathbb{Y}}$			$\ni, \not\ni$	$\ni, \not\ni$
y^+				$\equiv, \neq$
y^*				

all potential relationships by examining the Cartesian product of the six entries found in the upper triangular section of the table. This yields a total of 128 configurations, calculated as 4×2^5. Furthermore, we can decrease the configuration space by making a few straightforward observations.

- $y^* \equiv y^+$ can hold uniquely when the two vectors belong to the same sets;

- If $y^* \in \hat{\mathbb{Y}}$ and $y^* \in \overline{\mathbb{Y}}$ hold simultaneously then $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} \neq \varnothing$, and the same applies to y^+;

- $y^* \in \hat{\mathbb{Y}}$ and $y^* \notin \overline{\mathbb{Y}}$ are incompatible if $\hat{\mathbb{Y}} \subset \overline{\mathbb{Y}}$, and the same applies to y^+;

- $y^* \notin \hat{\mathbb{Y}}$ and $y^* \in \overline{\mathbb{Y}}$ are incompatible if $\hat{\mathbb{Y}} \supset \overline{\mathbb{Y}}$, and the same applies to y^+.

By taking these logical constraints into account, we identified 56 distinct legal combinations, whose listing is provided in Appendix A. While the number of possible combinations is not small, it is nevertheless finite, which can help with proofs of universally quantified statements (i.e. $\forall$ and $\not\exists$).

Now, let us examine each one-to-one relationship between these elements. With respect to the Hypothesis and Fair Space, the possible outcomes are as follows:

$\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} = \varnothing$ This scenario implies that it is not possible to learn a model that satisfies the fairness criteria. Although this is a rare occurrence, as most fairness metrics evaluate to zero on constant vectors (which can generally be represented by any machine learning model), it might still happen in certain situations. For example, this could be due to an excessively strict threshold imposed on the fairness constraint.

$\hat{\mathbb{Y}} \subseteq \overline{\mathbb{Y}}$ In this case, the machine learning model is said to be *fair-by-design* [8]. While achieving this is challenging in many practical cases, it can be attained by incorporating explicit rules into the model, ensuring that certain deontological fair principles are always upheld.

$\hat{\mathbb{Y}} \supseteq \overline{\mathbb{Y}}$ Here, the machine learning models can cover all existing fair outputs. This can be the case when employing powerful models like large neural networks.

$\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} \neq \varnothing$ This is the most common scenario encountered in practice. In this case, the goal of learning a fair model is to find an appropriate parameter configuration such that the output vector y belongs to the non-trivial intersection between the two sets, $\hat{\mathbb{Y}}$ and $\overline{\mathbb{Y}}$.

Similarly, when considering the relationships among vectors and sets, the following considerations can be made:

1. if y^+ and y^* coincide, it implies that the mapping $b: \mathcal{Y}^n \to \mathcal{Y}^n$ introduces no bias. However, this is an extremely rare scenario that leads to trivial solutions for any fairness task. In most real-world cases, the two vectors are not aligned, indicating a discrepancy between the information conveyed by y^+ and y^*. This misalignment suggests that the ground vector has been pushed away from the unbiased distribution to some extent;

2. if $y^+ \in \hat{\mathbb{Y}}$, it can be perfectly represented by the machine learning model, although this representation is not guaranteed to be fair unless y^+ is already in the Fair Space (as mentioned in Point 4 below). Conversely, when the model lacks the capacity to represent y^+ adequately, it will be trained to minimize the loss $\mathcal{L}$ between the labels and the model outputs;

3. the same considerations as in Point 2 apply to the relationship between y^* and $\hat{\mathbb{Y}}$. The only difference is that, in this case, the analysis is purely theoretical since no model can be trained on y^*, which is not observable in real-world scenarios;

4. if $y^+ \in \overline{\mathbb{Y}}$, it means that the ground vector aligns with the fairness criteria. This alignment can be due to various reasons, such as a weakly biased mapping that does not significantly deviate the ground vector from the unbiased distribution, or the fairness criteria being too permissive and allowing for a higher degree of fairness violation. Conversely, when y^+ is outside the Fair Space, learning a fair model becomes more challenging as it must explicitly account for the fairness constraints. This is the most common scenario in real-world settings;

5. Similar to Point 4, the gold vector can belong to the Fair Space or not. In most practical use cases, it does belong, indicating that the chosen fairness metric K aligns with the unbiased distribution and its threshold is well-tuned. However, if a misaligned metric is chosen or a too restrictive threshold is set, it is possible for $y^* \notin \overline{\mathbb{Y}}$.

