An Intuitive Generalisation of Information Geometry[⋆]

Martin Adamčík

Independent Researcher
maths38@gmail.com

Abstract. In this paper we recover some traditional results in geometry of probability distributions, and in particular the convergence of the alternating minimisation procedure, without actually referring to probability distributions. We will do this by discussing a new general concept of two types of points; admissible and agreeable, inspired by multi–agent uncertain reasoning. On one hand, this presents a unique opportunity to make traditional results accessible to a wider audience as no prior knowledge of the topic is required. On the other hand, it allows us to contemplate how a group of humans would seek an agreement without necessarily expressing it in terms of probability distributions, focusing instead on properties. Finally, we recover the traditional setting of probability distributions, including cross–entropy, in the appendix.

Keywords: Information geometry · Divergence · Uncertain reasoning · Cross–Entropy · Fixed point · Alternating minimisation

1 Intuition

> "A point is that which has no part."
>
> Euclid of Alexandria, [11]

Whenever we build a mathematical theory we need to consult our intuition. Should we not do it we may end up building a theory that little resembles the world we are living in, and which is equally inapplicable. In this section, we will start building an intuitive framework that deals with information. We will need to confer with our intuition in the form of our experience on how information is used and how conflicting statements are dealt with.

Our first notion will be indeed the *point*. As in the Euclidean definition that starts this section, it is a building block that is further indivisible. Our point is, however, introduced to represent information rather than the position in a three–dimensional world. We think of several different opinions on a particular matter; each different opinion can be represented as a point. We are not concerned with

[⋆] Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — 432308570.

what further constitutes the opinion and we disregard any knowledge concerning the origins of the opinion, it is simply an indivisible entity to us.

The points, which we have just introduced, can have any of the two following properties in this paper:

1. They can represent an admissible collective point of view of a given group of humans or collection of information sources, shortly called simply an *admissible point*,
2. and to represent an agreement of the group, shortly called an *agreeable point*.

Now, intuitively, an admissible point is meant only to represent the state of collective knowledge, individual members of the group could well disagree and there could be no, what we call, agreeable point. This collective framework of unceretain reasoning was pioneered by Wilmers [15], and we are directly extending it here. An illustration is in Figure 1, all figures can be found in the appendix.

Example. *To illustrate, one scientific study could suggest that the proportion of people that develop a particular disease is somewhere between* 10% *and* 30% *while the other study could indicate that this value is between* 20% *and* 50%. *One way of constructing a point is to specify an ordered pair of individually admissible proportions such as* $(25\%, 40\%)$, *where the first number is admissible according to the first study and the second number is admissible according to the second study. Agreeable admissible points in this particular representation will be the points* (x, x), $x \in [20\%, 30\%]$, *clearly representing the proportions on which the studies agree at the same time. There are other agreeable points that are not admissible, such as* $(35\%, 35\%)$, $(50\%, 50\%)$, $(0\%, 0\%)$ *and so on.*

The example above illustrates the kind of details we will need to go into before the intuitive concept that we develop here can be applied, but at this stage working out the details would only obstruct the general idea and the intuition behind it. We have therefore moved all technical examples and relevant references to the appendix. Here we only point out that our illustration fits Paris–Vencovská framework of uncertain reasoning as in [13].

The previous example was also straightforward enough in establishing agreeable points, but the following questions naturally arise:

1. What shall we do if admissible points contain no agreeable points?
2. How should we measure some kind of distance between an admissible point and an agreeable point in an effort to find a closest point of agreement?
3. Which intuitive principles such a notion of distance should satisfy?

We shall find the answers in this paper.

2 Information Divergence

In the previous section we saw the need for expressing some sort of information distance between two points, but we would not want to require much from this notion at this early stage. In particular, there is no apparent need for it to be a *metric*.

A metric is a symmetric distance between a pair of elements $\mathbf{x}, \mathbf{y}$ of a set. It assigns to each pair $(\mathbf{x}, \mathbf{y})$ a non–negative real number $d(\mathbf{x}, \mathbf{y})$, this number is independent on the order of elements, it is zero if and only if the elements are identical and it satisfies the triangular inequality; $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

Instead, we will consider a much weaker notion of *information divergence*, a mapping D that assigns an ordered pair of points a non–negative real number;

$$D(\mathbf{x}, \mathbf{y}) \geq 0.$$

We say that $D(\mathbf{x}, \mathbf{y})$ is the D *information divergence* from $\mathbf{x}$ to $\mathbf{y}$. Since the symmetry is not required, the D information divergence from $\mathbf{y}$ to $\mathbf{x}$ could be different and therefore we do not call it a distance but a divergence.

Now, let W be the set of all admissible points and V the set of agreeable points. Throughout the paper we will assume that they are both non–empty. Let $\mathit{\Delta}(W)$ be the set of all those agreeable points $\mathbf{v}$ that are such that $D(\mathbf{v}, \mathbf{w})$ is minimal subject to $\mathbf{v} \in V$ and $\mathbf{w} \in W$. In other words, we are looking here at all pairs $(\mathbf{v}, \mathbf{w})$, $\mathbf{v} \in V$ and $\mathbf{w} \in W$, establishing the minimal $D(\mathbf{v}, \mathbf{w})$ if it exits, and collecting all those $\mathbf{v}$ from V that give this minimal divergence into $\mathit{\Delta}(W)$. The purpose of the set $\mathit{\Delta}(W)$ is to determine those agreeable points that have the smallest D information divergence from them to admissible points and to use them as representatives of the set of all admissible points W. In other words, $\mathit{\Delta}(W) \subseteq V$ represents W; it is the agreement of a group of humans or collection of information sources. We will call the points in $\mathit{\Delta}(W)$ *representative points*. See Figure 2 for an illustration.

Intuitively, if $W \cap V \neq \emptyset$; i.e., there are agreeable admissible points, we expect the representation $\mathit{\Delta}(W)$ of W to be formed only by agreeable admissible points, although this ituition is not universally accepted [14]. The following property of D guarantees that this is the case:

Property 1 (Consistency). *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be any two points. Then*

$$D(\mathbf{v}, \mathbf{w}) = 0 \textit{ if and only if } \mathbf{v} = \mathbf{w}.$$

Observation 1. *Let D be such that it satisfies the consistency property. If there are agreeable admissible points then agreeable admissible points form all representative points;*

$$\textit{if } W \cap V \neq \emptyset \textit{ then } \mathit{\Delta}(W) = W \cap V.$$

Proof. First, if $\mathbf{v} \in W \cap V$ then by the consistency property $D(\mathbf{v}, \mathbf{v}) = 0$. We conclude that $\mathbf{v} \in \mathit{\Delta}(W)$ as $\mathbf{v}$ minimises $D(\mathbf{v}, \mathbf{w})$ subject to $\mathbf{v} \in V$ and $\mathbf{w} \in W$. (Note that $D(\mathbf{v}, \mathbf{v})$ cannot be smaller than zero by the definition.) Hence $\mathit{\Delta}(W) \supseteq W \cap V$.

