Annals of Mathematics and Artificial Intelligence (2023) 91:217–238
https://doi.org/10.1007/s10472-022-09807-0

On the use of Wasserstein distance in the distributional analysis of human decision making under uncertainty

Antonio Candelieri[1] · **Andrea Ponti**[1,2] · **Ilaria Giordani**[2,3] · **Francesco Archetti**[3]

Published online: 15 July 2022

Abstract
The key contribution of this paper is a theoretical framework to analyse humans' decision-making strategies under uncertainty, and more specifically how human subjects manage the trade-off between information gathering (exploration) and reward seeking (exploitation) in particular active learning in a black-box optimization task. Humans' decisions making according to these two objectives can be modelled in terms of Pareto rationality. If a decision set contains a Pareto efficient (dominant) strategy, a rational decision maker should always select the dominant strategy over its dominated alternatives. A distance from the Pareto frontier determines whether a choice is (Pareto) rational. The key element in the proposed analytical framework is the representation of behavioural patterns of human learners as a discrete probability distribution, specifically a histogram considered as a non-parametric estimate of discrete probability density function on the real line. Thus, the similarity between users can be captured by a distance between their associated histograms. This maps the problem of the characterization of humans' behaviour into a space, whose elements are probability distributions, structured by a distance between histograms, namely the optimal transport-based Wasserstein distance. The distributional analysis gives new insights into human behaviour in search tasks and their deviations from Pareto rationality. Since the uncertainty is one of the two objectives defining the Pareto frontier, the analysis has been performed for three different uncertainty quantification measures to identify which better explains the Pareto compliant behavioural patterns. Beside the analysis of individual patterns Wasserstein has also enabled a global analysis computing the WST barycenters and performing k-means Wasserstein clustering.

Keywords Active learning · Pareto analysis · Wasserstein distance · Barycenters · Uncertainty quantification · Human learning · Exploration-exploitation dilemma · Clustering

✉ Andrea Ponti
a.ponti5@campus.unimib.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

[2] Oaks s.r.l., Milan, Italy

[3] Department of Computer Science, Systems and Communication, Milan, Italy

1 Introduction

1.1 Motivation

Human decision-making under uncertainty requires balancing e*xploitation* – meaning the use of the *knowledge* collected so far to maximize immediate reward – and *exploration* – meaning investing resources to acquire more knowledge to update one's beliefs. This balance is usually referred to as the exploration-exploitation dilemma: decisions allowing for increasing knowledge do not necessarily lead to the greatest immediate reward [1, 2].

A number of researchers have been analysing how humans deal with uncertainty in their decision making. [3–5]: beside the relevance for cognitive sciences the issue is important also for machine learning: indeed, *human learners* are amazingly fast and effective at adapting to unfamiliar environments and incorporating upcoming knowledge and the analysis of their behaviour might be relevant in the design of algorithms of active learning. The reference task considered in this paper is the optimization problem:

$$\mathrm{x}^* = \underset{x \in \Omega \subset \Re^d}{\operatorname{argmax}} f(x) \tag{1}$$

with $f(x)$ black box, meaning that its analytical form is not given, no derivatives are available, and the value of $f(x)$ can be only known pointwise through expensive and noisy evaluations. Finally, Ω denotes the *search space*, usually box bounded.

Recently, the Bayesian Optimization framework (BO) [6, 7] has become one of the most efficient method for solving (1).BO is a sequential model based optimizer in which at a generic iteration n, the player/agent/algorithm chooses a location $x^{(n)}$ to query and observe/collect the associated function value, possibly perturbed by noise, that is $y^{(n)} = f(x^{(n)}) + \varepsilon$. The goal is to get close to the optimizer x^* within a limited number, N, of trials. The choice of $x^{(n)}$ is performed according to a so-called acquisition (or infill) function which, at each iteration, manages the balance between exploration and exploitation.

The key questions addressed in this paper are:

- How do humans solve problem (Eq. 1) which translates into how close the human generated sequence compares to the sequence generated by the optimization algorithm?
- How do humans solve the exploration-exploitation dilemma which underpins the question of how humans perceive uncertainty? Are humans' strategies sample efficient?
- Which analytical tools can give valuable insights about the analysis of human generated search sequences?
- Which quantification of uncertainty does better characterize human behaviour?
- Does the analysis of human behaviour offer valuable clues to the design of machine learning algorithms?

Have been arguing, based on empirical evidence, that strategies adopted by humans in solving global optimization problems have a much stronger association with BO than to other optimization algorithms [8, 9]. This conclusion matches with the fact that Gaussian Process (GPs) and Bayesian learning, first proposed in ([10, 11], have emerged in the cognitive sciences as central paradigms in modelling human learning. The GP model offers an evaluation of predictive uncertainty of the outcome of the next decision conditioned on previous decisions and observed outcomes. Fitting a GP requires to choose, a priori, a *kernel* as covariance function. Different kernels are available, each one implying a different

characterization for the approximation of $f(x)$. An important consideration reported in [12], is that *"GPs with standard kernels struggle on function extrapolation problems that are trivial for human learners"*. Moreover, [13] remarked that different quantifications of the uncertainty – as discussed later in the paper – are a key concept also in theories of cognition and emotion.

The key element in this paper is the representation of behavioural patterns of human learners, over different tasks, as a discrete probability distribution. To each human subject it is associated a histogram which is a sort of a signature of his/her behaviour. In this sense our approach is related to Symbolic Data Analysis (SDA) [14]. In this representation the similarity between users' behaviours can be captured by a distance between their associated histograms. This maps the problem of the characterization of humans' behaviour into a space whose elements are histograms: this space is structured by a distance between histograms, namely the Wasserstein distance. The simplest way is to compare a set of parametric features built from the probability distribution, such as the mean or higher moments. This approach would be limited as the effect of such parameters does not consider the whole distribution.

The Wasserstein (WST) distance is a field of mathematics which studies the geometry of probability spaces and provides a principled framework to compare and align probability distributions. The Wasserstein distance can be traced back to the works of Gaspard Monge [15] and Lev Kantorovich [16].

WST has evolved into a very rich mathematical structure whose complexity and flexibility are analysed in a landmark volume [17] and, in the discrete domain, in the tutorial [18].

The computation of the WST requires the solution of a constrained linear optimization problem which can have, a very large number of variables and constraints, and can be shown to be equivalent to a min-flow problem. Recently, several specialized computational approaches have drastically reduced the computational hurdles [19].

Specifically, we consider one instance of WST, called Earth Mover Distance (EMD), which is a natural and intuitive distance between discrete probability distributions and in particular histograms. WST requires a notion of distance between points in the underlying domain which is called the *ground distance* (usually Euclidean but can be any norm).

The main advantage of WST is that it is a cross binning distance, and it is not affected by different binning schemes. Moreover, WST matches naturally the perceptual notion of nearness and similarity. This is not the case of Kullback-Leibler (KL) and χ-square distances that account only for the correspondence between bins of the same index and do not use information across bins or distributions with different binning schemes, that is different support. An important element of the WST theory is the *barycenter* which offers a useful synthesis of a set of distributions. The barycenter allows for a standard clustering method like k-means to be generalized to WST spaces.

A key question is how humans manage the exploration-exploitation dilemma balancing, in the choice of the new point to evaluate, the expected improvement and the uncertainty. This is modelled as a bi-objective optimization problem which leads to the Pareto analysis. This in turn begets the question whether humans are Pareto-rational agents (i.e., take Pareto optimal decisions in the space of expected improvement and uncertainty). The analysis of computational results allows to formulate at least a tentative answer to why, or rather in which conditions, we observe deviations from (Pareto) "rationality" and switches towards "exasperated" exploration.

Since the uncertainty is one of the two objectives defining the Pareto frontier, the analysis has been performed for three different uncertainty quantification measures with the aim to identify the one making the humans' decisions sequences more compliant with the

Pareto-rationality model. The key result is an analytical framework to characterize how deviations from "rationality" depend on (i) individuals' features represented in the histograms and (ii) uncertainty quantification.

1.2 Contributions of this paper

The key contribution of the paper is the proposal of a distributional analysis of human search pattern based on the WST distance. This distributional analysis has been conducted at the individual level and an aggregate level computing barycenters and performing clustering in the WST space. It is also interesting to remark that while most of the previous works addressed how people assess the information value of possible queries, in this paper we rather address the issue of the perception of probabilistic uncertainty itself. Note that the BO algorithm is not actually executed: sequences of points are generated by humans and compared with optimal Pareto fronts generated analytically.

