Information geometry of influence diagrams and noncooperative games

Nils Bertschinger,1 David H. Wolpert,2 Eckehard Olbrich,1 and Jürgen Jost1,2

1Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
2Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
(Dated: January 3, 2014)

Abstract

What is the “value of information” in non-cooperative games with imperfect information? To answer this question, we propose to quantify information using concepts from Shannon’s information theory. We then relate quantitative changes to the information structure of a game to changes in the expected utility of the players. Our approach is based on the Multi-Agent Influence Diagram representation of games. We develop a generalization of the concept of marginal utility in decision scenarios to apply to infinitesimal changes of the channel parameters in noncooperative games. Using that framework we derive general conditions for negative value of information, and show that generically, these conditions hold in all games unless one imposes a priori constraints on the allowed changes to information channels. In other words, in any game in which a player values some aspect of the game’s specification beyond the information provided in that game, there will be an infinitesimal change to the parameter vector specifying the game that increases the information but hurts the player. Furthermore, we derive analogous results for $N > 1$ players, i.e., state general conditions for negative value of information simultaneously for all players. We demonstrate these results numerically on a decision problem as well as a leader-follower game and discuss their general implications.
I. INTRODUCTION

How a single agent (human, firm, animal, etc.) behaves typically depends on what information it has about its environment, and on its preferences. Accordingly, the joint behavior of multiple interacting agents can depend strongly on the information available to the separate agents, both about one another, and about external random variables. Precisely how the joint behavior depends on the information available to the agents is determined by the preferences of those agents. So in general there is a strong interplay among the preferences of all the agents, their behavior, and the information structure connecting them. This paper presents a novel approach based on information theory to study this interplay.

A. Previous work

In a seminal paper, Blackwell [6] has studied the case of a single player who has to take a decision based on a noisy signal about a relevant state of nature. He proved that a decision maker never prefers an increase in the noise of the information channel and formalized a corresponding partial order on the space of information channels meanwhile known as “garbling” channels. In game theory, i.e. multi-agent decision situations, such general results are largely missing. Instead, many researchers have illustrated that the players might prefer more or less information depending on the particular structure of the game, see [18] for an early example. The studied examples were showing that Blackwell’s result cannot easily be generalized to situations of strategic interactions. Interestingly, Blackwell has understood an information channel as a conditional probability distribution as is common in information theory. Game theory has mainly used the idea of an information partition which specifies which states are indistinguishable to an agent. In this case, more (less) information are formalized as refining (coarsening) an agent’s information partition.

Only recently, modern game theory makes a distinction between the “basic game” and the “information structure”: The basic game captures the available actions and the payoffs and the probability distribution over the states of nature, while the information structure specifies what the players belief about the game, the state of nature and each other (see for instance [5, 16]). More formally this is expressed in games of incomplete information having each player observing a signal, drawn from a conditional probability distribution, about the state of nature. In principle these signals are correlated. The effects of changes in the information structure were studied by
considering certain types of garbling nodes as by Blackwell. Lehrer et al. [16] showed that if two information structures are equivalent with respect to a specific garbling the game will have the same equilibrium outcomes. Thus, they characterized the class of changes to the information channels that leave the players indifferent with respect to a particular solution concept. Similarly, Bergemann and Morris [5] introduced a Blackwell like order on information structures called individual sufficiency that provides a notion of more and less informative structures in the sense that more information always shrinks the set of Bayes correlated equilibria. This result is in line with Gossner’s work [11] relating more knowledge of the players to an increase of their abilities, i.e. the set of possible actions available to them. Thus, more information can be seen to increase the number of constraints on a possible solution for the game.

A similar approach has been used in the work of influence diagrams which provide a graphical way to specify the information structure of a decision situation. Every random variable, denoting states of nature and moves of the players under a certain strategy respectively, is represented as a node with edges encoding conditional dependencies between them. Howard [12] defined the “value of information” of a particular parent node of a decision node as the difference between the maximum expected utility and the maximum expected utility when the edge from that parent to the decision node is removed.

B. Differential value of information

In general, when going from (single player) influence diagrams to a multi-agent influence diagram (MAID), Howard’s definition cannot be taken over directly. This is because now there are multiple equilibria, and even the number of equilibria may change when the edge is removed. So there is arbitrariness in which pairs of expected utility values from before and after the presence of the edge are compared.

One solution would be to fix the mixed strategies of all players other than player $i$ to their values at one particular equilibrium, i.e., replace them with states of Nature set to their form at that equilibrium. Player $i$ would then be faced with a conventional decision problem, and so the change in their maximal expected utility when an edge is removed would be well-defined. Replacing the other players with states of Nature is a drastic solution however, removing much of the strategic nature of the issue.

Another approach starts by noting that removing an edge is equivalent to replacing that edge
with a new one that is complete noise. This suggests a way to circumvent the problem of identifying an equilibrium from those before removal of an edge with an equilibrium from those after removal: Instead add an infinitesimal amount of noise to that edge, thereby moving a particular before-removal equilibrium an infinitesimal distance along a path that would ultimately end with complete noise, i.e., that would end a particular equilibrium from after removal of the edge. This replaces the notion of “value of information” of an edge with “differential value of information” of adding noise to that edge. Since it takes a vector to specify the kind of infinitesimal also, this also means that rather than a single scalar “value of information”, we need to work with a vector, of infinitesimal change in expected utility if we change information in a channel by moving in a particular direction in the space of channel parameters.

This approach of making infinitesimal changes to conditional distributions at chance nodes and examining the ramifications on expected utility is the approach we propose in this paper. For example, Howard’s definition of value of information only quantifies the value of an edge leading directly into a decision node. It cannot be applied to other edges, leading from one chance node to another. This is because how the conditional distribution at one chance node changes when an edge leading into that chance node is removed is undefined. However we we can add infinitesimal noise to any edge, and examine the ramifications on expected utility of a player. That means in particular we can add infinitesimal noise to an edge going from one chance node to another, and examine the resultant change in the expected utility of a player.

Ultimately, the core idea of this approach is to analyze the ramifications of infinitesimal changes to the parameter vector specifying the game. This can be applied to other kinds of infinitesimal changes besides those involving noise vectors in communication channels. For example, it can be applied to a change to the rationality exponent of a player at a quantal response equilibrium. As another example, it can be applied to a change in a utility function of a player. Indeed, we can even gradually change other characteristics of a game like the number of stages in the game. This flexibility allows us to extend Howard’s idea of “value of information” far beyond the scenarios he had in mind.

C. The need to use gradients and differential geometry to analyze value of information

Consider a player $i$ in a multi-stage noncooperative game who receives noisy signals at some stages of the game. Say we have also specified an equilibrium concept and associated refinement
So we have specified an equilibrium profile of player mixed strategies in the game, and therefore a joint probability distribution over all variables in the game.

Given this specification of a joint distribution, it is relatively clear how to quantify the value to player \( i \) of receiving one particular signal versus receiving another at a particular stage of the game; compare their expected utilities conditioned on those two signals under the specified strategy profile. But how should we quantify the value to the player of a change in the information available during the game? How much would they value having some information channel in the game made more informative? Note that answering this question requires answering how much they would value a change in the game specification itself.

To provide a quantitative answer to this question, first consider the simple case of a consumer whose preference function depends jointly on the quantities of all the goods they get in a market. Given some current bundle of goods, how much would we say that they value getting infinitesimally more of good \( j \)? The usual economic answer is that it is the marginal utility of good \( j \), i.e. derivative of their expected utility with respect to amount of good \( j \).