Second, assume that $W \cap V \neq \emptyset$ and $\mathbf{v} \in \mathit{\Delta}(W) \subseteq V$ is such that $\mathbf{v} \notin W$. Then $D(\mathbf{v}, \mathbf{w}) = 0$ for some $\mathbf{w} \in W$, which by the consistency principle gives $\mathbf{v} = \mathbf{w}$. Hence $\mathit{\Delta}(W) \subseteq W \cap V$. □

The consistency property above is formulated more strongly than it is needed to prove Observation 1. Rather than considering any points $\mathbf{v}$ and $\mathbf{w}$, we could have required it only for $\mathbf{v} \in V$ and $\mathbf{w} \in W$. The reason for our choice is that we will need the stronger version later on.

In contrast, if $\mathbf{v} = \mathbf{w}$ implies $D(\mathbf{v}, \mathbf{w}) = 0$ but there are $\mathbf{v} \neq \mathbf{w}$ such that $D(\mathbf{v}, \mathbf{w}) = 0$, it could be possible to have $W \cap V \neq \emptyset$ and $\Delta(W) \supsetneqq W \cap V$, so further weakening of the consistency property would be undesirable.

3 Projections

Our notion of an information divergence is too general to have further useful properties on its own; in particular, if there are no agreeable admissible points we cannot even say that the set of all representative points is always non–empty. We will keep adding assumptions concerning both D and sets of agreeable and admissible points V and W based on what appears rational to us in the context of information geometry. At some point, however, we will need to show that the list of our assumptions is consistent; we will need to find a particular information divergence, and sets W and V that satisfy all those assumptions.

In this section we will require D to have the following properties:

Property 2 (Projection). *Assume that* $\mathbf{v}$ *is an agreeable point. Then there is a unique admissible point* $\mathbf{w}$ *such that* $D(\mathbf{v}, \mathbf{w})$ *is minimal subject to* $\mathbf{w} \in W$.

The unique point $\mathbf{w}$ from the previous property will be denoted $\pi_W(\mathbf{v})$; it is the *D–projection* of $\mathbf{v}$ into W. An illustration is in Figure 3.

Property 3 (Conjugated Projection). *Assume that* $\mathbf{w}$ *is an admissible point. Then there is a unique agreeable point* $\mathbf{v}$ *such that* $D(\mathbf{v}, \mathbf{w})$ *is minimal subject to* $\mathbf{v} \in V$.

The unique point $\mathbf{w}$ from the previous property will be denoted $\widehat{\pi}_V(\mathbf{w})$; it is the *conjugated D–projection* of $\mathbf{w}$ into V. An illustration is in Figure 4.

Intuitively, if we present a group of humans with a point of agreement, we expect them to find a single point among those they consider admissible as their personal opinion in view of the presented agreement. On the other hand, we should be able to establish agreement regardless on which specific admissible point the group presents to us.

Taking this further, the following process taken from [1] and inspired by an earlier version of [15] could resemble a real life agreement seeking:

Example. *Consider a group of humans with their set of admissible points* W. *The group elects a committee whose task is to find a single agreeable point from the set* V. *Naturally, the committee presents the group with their personal opinion or any other provisional starting point* $\mathbf{v}_0$ *that they see appropriate. The group then decides which point from those they consider admissible must have been the case to reach the conclusion suggested by the committee; they project* *the committee's point to their set W. At this stage, being present with a single admissible point, it is now possible for the committee to determine the conjugated projection of that admissible point to the set V; finding the corresponding agreeable point $\mathbf{v}_1$ of the group. Now, it is not at all necessary that $\mathbf{v}_1 = \mathbf{v}_0$. Nevertheless, the committee would be compelled to iterate the whole process until the above process stabilises on a single agreeable point.*

The points of interest from the previous example, although at this stage it is not clear if they even exist, will be called *fixed points*. More explicitly, an agreeable point $\mathbf{v} \in V$ is a fixed point if

$$\widehat{\pi}_V(\pi_W(\mathbf{v})) = \mathbf{v}.$$

The set of all fixed points will be denoted $\varTheta(W)$. See Figure 5 for an illustration.

The example above is of course only one possible way of finding an agreement, although we argue that it is a rational one. An interesting question is how does this way relate to previously suggested information divergence D minimisation, which yields the set $\varDelta(W)$. There could be something:

Observation 2. *Let D be such that it satisfies the projection and conjugated projection properties. Then representative points are also fixed points;*

$$\varDelta(W) \subseteq \varTheta(W).$$

Proof. Let $\mathbf{v} \in \varDelta(W)$ and let d be the smallest D information divergence $D(\mathbf{v}, \mathbf{w})$ subject to $\mathbf{v} \in V$ and $\mathbf{w} \in W$. Such a real number exists by the definition of $\varDelta(W)$ and note that in this paper we always assume that both V and W are non–empty.

Clearly, $D(\mathbf{v}, \pi_W(\mathbf{v})) \geq d$. Now assume that $D(\mathbf{v}, \pi_W(\mathbf{v})) > d$ so there must be $\mathbf{w} \in W$ such that $D(\mathbf{v}, \pi_W(\mathbf{v})) > D(\mathbf{v}, \mathbf{w})$. But this contradicts the definition of $\pi_W(\mathbf{v})$. So it must be that

$$D(\mathbf{v}, \pi_W(\mathbf{v})) = d.$$

Now, assume that $\widehat{\pi}_V(\pi_W(\mathbf{v})) \neq \mathbf{v}$. Nevertheless,

$$D(\widehat{\pi}_V(\pi_W(\mathbf{v})), \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})) = d,$$

otherwise we would contradict the definition of $\widehat{\pi}_V(\pi_W(\mathbf{v}))$. Finally, the equation above implies that both $\mathbf{v}$ and $\widehat{\pi}_V(\pi_W(\mathbf{v}))$ minimise $D(\mathbf{v}, \pi_W(\mathbf{v}))$ subject to $\mathbf{v} \in V$. Such a minimiser is, however, by the conjugated projection property required to be unique, thus

$$\widehat{\pi}_V(\pi_W(\mathbf{v})) = \mathbf{v}.$$

□

It seems that after concluding this section we have more questions than answers:

1. What properties should we require from an information divergence D and sets W, V so that $\varDelta(W) = \varTheta(W)$? Is it even possible?
2. If we iterate the process from the example above; i.e., create a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$, where $\mathbf{v}_{i+1} = \widehat{\pi}_V(\pi_W(\mathbf{v}_i))$, what properties should we require from the information divergence D and sets W and V so that we find an agreement in that way?