The computational results and their analysis allow to formulate at least an insight into the following points:

- Do humans always make "rational" choices (i.e., Pareto optimal decisions between the improvement expected and uncertainty) or, in some cases, they "exasperate" exploration?
- Do different uncertainty quantification measures lead to different classifications of humans' decisions? And which uncertainty quantification measure make humans "more rational"?
- Does the distributional representation provide an efficient signature of the subject/task?
- Does the Wasserstein distance capture the difference between the behaviour of two subjects on a given task?
- What is the average behaviour on a given task as measured by the Wasserstein barycenter of the searches of all individuals?
- What is an index of the Pareto compliance of a human as measured by the Wasserstein barycenter of his/her behaviour over all tasks?
- How do different kernels and uncertainty quantifications impact on the Pareto compliance of an individual?

1.3 Related works

In Section 1.1 we have briefly introduced the issue of uncertainty quantification in humans and its relationship with learning and optimization and new analytical tools, based on the WST distance, to characterize humans' behaviour. Here we provide a more specific analysis of the prior work and significant recent results. The literature on the WST distance is now immense: general references have been already given in the introduction, other will be given in Section 4. Here we limit to very specific references mainly focused on computational issues related to barycenters and clustering. An early contribution is from [20], who proposes an EMD-based clustering to analyse mobility usage patterns which is shown to cluster meaningfully also sparse signatures. [21] uses a dimensionality reduction by Self-Organizing Maps (SOM) learning and then cluster data within a WST space. [22] proposes an Iterative Swapping Algorithm (ISA) for the computation of the barycenter which is shown to have a quadratic complexity. [23] proposes an approach based on the Alternating

Direction Method of Multipliers (ADMM) for WST clustering. [24] introduces a hybrid WST distance based on Gaussian approximations.

As far as cognitive science is concerned, an early contribution [25] analyses how humans manage the trade-off between exploration and exploitation in non-stationary environments. Successively, [2] demonstrates that humans use both *random* and *directed exploration*. [26] show how directed exploration in humans amounts to adding an *"uncertainty bonus"* to estimated reward values and how this brings to the *Upper Confidence Bound* (UCB) acquisition function in Multi Armed Bandits [27] and BO [28]. [4] distinguish between *irreducible uncertainty*, related to the reward stochasticity, and epistemic *uncertainty*, which can be reduced through information gathering. In the former the decision strategy is *random search* while in the latter is *directed exploration* which attaches an uncertainty bonus to each decision value. This distinction mirrors the one in Machine Learning between *aleatoric* uncertainty – due to the stochastic variability inherent in querying $f(x)$ – and *epistemic* uncertainty – due to the lack of knowledge about the actual structure of $f(x)$ – which can be reduced by collecting more information.

In Bayesian Optimization, the exploration-exploitation dilemma has recently modelled as a bi-objective optimization problem. [29] minimize the predictive mean (associated to exploitation) while maximizing uncertainty, typically the predictive standard deviation as in UCB (associated to exploration). [30, 31] show that taking a decision by randomly sampling from the Pareto frontier can outperform other acquisition functions. The main motivation is that the Pareto frontier offers a set of Pareto-efficient decisions wider than that allowed by "traditional" acquisition functions.

A recent important contribution is [32] which, given the observed search path generated by a human subject in the execution of a black box optimization task, infers the unknown acquisition function underlying the sequence. For the solution of this problem, referred to as Inverse Bayesian Optimization (IBO), a probabilistic framework for the non-parametric Bayesian inference of the acquisition function is proposed.

The issue of deviations from Pareto optimality in decision making has become mainstream economics under the name of behavioural economics and prospect theory [33], from the seminal work in [34] to [35] which identifies the most common causes for violations of dominance. A central question in economic theory is whether this analysis sits well with the Paretian expected utility theory or rather begets an entirely different approach as proposed in [36]. A basic conclusion of behavioural economics is that rather than being labelled "irrational", non-Pareto compliant behaviour is just not well described by the rational-agent model.

1.4 Outline of the paper

Section 2 introduces the definitions of Gaussian Process (GP) regression and different uncertainty quantifications. Section 3 develops a framework for the application of the Pareto analysis to the specific problem considering three different uncertainty quantification measures. Section 4 introduces the Wasserstein (WST) distance, both the basic notions and the computational issues. Section 5 introduces the experimental framework used for data collection, that is the decisions taken by the humans according to their personal search strategies, and the proposed analytical framework. Section 6 describes the relevant results obtained by the application of the analytical framework. Finally, Section 7 outlines the conclusions of this study and the perspective of future works.

2 Materials and methods

2.1 Gaussian process regression

A GP is a *random distribution over functions* $f : \Omega \subset \mathfrak{R}^d \rightarrow \mathfrak{R}$ denoted with $f(x) \sim GP(\mu(x), k(x, x^{'}))$ where $\mu(x) = \mathbb{E}(f(x)) : \Omega \rightarrow \mathfrak{R}$ is the mean function of the GP and $k\left(x, x^{'}\right) : \Omega \times \Omega \rightarrow \mathfrak{R}$ is the *kernel* or *covariance function*. One way to interpret a GP is as a collection of correlated random variables, any finite number of which have a joint Gaussian distribution, so $f(x)$ can be considered as a sample drawn from a multivariate normal distribution. In Machine Learning, GP modelling is largely used for both classification and regression tasks [37, 38], providing probabilistic predictions by conditioning $\mu(x)$ and $\sigma^2(x)$ on a set of available data/observations.

Let denote with $\mathrm{X}_{1:n} = \{x^{(i)}\}_{i=1,\ \ldots,\ n}$ a set of n locations in $\Omega \subset \mathfrak{R}^d$ and with $y_{1:n} = \{f(x^{(i)}) + \varepsilon\}_{i=1,\ldots,n}$ the associated function values, possibly noisy with ε a zero-mean Gaussian noise $\varepsilon \sim \mathcal{N}\left(0, \lambda^2\right)$. Then $\mu(x)$ and $\sigma^2(x)$ are the GP's posterior predictive mean and standard deviation, conditioned on $\mathrm{X}_{1:n}$ and $y_{1:n}$ according to the following equations:

$$\mu(x) = \mathrm{k}\left(x, \mathrm{X}_{1:n}\right) \left[\mathrm{K} + \lambda^2 I\right]^{-1} y_{1:n} \tag{2}$$

$$\sigma^2(x) = k(x, x) - \mathrm{k}\left(x, \mathrm{X}_{1:n}\right) \left[\mathrm{K} + \lambda^2 I\right]^{-1} \mathrm{k}\left(\mathrm{X}_{1:n}, x\right) \tag{3}$$

where $\mathrm{k}(x, \mathrm{X}_{1:n}) = \{k(x, x^{(i)})\}_{i=1,\ \ldots,\ n}$ and $\mathrm{K} \in \mathfrak{R}^{n \times n}$ with entries $\mathrm{K}_{ij} = k(x^{(i)}, x^{(j)})$.

The choice of the kernel establishes prior assumptions over the structural properties of the underlying (aka latent) function $f(x)$, specifically its smoothness. However, almost every kernel has its own hyperparameters to tune – usually via Maximum Log-likelihood Estimation (MLE) or Maximum A Posteriori (MAP) – for reducing the potential mismatches between prior smoothness assumptions and the observed data. Common kernels for GP regression – considered in this paper – are:

- Squared Exponential: $k_{SE}\left(x, x'\right) = e^{-\frac{\|x - x'\|^2}{2\ell^2}}$
- Exponential $k_{EXP}\left(x, x'\right) = e^{-\frac{\|x - x'\|}{\ell}}$
- Power-exponential $k_{PE}\left(x, x'\right) = e^{-\frac{\|x - x'\|^p}{\ell^p}}$
- Matérn3/2: $k_{M3/2}\left(x, x'\right) = \left(1 + \frac{\sqrt{3}\ \|x - x'\|}{\ell}\right) e^{-\frac{\sqrt{3}\ \|x - x'\|}{\ell}}$
- Matérn5/2: $k_{M5/2}\left(x, x'\right) = \left[1 + \frac{\sqrt{5}\ \|x - x'\|}{\ell} + \frac{5}{3}\left(\frac{\|x - x'\|}{\ell}\right)^2\right] e^{-\frac{\sqrt{5}\ \|x - x'\|}{\ell}}$

2.2 Uncertainty quantification and active learning

In decision making, uncertainty is usually associated to exploration: when the uncertainty is large it could be more profitable to bet on the upside and adopt an explorative behaviour to acquire more knowledge about $f(x)$. Global optimization methods differ one from another in how they generate the next decision (i.e., location) $x^{(n+1)}$. To do this, BO fits a GP according

to (2–3) and where $X_{1:n}=\{x^{(i)}\}_{i=1,\ldots,n}$ and $y_{1:n}=\{y^{(i)}\}_{i=1,\ldots,n}$ are the two sequences of, respectively, decisions made and associated observed outcomes. Then, an acquisition function, combining GP's $\mu(x)$ and $\sigma(x)$, is optimized to obtain $x^{(n+1)}$, while dealing with the exploration-exploitation trade-off. In this paper we shall consider 2 acquisition functions:

Expected Improvement (EI) [39] and GP Confidence Bound (i.e., Upper Confidence Bound, UCB, Lower Confidence Bound LCB for minimization) [28]:

$$EI(x) = \left(\mu(x) - y^{+}\right)\boldsymbol{\Phi}\left(\frac{\mu(x)-y^{+}}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{\mu(x)-y^{+}}{\sigma(x)}\right)$$
$$UCB(x) = \mu(x) + \sqrt{\beta}\sigma(x)$$

where $\boldsymbol{\Phi}$ and ϕ are the standard normal cumulative distribution function (cdf) and the standard normal probability density function (pdf).