3. Fairness Mitigation Through the Lens of GEOFFair

In this section, we will utilize the GEOFFair framework to analyze fairness mitigation techniques. In a previous work by Dutta et al. [3], it was demonstrated that maximizing accuracy solely based on the observed labels vector may not always be the optimal choice. They employed statistical distributions and mathematical tools from probability theory to establish this result. Rather than extending their findings, our objective is to employ our proposed geometric framework to support and validate them. By leveraging the GEOFFair framework, we aim to present similar conclusions in a more accessible and interpretable way and can bridge the gap between complex mathematical concepts and practical implications. This allows for a clearer comprehension of the challenges associated with fairness and the potential solutions that can be pursued.

3.1. Mitigation as Projection

Mitigation, in the AI fairness context, refers to the process of reducing unfairness by either transforming the biased distribution or by ensuring that the ML model behaviour is compatible with the fairness constraints. From a geometric point of view, such techniques can be viewed as projecting either the ground vector or the ML output onto the Fair Space. Analogously, training an ML model can be viewed as the problem of finding a vector in the Hypothesis Space that is closest to the ground vector in terms of the loss function, i.e. as projecting the ground vector onto the Hypothesis Space. Therefore, in the context of GEOFFair, projections provide a convenient lens through which we can study mitigation at pre-processing, training, and post-processing time in a uniform fashion.

We will focus our analysis on the more widespread case where learning a fair ML model is possible (i.e. $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} \neq \varnothing$). We start by introducing two additional vectors, i.e. the projections of the ground truths and the gold standard vector, respectively. These projections will be onto the intersection space between the Hypothesis and the Fair Space.

Definition 5 (Ground and Gold Fair Projections). *The optimal fair predictions p and z obtained from the ground (y^+) and gold (y^*) vectors, i.e.:*

$$p = \arg\min_{v}\{\mathcal{L}(v, y^+) \mid v \in \hat{\mathbb{Y}} \cap \overline{\mathbb{Y}}\} \tag{4}$$

$$z = \arg\min_{v}\{\mathcal{L}(v, y^*) \mid v \in \hat{\mathbb{Y}} \cap \overline{\mathbb{Y}}\} \tag{5}$$

Intuitively, p represents the outcome of training an ML model under fairness constraints, or equivalently of training an ML model over a ground distribution transformed so as to enforce the fairness restrictions. The z vector represents the best fair model that we could learn for the (typically unobservable) "unbiased" distribution.

It is worth noting that p and z might not be unique, as equally accurate outputs that are both fair and representable by the model can exist. Furthermore, for the purpose of our theoretical analysis, we will assume that p and z are obtained from exact and globally optimal algorithms. However, it is important to acknowledge that many machine learning models, especially larger ones, do not guarantee this optimality property in practice. Additionally, to avoid trivial cases, we assume that the *biased mapping* function $b: \mathcal{Y}^n \to \mathcal{Y}^n$ applies a modification to the input vector, i.e. that $y^* \not\equiv y^+$. This assumption narrows down our analysis to even fewer cases than those defined in Subsection 2.3, and let us draw the following conclusion:

$$\mathcal{L}(y^+, y^*) > 0 \tag{6}$$

where $\mathcal{L}$ is any non-negative loss function such that $\mathcal{L}(y^+, y^*) = 0$ iff $y^+ \equiv y^*$.