We shall find answers in the following sections.

4 Pythagorean Properties

The following property informally says that a group might establish the divergence of their agreement to an admissible point by adding their divergence to the projection of that point to the set of agreeable points and the divergence of the projection to the point concerned.

Property 4 (Pythagorean for Agreeable Points). *Let $\mathbf{v} \in V$ be an agreeable point and $\mathbf{w} \in W$ be an admissible point. Then*

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w}).$$

This property is counter–intuitive from the point of view of the classical Euclidean distance. Although it does not violate the triangular inequality, it is certainly not a property of a distance we are used to. On the other hand, it quite closely resembles how squares taken over the sides of a right–angled triangle behave in the Euclidean geometry (hence the name), see Figures 7 and 8 for an illustration.

Intuitively, using an analogy from the Euclidean geometry, we expect the set of agreeable points in respect to the conjugated D–projection to behave as a flat space into which we projects admissible points. This is quite a strong requirement, we would not want to be so harsh on the set of admissible points. The following property will make admissible points to behave as a convex set.

Property 5 (Pythagorean for Admissible Points). *Let $\mathbf{v} \in V$ be an agreeable point and $\mathbf{w} \in W$ be and admissible point. Then*

$$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}) \leq D(\mathbf{v}, \mathbf{w}).$$

This property is similar to the Pythagorean property for agreeable points but it is weaker. And if the inequality from the statement actually holds in some case for a particular D then this information divergence D is not a metric. See Figures 9 and 10 for an illustration.

The following observation gives us something that also follows from the consistency property on Page , but without assuming it.

Observation 3. *Let D be such that it satisfies the projection and conjugated projection properties, and the Pythagorean properties for agreeable and admissible points. If $\mathbf{v} \in V$ is a fixed point then $D(\mathbf{v}, \mathbf{v}) = 0$ and $D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) = 0$.*

Proof. By the Pythagorean property for agreeable points

$$D(\mathbf{v}, \widehat{\pi}_V(\pi_W(\mathbf{v}))) + D(\widehat{\pi}_V(\pi_W(\mathbf{v})), \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})).$$

But since $\mathbf{v}$ is fixed the above is equivalent to

$$D(\mathbf{v}, \mathbf{v}) + D(\mathbf{v}, \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})),$$

which is possible only if $D(\mathbf{v}, \mathbf{v}) = 0$.

Similarly, by the Pythagorean property for admissible points

$$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) \leq D(\mathbf{v}, \pi_W(\mathbf{v})),$$

which is possible, due to non–negativity of information divergence, only if

$$D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) = 0.$$

□

5 Fixed Points are Representative Points

The following natural property says that the D information divergence from one admissible point to another admissible point should not be smaller than the D information divergence from and to the corresponding agreeable points. Intuitively, seeking an agreement should not take us further apart, see Figure 11.

Property 6 (Convexity). *Let* $\mathbf{w}, \mathbf{u} \in W$. *Then*

$$D(\mathbf{w}, \mathbf{u}) \geq D(\widehat{\pi}_V(\mathbf{w}), \widehat{\pi}_V(\mathbf{u})).$$

We have now all the tools sufficient to prove that fixed points are also representative points, if there is actually a representative point.

Theorem 1 (Characterisation). *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the convexity property. If a representative point exists then the set of fixed points and the set of representative points are equal;*

$$\mathit{\Delta}(W) = \mathit{\Theta}(W).$$

Proof. By Observation 2 on Page we already know that $\mathit{\Delta}(W) \subseteq \mathit{\Theta}(W)$ so it is sufficient to show that $\mathit{\Delta}(W) \supseteq \mathit{\Theta}(W)$.

Because we assumed that a representative point exists and we already know that every representative point is also a fixed point, we may assume that $\widehat{\pi}_V(\mathbf{w}) \in \mathit{\Delta}(W)$, for some $\mathbf{w} \in W$. To make the argument, we now also assume that $\mathbf{v} \in \mathit{\Theta}(W)$ and show that $\mathbf{v} \in \mathit{\Delta}(W)$ in what follows.

The Pythagorean property for representative points

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w})$$

and the Pythagorean property for admissible points

$$D(\mathbf{v},\mathbf{w}) \geq D(\mathbf{v},\pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}),\mathbf{w})$$

give

$$D(\mathbf{v},\widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}),\mathbf{w}) \geq D(\mathbf{v},\pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}),\mathbf{w}), \tag{1}$$

see Figure 12 for an illustration.

Since $\mathbf{v}$ is a fixed point and hence $\mathbf{v} = \widehat{\pi}_V(\pi_W(\mathbf{v}))$, by the convexity property

$$D(\mathbf{v},\widehat{\pi}_V(\mathbf{w})) \leq D(\pi_W(\mathbf{v}),\mathbf{w}). \tag{2}$$

Now, Equations 1 and 2 give

$$D(\widehat{\pi}_V(\mathbf{w}),\mathbf{w}) \geq D(\mathbf{v},\pi_W(\mathbf{v})).$$

Since $\mathbf{w} \in \mathit{\Delta}(W)$ the above must hold with equality and therefore $\mathbf{v} \in \mathit{\Delta}(W)$. □

The proof above was based on ideas from [2].

Observation 4. *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the convexity property. Let $\mathbf{v},\mathbf{u} \in \mathit{\Delta}(W) = \mathit{\Theta}(W)$. Then*

$$D(\mathbf{v},\mathbf{u}) = D(\pi_W(\mathbf{v}),\pi_W(\mathbf{u})).$$

Proof. Looking at (1) in the previous proof, which employed the identical assumptions, and taking $\mathbf{u} = \widehat{\pi}_V(\mathbf{w})$, we obtain

$$D(\mathbf{v},\mathbf{u}) + D(\mathbf{u},\mathbf{w}) \geq D(\mathbf{v},\pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}),\mathbf{w}).$$

Since $\mathbf{v},\mathbf{u} \in \mathit{\Delta}(W)$ we have $D(\mathbf{u},\mathbf{w}) = D(\mathbf{v},\pi_W(\mathbf{v}))$ and the above becomes

$$D(\mathbf{v},\mathbf{u}) \geq D(\pi_W(\mathbf{v}),\mathbf{w}).$$

Finally, by the convexity property, the above is possible only with the equality. □

6 Enter Metric Topology

> "Every reasonable non–pathological space in topology will turn out to be a metric space. On the other hand, developments (. . .) showed there was a need to study a more general class of spaces than merely Euclidean spaces."
>
> Donal W. Kahn, [12]

Thus far we have avoided the need to introduce any topological structure on the set of all points, but this is going to change in this section. First, let us finally introduce a symbol for the set of points here considered as X. Then, let us equip the set of points X with a metric $d(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are any points. Recall that the notion of metric was discussed on Page .