Let $\mathcal{K}$ denotes the set of kernels to choose as GP's prior. In this study $\mathcal{K} = \left\{k_{SE}, k_{EXP}, k_{PE}, k_{M3/2}, k_{M5/2}\right\}$.

Let $\zeta(x)$ denotes the improvement expected by querying the objective function at location x, depending on the GPs' posterior (i.e., one GP for each kernel in $\mathcal{K}$). Formally, $\zeta(x)=\mu(x)-y^{+}$, where $y^{+} = \max_{i=1,\ldots,n}\left\{y^{(i)}\right\}$ because we are considering $\max_{x\in\Omega\subset\mathbb{R}^d} f(x)$.

Let denote with $\mathcal{U}$ the set of possible uncertainty quantification measures.

In this paper we consider the following three alternatives:

- GP's predictive standard deviation, namely $\sigma(x)$.
- GP's differential entropy. For a GP it is given by $H\left(y|X_{1:n}\right) = \frac{1}{2}\log\det(\mathrm{K}) + \frac{d}{2}\log\det(2\pi e)$, where $\mathrm{K}\in\mathfrak{R}^{n\times n}$ with entries $\mathrm{K}_{ij}=k(x^{(i)},x^{(j)}), \forall x^{(i)}, x^{(j)}\in X_{1:n}$ [37].
- Distance from previous decisions, inspired from [40] and denoted by $z(x)$:

$$z(x) = \begin{cases} 0 \text{ if} \exists x^{(i)} \in \mathbf{X}_{1:\mathrm{n}} : \left\|x - x^{(i)}\right\|_2^2 = 0 \\ \frac{2}{\pi}\tan^{-1}\left(\frac{1}{\sum_{j=1}^{n} w_j(x)}\right) \text{otherwise} \end{cases} \tag{4}$$

with $w_j(x) = \frac{e^{-\left\|x-x^{(j)}\right\|_2^2}}{\left\|x-x^{(j)}\right\|_2^2}$.

3 Pareto analysis and Pareto compliance

Given the GP conditioned on the decisions performed so far, it is possible to map the next decision $x^{(n+1)}\in\Omega$ as the solution of a bi-objective choice, with objectives $\zeta(x)$ and $u(x)\in\mathcal{U}$ (both to be maximized).

Pareto rationality is the theoretical framework to analyse multi-objective optimization problems where q objective functions $\gamma_1(x)$, …, $\gamma_q(x)$ where $\gamma_i(x):\Omega\to\mathbb{R}$ are to be simultaneously optimized in $\Omega\subseteq\mathbb{R}^d$. We use the notation $\boldsymbol{\gamma}(x)=(\gamma_1(x),\cdots,\gamma_q(x))$ to refer to the vector of all objectives evaluated at a location $\mathbf{x}$. The goal in multi-objective optimization is to identify the Pareto frontier of $\boldsymbol{\gamma}(x)$.

To do this we need an ordering relation in $\mathbb{R}^q$: $\boldsymbol{\gamma} = \left(y_1,\ldots,y_q\right) \preccurlyeq \boldsymbol{\gamma}' = \left(y_1',\ldots,y_q'\right)$ if and only if $\gamma_i \le \gamma_i'$. This ordering relation induces an order in $\Omega: x \preccurlyeq x'$ if and only if $\boldsymbol{\gamma}(x)\preccurlyeq\boldsymbol{\gamma}(x')$.

We also say that γ' dominates γ (strongly) and $\boldsymbol{\gamma}\preccurlyeq\boldsymbol{\gamma}'$ if $\exists i=1,\ldots,q$ for which $\gamma_i < \gamma_i'$. The optimal non-dominated solutions define the so-called Pareto frontier.

The interest in finding locations x having the associated $\boldsymbol{\gamma}(x)$ on the Pareto frontier is clear: they represent an optimal trade-off between conflicting objectives and are the only ones, according to the Pareto rationality, to be considered. In this paper $q=2$, with $\gamma_1(x)=\zeta(x)$ and $\gamma_2(x) = u(x) \in \mathcal{U}$. As both the objectives are not expensive to evaluate the Pareto frontier can be approximated by sampling a grid of m points in Ω, denoted by $\hat{\mathbf{X}}_{1:m} = \left\{x^{(j)}\right\}_{j=1,\ldots,m}$, and then computing the associated pairs $\Psi_{1:m}=\{(\zeta(x^{(j)}),u(x^{(j)}))\}_{j=1,\ldots,m}$.

The Pareto frontier can be approximated as:

$$\mathcal{P}\left(\Psi_{1:m}\right) = \left\{\psi \in \Psi_{1:m} : \forall \psi' \in \Psi_{1:m}\ \psi \succ \psi'\right\} \tag{5}$$

where $\psi=(\zeta(x),u(x))$ and $\psi'=(\zeta(x'),u(x'))$, and $\psi \succ \psi' \Longleftrightarrow \zeta(x)>\zeta(x') \wedge u(x)>u(x')$.

The only way to analyse how different uncertainty quantification measures can lead to completely different decisions – even if anyway Pareto rational – is to localize, within the search space $\Omega \subset \mathfrak{R}^d$, the locations whose associated objectives lay on the Pareto frontier (namely, the Pareto set). According to results reported in [41], the Pareto-rational decisions (i.e., Pareto set) do not significantly depend on kernel. Instead, an evident difference arises with respect to the uncertainty quantification measure: one of the three considered in the study allows for accounting, as Pareto rational, choices which are instead explorative for the other two measures.

Summarizing, the trade-off mechanism between exploration and exploitation is associated to Pareto-rationality: humans deal with this trade-off by making decisions whose expected value and uncertainty lay on the Pareto front. However, our hypothesis is that humans could also take non-Pareto-rational decisions, which therefore go beyond all the possible exploration-exploitation trade-offs allowed by this model. Thus, it is important to measure how much a decision can be considered "far from a Pareto-rational one".

Every next decision, $x^{(n+1)}$, can be analysed according to the distance of its "image" $(\zeta(x^{(n+1)}),u(x^{(n+1)}))$ from the Pareto frontier, computed as follows:

$$d\left(\overline{\psi}, \overline{\mathcal{P}}\right) = \min_{\psi \in \overline{\mathcal{P}}}\left\{\left\|\overline{\psi} - \psi\right\|_2^2\right\} \tag{6}$$

where $\overline{\psi} = \left(\zeta\left(x^{(n+1)}\right), u\left(x^{(n+1)}\right)\right)$ and $\overline{\mathcal{P}} = \mathcal{P}\left(\Psi_{1:m}\right) \cup \left\{\overline{\psi}\right\}$.

This distance is computed for every choice among the five kernels and the three uncertainty quantification measures previously presented.

The most relevant result [41] is that, in some cases, it is possible to observe a shift from Pareto-rationality to not-Pareto-rationality, whichever is the uncertainty quantification measure adopted, including that maximizing the number of Pareto-rational decisions. This means that, in the case where there is not evident chance to exploit, and there is not any exploration-exploitation trade-off compliant to the Pareto-rational model, humans move towards "exasperated" exploration, where with the term "exasperated" we want to remark the fact that the decision is even more explorative than the pure exploration offered by the Pareto-rational model.

4 The Wasserstein distance – Basic notions and numerical approximation

Measuring the distance between distributions can be accomplished by many alternative methods. A general class of distances, known as *f*-divergences, is based on the expected value of a convex function of the ratio of two distributions. If P and Q are two probability distributions over $\mathbb{R}^d$ and f is a convex function such that $f(0)=1$ the f-divergence is given by:

$$D_f(P,Q) = \mathbb{E}_Q f\left(\frac{P}{Q}\right) \tag{7}$$

According to the choice of f the above formula yields specific distances including Kullback-Leibler (and its symmetrized version Jensen-Shannon), Hellinger, total variation and χ-square divergence.