Basic Properties of Fair Projections Let us consider the optimization problems defined in Equations (4)-(5) and examine the behaviour of p and z in terms of fairness based on the position of y^+ and y^*, respectively. We will rely on the formulation of the Fair Space based on a fairness metric $K(\cdot)$ that we introduced in Section 2, i.e.:

$$\overline{\mathbb{Y}} = \{y \in \mathcal{Y}^n \mid K(y) \leq t\} \tag{7}$$

Before establishing a fundamental property of fair projections, let us introduce some notation to describe the concept of *Fair Space Frontier*. It can be described as:

Notation 3 (Fair Frontier).

$$\partial \overline{\mathbb{Y}} = \{y \in \mathcal{Y}^n \mid K(y) = t\} \tag{8}$$

The Fair Space Frontier represents the boundary of the Fair Space, namely the region of the space containing all the vectors exhibiting a threshold-level fairness. Likewise, we can introduce the concept of *Internal Fair Set*, which encompasses the vectors within the Fair Space but not on the Fair Frontier:

Notation 4 (Internal Fair Set).

$$\Delta \overline{\mathbb{Y}} = \overline{\mathbb{Y}} \setminus \partial \overline{\mathbb{Y}} = \{y \in \mathcal{Y}^n \mid K(y) < t\} \tag{9}$$

Property 1 (Fair Projections). *Given a vector y and its projection y' onto the Fair Space as defined in Equation (7), we know that:*

$$y \in \overline{\mathbb{Y}} \implies y' \equiv y \implies K(y') = K(y) \tag{10}$$

$$y \notin \overline{\mathbb{Y}} \implies y' \in \partial \overline{\mathbb{Y}} \implies K(y') = t \tag{11}$$

meaning that any vector lying within the Fair Space will be projected onto itself (thus exhibiting the same fairness level); conversely, if the vector is outside the Fair Space, its projection will be on the boundary of the Fair Space, resulting in threshold-level fairness.

This is a well-known property in both convex and non-convex optimization, whose proof can be found in [9]. Now, if we take into account the capabilities of the ML model, we can extend Property 1 as follows:

Property 2 (Representable Fair Projections). *Given a vector y and its projection y' onto the intersection between the Fair and Hypothesis Space, we know that:*

$$y \in \overline{\mathbb{Y}} \lor \hat{\mathbb{Y}} \subseteq \overline{\mathbb{Y}} \implies K(y') \leq t \tag{12}$$

$$y \notin \overline{\mathbb{Y}} \land \hat{\mathbb{Y}} \supseteq \overline{\mathbb{Y}} \implies K(y') = t \tag{13}$$

It is important to note that when the Fair Space and the Hypothesis Space have a non-trivial intersection – i.e. neither space is a subset of the other –, we cannot draw conclusions about $K(y')$ since points in the boundary of the intersection can exhibit different fairness levels.

3.2. Possible Cases Configuration

Based on the properties and assumptions discussed in the previous subsection, we can now outline five distinct cases that summarize the different combinations arising from the positions of the input vectors (y^+ and y^*) and their projections (p and z). Each case description is accompanied by illustrative figures, where blue stripes represent the region where the ground projection p can fall, and green stripes indicate the region where the gold projection z can fall; additionally, the figures depict two possible ground and gold vectors, along with their projections. On a final note, we underline that these cases are mutually exclusive, meaning that *the conditions of each subsequent case implicitly exclude the conditions of the previous ones.*
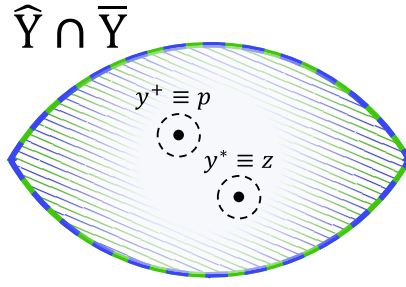
Figure 1: Graphical representation of *Case 1*. Projections p and z cannot coincide, although they can both fall in the same region as highlighted by the green and blue stripes.