Note that the amount of good \( j \) is a scalar-valued function of the vector \( v \) of quantities of all goods — it is just the \( j \)’th component of that vector. However in general the consumer might be asked to value an infinitesimal change to some other scalar-valued function of \( v \). For example, they might be asked how much they value not infinitesimally more of good \( j \), but rather infinitesimally more of some linear combination of \( j \) and a different good \( j' \). More elaborately, they could be asked how much value they would assign to infinitesimally more of any real-valued function of \( v \), even nonlinear ones.

We can use the simple example of quantifying the value of an amount of a single good to motivate a way to quantify value of this more general type of function of \( v \). Write the player’s expected utility (at a particular equilibrium) as \( U(v) \) and the amount of good \( j \) as the function \( g(v) \equiv v_j \). Note that how much they value good \( j \) is the directional derivative of their expected utility in the direction of maximal gain in the amount of good \( j \), i.e. the projection of gradient \( U(v) \) on the gradient of \( g(v) \). So to quantify how much player \( i \) values an infinitesimal change to the value of some other scalar-valued function \( f(v) \), it is natural to use the projection of the gradient \( U \) on the direction of gradient \( f \).

Loosely speaking, this projection is how much expected utility would change if the value of \( f \) were changed infinitesimally, but to first order no other degree of freedom aside from the value of \( f \) were changed. More formally, it is given by the dot product between the two gradients of
$U$ and $f$ where the gradient of $f$ is normalized to unit length. We have to be a bit more careful than this though, due to unit considerations. To be consistent with conventional terminology, we would like to define how much a player values an infinitesimal change to $f$ expressed per unit of $f$. Indeed, we would typically say that how much player $i$ values a change to good $j$ is given by the associated change in utility divided by the amount of change in good $j$. (After all, that tells us change in utility per unit change in the good.) Based on this reasoning, we propose to measure the value of an infinitesimal change of $f$ as

$$\frac{\langle \nabla U, \nabla f \rangle}{||\nabla f||^2}$$

This says that if a small change in the value of $f$ leads to a big change in expected utility, $f$ is more valuable than if the same change in expected utility required a bigger change in the value of $f$. However while this quantity may be accord with common usage, it also has the disadvantage that it may be infinite, depending on the current $v$ and the form of $f$. Thus, it might also be useful to just study the dot product between the gradient of expected utility and the gradient of $f$.

Whichever of these expressions involving dot products of gradients we choose to use to quantify the “differential value” to the player of a change in the information of the game, we must still specify what information, i.e. of which signal and about what other quantity, we are interested in. In this paper we focus on various measures that are provided by Shannon’s information theory, for reasons described below. These include in particular the mutual informations among various pairs of random variables, various conditional entropies, and various information capacities, and more generally quantities like transfer entropy, multi-information, etc. (which we defer to later work). Each of these measures of information is a scalar-valued function of the equilibrium profile (joint probability distribution) over player and Nature moves, and thus of the parameters specifying the game. Given the foregoing discussion, this leads to our overall approach: Identify the game parameters with $v$, and identify the information of interest in the game equilibrium induced by a particular game $v$ as a scalar-valued function $f(v)$ of the equilibrium profile that is implicitly specified by $v$. (Shannon information theory provides multiple choices for $f$, depending on precisely what type of information one wishes to measure.) Then use the projection of gradient $f(v)$ on the gradient of expected utility of some player $i$ of the game to quantify the value of that type of information to player $i$. We can also use the dot products for more than one type of information measure, to assign relative values of different types of information, e.g., concerning different random variables.
We have to make several choices every time we use this approach. First we must decide what Shannon information theory measures we wish to analyze, e.g. the capacity of a certain information channel, the information between moves of the players to investigate how they value coordination, etc. Second we must choose what parameters of the game to vary. And finally we must choose what precise function of the dot products of gradients to use, e.g. whether to consider normalized ratios or not.

Taken together these fix what economic question we are considering. However whatever such choices we make, we immediately confront an additional question, that does not is of economical relevance, and that we do not want to affect our answer. This is the question of what coordinate system to use to evaluate dot products. The difficulty is that changing the coordinate system changes the values of dot products in general. So different choices of coordinate system would give different values of information. This is quite undesirable; we don’t want our analysis to change if we change how we parametrize the noise level in a communication channel, for example.

Formally, to avoid this issue we must use inner products, defined in terms of a metric tensor, rather than dot products. Calculations of inner products are guaranteed to be “covariant”, not changing as we change our choice of coordinate system. The associated metric tensor also tells us how to measure distance, i.e., how to normalize basis vectors. So we see that very elementary considerations force us to to use tensor calculus with an associated metric to analyze the value of information in games.

For many economic questions, there is no clearly preferred distance measure, and no clearly preferred way of defining inner products. For such questions, the precise metric tensor we use should not matter, only the fact that we use some metric tensor. The analysis bears this out. In particular, the existence / impossibility theorems we prove below do not depend on which metric tensor we use, only that we use one.

However for other purposes the choice of tensor is important. For example, this is true when we evaluating precise values of information, plotting vector fields of gradients of mutual information, etc. When we need to specify a precise choice of such a metric, for reasons discussed below in

---

1 To give a simple example, consider the gradient of the function from $\mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = x^2 + y^2$ in Cartesian coordinates. The vector of partial derivatives of $f$ in Cartesian coordinates is the (Cartesian) vector $2(x,y)$. However if we express $f$ in polar coordinates, and evaluate the vector of partial derivatives with respect to those coordinates, we get $(2r,0)$, which when transformed back to Cartesian coordinates is the vector $(2x,0)$. So the gradients change, as claimed.

To give a simple example that also the dot product changes, consider the two Cartesian vectors $(1,0)$ and $(0,1)$. Their dot product in Cartesian coordinates equals 0. However if we translate those two vectors into polar coordinates...
we restrict attention to the Fisher information metric \cite{1, 8}, since it is based on Shannon information theory. Similar analysis could be done using other choices.

D. Roadmap

In Sec. \[\text{II}\] we motivate our formal framework and review basic information theory as well as information geometry. Then, in Sec. \[\text{III}\] we introduce multi-agent influence diagrams and explain why they are especially suited to study information in games. Next, we introduce quantal response equilibria and show how to calculate partial derivatives of the resulting equilibrium strategies. Based on these, we define the differential value of information and prove general conditions for the existence of negative value of information. We also discuss, how to identify parameter settings such that all players prefer less information. Finally, in Sec. \[\text{V}\] we illustrate our proposed definitions and results in a simple decision situation as well as a two-player leader follower game.

II. FORMAL FRAMEWORK: INFORMATION GEOMETRY OF GAMES

A. Motivation of formal framework

Bayes nets \cite{14} provide a very concise, powerful way to model scenarios where there are multiple interacting Nature players (either automata or inanimate natural phenomena), but no human players. They do this by representing the information structure of the scenario in terms of a Directed Acyclic Graph (DAG) with conditional probability distributions at the nodes of the graph. In particular, the use of conditional distributions rather than information set partitions greatly facilitates the analysis and associated computation of the role of information in such systems. As a result they have become very wide-spread in machine learning and information theory in particular, and in computer science and the physical sciences more generally.

Influence Diagrams (IDs \cite{12}) were introduced to extend Bayes nets to model scenarios where there is a (single) human player interacting with Nature players. There has been much analysis of how to exploit the graphical structure of the ID to speed up computation of the optimal behavior assuming full rationality. More recently, Multi-Agent Influence Diagrams (MAIDs \cite{15}) and their variants like Interactive POMDP’s \cite{9} have extended IDs to model games involving arbitrary numbers of players. As such, the work on MAIDs can be viewed as an attempt to create a new game
theory representation of multi-stage games based on Bayes nets, in addition to strategic form and extensive form representations.