We say that a sequence $\{\mathbf{v}_i\}_{i=1}^{\infty}$ of points *converges* to a point $\mathbf{v}$ if for any real number $\epsilon > 0$ there is j such that $d(\mathbf{v}_i, \mathbf{v}) < \epsilon$ for all $i > j$. We call such a $\mathbf{v}$ a *limit point*.

What we need to establish now is a connection between the metric d and the divergence D, which is a mapping from a Cartesian product $X \times X$ to $\mathbb{R}$;

$$D : X \times X \to \mathbb{R}.$$

Therefore, we need to have a product metric

$$d_p((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \Big([d(\mathbf{x}_1, \mathbf{x}_2)]^p + [d(\mathbf{y}_1, \mathbf{y}_2)]^p\Big)^{\frac{1}{p}},$$

where $1 \leq p < \infty$, in place. Then we can define that a mapping $f : X \times X \to \mathbb{R}$ is *continuous*, if for any real number $\epsilon > 0$ there is $\delta > 0$ such that whenever $d_p((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) < \delta$ we have $|f(\mathbf{x}_1, \mathbf{y}_1) - f(\mathbf{x}_2, \mathbf{y}_2)| < \epsilon$. The last expression is just the standard metric on $\mathbb{R}$, and our definition follows the usual definition of continuity of a mapping between metric spaces. The connection we were looking for is then the following.

Property 7 (Continuity). *D is continuous.*

Intuitively, the property above says that if two pairs of points are close to each other in the product metric, then D does not rip them apart in $\mathbb{R}$.

The following is a straightforward and intuitive consequence of D being continuous, and it is how we will employ continuity to obtain future the results.

Observation 5. *Let D satisfy the continuity property. Assume that a sequence of agreeable points $\{\mathbf{v}_i\}_{i=1}^{\infty}$ converges to $\mathbf{v}$ and a sequence of admissible points $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$. Then the sequence*

$$\{D(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{\infty}$$

converges to $D(\mathbf{v}, \mathbf{w})$.

Proof. For any $\epsilon > 0$ we are tasked with finding j such that $|D(\mathbf{v}_i, \mathbf{w}_i) - D(\mathbf{v}, \mathbf{w})| < \epsilon$ for all $i > j$. Since D is continuous, for any $\epsilon > 0$ there is $\delta > 0$ such that whenever

$$\Big([d(\mathbf{v}_i, \mathbf{v})]^p + [d(\mathbf{w}_i, \mathbf{w})]^p\Big)^{\frac{1}{p}} < \delta$$

we have $|D(\mathbf{v}_i, \mathbf{w}_i) - D(\mathbf{v}, \mathbf{w})| < \epsilon$. Now we simply select j so that $[d(\mathbf{v}_i, \mathbf{v})]^p + [d(\mathbf{w}_i, \mathbf{w})]^p < \delta^p$ for all $i > j$. This is always possible since $\{\mathbf{v}_i\}_{i=1}^{\infty}$ converges to $\mathbf{v}$ and $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$. □

Since we operate in a metric space, we can define that a subset of points X is *compact* if every sequence that can be constructed from its elements has a convergent subsequence and the limit point of this convergent subsequence lies in this subset. In other words, it has Bolzano–Weierstrass property, which in metric spaces is equivalent to compactness.

Observation 6. *If V and W are compact, and D satisfies the continuity property then a representative point exists.*

Proof. Consider the set of all real numbers $D(\mathbf{v},\mathbf{w})$ such that $\mathbf{v}\in V$ and $\mathbf{w}\in W$. This set is bounded from below so it has also the greatest lower bound (a basic property of real numbers). Let us denote it b.

Now, for every $\epsilon>0$ there are $\mathbf{v}\in V$ and $\mathbf{w}\in W$ such that

$$\epsilon+b>D(\mathbf{v},\mathbf{w})\geq b,$$

otherwise b would not be the greatest lower bound. Therefore, we can construct a sequence $\{D(\mathbf{v}_i,\mathbf{w}_i)\}_{i=1}^{\infty}$ that converges to b. Now, due to the compactness of V the sequence $\{\mathbf{v}_i\}_{i=1}^{\infty}$ has a convergent subsequence, say $\{\mathbf{v}_{i_j}\}_{j=1}^{\infty}$. Let $\{\mathbf{w}_{i_j}\}_{j=1}^{\infty}$ be the corresponding sequence in W, which is also compact, so it has also a convergent subsequence. Let $\mathbf{w}\in W$ be its limit point and let $\mathbf{v}\in V$ be the limit point of $\{\mathbf{v}_{i_j}\}_{j=1}^{\infty}$. Then due to Observation 5

$$D(\mathbf{v},\mathbf{w})=b$$

so $\mathbf{v}$ must be a representative point. □

Looking now at the statement of Theorem 1 on Page we can replace the requirement for existence of a representative point by requiring compactness of V and W, and asking D to satisfy the continuity property.

7 Convergence

The following property will be needed to prove that a representative point can be reached by an iterative process.

Property 8 (Four Points). *Let $\mathbf{w},\mathbf{u}\in W$ and $\mathbf{v}\in V$. Then*

$$D(\widehat{\pi}_V(\mathbf{w}),\mathbf{u})\leq D(\mathbf{w},\mathbf{u})+D(\mathbf{v},\mathbf{u}).$$

The four–point property is illustrated in Figure 13.

Thus far we had one property that linked the concept of divergence D and the metric topology given by d; it was the continuity property. Here we provide another one, which somewhat goes in the opposite direction.

Property 9 (Connectivity). *If $\{D(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty}$ converges to zero then so does $\{d(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty}$.*

Naturally, the property above implies that if $\{D(\mathbf{v}_i, \mathbf{v})\}_{i=1}^{\infty}$ converges to zero then $\mathbf{v}$ is the limit point of $\{\mathbf{v}_i\}_{i=1}^{\infty}$. We will use this in the following theorem.