Case 1: $y^+, y^* \in \hat{\mathbb{Y}} \cap \overline{\mathbb{Y}}$. This is a trivial scenario in which both vectors are already representable and satisfy the fair condition. As shown in Figure 1, in this case, the projections coincide with the original vectors, therefore we know that:

$$\mathscr{L}(p, z) = \mathscr{L}(y^+, y^*) > 0 \tag{14}$$

where the strict inequality follows from having assumed a non-trivial biased mapping, i.e. from Equation (6). The existence of a gap even in this simple scenario shows that maximizing accuracy based on the biased data (under fairness constraints) may not always yield the best solution. Furthermore, we have that:

$$K(p) \lesseqgtr K(z) \tag{15}$$

In other words, there is also no guarantee that the gold projection is fairer than the ground projection. Such a variety of possible outcomes is due to the fact that this case captures situations where fairness constraints are not particularly restrictive, so applying mitigation techniques is not really meaningful.

Case 2: $y^+, y^* \in \Delta\overline{\mathbb{Y}} \vee \hat{\mathbb{Y}} \subseteq \overline{\mathbb{Y}}$. We can identify two sub-cases within this scenario. In the first sub-case, both the gold vector and the ground vector satisfy the fairness criteria, even if at least one of them does not belong to the Hypothesis Space – this distinction is necessary to avoid falling back into *Case 1*. In the second one, the ML model is *fair-by-design*, meaning that any possible output is guaranteed to be within the Fair Space. Intuitively, the former sub-case reflects another situation where fairness constraints are not particularly restrictive; in the latter, fairness issues have already been addressed by acting on the model architecture. Although these two sub-cases might look very different, the resulting projections exhibit the same behaviour. In fact, similar to *Case 1*, p and z may be arbitrarily close or far, depending on the positions of the two original vectors, and no mutual information on $K(p)$ and $K(z)$ can be obtained. This observation stresses the potential impact of the chosen class of models (the Hypothesis Space) on the outcome of mitigation approaches.

Case 3: $y^+ \notin \Delta\overline{\mathbb{Y}} \wedge y^* \in \Delta\overline{\mathbb{Y}}$. In this case, the ground target y^+ is either outside or at the frontier of $\overline{\mathbb{Y}}$, meaning that there is a non-null gap in terms of the fairness metric between itself

(a) $y^+, y^* \in \Delta\overline{\mathbb{Y}}$.

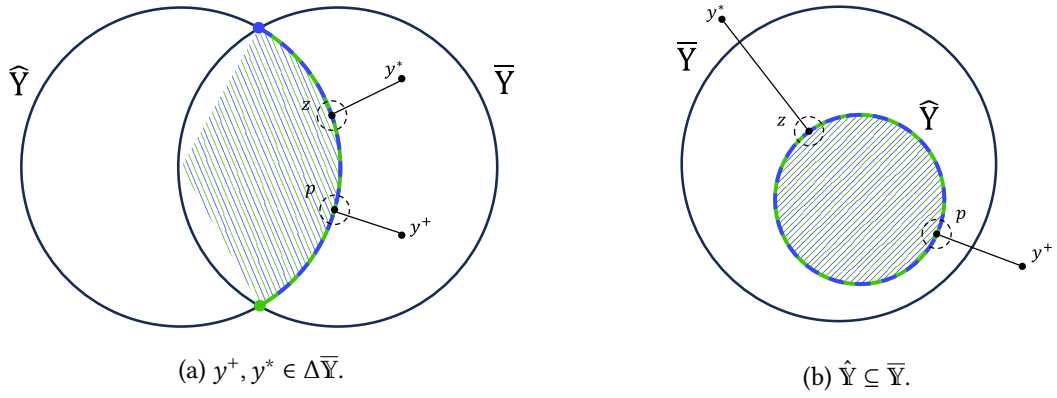(b) $\hat{\mathbb{Y}} \subseteq \overline{\mathbb{Y}}$.