Whether we use MAIDs or extensive form representations, we need a way to quantify information. Game theory has traditionally studied information in multi-stage games using the extensive form representation and associated information partitions. This allows the analysis of some aspects of how changes to a game specification affect information within the game. In particular, “less information” to a player is often taken to mean coarsening their information partition or inserting garbling nodes, as in Blackwell’s results concerning play against Nature [6].

However these conventional approaches only provide a partial order over the space of the conditional distributions at the chance nodes in a MAID. And even if one restricts attention to the distributions lying in a chain of that partial order, these conventional approaches do not quantify amounts of information numerically; they only allow us to investigate information from an ordinal rather than cardinal perspective.

Information theory [8, 19] is an alternative approach to quantifying and analyzing information. As discussed below, it has some very compelling first-principles justifications. It also has many practical advantages for our purposes. In contrast to an approach based on coarsening of information partitions and/or insertion of garbling nodes, it allows us to assign quantitative values to amounts of information, to channel capacities, etc.

Moreover, as described in Sec. [11] using Shannon’s information theory provides a natural choice for what metric tensor to use with our approach, namely the Fisher information metric. (Differential geometry for this choice of metric is called “information geometry” [11].) In addition, the tools of information theory provide a rich set of ways of analyzing stochastic systems. For example, the equilibrium value of the mutual information between the value at a decision node of a player in a MAID and the value of one of the parents of that decision node is a measure of how much that player depends on or coordinates with that parent node at that equilibrium.

Combining all these elementary considerations leads us to adopt MAIDs as our game form representation, Shannon’s information theory to quantify information, and therefore the Fisher information metric to specify the differential geometry we use to evaluate inner products between gradients. In this way elementary considerations lead us to analyze the information geometry of noncooperative games in order to understand the value of information in games.
B. Information theory

We will use notation that is a combination of standard game theory notation [10] and standard Bayes net notation [14]. (See [15] for a good review of Bayes nets for game theoreticians.)

The probability simplex over a space $X$ is written as $\Delta_X$. $\Delta_{X|Y}$ is the space of all possible conditional distributions of $x \in X$ conditioned on a value $y \in Y$. For ease of exposition, this notation is adopted even if $X \cap Y \neq \emptyset$. We use uppercase letters to indicate a random variable or its domain, with the context making the choice clear. We use lowercase letters to indicate a particular element of the associated random variable’s range, i.e., a particular value of that random variable. In particular, $p(X) \in \Delta_X$ always means an entire probability distribution vector over all $x \in X$, whereas $p(x)$ will typically refer instead to the value of $p(.)$ at the particular argument $x$. Here, we couch the discussion in terms of countable spaces, but much of the discussion carries over to the uncountable case.

We start by presenting some basic information theory. For any random variable $X$ taking values $x \in X$ with probability $p(x)$, the Shannon entropy of that variable is

$$H(X) \triangleq -\sum_{x \in X} p(x) \ln p(x)$$

[8, 19, 20] The entropy of a bounded random variable is largest when the associated distribution is uniform, and it equals 0 if it is concentrated on a single value of $X$. So entropy measures “how spread out” a distribution is, which corresponds to the uncertainty of the outcome of a measurement of $X$.

Similarly, if we also have a random variable $Y$ taking values $y \in Y$, then the conditional entropy of $X$ given $Y$ is

$$H(X \mid Y) \triangleq -\sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y) \ln p(x \mid y)$$

$$= -\sum_{x,y} p(x,y) \ln p(x \mid y)$$

(1)

(Unless explicitly stated otherwise, all logarithms in entropies are assumed to be base 2.) This is a measure of how much knowing $y$ tells you about what $x$ is, averaged according to the probability of the various $y$’s. More precisely, $H(X \mid Y)$ can be viewed as the average uncertainty about $X$ that

---

2 If $x$ is uncountable an integral replaces the sum and a base measure $\mu$ is introduced into the $\ln$. See [8, 19] for details.
remains after $Y$ has been observed. Note that

$$H(X \mid Y) = H(X, Y) - H(Y)$$

where $H(X, Y) = -\sum_{x,y} p(x, y) \ln p(x, y)$ quantifies the total uncertainty of $X$ and $Y$ when taken together.

The **mutual information** between $X$ and $Y$ is a measure of the “dependency” between the random variables $X$ and $Y$. It is defined as

$$I(X; Y) = H(X) - H(X \mid Y)$$

$$= H(X) + H(Y) - H(X, Y)$$

$$= I(Y; X).$$

Considering the conditional distribution $p(y|x)$ as an information channel with input $X$ and output $Y$, the mutual information between $X$ and $Y$ quantifies how much the uncertainty of the channel input $x$ is reduced when the channel output $y$ can be observed. Due to its symmetry, that mutual information also tells you how extra much knowing the channel input $x$ tells you about the channel output $y$.

The mutual information $I(X; Y)$ vanishes if and only if $X$ and $Y$ are independent. Intuitively, mutual information can be thought of as a nonlinear generalization of the correlation in the case of linear relationships between the data. As an example, let $X$ and $Y$ be real-valued random variables, where $p(x, y) \propto \delta(x^2 + y^2 - 1)$ where $\delta(.)$ is the Dirac delta function. Then the correlation of those two variables is zero, even though knowing the value of $x$ tells you a huge amount concerning the value of $y$ (and vice-versa). In contrast, their mutual information is very large, reflecting the tight relation of the two random variables.

Moreover, the mutual information of two variables is perfectly well-defined even if the variables are categorical rather than numeric or at least ordinal. For these reasons, mutual information is a natural choice for how to measure how closely two random variables are related.

Both, entropy and mutual information are key concepts in the modern telecommunication theory. Whereas entropy lower bounds the length of an optimal code that is used to compress an information source, mutual information is used to define the **information capacity** of a channel.

---

\[3\text{ Strictly speaking, to make the definition of mutual information for this scenario precise, one has to choose a measure over } \mathbb{R}^2. \text{ Or more simply, one could simply restrict the values of } x \text{ and } y \text{ to lie in a fine-grained discretization of } \mathbb{R}^2, \text{ so that the definition of mutual information involves sums rather than integrals.}\]
The information capacity is the maximal information that can be transferred from the input \( x \) to the output \( y \) of a channel \( p(y \mid x) \). Formally, it is obtained by maximizing the mutual information \( I(X; Y) \) over all possible input distributions \( p(x) \), i.e. \( \sup_{p(x)} I(X; Y) \). (Note that once \( p(x) \) is specified, the joint distribution \( p(x,y) \), which gives the mutual information, is fixed by the channel’s distribution \( p(y \mid x) \).)

In general, solving the required maximization cannot be done analytically, and correspondingly closed formulas for the channel capacity are only known for special cases. This means, that also the partial derivatives of the channel capacity with respect to the channel parameters are difficult to calculate. In the examples in this paper, we will use a binary (asymmetric) channel the capacity of which is known \([2]\). Another important class of information channels with known capacity are the so called symmetric channels \([8]\). In this case, the noise is symmetric in the sense that it does not depend on a particular input, i.e. the channel is invariant under relabeling of the inputs. This class is rather common in practice and excludes channels with continuous input, e.g. the Gaussian channel.