In this paper we focus on the WST distance whose basic notions are given in Section 4.1 while Section 4.2 is devoted to the computation of the barycenter between distributions and the extension of k-means clustering to the Wasserstein space. It is important to remark that the presentation is quite basic omitting any mathematical characterization of WST for which the reader is referred to ([17, 19].

The WST metric is based on the solution of an optimal transport (OT) problem. WST enables to synthetizes the comparison between two multi-dimensional distributions through a single metric using all information in the distributions. Moreover, WST distance is generally well defined and provide an interpretable metric between distributions.

The WST distance can be traced back to the works of Gaspard Monge [15] and Lev Kantorovich [16]. Recently, also under the name of Earth Mover Distance (EMD) it has been gaining increasing importance in several fields like Imaging [42], Natural Language Processing (NLP) [43] and the generation of adversarial networks [44].

4.1 Basic notions

The WST distance between continuous probability distributions is:

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\inf_{\gamma \in \Gamma\left(P^{(1)}, P^{(2)}\right)} \int_{X \times X} d\left(x^{(1)}, x^{(2)}\right)^p d\gamma\left(x^{(1)}, x^{(2)}\right)\right)^{\frac{1}{p}} \tag{8}$$

where $d(x^{(1)},x^{(2)})$ is also called *ground distance* (usually it is the Euclidean norm), $\Gamma(P^{(1)},P^{(2)})$ denotes the set of all joint distributions $\gamma(x^{(1)},x^{(2)})$ whose marginals are respectively $P^{(1)}$ and $P^{(2)}$, and p is an index. The Wasserstein distance is also called the Earth Mover Distance (EMD). The EMD is the minimum energy cost of moving and transforming a pile of sand in the shape of $P^{(1)}$ to the shape of $P^{(2)}$. The cost is quantified by the amount of sand moved times the moving distance $d(x^{(1)},x^{(2)})$. The EMD then is the cost of the optimal transport plan.

There are some specific cases, very relevant in applications, where WST can be written in an explicit form. Let $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ be the cumulative distribution for one-dimensional distributions $P^{(1)}$ and $P^{(2)}$ on the real line and $\left(\hat{P}^{(1)}\right)^{-1}$ and $\left(\hat{P}^{(2)}\right)^{-1}$ be their quantile functions.

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\int_0^1 \left|\left(\hat{P}^{(1)}\right)^{-1}\left(x^{(1)}\right) - \left(\hat{P}^{(2)}\right)^{-1}\left(x^{(2)}\right)\right|^p dx\right)^{\frac{1}{p}} \tag{9}$$

Let's now consider the case of a discrete distribution P specified by a set of support points x_i with $i=1, \ldots, m$ and their associated probabilities w_i such that $\sum_{i=1}^m w_i = 1$ with $w_i \geq 0$ and $x_i \in M$ for $i=1, \ldots, m$.

Usually, $M=\mathbb{R}^d$ is the d-dimensional Euclidean space with the l_p norm and x_i are called the support vectors. M can also be a symbolic set provided with a symbol-to-symbol similarity. P can also be written using the notation:

$$P(x) = \sum_{i=1}^{m} w_i \delta\left(x - x_i\right) \tag{10}$$

where $\delta(\cdot)$ is the Kronecker delta.

The WST distance between two distributions $P^{(1)} = \left\{ w_i^{(1)}, x_i^{(1)} \right\}$ with $i=1, \ldots, m_1$ and $P^{(2)} = \left\{ w_i^{(2)}, x_i^{(2)} \right\}$ with $i=1, \ldots, m_2$ is obtained by solving the following linear program:

$$W\left(P^{(1)}, P^{(2)}\right) = \min_{\gamma_{ij} \in \mathbb{R}^+} \sum_{i \in I_1, j \in I_2} \gamma_{ij}\, d\left(x_i^{(1)}, x_j^{(2)}\right) \tag{11}$$

The cost of transport between $x_i^{(1)}$ and $x_j^{(2)}$, $d\left(x_i^{(1)}, x_j^{(2)}\right)$, is defined by the p-th power of the norm $\left\| x_i^{(1)}, x_j^{(2)} \right\|$ (usually the Euclidean distance).

We define two index sets $I_1 = \{1, \ldots, m_1\}$ and I_2 likewise, such that

$$\sum_{i \in I_1} \gamma_{ij} = w_j^{(2)}, \forall j \in I_2 \tag{12}$$

$$\sum_{j \in I_2} \gamma_{ij} = w_i^{(1)}, \forall i \in I_1 \tag{13}$$

Equations (12) and (13) represent the in-flow and out-flow constraint, respectively. The terms γ_{ij} are called matching weights between support points $x_i^{(1)}$ and $x_j^{(2)}$ or the optimal coupling for $P^{(1)}$ and $P^{(2)}$.

The discrete version of the WST distance is usually called Earth Mover Distance (EMD). For instance, when measuring the distance between grey scale images, the histogram weights are given by the pixel values and the coordinates by the pixel positions. In the specific case of histograms, the entries γ_{ij} denote the amount of weight of the bin i (source) that has to be moved to bin j (sink).

The basic computation of OT between 2 discrete distributions involves solving a network flow problem whose computation scales typically cubically in the size of the measures.

In the case of one-dimensional histograms, which will be considered in this paper, the computation of WST can be performed by a simple sorting and the application of the following equation

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\frac{1}{n} \sum_{i}^{n} \left| x_i^{(1)*} - x_i^{(2)*} \right|^p \right)^{\frac{1}{p}} \tag{14}$$

where $x_i^{(1)*}$ and $x_i^{(2)*}$ are the sorted samples.

4.2 The barycenter and clustering

Consider a set of N discrete distributions, $\mathbf{P} = \{P^{(1)}, \ldots, P^{(N)}\}$, with $P^{(k)} = \left\{ \left(w_i^{(k)}, x_i^{(k)} \right) : i = 1, \ldots, m_k \right\}$ and $k=1, \ldots, N$, then, the associated barycenter, denoted with $\overline{P} = \left\{ \left(\overline{w}_1, x_1 \right), \ldots, \left(\overline{w}_m, x_m \right) \right\}$, is computed as follows:

$$\overline{P} = \underset{P}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{N} \lambda_k W\left(P, P^{(k)}\right) \tag{15}$$

where the values λ_k are used to weight the different contributions of each distribution in the computation. Without loss of generality, they can be set to $\lambda_k = 1/N \forall k = 1, \ldots, N$.

The synthesis through a barycenter of a set of distributions have several advantages, among which: the WST barycenter – also called the Fréchet mean of distributions – appears to be a meaningful feature to represent the mean variation of a set of distributions, and offers a useful synthesis of the structure of probability distributions, in particular:

- it is sensitive to the underlying geometry. Consider 3 distributions $P^{(1)}=\delta_0$, $P^{(2)}=\delta_\varepsilon$ and $P^{(3)}=\delta_{100}$. $W(P^{(1)},P^{(2)})\approx 0$, $W(P^{(1)},P^{(3)})\approx W(P^{(2)},P^{(3)})\approx 100$. The distances Total variation, Hellinger and Kullback-Leibler take the value 1, thus they fail to capture our intuition that $P^{(1)}$ and $P^{(2)}$ are close to each other while they are far away from $P^{(3)}$.
- it is *shape preserving*. Denote $P^{(1)}$, …, $P^{(N)}$ and assume that each $P^{(j)}$ can be written as a location shift of any other $P^{(i)}$, with $i \neq j$. Suppose that each $P^{(j)}$ is defined as $P^{(j)} = \mathcal{N}(\mu_j, \Sigma)$, then the barycenter has the closed form:

$$\overline{P} = \mathcal{N}\left(\frac{1}{N}\sum_{j=1}^{N}\mu_j, \Sigma\right) \tag{16}$$

in contrast to the (Euclidean) average of the $\frac{1}{N}\sum_{j=1}^{N} P^{(j)}$.

Therefore, the concept of barycenter enables clustering among distributions, in a space whose metric is the WST distance. More simply, the barycenter in a space of distributions is the analogue of the centroid when the clustering takes place in a Euclidean space. The most common and well-known algorithm for clustering data in the Euclidean space is k-means. Since it is an iterative distance-based (aka representative based) algorithm, it is easy to propose variants of k-means by simply changing the distance adopted to create clusters, such as the Manhattan distance (leading to k-medoids) or any kernel allowing for non-spherical clusters (i.e., kernel k-means). The crucial point is that only the distance is changed, while the overall iterative two-step algorithm is maintained. This is also valid in the case of the WST k-means, where the Euclidean distance is replaced by WST and centroids are replaced by barycenters:

- Step 1 – Assign. Given the current k barycenters at iteration t, namely $\overline{P}_t^{(1)}, \ldots, \overline{P}_t^{(k)}$, clusters $C_t^{(1)}, \ldots, C_t^{(k)}$ are identified by assigning each one of the distributions $P^{(1)}$, …, $P^{(N)}$ to the closest barycenter:

$$C_t^{(i)} = \left\{ P^{(j)} \in \mathbf{P} : \overline{P}_t^{(i)} = \underset{Q=\left\{\overline{P}_t^{(1)},\ldots,\overline{P}_t^{(k)}\right\}}{\operatorname{argmin}} W\left(Q, P^{(j)}\right) \right\}, \forall i = 1, \ldots, k \tag{17}$$

- Step 2 – Optimize. Given the updated composition of the clusters, update the barycenters:

$$\overline{P}_{t+1}^{(i)} = \underset{Q}{\operatorname{argmin}} \frac{1}{\left|C_t^{(i)}\right|} \sum\nolimits_{P \in C_t^{(i)}} W(Q, P) \tag{18}$$

that comes directly from Eq. (15).