Figure 2: Graphical representation of *Case 2*. The represented vectors show an illustrative scenario where both the vectors are outside the Hypothesis Space, although at most one of them can belong to the intersection of the two sets.
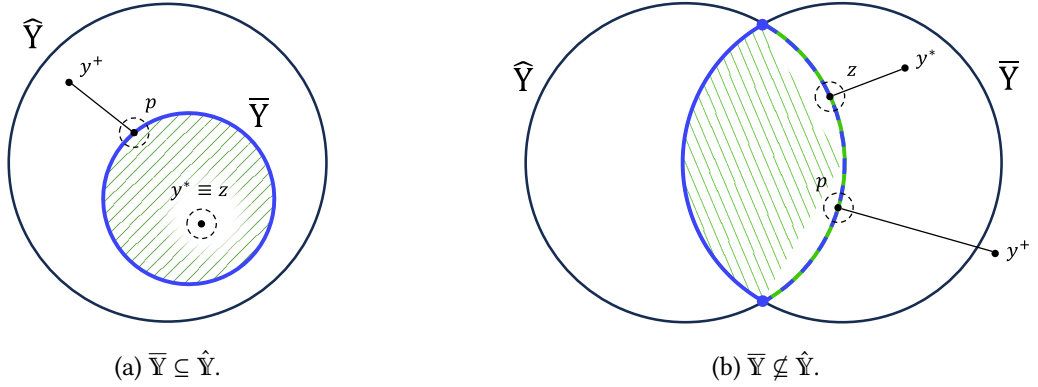

(a) $\overline{\mathbb{Y}} \subseteq \hat{\mathbb{Y}}$.

(b) $\overline{\mathbb{Y}} \nsubseteq \hat{\mathbb{Y}}$.

Figure 3: Graphical representation of *Case 3*. In this case, we are guaranteed that a fairer and more accurate solution would exist if the Fair Space is completely contained in the Hypothesis Space (Figure a). Whenever this does not happen (Figure b), such a property cannot be formally proven.

and the gold vector (which is in the Fair Space).

Again, we can identify two main sub-cases, depending on the position of the Hypotheses Space with respect to the Fair Space. Since we are assuming that the intersection between the two sets is not empty, and the scenario in which $\hat{\mathbb{Y}}$ is a subset of $\overline{\mathbb{Y}}$ has been already covered in *Case 2*, the possible outcomes are: (a) $\overline{\mathbb{Y}} \subseteq \hat{\mathbb{Y}}$, or (b) $\overline{\mathbb{Y}} \nsubseteq \hat{\mathbb{Y}}$. In the former (Figure 3a), mitigation has a beneficial effect in terms of fairness, but a guaranteed gap will remain wrt the best possible fair project, since Property 2 guarantees that $K(p) = t > K(z)$. In this situation, as long as the fairness metric is sufficiently aligned with the unbiased distribution (i.e. $y^* \in \overline{\mathbb{Y}}$), better results can be obtained by simply making the fairness constraints more restrictive.

On the contrary, in the latter sub-case (Figure 3b), nothing can be said on the fairness level of p as Property 2 does not cover the case of non-trivial intersection. This suggests that using sufficiently expressive model classes (e.g. larger neural networks) in mitigation approaches may

lead to more consistent outcomes, at least as long as overfitting is successfully prevented.

As a final consideration, we underline that *Case 3 is the most common real-world scenario*, as it assumes that a non-trivial, strong enough biased mapping is applied to y^*, and that the fairness metric is both aligned and correctly calibrated on it.

Case 4: $y^+ \in \Delta \overline{\mathbb{Y}} \wedge y^* \notin \Delta \overline{\mathbb{Y}}$. Contrarily to the previous case, here the positions of the ground and gold vectors are swapped. This is a very unlikely scenario, which is implied by the adoption of a wrong fairness metric that is aligned with the biased data but not with the unbiased one. As for *Case 3*, we can distinguish among two different sub-cases. When $\overline{\mathbb{Y}} \subseteq \hat{\mathbb{Y}}$ (Figure 4a), we are guaranteed that there could be a better solution with respect to the gold vectors, although in this case this solution would exhibit a *higher degree of unfairness* due to Property 2. On the contrary (Figure 4b), nothing can be proven, but the same considerations of *Case 3* remain.