C. Information geometry

Consider a (joint) distribution \( p(x) \) that is parametrized with \( d \) parameters \( \theta = \theta^1, \ldots, \theta^d \) living in a \( d \)-dimensional differentiable manifold \( \Theta \), i.e. \( p(x; \theta) \). Here, we use the convention of differential geometry to denote components of contra-variant vectors by upper and components of co-variant vectors by lower indices (see appendix VII for details). In general, expected utilities and information quantities depend on the parameters, either directly such as the capacity of a channel with certain noise parameters or indirectly by the players adjusting their strategies to changes in the channel noise. Here, we assume that all functions of interest are differentiable in the interior of \( \Theta \).

The above assumptions allow us to evaluate the partial derivatives \( \frac{\partial}{\partial \theta^i} \) with respect to the chosen parameters. In order to obtain results, that are independent of the chosen parametrization, we need a metric on the space of information channels. A suitable choice is the Fisher information metric, which is given by

\[
g_{kl}(\theta) = \sum_x p(x; \theta) \frac{\partial \log p(x; \theta)}{\partial \theta^k} \frac{\partial \log p(x; \theta)}{\partial \theta^l} \tag{3}
\]

where \( p(x; \theta) \) is a probability distribution parametrized by \( \theta \).

The statistical origin of the Fisher metric lies in the task of estimating a probability distribution
from a family parametrized by $\theta$ from observations of the variable $x$. The Fisher metric expresses the sensitivity of the dependence of the family on $\theta$, that is, how well observations of $x$ can discriminate among nearby values of $\theta$.

With this metric, and using the Einstein summation convention (see appendix VII again), we can form the scalar product of two tangent vectors $v = v^i \frac{\partial}{\partial \theta_i}, w = w^j \frac{\partial}{\partial \theta_j}$ as

$$\langle v, w \rangle_g = g^{ij}v^i w^j \quad (4)$$

The norm of a vector $v$ is then given as $\|v\|_g = \langle v, v \rangle_g^{\frac{1}{2}}$.

The gradient of any functional $f : \Delta_x(\theta) \rightarrow \mathbb{R}$ can then be obtained from the partial derivatives as follows:

$$\text{grad}(f)^i = g^{ij} \frac{\partial f}{\partial \theta_j}$$

where $g^{ij}$ denotes the inverse of the metric $g_{ij}$ and we have again used Einstein summation for the index $j$. Thus, the gradient is a contra-variant vector with entries $(\text{grad}(f)^1, \ldots, \text{grad}(f)^d)$.

As a motivating example, consider a binary asymmetric channel $p(y|x; \theta)$ with input distribution

$$p(x) = \begin{cases} 
q & \text{if } x = 0 \\
1 - q & \text{if } x = 1
\end{cases}$$

and parameters $\theta = (\epsilon^1, \epsilon^2)$ for transmission errors

$$p(y|x; \theta) = \begin{cases} 
1 - \epsilon^1 & \text{if } x = 0, y = 0 \\
\epsilon^1 & \text{if } x = 0, y = 1 \\
\epsilon^2 & \text{if } x = 1, y = 0 \\
1 - \epsilon^2 & \text{if } x = 1, y = 1
\end{cases} \quad (5)$$

Then, the Fisher information metric of $p(x,y; \epsilon^1, \epsilon^2)$ is a diagonal $2 \times 2$ matrix with entries

$$g(\epsilon^1, \epsilon^2) = \begin{pmatrix}
\frac{q}{\epsilon^1(1-\epsilon^1)} & 0 \\
0 & \frac{1-q}{\epsilon^2(1-\epsilon^2)}
\end{pmatrix}$$

III. MULTI-AGENT INFLUENCE DIAGRAMS

Games are commonly specified in normal- or extensive form. Multi-agent influence diagrams (MAIDs) \cite{15} are an alternative representation which most clearly expresses the interaction structure, i.e., which information is available to each player in a certain state. MAIDs are a generalization of Bayesian networks which are widely used in machine learning and information theory. Thus, they are a natural starting point when studying the role of information in games.
Definition 1. An $n$-player **MAID** is defined as a tuple $(G, \{X_v\}, \{p(x_v | x_{pa(v)})\}, \{u_i\})$ of the following elements:

- A directed acyclic graph $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \mathcal{D} \cup \mathcal{N}$ is partitioned into
  - a set of nature or chance nodes $\mathcal{N}$ and
  - a set of decision nodes $\mathcal{D}$ which is further partitioned into $n$ sets of decision nodes $\mathcal{D}_i$, one for each player $i = 1, \ldots, n$,
- a set $X_v$ of states for each $v \in \mathcal{V}$,
- a conditional probability distribution $p(x_v | x_{pa(v)})$ for each nature node $v \in \mathcal{N}$, where $pa(v) = \{u : (u, v) \in \mathcal{E}\}$ denotes the parents of $v$ and $x_{pa(v)}$ is their joint state.
- a family of utility functions $\{u_i : \prod_{v \in \mathcal{V}} X_v \to \mathbb{R}\}_{i=1, \ldots, n}$.

A one-person MAID is called an **influence diagram** (ID [12]). (Historically, ID’s were first investigated decades before MAIDs, which is the reason for the terminology.)

In the following, the states $x_v \in X_v$ of a decision node $v \in \mathcal{D}$ will usually be called *actions* or *moves*, and sometimes will be denoted by $a_v \in X_v$. We adopt the convention that “$p(x_v | x_{pa(v)})$” means $p(x_v)$ if $v$ is a root node, so that $pa(v)$ is empty. We write elements of $X$ as $x$. We define $X_{\mathcal{A}} \equiv \prod_{v \in \mathcal{A}} X_v$ for any $\mathcal{A} \subseteq \mathcal{V}$, with elements of $X_{\mathcal{A}}$ written as $x_{\mathcal{A}}$. So in particular, $X_{\mathcal{D}} \equiv \prod_{v \in \mathcal{D}} X_v$, and $X_{\mathcal{N}} \equiv \prod_{v \in \mathcal{N}} X_v$, and we write elements of these sets as $x_{\mathcal{D}}$ (or $a_{\mathcal{D}}$) and $x_{\mathcal{N}}$, respectively.

We will sometimes write an $n$-player MAID as $(G, X, p, \{u_i\})$, with the decompositions of those variables and associations among them implicit. (So for example the decomposition of $G$ in terms of $\mathcal{E}$ and a set of nodes $\bigcup_{i=1, \ldots, n} \mathcal{D}_i \cup \mathcal{N}$ will sometimes be implicit.)

A **solution concept** is a map from any MAID $(\mathcal{P}, G, X, p, \{u_i\})$ to a set of conditional distributions $\{\sigma_i(x_v | x_{pa(v)}) : v \in \mathcal{D}_i, i = 1, \ldots, n\}$. We refer to the set of distributions $\{\sigma_i(x_v | x_{pa(v)}) : v \in \mathcal{D}_i\}$ for any particular player $i$ as that player’s *strategy*. We refer to the full set $\{\sigma_i(x_v | x_{pa(v)}) : v \in \mathcal{D}_i, i = 1, \ldots, n\}$ as the *strategy profile*. We sometimes write $\sigma_v$ for a $v \in \mathcal{D}_i$ to refer to one distribution in a player’s strategy and use $\sigma$ to refer to a strategy profile.

The intuition is that each player can set the conditional distribution at each of their decision nodes, but is not able to introduce arbitrary dependencies between actions at different decision
nodes. In the terminology of game theory, this is called the agent representation. The rule for how the set of all players jointly set the strategy profile is the solution concept.