As in k-means, a key point of WST k-means is the initialization of the barycenters. In the case that all the distributions in **P** are defined on the same support, then they can be randomly initialized, otherwise, a possibility is to start from k distributions randomly chosen among

those in **P**. Finally, termination of the iterative procedure occurs when the result of the assignment step does not change any longer or a prefixed maximum number of iterations is achieved.

The major computational issue is the polynomial complexity of the linear programming solvers commonly used to compute WST. Starting from the consideration that the variables in w are more important than the matching weights, approximate solvers have been proposed, specifically Sinkhorn solvers, which will be detailed later. Here it is just important to remark that they allow to manage the trade-off between accuracy and computational cost through a regularization hyperparameter. Another approach is taken in [23] based on ADMM. Entropic regularization enables scalable computations, but large values of the regularization parameter could induce an undesirable smoothing effect while low values not only reduce the scalability but might induce several numeric instabilities.

5 Experimental results and their analysis

This section contains the results of the WST based analysis of the humans' search data, collected through a gaming application based on the implementation used in [9]. Figure 1 shows the web-based Graphical User Interface of the gaming app, with a game play example reporting: (a) the game field with previous selected locations and associated scores, (b) the current accumulated score, and (c) the remaining trials. The game target is searching for the location having the highest score.

Fourteen volunteers have been enrolled (among colleagues and friends), asking for solving ten different tasks, each one referring to a global optimization test function (reported in the Appendix). For each task, every player has a maximum number of 20 clicks (decisions) available.

In Section 5.1 the root variable is the specific task, in Section 5.2 the individual subject. The data has been analysed both locally, comparing different behaviours, and locally computing the barycenter and performing a WST k-means clustering.

5.1 Wasserstein analysis of the test functions

The histogram we use is based on the notion of decile. A decile rank arranges the data in order from the lowest to the highest and it is done on a scale of one to ten where each successive decile corresponds to an increase of 10% of the points. The basic histogram therefore has 10 bins corresponding to deciles in the distributions, with weights representing the number of players whose decisions were Paretian for that decile.

A first insight can be performed by visual inspection: clearly "stybtang" and "bukin" are difficult in that they generate fewer Paretian decisions than "schwefel". The ideal distribution, that is a fully Pareto compliant distribution, is a useful target which enables an intuitive yet quantitative evaluation of the "Pareto value" of each histogram as represented in Fig. 2 through the WST distance between each histogram and the ideal one. A further analysis, using WST, can be conducted according to the distance between a function or a subject histogram and its Paretian ideal. This is reported in Table 1.

The data can be evaluated globally using the barycenter calculated according to formulas (15) and shown in Fig. 3 (the distance from the barycenter to the ideal histogram is in the last row of Table 1).

Also clustering can be performed in the WST space. Specifically, we have used WST k-means. Since our main objective is to partition the behavioural patterns into Pareto

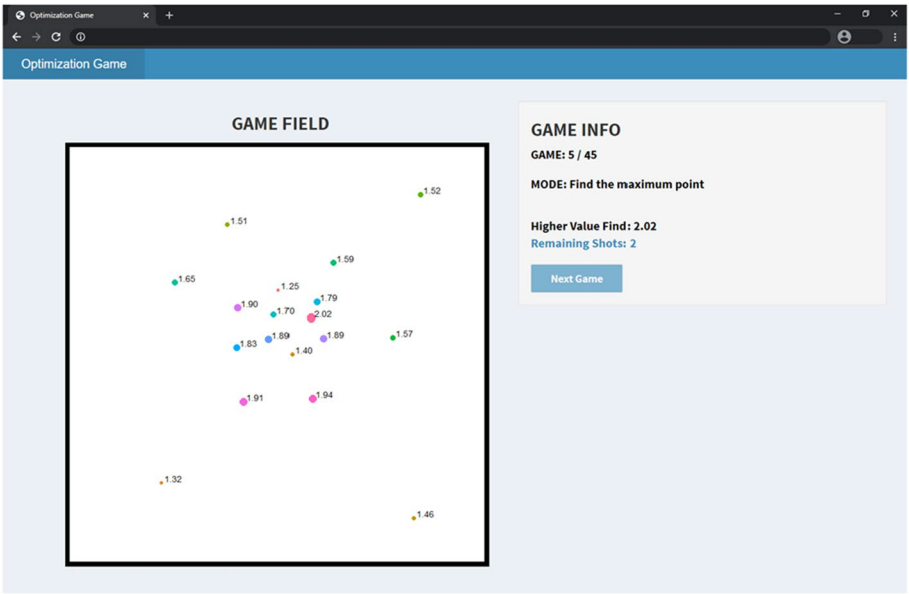


Fig. 1 Gamification app: locations selected so far and associated score

and not-Pareto decisions, k = 2 is a reasonable choice. The results depend on the uncertainty measure and are:

- For the entropy-based uncertainty quantification:

$$C_1^h = (griewank, levy, rastr, schwef, ackley),$$
$$C_2^h = (stybtang, goldpr, beale, bukin6, branin)$$

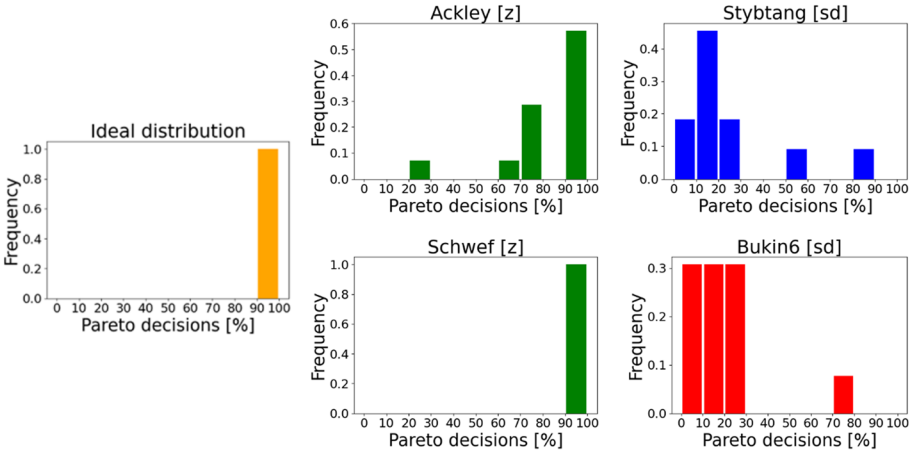


Fig. 2 Number of players with respect to percentage of decisions classified as Pareto rational, separately for the three uncertainty quantification measures. One chart for each test problem

Table 1 Wasserstein distances from the ideal distribution

Test function	H	SD	Z
Ackley	3.57	4.79	1.29
Beale	6.77	7.23	7.15
Branin	6.43	7.73	6.92
bukin6	7.23	7.54	7.25
Goldpr	6.21	4.36	5.50
griewank	4.00	0.57	3.86
Levy	4.79	6.43	2.86
Rastr	3.64	4.21	3.79
schwef	4.71	5.23	0.00
stybtang	6.92	7.00	7.08
Barycenter	5.33	5.52	4.58

- For the σ-based uncertainty quantification:

$$C_1^{sd} = (griewank, rastr, schwef, ackley, goldpr)$$
$$C_2^{sd} = (stybtang, levy, beale, bukin6, branin)$$

- For the inverse distance-based uncertainty quantification:

$$C_1^{z} = (griewank, rastr, schwef, ackley, levy)$$
$$C_2^{z} = (stybtang, goldpr, beale, bukin6, branin)$$

The functions which were visually singled out as hard and easy are correctly assigned to two different clusters under the uncertainty measure h and z. The cluster quality metric is Dunn's. We have computed also $k=3$ and $k=4$, the interpretation is less natural, and the metric value accordingly worsens.

The values in Table 1 of the "distance from the ideal", are summarized in Fig. 4. Two relevant results have to be remarked: (i) the two identified clusters always capture also the difference in terms of "distance from ideal" independently of the uncertainty quantification measures, and (ii) the distinction between the two clusters is maximized for the uncertainty quantification measure z.