(a) $\overline{\mathbb{Y}} \subseteq \hat{\mathbb{Y}}$.   (b) $\overline{\mathbb{Y}} \nsubseteq \hat{\mathbb{Y}}$.

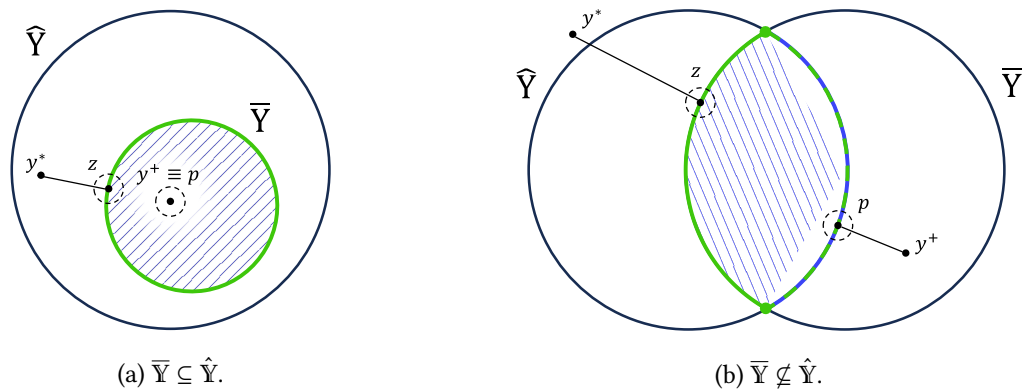**Figure 4:** Graphical representation of *Case 4*. This is the opposite of Figure 3, where a misaligned fairness metric makes the gold vector more unfair than the ground one.

(a) $\overline{\mathbb{Y}} \not\subset \hat{\mathbb{Y}}$.   (b) $\overline{\mathbb{Y}} \subset \hat{\mathbb{Y}}$.
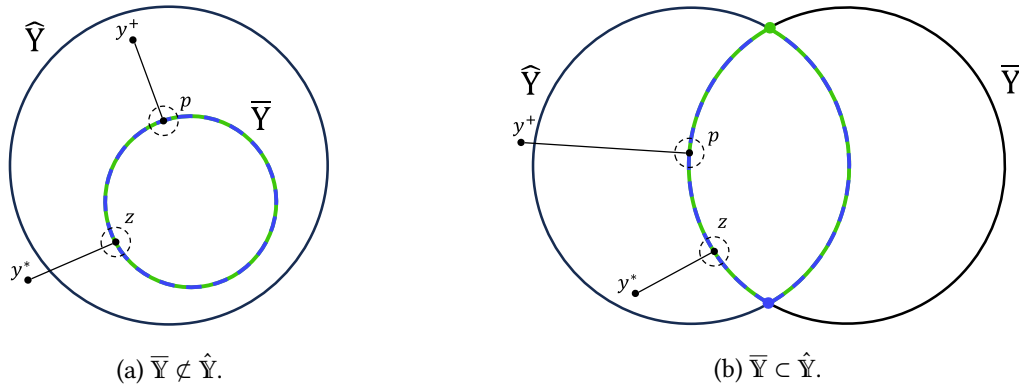
Figure 5: Graphical representation of *Case 5*. In this case, the misaligned fairness metric makes both the gold and ground vectors unfair.

Case 5: $y^+, y^* \notin \Delta \overline{\mathbb{Y}}$. Finally, we consider the case in which the fairness metric is either misaligned or wrongly calibrated for both vectors. The usual sub-cases (Figure 5a: $\overline{\mathbb{Y}} \subseteq \hat{\mathbb{Y}}$, and Figure 5b: $\overline{\mathbb{Y}} \not\subseteq \hat{\mathbb{Y}}$) can be identified, with similar conclusions as for *Case 3* and *Case 4*. In fact, in the former one, we are guaranteed that the level of fairness of p and z will coincide independently from the position of the original vectors since they will both fall on the Fair Space Frontier. On the contrary, no property can be proven for the latter sub-case.