In addition, we allow the solution concept to depend on parameters. Typically there will be one set of parameters associated with each player. When that is the case we sometimes write the strategy of each player \( i \) that is produced by the solution concept as \( \sigma_i(a_v \mid x_{pa(v)}; \beta) \) where \( \beta \) is the set of parameters that specify how \( \sigma_i \) was determined via the solution concept.

The combination of a MAID \( (G, X, p, \{u_i\}) \) and a solution concept specifies the conditional distributions at all the nodes of the DAG \( G \). Accordingly it specifies a joint probability distribution

\[
p(x_V) = \prod_{v \in N} p(x_v \mid x_{pa(v)}) \prod_{i=1}^n \prod_{v \in D_i} \sigma_i(a_v \mid x_{pa(v)})
\]

(6)

\[
= \prod_{v \in V} p(x_v \mid x_{pa(v)})
\]

(7)

where we abuse notation and denote \( \sigma_i(a_v \mid x_{pa(v)}) \) by \( p(x_v \mid x_{pa(v)}) \) whenever \( v \in D_i \).

In the usual way, once we have such a joint distribution over all variables, we have fully defined the joint distribution over \( X \) and therefore defined conditional probabilities of the states of one subset of the nodes in the MAID, \( A \), given the states of another subset of the nodes, \( B \):

\[
p(x_A \mid x_B) = \frac{p(x_A, x_B)}{p(x_B)} = \frac{\sum_{x_V \setminus (A \cup B)} p(x_{A \cup B}, x_V \setminus (A \cup B))}{\sum_{x_V \setminus B} p(x_B, x_V \setminus B)}
\]

(8)

Similarly the combination of a MAID and a solution concept fully defines the conditional value of a scalar-valued function of all variables in the MAID, given the values of some other variables in the MAID. In particular, the conditional expected utilities are given by

\[
\mathbb{E}(u_i \mid x_A) = \sum_{x_V \setminus A} p(x_V \setminus A \mid x_A)u_i(x_V \setminus A, x_A)
\]

(9)

We will sometimes use the term “information structure” to refer to the graph of a MAID and/or the conditional distributions at its Nature nodes. (Note that this is a slightly different use of the term from that used in extensive form games.) In order to study the effect of changes to the information structure of a MAID, we will assume that the probability distributions at the nature nodes are parametrized by a set of parameters \( \theta \), i.e., \( p_v(x_v \mid x_{pa(v)}; \theta) \). We are interested in how infinitesimal changes to \( \theta \) (and other parameters of the MAID like \( \beta \)) affect \( p(x_V) \), expected utilities, mutual information among nodes in the MAID, etc.
A. Quantal response equilibria of MAIDs

A solution concept for a game specifies how the actions of the players are chosen. A popular model for bounded rational players is the quantal response equilibrium (QRE). Under a QRE, each player $i$ chooses his actions at the decision node $v \in \mathcal{D}_i$ from a Boltzmann distribution over his move-conditional expected utilities:

$$
\sigma_i(a_v | x_{pa}(v)) = \frac{1}{Z_i(x_{pa}(v))} e^{\beta_i \mathbb{E}(u_i|a_v, x_{pa}(v))}
$$

for all $a_v \in X_v$ and $x_{pa}(v) \in \prod_{u \in Pa(v)} X_u$. In this expression $Z_i(x_{pa}(v)) = \sum_{a \in X_{pa}(v)} e^{\beta_i \mathbb{E}(u_i|a, x_{pa}(v))}$ is a normalization constant, $\mathbb{E}(u_i|a_v, x_{pa}(v))$ denotes the conditional expected utility as defined in eq. (9) and $\beta_i$ is a parameter specifying the “rationality” of player $i$.

As shorthand, we denote the expected utility of player $i$ at some equilibrium $\{\sigma_i\}_{i=1,...,n}$, $\mathbb{E}(\sigma_i)_{i=1,...,n}(u_i)$, by $\hat{V}_i$.

B. Partial derivatives of QREs of MAIDs with respect to game parameters

In general, there can be multiple equilibria, i.e. strategy profiles $\{\sigma_i\}_{i=1,...,n}$ that simultaneously solves eq. (10) for all players. In this case, we have to choose a particular equilibrium branch at which to calculate partial derivatives. Depending on the equilibrium branch chosen, not only the strategies of the players, but also their partial derivatives will be different. This in turn means, that players will value changes to the parameters of the game differently on different equilibrium branches. All calculations in this section are general and apply to any equilibrium branch. Thus, in the following we implicitly assume that we have chosen an equilibrium branch on which we want to investigate the value of information.

For computations involving the partial derivatives of the players strategies at a QRE (branch) it can help to explicitly introduce the normalization constants as an auxiliary variable. The QRE condition from eq. (10) is then replaced by the following conditions

$$
\sigma_i(a_v | x_{pa}(v); \beta_i, \theta) - \frac{e^{\beta_i \mathbb{E}(u_i|a_v, x_{pa}(v); \beta, \theta)}}{Z_i(x_{pa}(v); \beta_i, \theta)} = 0
$$

$$
Z_i(x_{pa}(v); \beta_i, \theta) - \sum_{a \in X_v} e^{\beta_i \mathbb{E}(u_i|a, x_{pa}(v); \beta, \theta)} = 0
$$

4 This interpretation is based on the observation that a player with $\beta = 0$ will choose her actions uniformly at random, whereas $\beta \to \infty$ will choose the action(s) with highest expected utility, i.e., corresponds to the rational action choice.
for all players $i$, decision nodes $v \in D_i$ and all states $a_v \in X_v, x_v \in \prod_{u \in Pa(v)} X_u$. (Here and throughout this section, subscripts on $\sigma, Z$, etc. should not be understood as specifications of coordinates as in the Einstein summation convention.)

Overall, this gives rise to a total of $M$ equations for $M$ unknown quantities $\sigma_i(a_v|x_{pa(v)}), Z_i(x_{pa(v)})$. Using a vector valued function $f$ we can abbreviate the above by the following equation:

$$f(\sigma_{\beta, \theta}, Z_{\beta, \theta}, \beta, \theta) = 0$$

(11)

where $\sigma_{\beta, \theta}$ is a vector of all strategies

$$\{\sigma_i(a_v \mid x_v; \beta, \theta) : i \in \mathcal{P}, v \in D_i, a_v \in X_v, x_v \in \prod_{u \in Pa(v)} X_u\},$$

$Z_{\beta, \theta}$ collects all normalization constants, and $\mathbf{0}$ is the $M$-dimensional vector of all 0’s. Note that in general, even once the distributions at all decision nodes have been fixed, the distributions at chance nodes affect the value of $\mathbb{E}(u_i \mid a_v, x_{pa(v)}; \beta, \theta)$. Therefore they affect the value of the function $f$. This is why $f$ can depend explicitly on $\theta$, as well as depend directly on $\beta$.

The (vector-valued) partial derivative of the position of the QRE in $(\sigma_\theta, Z_\theta)$ with respect to $\theta$ is then given by implicit differentiation of eq. (11):

$$\begin{bmatrix}
\frac{\partial \sigma_\theta}{\partial \theta} \\
\frac{\partial Z_\theta}{\partial \theta}
\end{bmatrix} = - \left[ \frac{\partial f}{\partial \sigma_\theta} \frac{\partial f}{\partial Z_\theta} \right]^{-1} \frac{\partial f}{\partial \theta}$$

(12)

where the dependence on $\beta$ is hidden for clarity, all partial derivatives are evaluated at the QRE, and we assume that the matrix $\left[ \frac{\partial f}{\partial \sigma_\theta} \frac{\partial f}{\partial Z_\theta} \right]$ is non-singular at the point $\theta$ at which we are evaluating the partial derivatives.