Theorem 2 (Convergence). *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the consistency, convexity, continuity, four–point and connectivity properties. Let $\mathbf{v}_0 \in V$. Define a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ recursively by $\mathbf{v}_{i+1} = \widehat{\pi}_V(\pi_W(\mathbf{v}_i))$. If V and W are compact then the sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ converges to a fixed point.*

Proof. First notice that

$$D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) \geq D(\pi_W(\mathbf{v}_i), \widehat{\pi}_V(\pi_W(\mathbf{v}_i))) \geq$$

$$\geq D(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)), \pi_W(\widehat{\pi}_V(\pi_W(\mathbf{v}_i))))$$

so the sequence of non–negative real numbers $D(\mathbf{v}_i, \pi_W(\mathbf{v}_i))_{i=0}^{\infty}$ converges and its limit point exists (the closed interval $[0, D(\mathbf{v}_0, \pi_W(\mathbf{v}_0))]$ is compact in $\mathbb{R}$ equipped with the standard metric). We will denote this limit information divergence λ.

Furthermore, due to the compactness of V and W the sequences $\{\mathbf{v}_i\}_{i=0}^{\infty}$ and $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ have both a convergent subsequence with a corresponding limit point, we denote these limits points $\mathbf{v} \in V$ and $\mathbf{w} \in W$ respectively. Therefore, by Observation 5,

$$D(\mathbf{v}, \mathbf{w}) = \lambda.$$

What we need to prove at this stage is that the whole sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$, not just its subsequence, converges to $\mathbf{v}$. We will do this considering Figure 14.

By the four–point property

$$D(\mathbf{v}_i, \mathbf{w}) \leq D(\pi_W(\mathbf{v}_{i-1}), \mathbf{w}) + D(\mathbf{v}, \mathbf{w}).$$

and by the Pythagorean property for admissible points

$$D(\pi_W(\mathbf{v}_i), \mathbf{w}) + D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) \leq D(\mathbf{v}_i, \mathbf{w}).$$

Since

$$D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) \geq D(\mathbf{v}, \mathbf{w})$$

it follows that

$$D(\pi_W(\mathbf{v}_i), \mathbf{w}) \leq D(\pi_W(\mathbf{v}_{i-1}), \mathbf{w}).$$

However, we already know that a subsequence of $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ converges to $\mathbf{w}$, so this means that $\{D(\pi_W(\mathbf{v}_i), \mathbf{w})\}_{i=0}^{\infty}$ converges, by Observation 5, to $D(\mathbf{w}, \mathbf{w})$, which is by the consistency property 0. Finally, using the connectivity property, the whole sequence $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ must converge to $\mathbf{w}$.

By the convexity property $D(\pi_W(\mathbf{v}_i), \mathbf{w}) \geq D(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)), \mathbf{v})$ for all i so also

$$\{D(\mathbf{v}_i, \mathbf{v})\}_{i=1}^{\infty}$$

converges to zero which in turn means, making the same argument as above, that $\{\mathbf{v}_i\}_{i=0}^{\infty}$ converges to $\mathbf{v}$ as desired.

However, in order to apply the convexity property above, we need to first establish that $\widehat{\pi}_V(\mathbf{w}) = \mathbf{v}$. For a contradiction let us assume that $\mathbf{v} \neq \widehat{\pi}_V(\mathbf{w})$. By the Pythagorean property for agreeable points

$$D(\widehat{\pi}_V(\mathbf{w}), \widehat{\pi}_V(\mathbf{w}_i)) + D(\widehat{\pi}_V(\mathbf{w}_i), \mathbf{w}_i) = D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}_i),$$

for all i. Since $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$, and $\{\mathbf{v}_i\}_{i=1}^{\infty}$ has a subsequence converging to $\mathbf{v}$ (so we focus only on it), and by Observation 5, we can also write

$$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{v}) + D(\mathbf{v}, \mathbf{w}) = D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}).$$

By the assumption and the consistency property $D(\widehat{\pi}_V(\mathbf{w}), \mathbf{v}) > 0$, so we have that $D(\mathbf{v}, \mathbf{w}) < D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$. But this is not possible, a contradiction.

Similarly we can establish a contradiction with the uniqueness of the D–projection should $\mathbf{w} \neq \pi_W(\mathbf{v})$ utilising the Pythagorean property for admissible points. Therefore $\mathbf{v} = \widehat{\pi}_V(\pi_W(\mathbf{v}))$ and $\mathbf{v}$ is a fixed point. □

Finally, considering Theorem 1 we may claim that the fixed point from the theorem above is also a representative point.

The algorithm (and in fact the idea of the proof presented above) is due to Csiszár and Tusnády [8], who developed it for a particular information divergence and it is known as an *alternating minimisation procedure*. The algorithm was then generalised many times in the literature, see e.g. [6], and the version above can be considered as a further step. Nevertheless, it is still the same idea developed in 1984.

8 Conclusion

We have now achieved the goal as initially stated; we have introduced information geometry without actually specifying the exact nature of admissible and agreeable points we worked with. However, the paper is far from finished. First, in the appendix the classical setting will be formally established and staples of inductive logic; discrete probability distributions and cross–entropy, will be discussed.

Second, as this aspired to be an actual generalisation of information geometry, the future development should be aimed to find a non–trivial and different formalisation of the intuitive concept than the one from the appendix, which is one usually used in inductive logic. This is exciting as one could hope to recover information geometry on mathematical objects originally meant to capture something else, leading to entirely new connections and applications.

Finally, we should also admit that the results presented here were somewhat easy; we placed in enough properties so that the proofs of the desired results went through. The hard job is the opposite: What properties are necessary? This question remains open for now.