3.3. Fairness Threshold Analysis

The threshold t plays a significant role in distinguishing different cases among those mentioned earlier. For example, by analyzing the effect of decreasing t in *Case 1* we observe that when $\overline{\mathbb{Y}}$ is more aligned to y^* than y^+, it results in *Case 3*; conversely, if it is more aligned with y^+ than y^*, it may result in *Case 4*. Continuously lowering t eventually leads to *Case 5*, and while we can intuitively observe that decreasing the threshold results in closer projections p and z, we yet have no formal proof of the relationship between t and the loss $\mathscr{L}(p, y^*)$.

4. Conclusion

This study introduces GEOFFair – a GEOmetric Framework for Fairness – which leverages geometrical concepts to provide a rigorous and intuitive understanding of fairness in AI. By representing distributions, ML models, fairness constraints, and hypothesis spaces as vectors and sets, GEOFFair allows for visualizing mitigation techniques and constructing proofs-by-witness. The framework facilitates the exploration of various fairness properties, including geometrical distances between fairness vectors, relative fairness comparisons, and the analysis of symmetries, invariances, and trade-offs between fairness metrics.

Through the lens of GEOFFair, we conducted a theoretical analysis of mitigation techniques, leading to the identification of five distinct cases that are essential for analyzing different fairness scenarios. These cases provide valuable insights into the relationship between input vectors, their projections, and the fairness level achieved.

Future work will focus on applying GEOFFair to analyze well-known fairness problems. Geometrical reasoning and projection might also prove very effective for understanding how properties of the loss function and fairness metrics (e.g. convexity, triangle inequality) affect the effectiveness of mitigation techniques.

Finally, exploring the generation of biased data to assess the fairness of AI applications through the lens of GEOFFair will be an important avenue for future research. Overall, the adoption of the GEOFFair framework holds promise for advancing the understanding and development of fair AI systems.

Acknowledgments

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[2] M. Srivastava, H. Heidari, A. Krause, Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2459–2468.

[3] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, K. Varshney, Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing, in: International Conference on Machine Learning, PMLR, 2020, pp. 2803–2813.

[4] I. Kansizoglou, L. Bampis, A. Gasteratos, Deep feature space: A geometrical perspective, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 6823–6838.

[5] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond euclidean data, IEEE Signal Processing Magazine 34 (2017) 18–42.

[6] O. Shahmirzadi, A. Lugowski, K. Younge, Text similarity in vector space models: a comparative study, in: 2019 18th IEEE international conference on machine learning and applications (ICMLA), IEEE, 2019, pp. 659–666.

[7] S. Aghaei, M. J. Azizi, P. Vayanos, Learning optimal and fair decision trees for non-discriminative decision-making, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 1418–1426. URL: https://doi.org/10.1609/aaai.v33i01.33011418. doi:10.1609/aaai.v33i01.33011418.

[8] V. Nurock, R. Chatila, M.-H. Parizeau, What does "ethical by design" mean?, Reflections on Artificial Intelligence for Humanity (2021) 171–190.

[9] P. Jain, P. Kar, Non-convex Optimization for Machine Learning, 2017. doi:10.1561/9781680833690.

A. Possible Scenarios

Let $\mathbb{Z}$ be $\mathcal{Y} \setminus (\hat{\mathbb{Y}} \cup \overline{\mathbb{Y}})$.

Scenario $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} = \varnothing$

Cases in which $y* \in \overline{\mathbb{Y}}$:

   1) $y^*, y^+ \in \overline{\mathbb{Y}}$ and $y^* \equiv y^+$.

   2) $y^*, y^+ \in \overline{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

   3) $y^* \in \overline{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

   4) $y^* \in \overline{\mathbb{Y}}$ and $y^+ \in \hat{\mathbb{Y}}$.

Cases in which $y* \notin \overline{\mathbb{Y}}$:

   5) $y^*, y^+ \in \mathbb{Z}$ and $y^* \equiv y^+$.

   6) $y^*, y^+ \in \mathbb{Z}$ and $y^* \not\equiv y^+$.