These equations give the partial derivatives of the mixed strategy profile. They apply to any MAID, and allow us to write the partial derivatives of other quantities of interest. In particular, the partial derivative of the expected utility of any player $i$ is

$$\frac{\partial \mathbb{E}(u_i)}{\partial \theta} = \sum_x u_i(x) \frac{\partial p(x; \theta)}{\partial \theta}$$

$$= \sum_x u_i(x) \frac{\partial \prod_v p(x_v \mid x_{pa(v)}; \theta)}{\partial \theta}$$

$$= \sum_x u_i(x) \sum_{v' \neq v} \frac{\partial p(x_v \mid x_{pa(v)}; \theta)}{\partial \theta} \prod_{v' \neq v} p(x_{v'} \mid x_{pa(v')}; \theta)$$

(13)
where each term \( \frac{\partial p(x_{v'} | x_{pa(v')}; \theta)}{\partial \theta} \) is given by the appropriate component of Eq. (12) if \( v' \) is a decision node. (For the other, chance nodes, \( \frac{\partial p(x_{v'} | x_{pa(v')}; \theta)}{\partial \theta} \) can be calculated directly).

Similarly, the partial derivatives of other functions of interest such as mutual informations between certain nodes of the MAID can be calculated. Instead of evaluating those derivatives and the additional ones needed for the Fisher metric by hand, we used automatic differentiation to obtain numerical results for certain parameter settings and equilibrium branches. Note that automatic differentiation is not a numerical approximation, e.g. via finite differences, but uses the chain rule to evaluate the derivative alongside the value of the function.

IV. INFORMATION GEOMETRY OF MAIDS

As explained above, to obtain results that are independent of a particular parametrization, we need to work with gradients instead of partial derivatives. Here, we consider gradients with respect to the Fisher metric on \( p(x_v; \theta, \beta) \). Thus, we take into account how sensitive the distribution on the MAID depends on the parameters of the game. Note that this includes the implicit effect on the equilibrium by the players reacting to changing parameters.

Throughout we assume that any such space of parameters of a game is considered under a coordinate system such that the associated metric is full rank\(^5\). Note that due to the multi-branch nature of QREs for multiple players, we are evaluating the gradient locally at a particular branch of the QRE surface.

Based on this geometric structure, we now define several ways to quantify the differential value of parameter changes in certain directions as well as the differential value of some function \( f \). Furthermore, we state general results (that are independent of the metric) about negative values and illustrate the possible results with several examples.

A. Differential value

In general, even if we fix all distributions at nature nodes in a MAID except for some \( p(x_{v} | x_{pa(v)}) \), the expected utility of some player \( i \) in that MAID is not a single-valued function of the mutual information \( I(X_v; X_{pa(v)}) \). First, the same value of \( I(X_v; X_{pa(v)}) \) can occur for different conditional distributions \( p(x_v | x_{pa(v)}) \), and therefore that value of \( I(X_v; X_{pa(v)}) \) can correspond to multiple

\(^5\) This means that the parameters \( \theta^j \) are non-redundant in the sense that the family of probability distributions parametrized by \( (\theta^1, \ldots, \theta^d) \) is locally a non-singular \( d \)-dimensional manifold.
values of expected utility in general. Second, there might be several equilibria, i.e. strategy profiles, which solve the QRE condition, but correspond to different distributions at the decision nodes of the MAID. Thus, if v is a chance node in a MAID, there is no unambiguously defined “differential value of mutual information” to a player i in that MAID of the information channel from pa(v) to v. We can only talk about differential value of mutual information at a particular joint distribution of the MAID that corresponds to an equilibrium on a particular branch.

At such a point in the space of allowed joint distributions we can quantify the “alignment” between changes in expected utility and changes in mutual information. We propose to do this by taking the scalar product between \( \frac{\partial}{\partial \theta} V \) and \( \frac{\partial}{\partial \theta} I(X; S) \). To obtain a result that is independent of the parametrization chosen, the scalar product is taken with respect to the Fisher information metric, i.e., we choose

\[
\langle \nabla V, \nabla I \rangle \equiv \sum_{kl} g^{kl}(\theta) \frac{\partial}{\partial \theta^k} V \cdot \frac{\partial}{\partial \theta^l} I
\]

where \( g^{kl}(\theta) \) denotes the inverse of the Fisher information matrix \( g_{kl}(\theta) \) as defined in eq. (3).

Similarly, we choose the contravariant vector norm \(|v| \equiv \sqrt{v^k g_{kl} v^l}\) and similarly for covariant vectors.

To begin we consider the differential value of making an arbitrary infinitesimal change in parameter space, not necessarily a change along \( \nabla I(X; S) \):

**Definition 2.** Fix a set of MAIDs indexed by an d-dimensional differentiable manifold \( \Theta \) of game parameter vectors. Let \( \theta \) be a point in the interior of \( \Theta \) and \( \delta \theta \in \mathbb{R}^d \). The (differential) value of direction \( \delta \theta \) at \( \theta \) is defined as

\[
\mathcal{V}_{\delta \theta}(\theta) = \frac{\langle \nabla V, \delta \theta \rangle}{|\delta \theta|}
\]

Note that this is the length of the projection of \( \nabla V \) in the unit direction \( \delta \theta \). Thus, the direction \( \delta \theta \) is valuable to the extend that \( V \) increases in this direction (When \( V \) decreases in this direction, the value is negative).

In general, a mixed-index metric like \( g(\theta)^k_l \) must be the Kronecker delta function (regardless of the choice of metric g). Therefore we can expand

\[
\mathcal{V}_{\delta \theta}(\theta) = \frac{\frac{\partial}{\partial \theta^k} V \cdot g(\theta)^k_l \cdot \delta \theta^l}{\sqrt{\delta \theta^k \cdot g(\theta)_{kl} \cdot \delta \theta^l}}
\]

\[
= \frac{\frac{\partial}{\partial \theta^k} V \cdot \delta \theta^k}{\sqrt{\delta \theta^k \cdot \delta \theta_k}}
\]
The absence of the metric in the numerator in Eq. (14) reflects the fact that the vector of partial derivatives is a covariant vector, whereas $\delta \theta$ is a contravariant vector.

One important class of directions at a given game vector $\theta$ is the gradients of information-theoretic quantities at $\theta$. Values of information in these directions are quite important, as discussed below. However even when the direction we are considering is not parallel to the gradient of a mutual information, capacity, rationality, or some other function $f(\theta)$, we will often be concerned with quantifying quantities like “value of $f$” (e.g., “value of information capacity”) in that direction. We can do this with the following definition.

**Definition 3.** Fix a set of MAIDs indexed by an $d$-dimensional differentiable manifold $\Theta$ of game parameter vectors. Let $\theta$ be a point in the interior of $\Theta$, $\delta \theta \in \mathbb{R}^d$, and fix a smooth function $f : \Theta \to \mathbb{R}$. The (differential) value of $f$ in direction $\delta \theta$ at $\theta$ is defined as:

$$V_{f,\delta \theta} = \frac{\langle \text{grad}(V), \delta \theta \rangle}{\langle \text{grad}(f), \delta \theta \rangle} = \frac{\langle \text{grad}(V), \delta \theta \rangle}{\langle \text{grad}(f), \delta \theta \rangle}$$

Thus we consider how $V$ and $f$ change, when moving in the direction $\delta \theta$. If the sign of the differential value of $f$ in direction $\delta \theta$ at $\theta$ is positive, then an infinitesimal step in in direction $\delta \theta$ at $\theta$ will either increase both $V$ and $f$ or decrease both of them. If instead the sign is negative, then such a step will have opposite effects on $V$ and $f$. The size of the differential value of $f$ in direction $\delta \theta$ at $\theta$ gives the rate of change in $V$ per unit of $f$, for movement in that direction.