A way of visualizing the clustering results is through the box-plot representation, which shows that cluster 1, associated to Pareto decisions, has a significantly smaller WST distance than cluster 2, even more significant in the case that z is used as uncertainty quantification measure.

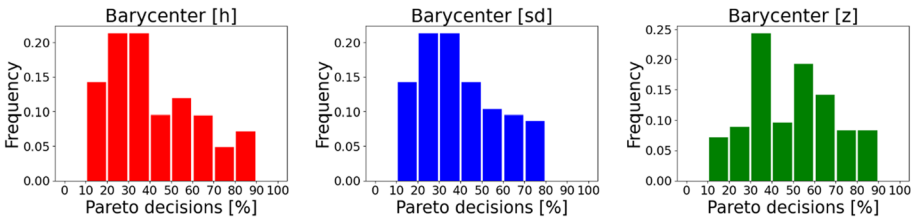


Fig. 3 Wasserstein barycenters of the histograms related to the test functions

5.2 Wasserstein analysis of the users

Figure 5 shows the same histogram as in Section 5.1 but referred to each subject: the weights are the distribution of the number of test problems with respect to the percentage of Pareto rational decisions. A histogram is provided for each subject, comparing the distributions obtained considering each one of the three uncertainty quantification measures. A stacked histogram is provided for each subject, comparing the distributions obtained for each one of the three uncertainty quantification measures. The full set of charts is reported in the supplementary material. Again, we can notice, by visual inspection, two relatively Paretian players (U01 and U13) and two not-Paretian (U05 and U14). Fig. 6 shows that the highest percentages of Pareto rational decisions (90% or 100%) are associated to $z(x)$. The ideal distribution, that is a fully Pareto compliant distribution, is useful target which enables an intuitive yet quantitative evaluation of the "Pareto value" of each histogram as represented in Table 2 through the WST distance between each histogram and the ideal one.

This effect can be evaluated globally using the barycenter (Fig. 6) computed according to the formula and the distance from the barycenter to the ideal situation (last row of Table 2).

Since our main objective is to partition the behavioural patterns into Paretian and non-Paretian, K = 2 seems a reasonable choice Clustering in WST space show that the two clusters can capture most of the difference in terms of distance from ideal, independently of the adopted uncertainty quantification measure, but now the difference is more relevant in the case of the entropy-based uncertainty quantification measure, instead of the inverse distance based one. The subjects which were visually singled out as Paretian and non-Paretian are correctly assigned to two different clusters under the uncertainty measure h and z. The cluster quality metric is Dunn's. We have also computed $k = 3$ and $k = 4$, the interpretation is less natural, and the metric value accordingly worsens.

A way of visualizing the clustering results is through the box-plot representation which shows that cluster 1 (Fig. 7) has a significantly smaller WST distance from the ideal than cluster 2, more significant for z.

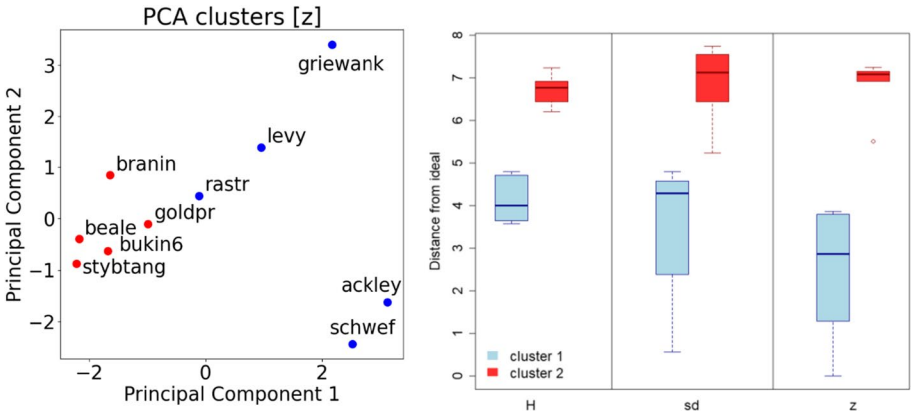


Fig. 4 Functions' clustering represented using PCA (left). Box plot of the distance between functions' histograms and the ideal one (right)

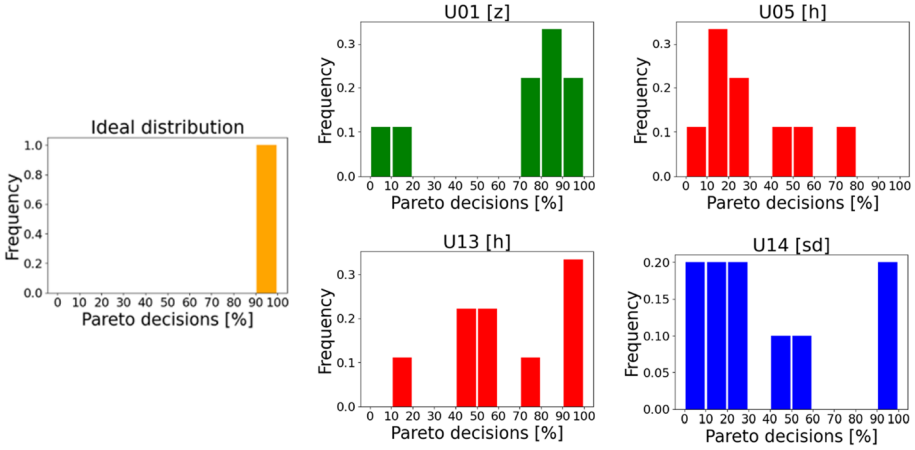


Fig. 5 Number of test problems with respect to percentage of decisions classified as Pareto rational, separately for the three uncertainty quantification measures. One chart for each player

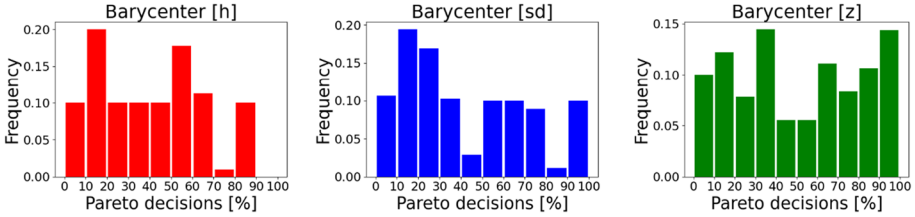


Fig. 6 Wasserstein barycenters of the histograms related to the users

Table 2 Wasserstein distances from the ideal distribution

User	H	SD	Z
U01	3.89	4.11	2.67
U02	5.50	5.22	4.89
U03	6.10	5.00	5.10
U04	5.80	5.90	4.56
U05	6.44	5.89	5.00
U06	5.30	5.20	3.00
U07	5.30	5.78	5.00
U08	5.50	5.40	3.80
U09	5.33	4.14	4.50
U10	5.80	5.60	4.78
U11	5.60	6.10	4.30
U12	5.80	6.80	5.80
U13	3.11	4.20	3.50
U14	5.60	5.70	4.78
Barycenter	5.47	5.34	4.40

6 Learning with distances

Beside the specific issue of analysing human behaviour, distributional inputs can occur in a number of practical situations as physical simulations or computer experiments. The simplest method is to compare a set of parametric features such as the mean or higher moments, but these parameters do not take the whole distribution into account. Commonly used kernels depend on the Euclidean distance between points. In the case of distributional inputs, we want to construct positive definite kernels on sets of probability measures. The research on this topic research has branched in two directions.

6.1 Wasserstein induced kernel

Bayesian learning has largely focused on Euclidean and categorical domains which are suitable for optimizing scalar hyper-parameters of machine learning algorithms. This is no longer sufficient for important applications as neural architecture search. To this effect [45] proposes to use a distance metric in the space of neural network architectures based on optimal transport and a Bayesian Optimization framework whose kernel is induced by Wasserstein distance.

Using kernels limits the choice of distribution distances as the resulting kernel has to be positive definite: a widely used distance as Kullback-Leibler does not qualify. The key difficulty in "kernelizing" WST is that the kernel obtained computing the exponential of the square WST distance between distributional inputs does not lead to a positive definite kernel. As shown in Bachoc [46] many eigenvalues of the Wasserstein induced covariance matrix are negative. Three recent papers [47–49] have analysed the specific conditions in which the exponentiation of WST yields a positive definite kernel.

A general solution to the problem in the setting of Hilbert spaces has been provided in [50]. and [51]. Specific positive definite kernels are designed in order to map distributions into a Reproducing Kernel Hilbert Space (RKHS) and extend kernel methods to probability measures. Results are strongly dependent of d and p. In the first paper the problem has been solved for $d = 1$. The positive definite kernel provided for $d = 1$ is not positive any longer when it is extended to $d > 1$.