   7) $y^* \in \mathbb{Z}$ and $y^+ \in \overline{\mathbb{Y}}$.

   8) $y^* \in \mathbb{Z}$ and $y^+ \in \hat{\mathbb{Y}}$.

   9) $y^*, y^+ \in \hat{\mathbb{Y}}$ and $y^* \equiv y^+$.

  10) $y^*, y^+ \in \hat{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

  11) $y^* \in \hat{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

  12) $y^* \in \hat{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}}$.

Scenario $\overline{\mathbb{Y}} \subset \hat{\mathbb{Y}}$

Cases in which $y* \in \overline{\mathbb{Y}}$:

  13) $y^*, y^+ \in \overline{\mathbb{Y}}$ and $y^* \equiv y^+$.

  14) $y^*, y^+ \in \overline{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

  15) $y^* \in \overline{\mathbb{Y}}$ and $y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$.

  16) $y^* \in \overline{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

Cases in which $y* \notin \overline{\mathbb{Y}}$:

  17) $y^*, y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^* \equiv y^+$.

  18) $y^*, y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

19) $y^* \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}}$.

20) $y^* \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

21) $y^*, y^+ \in \mathbb{Z}$ and $y^* \equiv y^+$.

22) $y^*, y^+ \in \mathbb{Z}$ and $y^* \not\equiv y^+$.

23) $y^* \in \mathbb{Z}$ and $y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$.

24) $y^* \in \mathbb{Z}$ and $y^+ \in \overline{\mathbb{Y}}$.

Scenario $\hat{\mathbb{Y}} \subset \overline{\mathbb{Y}}$

Cases in which $y* \in \overline{\mathbb{Y}}$:

25) $y^*, y^+ \in \hat{\mathbb{Y}}$ and $y^* \equiv y^+$.

26) $y^*, y^+ \in \hat{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

27) $y^* \in \hat{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$.

28) $y^* \in \hat{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

29) $y^*, y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^* \equiv y^+$.

30) $y^*, y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

31) $y^* \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^+ \in \hat{\mathbb{Y}}$.

32) $y^* \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

Cases in which $y* \notin \overline{\mathbb{Y}}$:

33) $y^*, y^+ \in \mathbb{Z}$ and $y^* \equiv y^+$.

34) $y^*, y^+ \in \mathbb{Z}$ and $y^* \not\equiv y^+$.

35) $y^* \in \mathbb{Z}$ and $y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$.

36) $y^* \in \mathbb{Z}$ and $y^+ \in \hat{\mathbb{Y}}$.

Scenario $\hat{\mathbb{Y}} \cap \overline{\mathbb{Y}} \neq \emptyset$ and not a previous case

Cases in which $y* \in \overline{\mathbb{Y}}$:

37) $y^*, y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^* \equiv y^+$.

38) $y^*, y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

39) $y^* \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$.

40) $y^* \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$.

41) $y^* \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

42) $y^*, y^+ \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$ and $y^* \equiv y^+$.

43) $y^*, y^+ \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

44) $y^* \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$ and $y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$.

45) $y^* \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$.

46) $y^* \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

Cases in which $y* \notin \overline{\mathbb{Y}}$:

47) $y^*, y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^* \equiv y^+$.

48) $y^*, y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^* \not\equiv y^+$.

49) $y^* \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$.

50) $y^* \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$.

51) $y^* \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$ and $y^+ \in \mathbb{Z}$.

52) $y^*, y^+ \in \mathbb{Z}$ and $y^* \equiv y^+$.

53) $y^*, y^+ \in \mathbb{Z}$ and $y^* \not\equiv y^+$.

54) $y^* \in \mathbb{Z}$ and $y^+ \in \hat{\mathbb{Y}} \setminus \overline{\mathbb{Y}}$.

55) $y^* \in \mathbb{Z}$ and $y^+ \in \overline{\mathbb{Y}} \cap \hat{\mathbb{Y}}$.

56) $y^* \in \mathbb{Z}$ and $y^+ \in \overline{\mathbb{Y}} \setminus \hat{\mathbb{Y}}$.