Now, to measure the “value of $f$” as such, we simply consider how $V$ changes when stepping into the direction of $\text{grad}(f)$, i.e. the direction corresponding to the steepest increase in $f$. This is captured by the following definition:

**Definition 4.** Fix a set of MAIDs indexed by an $d$-dimensional differentiable manifold $\Theta$ of game parameter vectors. Let $\theta$ be a point in the interior of $\Theta$, $\delta \theta \in \mathbb{R}^d$, and fix a smooth function $f : \Theta \to \mathbb{R}$. The (differential) value of $f$ at $\theta$ is defined as:

$$V_f(\theta) = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{|\text{grad}(f)|^2}$$

$$= \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(f), \text{grad}(f) \rangle}$$

(15)
Note that in contrast to $\mathcal{V}_{f, \theta}$, the value of $f$, i.e. $\mathcal{V}_f$ does depend on the metric, as can be seen by expanding the above definition.

$$\mathcal{V}_f(\theta) = \frac{\langle \text{grad}(V), \text{grad}(f) \rangle}{\langle \text{grad}(V), \text{grad}(f) \rangle} = \frac{\text{grad}(V)^k g_{kl} \text{grad}(f)^l}{\text{grad}(f)^k g_{kl} \text{grad}(f)^l} = \frac{\partial V}{\partial \theta^i} g_{ik} g_{lj} \frac{\partial f}{\partial \theta^j}$$

where we have used that $g^{ij}$ is the inverse of $g_{ij}$.

By the Cauchy-Schwarz inequality, so long as $|\text{grad}(f)(\theta)|^2 \neq 0$, $\mathcal{V}_f(\theta) \leq |\text{grad}(V)| / |\text{grad}(f)|$, with equality if and only if $\text{grad}(f) \propto \text{grad}(V)$. Note that not only does the bit of whether $\mathcal{V}_f(\theta)$ is maximal or not have the same value in all coordinate systems, it is also independent of the metric.

If $\mathcal{V}_f(\theta)$ is less than maximal, and the player is allowed to make any change to the current $\theta$ that has a given (infinitesimal) magnitude, they would not allocate all of that change only to improve the value $f$. They would prefer use some of it to improve other aspects of the game’s parameter vector. Intuitively, so long as they value anything other than $f$, $\mathcal{V}_f(\theta)$ is less than maximal.

As an illustration, in our example of a decision situation, the “differential value of mutual information” (between $X$ and $S$) is

$$\mathcal{V}_{I(X;S)}(\theta) = \frac{\partial}{\partial \theta^k} V \cdot g(\theta)^{kl} \cdot \frac{\partial}{\partial \theta^l} I(X;S) \cdot \text{grad}(I(X;S))^k \cdot g(\theta)^{kl} \cdot \text{grad}(I(X;S))^l.$$  

This is the amount that the player would value a change in the mutual information between $X$ and $S$, measured per unit of that mutual information.

To get an intuition for differential value of $f$, consider a locally invertible coordinate transformation at $\theta$ that makes the normalized version of $\text{grad}(f)$ be one of the basis vectors, $\hat{e}$, as discussed in the introduction. When we evaluate “(differential) value of $f$ at $\theta$”, we are evaluating the partial derivative of expected utility with respect to the new coordinate associated with that $\hat{e}$. (This is true no matter what we choose for the other basis vectors of the new coordinate system.)

---

6 N.b., at some $\theta$’s this transformation would be singular — exclude those from the current discussion.
Now since the coordinate transformation is locally invertible, moving in the direction \( \hat{e} \) in the new coordinate system induces a change in the position in the original game parameter coordinate system, i.e., a change in \( \theta \). This change in turn induces a change in the equilibrium profile \( \sigma \). Therefore it induces a change in the expected utilities of the players. It is precisely the outcome of this chain of effects that “value of \( f \)” measures.

Changing the original coordinate system \( \Theta \) will not change the outcome of this chain of effects — differential value of \( f \) is a covariant quantity. However changing the underlying space of game parameters, i.e. what properties of the game are free to vary, (as opposed to the choice of coordinate system of that space) will modify the outcome of this chain of effects. In other words, changing the parametrized family of games that that we are considering will change the value of the differential value of \( f \). So we must be careful in choosing the game parameter space; in general, we should choose it to be exactly those attributes of the game that we allow to vary. For example, if we suppose that some channels are free to vary, their specification must be included. Similarly, if we choose a model in which an overall multiplicative factor equally affecting all utility functions (i.e., a uniform tax rate) is free to vary, then we must also include that factor in our game parameter space. Conversely, if we choose a model in which there is no tax, then we must not include such a rate in our game parameter space. All of these choices will affect the dimensionality and structure of the parameter space and thus the formula we use to evaluate value of \( f \).

**B. Properties of differential value**

We now present some general results concerning value of a function \( f \), in particular conditions for negative values.

Throughout this section, we assume that both \( f \) and \( V \) are twice continuously differentiable.

The following observation may be helpful for understanding the geometric essence of the sequel. We shall work in some tangent space of a parameter manifold that is equipped with the Fisher metric \( \langle \cdot, \cdot \rangle \). Now, whenever we have such a scalar product on a (finite dimensional) vector space, then we can perform a linear transformation of that vector space to turn that scalar product into the standard Euclidean one. Thus, in the sequel, we shall essentially do elementary Euclidean geometry. (The nonlinear nature of the Fisher or any other Riemannian metric only comes into play when we look at the tangent spaces of several points simultaneously. The linear coordinate transformation turning the Fisher metric into the Euclidean one will depend on the particular tan-
gent space we are working, and in general, there will be no such transformation working for all tangent spaces simultaneously.)

We introduce some convex cones for our analysis of differential value of $f$ for a single player:

**Definition 5.** Fix a set of MAIDs indexed by an $d$-dimensional differentiable manifold $\Theta$ of game parameter vectors. Let $\theta$ be a point in the interior of $\Theta$, and fix a function $f : \Theta \to \mathbb{R}$. Define two hyperplanes

$$H_{0\theta} \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle = 0 \}$$
$$H_{00} \equiv \{ \delta \theta : \langle \text{grad}(f), \delta \theta \rangle = 0 \}$$

and four cones

$$C_{++}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle > 0, \langle \text{grad}(f), \delta \theta \rangle > 0 \}$$
$$C_{+-}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle > 0, \langle \text{grad}(f), \delta \theta \rangle < 0 \}$$
$$C_{-+}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle < 0, \langle \text{grad}(f), \delta \theta \rangle > 0 \}$$
$$C_{--}(\theta) \equiv \{ \delta \theta : \langle \text{grad}(V), \delta \theta \rangle < 0, \langle \text{grad}(f), \delta \theta \rangle < 0 \}.$$

We also put

$$C_{\pm}(\theta) \equiv C_{+-}(\theta) \cup C_{-+}(\theta).$$

The two hyperplanes $H_{0\theta}, H_{00}$ separate the tangent space at $\theta$ into the four disjoint convex cones $C_{++}, \ldots$. The $C$’s are cones because whenever $v \in C$, then also $\alpha v \in C$ for $\alpha > 0$. They are convex because whenever $v, w \in C$, then also $\lambda v + (1 - \lambda)w \in C$ for $0 \leq \lambda \leq 1$. They are obviously disjoint. Also $C_{--} = -C_{++}$ and $C_{-+} = -C_{+-}$. Since each of them is contained in an open halfspace, whenever $v \in C$, then $-v \not\in C$. Thus, they are pointed, meaning that they do not contain any entire straight line.