References

1. Adamčík, M.: Collective Reasoning under Uncertainty and Inconsistency. Phd thesis, University of Manchester (2014)
2. Adamčík, M.: The information geometry of Bregman divergences and some applications in multi–expert reasoning. Entropy **16**, 6338–6381 (2014)
3. Adamčík, M.: On the applicability of the 'number of possible states' argument in multi–expert reasoning. Journal of Applied Logic **19**, 20–49 (2016)
4. Adamčík, M.: A logician's approach to meta–analysis with unexplained heterogeneity. Journal of Biomedical Informatics **71**, 110–129 (2017)
5. Adamčík, M.: A note on how Rényi entropy can create a spectrum of probabilistic merging operators. Kybernetika **55**, 605–617 (2019)
6. Bauschke, H.H., Combettes, P.L., Noll, D.: Joint minimization with alternating Bregman proximity operators. Pacific Journal of Optimization **2**, 401–524 (2006)
7. Bregman, L.M.: The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **1**, 200–217 (1967)
8. Csiszár, I., Tusnády, G.: Informational geometry and alternating minimization procedures. Statistic and Decisions **1**, 205–237 (1984)
9. Genest, C., Wagner, C.G.: Further evidence against independence preservation in expert judgement synthesis. Aequationes Mathematicae **32**, 74–86 (1987)
10. Jaynes, E.T.: Where do we stand on maximum entropy? In: Levine, R.D., Tribus, M. (eds.) The Maximum Entropy Formalism. pp. 15–118. M.I.T. Press (1979)
11. Joyce, D.E.: Online eddition of Euclid's Elements. http://aleph0.clarku.edu/ djoyce/java/elements (1998), [Online; accessed 12–December–2016]
12. Kahn, D.W.: Topology. Dover Publications, New York (1995)
13. Paris, J.B.: The uncertain reasoner companion. Cambridge University Press, Cambridge (1994)
14. Williamson, J.: Deliberation, judgement and the nature of evidence. Economics and Philosophy **31**, 27–65 (2015)
15. Wilmers, G.M.: A foundational approach to generalising the maximum entropy inference process to the multi–agent context. Entropy **17**, 594–645 (2015)

Appendix

> "One could not see the forest for the trees."
>
> A Common Proverb

In this paper we have accumulated a large number of properties that we require from an information divergence D and from the sets of agreeable and admissible points. Naturally we should ask the following question: Is it actually possible to satisfy them all? In this section we show particular examples that satisfy all the properties, but we will need some additional notions to define them.

Obdurate Committee

> "The point is that we are not ignoring the dynamics, and we are not getting something from nothing, (...) for these all circumstances that are not under the experimenter's control must, of necessity, be irrelevant. (...) Solution by the Maximum Entropy Principle is so unbelievably simple just because it eliminates those irrelevant details right at the beginning of the calculation by averaging over them."
>
> Edwin T. Jaynes, [10]

Here we consider an obdurate committee who stubbornly refuses to iterate the process $\mathbf{v}_1 = \widehat{\pi}_V(\pi_W(\mathbf{v}_0))$. This will help us to further illustrate the setting, toy with it, but foremost illustrate some singular points of information geometry.

First, we postulate existence of the *most uninformative point* $\mathbf{u}$ in the set of agreeable points V. Second, the committee finds $\pi_W(\mathbf{u})$, a unique agreeable point that has the smallest D–divergence from $\mathbf{u}$. If we wanted to represent W by a single point, this is the most natural option as, in respect to D, it has the least added 'information' to it among the agreeable points.

This generalises the concept of the famous *most entropic point*; we recover the usual concept if we choose a specific information divergence D and a specific concept of the point. We will elaborate the details later in this appendix. We only mention that the committee is not ignoring the dynamics of the set of admissible points W by selecting that single point there as well as the experimenter is not doing so in the citation above. If the dynamics were laboriously worked out, we would have obtained this solution anyway.

Let us denote $\pi_W(\mathbf{u})$ by $\mathbf{ME}_D(W)$, and call it the most entropic point in W (in respect to D). Now, the committee wishes to find the agreeable point (if it is not already and agreeable admissible point). To that end, $\widehat{\pi}_V(\mathbf{ME}_D(W))$ is picked, and we denote $O(W) = \{\widehat{\pi}_V(\mathbf{ME}_D(W))\}$ the singleton containing it. An illustration is in Figure 6.

This *obdurate point* need not be a fixed point; and even less a representative point, considering Observation 2. It would indeed be a stubborn committee not to iterate the process further, but be content with it. The committee would argue that the advantage of O is that it contains a single point. We would point out that $O(W) \neq W \cap V = \varDelta(W)$, if $W \cap V \neq \emptyset$ and $W \cap V$ has at least two elements (given D satisfies the consistency property), as shown in Observation 1. Nevertheless, starting the whole iteration process from the most uninformative point appears a well justified idea that indeed lead to a unique point as investigated in Section 7.

Finally, let us point out the following obvious statements.

Observation 7. *If W is a singleton, then*

$$O(W) = \varDelta(W).$$

Observation 8. *If $W \subseteq V$, then*

$$O(W) \subseteq \varDelta(W).$$

The prior follows from Property 3, while the latter from Observation 1.

Points

To provide an actual example of what was discussed in the paper, we start with the J–dimensional *Euclidean space*, which is a set of all ordered J–tuples

$$\mathbf{v} = (v_1, \ldots, v_J),$$

where every v_j is a real number. In other words, $\mathbf{v} \in \mathbb{R}^J$. A $(J-1)$–dimensional *probabilistic simplex* $\mathbb{D}^J$, $J \geq 2$, is a subspace of the J–dimensional Euclidean space defined as those $\mathbf{v} \in \mathbb{R}^J$ that satisfy

$$\sum_{j=1}^{J} v_j = 1.$$

We will confine ourselves to the case when $v_j > 0$, for all $1 \leq j \leq J$, to avoid any pathological cases, which makes $\mathbb{D}^J$ an open set. Such a defined *discrete probability distribution* $\mathbf{v}$ could perhaps represent a probabilistic opinion that an individual may have about the world, and thus it plays a central role in inductive logic and uncertain reasoning [13, 15]. More recently, in [4] they have been used to represent results of individual medical studies.

We say that a subset W of points in $\mathbb{R}^I$ is *convex* if for any two $\mathbf{v}, \mathbf{w} \in W$ we have that also

$$(\lambda \cdot v_1 + (1-\lambda) \cdot w_1, \ldots, \lambda \cdot v_I + (1-\lambda) \cdot w_I) \in W,$$

for all $\lambda \in [0,1]$. We say that a subset W of points in $\mathbb{R}^I$ is *closed* if the limit point of every convergent sequence constructed from the elements of W has its limit inside W, in respect to the standard Euclidean metric.

Now, let us consider a closed convex set of points

$$W \subseteq \underbrace{\mathbb{D}^J \times \ldots \times \mathbb{D}^J}_{n}.$$

Note that $I = Jn$ (in the definition of convexity above) and $\mathbf{w} \in W$ is of the form $\mathbf{w} = (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})$, where each $\mathbf{v}^{(i)} \in \mathbb{D}^J$ is a probability distribution admissible by the member i of a group of n individuals. This set W will be an example of a set of admissible points discussed earlier in the paper.