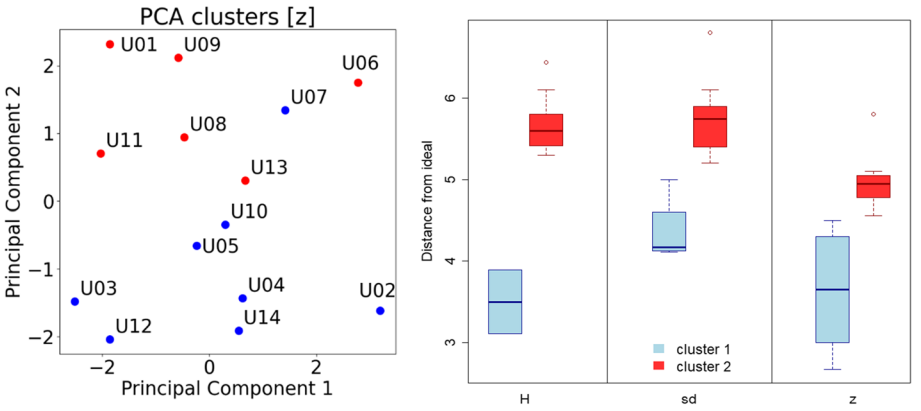


Fig. 7 Users' clustering represented using PCA (left). Box plot of the distance between users' histograms and the ideal one (right)

The problem is solved for $d>1$ in the second paper, in which GPs with multidimensional distribution inputs are embedded using optimal transport. The kernels are of the form $K(u,v)=F(\|u-v\|)_H$ for all u, $v\in H$. F provides a positive definite kernel on any Hilbert space iff it does so in finite Euclidean spaces. Any finite nonnegative Borel measure v on $[0,\infty)$defines a function $F(t)=\int_{\mathbb{R}} e^{-ut^2 v(u)}$ resulting in a definite positive kernel. It is shown in Bachoc [46] that special cases of $F(t)$ result in common kernels as squared exponential, Matérn and power exponential. In some special cases, notably for Gaussian distributions the optimal transport map, and the related kernel, can be computed explicitly [52].

6.2 Using directly the distance

A very interesting perspective opened by the results analysed in Section 6 stems from the observation that the locations in the space $(\psi=(\zeta(x),u(x)))$ associated to Pareto-rational decisions and the degree of Pareto compliance do not change so much depending on the kernel, but rather on the uncertainty quantification: $\sigma(x)$, $h(x)$ or $z(x)$. Moreover since $z(x)$ is most closely related to Pareto behaviour and it's also the one non kernel based this result could also be regarded as an indication that distance could be embedded directly in the learning process. In this way we could design learning algorithms that are built from pairwise dissimilarity measures between distributions. This perspective has been first analysed in [53] which introduces the concept of "good" dissimilarity functions that map a distribution into a space. It can be also shown that there exists in that space a linear separator that produces low errors. These results have been extended in [54] to discrete distributions including the theoretical guarantees if the number of distributions is large enough and enough samples are obtained for each distribution. Based on the Wasserstein distance the paper contains a performance analysis of kernel versus distance-based classifiers. It shows that Wasserstein distance embedding performs better than kernel mean embeddings and computing WST is more tractable than estimating f-divergences of empirical distribution.

A distributional distance-based learning has been shown to be very effective also in simulation-optimization problems over discrete structures ([55, 56]: the Multi Objective Evolutionary Algorithm based on the Wasserstein distance (MOEA/WST) has been shown to be more sample efficient than benchmark evolutionary approaches. This method also compares favourably, in terms of wall clock time, with the BoTorch implementation of multi-objective BO using Expected Improvement and the SE kernel.

7 Conclusions and perspectives

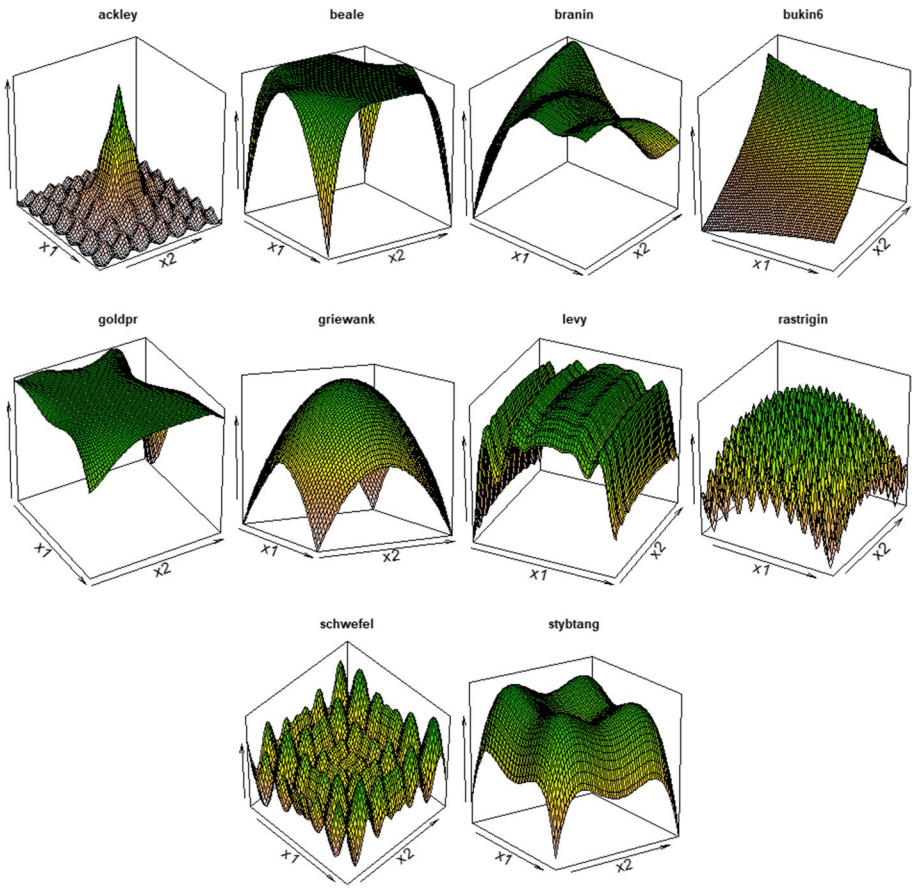
The Pareto analysis of human search data offers insights as whether humans are Pareto rational in performing search tasks. The Wasserstein space has been shown to offer a meaningful representation of how different uncertainty quantifications are implied by human behaviour when solving optimization tasks. The characterization of human behaviour can be performed both at the individual level (single user/single task) and an aggregate level computing barycenters and performing clustering in the Wasserstein space. The experimental results show that from gamification experiments something can be gleaned about how humans take to uncertainty searching for a goal and most humans are Pareto compliant under any uncertainty representation. Still there is a sizable minority whose behaviour does not follow Paretian expected utility theory and would beget a different approach.

Research in Machine Learning has focused mostly on how to assess the informational utility of possible queries, we rather addressed the issue of the perception and quantification of probabilistic uncertainty itself. Different uncertainty quantification measures lead to different classifications of humans' decisions. The z uncertainty quantification measure makes humans "more rational". These results agree with very recent results as to embed non-Euclidean distance in the learning process and offer new insights into the design of algorithms of machine learning and optimization. This problem is still an open question in Machine Learning and cognitive sciences and neither our results nor those prevailing in the rich literature about this issue provide unequivocal evidence about the algorithms underpinned by humans' decisions.

Acknowledgements We greatly acknowledge the DEMS Data Science Lab, Department of Economics Management and Statistics (DEMS), for supporting this work by providing computational resources.

Appendix The ten global optimization test functions used in this study, including their analytical formulations, search spaces and information about optimums and optimizers, can be found at the following link: https://www.sfu.ca/~ssurjano/optimization.html

Since they are minimization test functions, we have returned $-f(x)$ as score in order to translate them into the maximization problems depicted in the following.



Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

Data availability Data are those already used in [9] and available at the following link: https://github.com/acandelieri/humans_strategies_analysis.

Declarations

Conflict of interest Authors declare that they do not have any conflicts of interests or competing interests.

Ethics approval Informed consent was given in accordance with the university's procedure and the Helsinki declaration.