Finally, let

$$V \subseteq \underbrace{\mathbb{D}^J \times \ldots \times \mathbb{D}^J}_{n}$$

be such that in each $\mathbf{v} \in V$ all members are in agreement; $\mathbf{v} = (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})$, where $\mathbf{v}^{(1)} = \ldots = \mathbf{v}^{(n)}$. This set V is not closed (because $\mathbb{D}^J$ is not), but we can fix a sufficiently small $\epsilon > 0$ and ask every $v_j^{(i)} > \epsilon$, $1 \leq i \leq n$, $1 \leq j \leq J$. A suitable ϵ exists (in a sense that $W \subseteq V$ must be possible), since W is assumed closed. Such a set V will be an example of a set of agreeable points discussed earlier in the paper.

Clearly, it could be that there are some agreeable points in W, but V and W could be as well disjoint. In any case, W is assumed non–empty, while V is non–empty by definition. Both W and V are defined closed and bounded, and hence they are both compact. Note that compactness was required in Observation 6 and Theorem 2.

Divergences

After we have introduced the points, let us now define a divergence from one point to another. In [5], the following divergence from $\mathbf{v} \in V$ to $\mathbf{w} \in W$ based on the Rényi entropy was defined:

$$D_r(\mathbf{v}, \mathbf{w}) = \frac{1}{n} \sum_{j=1}^{Jn} [(w_j)^r - (v_j)^r - r(w_j - v_j)(v_j)^{r-1}],$$

where $2 \geq r > 1$. For $r = 2$ this divergence becomes the well known *squared Euclidean distance*

$$E(\mathbf{v}, \mathbf{w}) = \frac{1}{n} \sum_{j=1}^{Jn} (v_j - w_j)^2,$$

exceptionally a symmetric divergence. The proof that the set of representative points $\varDelta^{D_r}(W)$ based on the Rényi entropy is well defined is in [1].

Another way to define the divergence D from $\mathbf{v} \in V$ to $\mathbf{w} \in W$ is to take the *Kullback–Leibler divergence* (also known as cross–entropy)

$$\mathrm{KL}(\mathbf{v}, \mathbf{w}) = \frac{1}{n} \sum_{j=1}^{Jn} w_j^{(i)} \log \frac{w_j^{(i)}}{v_j}.$$

A limit theorem relating the set of representative points $\varDelta^{D_r}(W)$ based on the Rényi entropy to the set of representative points $\varDelta^{\mathrm{KL}}(W)$ based on the Kullback–Leibler divergence has been proven in [5];

$$\emptyset \neq \lim_{r \searrow 1} \varDelta^{D_r}(W) \subseteq \varDelta^{\mathrm{KL}}(W).$$

Whether or not the above holds with equality is an open problem.

The proofs that the divergences defined above satisfy all Properties 1 to 9 discussed in this paper are scattered in [1] and [2], and they are all special cases of a general convex Bregmann divergence [7].

What we discussed in this paper now gives us

$$\varDelta^{D_r}(W) = \varTheta^{D_r}(W), \text{ and } \varDelta^{\mathrm{KL}}(W) = \varTheta^{\mathrm{KL}}(W),$$

the representative and fixed points are the same points, and we can get a representative point by iterating projections and conjugated projections.

Discussion

The technical nature of this appendix obscured how natural and simple these examples actually are. We will mention some singular points in what follows.

Regardless on which of the above mentioned divergences is taken for D, the conjugated D–projection of an admissible point $\mathbf{w} = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in W$ to the set of agreeable points $(\underbrace{\mathbf{v}, \dots, \mathbf{v}}_{n}) \in V$ in fact gives

$$\mathbf{v} = \Big(\frac{1}{n}\sum_{i=1}^{n} w_1^{(i)}, \dots, \frac{1}{n}\sum_{i=1}^{n} w_J^{(i)}\Big).$$

So we have here only the ordinary arithmetic mean applied to J coordinates respectively. In the literature this operator is known as the *linear pooling operator*, see [9]. It is a common choice of representing different opinions $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ of n individuals as a single point in $\mathbb{D}^J$, a natural agreeable point.

Defining the most uninformative point $\mathbf{u} = (\underbrace{\mathbf{v}, \dots, \mathbf{v}}_{n}) \in V$ using the *uniform probability distribution*

$$\mathbf{v} = \Big(\frac{1}{J}, \dots, \frac{1}{J}\Big) \in \mathbb{D}^J,$$

$\mathbf{ME}_{\mathrm{KL}}(W)$, defined as the KL–projection of $\mathbf{u}$ into W, is the usual most entropic point in W. It is defined as that $\mathbf{w}$ that maximises the Shannon entropy

$$-\sum_{j=1}^{Jn} w_j \log w_j.$$

An obdurate committee would then take this most entropic point and find the conjugated KL–projection in the set of agreeable points V, which we now know to be equivalent to applying a linear pooling operator, and be content with it.

We suggest that a rational committee would iterate the whole process endlessly until a representative point in $\varDelta^{\mathrm{KL}}(W) = \varTheta^{\mathrm{KL}}(W) \subseteq V$ is reached. A combinatorial argument in favour of using $\varDelta^{\mathrm{KL}}(W)$ in a specific context was presented in [3].

Should there be only one individual, $n = 1$, then $W = \varDelta^{\mathrm{KL}}(W) = \varTheta^{\mathrm{KL}}(W) \subseteq V$, so there would be no need to iterate the process as $\mathbf{ME}_{\mathrm{KL}}(W)$ would be trivially, see Observation 8, a fixed point. This would correspond to the classical most entropic solution when there are no conflicting sources of information.

Figures

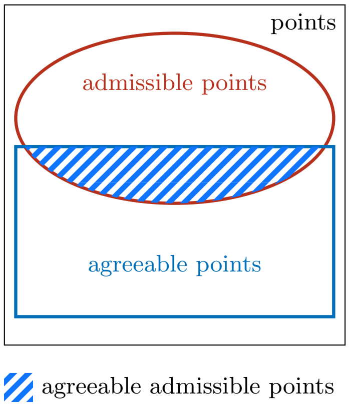


Fig. 1. An illustration of the set of all points.

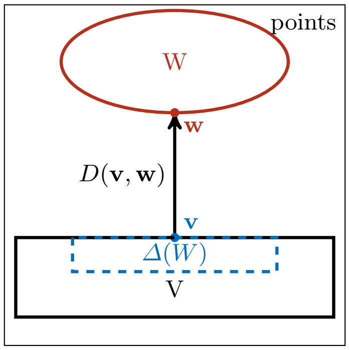


Fig. 2. An illustration of the representative points.

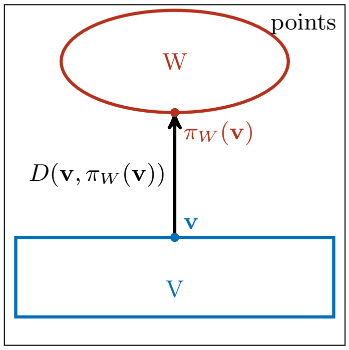


Fig. 3. An illustration of the D–projection.

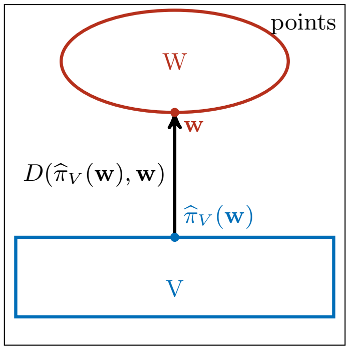

points
W
$\mathbf{w}$
$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$
$\widehat{\pi}_V(\mathbf{w})$
V

Fig. 4. An illustration of the conjugated D–projection.

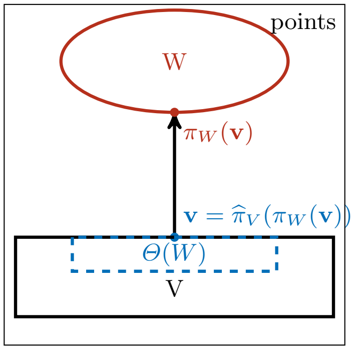


Fig. 5. An illustration of the fixed points.

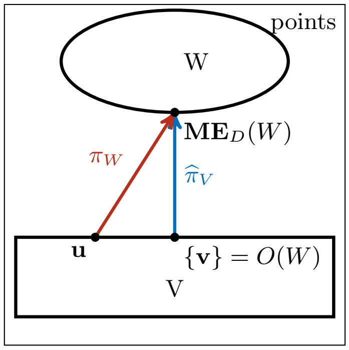
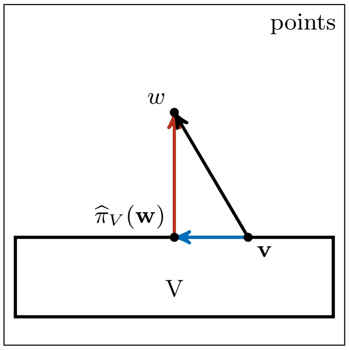
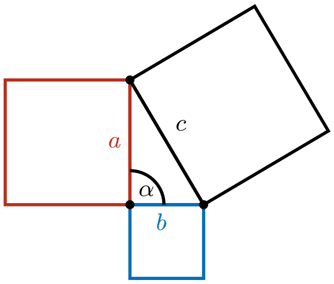


Fig. 6. An illustration of an obdurate committee.



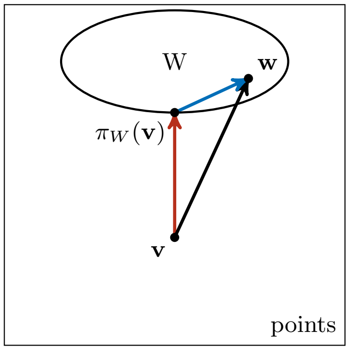
$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w})$$

Fig. 7. An illustration of the Pythagorean property for agreeable points.



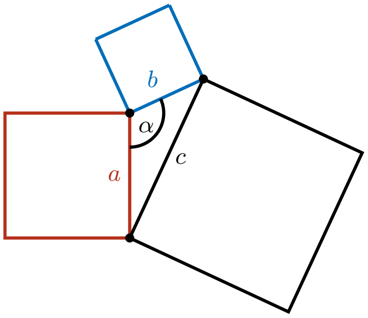
if $\alpha = 90^\circ$ then $a^2 + b^2 = c^2$

Fig. 8. How squares behave in the Euclidean geometry.



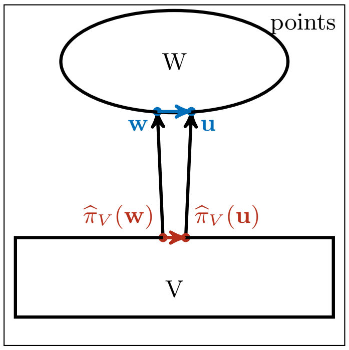
$$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}) \leq D(\mathbf{v}, \mathbf{w})$$

Fig. 9. An illustration of the Pythagorean property for admissible points.



if $90^\circ \leq \alpha \leq 180^\circ$ then $a^2 + b^2 \leq c^2$

Fig. 10. How squares behave in the Euclidean geometry.



$$D(\mathbf{w}, \mathbf{u}) \geq D(\widehat{\pi}_V(\mathbf{w}), \widehat{\pi}_V(\mathbf{u}))$$

Fig. 11. An illustration of the convexity property.

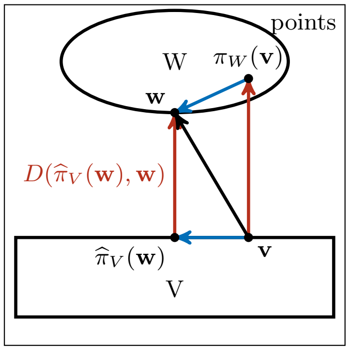
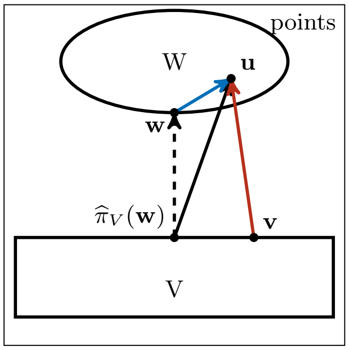

points
W
$\pi_W(\mathbf{v})$
$\mathbf{w}$
$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$
$\widehat{\pi}_V(\mathbf{w})$
$\mathbf{v}$
V

Fig. 12. An illustration of the proof for Theorem 1.



$$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{u}) \leq D(\mathbf{w}, \mathbf{u}) + D(\mathbf{v}, \mathbf{u})$$

Fig. 13. An illustration of the four points property.

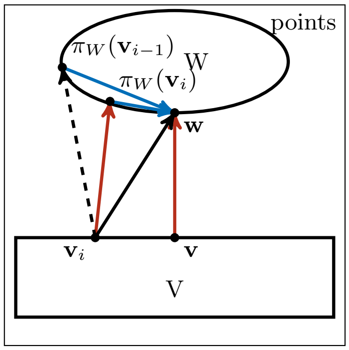


Fig. 14. An illustration of the proof of Theorem 2.