References

1. Wilson, R.C., Bonawitz, E., Costa, V.D., Ebitz, R.B.: Balancing exploration and exploitation with information and randomization. Curr. Opin. Behav. Sci. **38**, 49–56 (2020)
2. Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., Cohen, J.D.: Humans use directed and random exploration to solve the explore–exploit dilemma. J. Exp. Psychol. Gen. **143**(6), 2074 (2014)
3. Gershman, S.J.: Deconstructing the human algorithms for exploration. Cognition. **173**, 34–42 (2018)
4. Schulz, E., Gershman, S.J.: The algorithmic architecture of exploration in the human brain. Curr. Opin. Neurobiol. **55**, 7–14 (2019)
5. Schulz, E., Tenenbaum, J.B., Reshef, D.N., Speekenbrink, M., Gershman, S.: Assessing the Perceived Predictability of Functions. In: CogSci, vol. 6 (2015, November)
6. Archetti, F., Candelieri, A.: Bayesian Optimization and Data Science. Springer International Publishing (2019)
7. Frazier, P.I.: Bayesian optimization. In: Recent Advances in Optimization and Modeling of Contemporary Problems, pp. 255–278. INFORMS (2018)
8. Borji, A., Itti, L.: Bayesian Optimization Explains Human Active Search. Adv. Neural Inform. Process. Syst. **26**, 55–63 (2013)
9. Candelieri, A., Perego, R., Giordani, I., Ponti, A., Archetti, F.: Modelling human active search in optimizing black-box functions. Soft. Comput. **24**, 17771–17785 (2020). https://doi.org/10.1007/s00500-020-05398-2
10. Griffiths, T.L., Kemp, C., Tenenbaum, J.B.: Bayesian models of cognition. In: Sun, R. (ed.) Cambridge Handbook of Computational Cognitive Modelling. Cambridge University Press, Cambridge (2008)
11. Kruschke, J.K.: Bayesian approaches to associative learning: from passive to active learning. Learn. Behav. **36**(3), 210–226 (2008)
12. Wilson, A.G., Dann, C., Lucas, C., Xing, E.P.: The human kernel. Adv. Neural Inform. Process. Syst. **28**, 2854-2862 (2015)
13. Gershman, S.J.: Uncertainty and exploration. Decision. **6**(3), 277 (2019)
14. Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer Science & Business Media (2012)
15. Monge, G.: Mémoire sur la théorie des déblais et des remblais. Mem. Math. Phys. Acad. Royale Sci. 666–704 (1781)
16. Kantorovich, L.: On the Transfer of Masses (in Russian). In: Doklady Akademii Nauk. pp. 227–229 (1942)
17. Villani, C.: Optimal Transport: Old and New, vol. 338, p. 23. Springer, Berlin (2009)

18. Solomon, J., Rustamov, R., Guibas, L., & Butscher, A.: Wasserstein propagation for semi-supervised learning. In: International Conference on Machine Learning, pp. 306–314. PMLR (2014)
19. Peyré, G., Cuturi, M.: Computational optimal transport: with applica-tions to data science. Foundations and trends®. Mach. Learn. **11**(5–6), 355–607 (2019)
20. Applegate, D., Dasu, T., Krishnan, S., Urbanek, S.: Unsupervised clustering of multidimensional distributions using earth mover distance. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 636–644. (2011, August)
21. Cabanes, G., Bennani, Y., Verde, R., Irpino, A.: On the use of Wasserstein metric in topological clustering of distributional data. arXiv preprint arXiv:2109.04301 (2021)
22. Puccetti, G., Rüschendorf, L., Vanduffel, S.: On the computation of Wasserstein barycenters. J. Multivar. Anal. **176**, 104581 (2020)
23. Ye, J., Wu, P., Wang, J.Z., Li, J.: Fast discrete distribution clustering using Wasserstein barycenter with sparse support. IEEE Trans. Signal Process. **65**(9), 2317–2332 (2017)
24. Verdinelli, I., Wasserman, L.: Hybrid Wasserstein distance and fast distribution clustering. Electron. J. Stat. **13**(2), 5088–5119 (2019)
25. Cohen, J.D., McClure, S.M., Yu, A.J.: Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Phil. Trans. R. Soc. B Biol. Sci. **362**(1481), 933–942 (2007)
26. Gershman, S.J., Uchida, N.: Believing in dopamine. Nat. Rev. Neurosci. **20**(11), 703–714 (2019)
27. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2), 235–256 (2002)
28. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE Trans. Inf. Theory. **58**(5), 3250–3265 (2012)
29. Žilinskas, A., Calvin, J.: Bi-objective decision making in global optimization based on statistical models. J. Glob. Optim. **74**(4), 599–609 (2019)
30. De Ath, G., Everson, R.M., Rahat, A.A., Fieldsend, J.E.: Greed is good: exploration and exploitation trade-offs in Bayesian optimisation. arXiv preprint arXiv:1911.12809 (2019)
31. De Ath, G., Everson, R.M., Fieldsend, J.E., Rahat, A.A.: ϵ-shotgun: ϵ-greedy batch bayesian optimisation. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference, pp. 787–795. (2020)
32. Sandholtz, N. Modeling Human Decision-Making in Spatial and Temporal Systems (Doctoral Dissertation, Science: Department of Statistics and Actuarial Science) (2020)
33. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2011)
34. Tversky, A., Kahneman, D.: Rational Choice and the Framing of Decisions. In Multiple Criteria Decision Making and Risk Analysis Using Microcomputers, pp. 81–126. Springer, Berlin, Heidelberg (1989)
35. Kourouxous, T., Bauer, T.: Violations of dominance in decision-making. Bus. Res. **12**(1), 209–239 (2019)
36. Peters, O.: The ergodicity problem in economics. Nat. Phys. **15**(12), 1216–1221 (2019)
37. Williams, C. K., & Rasmussen, C. E. (2006). Gaussian Processes for Machine Learning (Vol. 2, No. 3, p. 4). Cambridge: MIT Press
38. Gramacy, R. B.: Surrogates: Gaussian process modeling, design, and optimization for the applied sciences. Chapman and Hall/CRC (2020)
39. Močkus, J. On Bayesian Methods for Seeking the Extremum. In: Optimization Techniques IFIP Technical Conference (pp. 400–404). Springer, Berlin, Heidelberg (1975)
40. Bemporad, A.: Global optimization via inverse distance weighting and radial basis functions. Comput. Optim. Appl. **77**(2), 571–595 (2020)
41. Candelieri, A., Ponti, A., Archetti, F.: Uncertainty quantification and exploration–exploitation trade-off in humans. J. Ambient. Intell. Humaniz. Comput. 1–34 (2021)
42. Bonneel, N., Peyré, G., Cuturi, M.: Wasserstein barycentric coordinates: histogram regression using optimal transport. ACM Trans. Graph. **35**(4), 71–71 (2016)
43. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K. From Word Embeddings to Document Distances. In: International conference on machine learning (pp. 957–966). PMLR (2015, June)
44. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning, pp. 214–223. PMLR (2017, July)
45. Kandasamy, K., Neiswanger, W., Schneider, J., Poczos, B., Xing, E.: Neural architecture search with bayesian optimisation and optimal transport. arXiv preprint arXiv:1802.07191 (2018)
46. Bachoc, F.: Advances in Gaussian Process. (2019)
47. De Plaen, H., Fanuel, M., Suykens, J.A.: Wasserstein Exponential Kernels. In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2020, July)

48. Le, T., Yamada, M., Fukumizu, K., Cuturi, M.: Tree-sliced variants of wasserstein distances. arXiv preprint arXiv:1902.00342 (2019)
49. Oh, J.H., Pouryahya, M., Iyer, A., Apte, A.P., Tannenbaum, A., Deasy, J.O.: Kernel wasserstein distance. arXiv preprint arXiv:1905.09314 (2019)
50. Bachoc, F., Gamboa, F., Loubes, J.M., Venet, N.: A Gaussian process regression model for distribution inputs. IEEE Trans. Inf. Theory. **64**(10), 6620–6637 (2017)
51. Bachoc, F., Suvorikova, A., Ginsbourger, D., Loubes, J.M., Spokoiny, V.: Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding. Electron. J. Stat. **14**(2), 2742–2772 (2020)
52. Mallasto, A., Gerolin, A., Minh, H.Q.: Entropy-regularized 2-Wasserstein distance between Gaussian measures. Inf. Geom. 1–35 (2021)
53. Balcan, M.F., Blum, A., Srebro, N.: A theory of learning with similarity functions. Mach. Learn. **72**(1), 89–112 (2008)
54. Rakotomamonjy, A., Traoré, A., Berar, M., Flamary, R., Courty, N.: Distance measure machines. arXiv preprint arXiv:1803.00250 (2018)
55. Ponti, A., Candelieri, A., Archetti, F.: A new evolutionary approach to optimal sensor placement in water distribution networks. Water. **13**(12), 1625 (2021a)
56. Ponti, A., Candelieri, A., Archetti, F.: A Wasserstein distance based multiobjective evolutionary algorithm for the risk aware optimization of sensor placement. Intell. Syst. Appl. **10**, 200047 (2021b)