By the definition of the differential value of $f$ in the direction $\delta \theta$, it is negative for all $\delta \theta$ in either $C_{+-}(\theta)$ or $C_{-+}(\theta)$, that is, in $C_{\pm}(\theta)$. Moreover, it is positive on the other two cones, 0 on $H_{0\theta}$ and infinite on $H_{00}$.

In principle, either the two cones $C_{++}$ and $C_{--}$ or the two cones $C_{+-}$ and $C_{-+}$ could be empty. We now observe that the latter ones are nonempty — so there are directions in which the value of $f$ is negative — iff the value of $f$ is less than its maximum:
Proposition 1. Fix a set of MAIDs indexed by an $d$-dimensional differentiable manifold $\Theta$ of game parameter vectors. Let $\theta$ be a point in the interior of $\Theta$, and fix a function $f : \Theta \to \mathbb{R}$. Assume that $\text{grad}(V)$ and $\text{grad}(f)$ are both nonzero at $\theta$. Then $C_+^-$ and $C_-^+$ are nonempty iff

$$\mathcal{V}_f(\theta) < \frac{|\text{grad}V(\theta)|}{|\text{grad}f(\theta)|}$$

(16)

Proof. (16) is equivalent to

$$\langle \text{grad}(V), \text{grad}(f) \rangle < |\text{grad}(V)| |\text{grad}(f)|,$$

that is, the two vectors $\text{grad}(V)$ and $\text{grad}(f)$ are not positively collinear, that is, they are not positive multiples of each other, or in other words, they are not positively proportional to each other. But elementary linear algebra tells us that two (nonzero) vectors $V_1, V_2$ are not positively collinear iff there exists a third vector $W$ with

$$\langle V_1, W \rangle > 0, \quad \langle V_2, W \rangle < 0.$$  

(17)

With $V_1 = \text{grad}(V), V_2 = \text{grad}(f)$, this means that $C_+^- \neq \emptyset$, and this in turn is equivalent to $C_-^+ = -C_+^- \neq \emptyset$. □

Recall that $\mathcal{V}_f(\theta) < |\text{grad}V(\theta)| / |\text{grad}f(\theta)|$ iff $\text{grad}V(\theta) \not\propto \text{grad}f(\theta)$. So Prop. [1] identifies the question of whether $\text{grad}V(\theta)$ is positively proportional to $\text{grad}f(\theta)$ with the question of whether $C_{\pm}(\theta)$ is empty.

As an example of Prop. [1] say that we have a game specified by a parameter vector $\theta$ and that at the associated QRE player $i$ would prefer an increase in the information capacity of a particular information channel, everything else being equal. Suppose though that $i$ does not value such an increase the maximal possible amount. Then we know there is a way of changing the conditional distribution that defines that channel which simultaneously increases its information capacity and reduces the expected utility of player $i$.

We now assume that $f$ is a monotonically increasing function of $V$ across a compact $S \subseteq \Theta$. This means that $f$ and $V$ have the same level hypersurfaces (although, of course the values of $V$ and $f$ on any such common level hypersurface will in general be different) across $S$. The monotonicity implies that the linear order induced by values $f(\theta)$ relates the level hypersurfaces in the same way that the linear order induced by values $V(\theta)$ relates those level hypersurfaces. We also assume
that neither \( \nabla f \) nor \( \nabla V \) equals 0 anywhere in \( S \). The implicit function theorem then implies that those level hypersurfaces are smooth throughout \( S \) and in particular have no self-intersections. So \( \nabla V \) and \( \nabla f \) are proportional to one another throughout \( S \) (although the proportionality constant may change). This means that \( V_f(\theta) \) is maximal throughout \( S \) and so by Prop. 1 for no \( \theta \) in \( S \) is there a direction \( \delta \theta \) such that \( V_f,\delta \theta(\theta) < 0 \).

In general though level hypersurfaces will not match up throughout a region, as the condition that \( \nabla V \) and \( \nabla f \) be proportional is very restrictive and special. When they do not, we have points in that region that have directions with negative value of \( f \). We now derive a criterion involving both the gradients and the Hessians of \( V \) and \( f \) to identify such a mismatch.

Proposition 2. Fix a set of MAIDs indexed by an d-dimensional differentiable manifold \( \Theta \) of game parameter vectors. Let \( \theta \) be a point in the interior of \( \Theta \), fix a function \( f : \Theta \to \mathbb{R} \), and assume that both \( f \) and \( V \) are analytic at \( \theta \) with nonzero gradients. Define

\[
\nu \equiv [\nabla V](\theta), \\
\phi \equiv [\nabla f](\theta).
\]

Then for \( \epsilon > 0 \),

\[
\forall \ \theta' \in B_\epsilon(\theta) : \ C_+(-\theta') = \emptyset \quad \Leftrightarrow \quad \forall \ i, j : \ D^2 F - |\phi| |\nu|^2 D^2 V - \frac{|\phi|}{|\nu|} D_{ij} V \text{ in } B_\epsilon(\theta)
\]

Proof. We start with an observation. Whereas the two vectors \( \nabla V \) and \( \nabla f \) depend on the Fisher metric, the question whether two vectors are collinear in a vector space does not depend on the choice of any metric (this is in contrast to the question whether they are orthogonal which needs a metric). As, however, these two vectors depend on the metric, and as we shall have to take derivatives and therefore cannot stay in a single tangent space, we shall have to employ the geometric calculus established in the Appendix.

From what we have discussed, \( C_+(-\theta') = \emptyset \) iff \( \nabla V(\theta') \) and \( \nabla f(\theta') \) are positively proportional, that is,

\[
\nabla f(\theta') = \rho(\theta') \nabla V(\theta')
\]
for some positive function $\rho$. This yields

$$\rho = \frac{|\phi|}{|\nu|}. \quad (21)$$

If (19) holds in some open ball around $\theta$, we can differentiate it there to derive the following identity

$$D_{ij}f - \rho D_{ij}V = \frac{\partial \rho}{\partial x^i} \langle \nabla(V), e^j \rangle \quad (22)$$

where we use local coordinates $(x^1, \ldots, x^m)$ with $e^i = \frac{\partial}{\partial x^i}$ and the second covariant derivatives

$$D_{ij}h = \langle \nabla_v \nabla(h), e^j \rangle \quad (23)$$

for a twice differentiable function $h$. Since this Hessian $D_{ij}h$ is symmetric w.r.t. $i, j$, we conclude by symmetrization that (22) can be written as

$$D_{ij}f - \rho D_{ij}V = k \langle \nabla(V), e^i \rangle \langle \nabla(V), e^j \rangle \quad (24)$$

for some function $k$. Taking norms yields

$$k = \left| \frac{D^2 F - \frac{|\phi|}{|\nu|} D^2 V}{|\nu|^2} \right|, \quad (25)$$

whence (18). The converse also follows, essentially by reversing the steps of this argument, but since we shall not utilize it here, we spare us the details.

\[\square\]

We will refer to the matrix with entries

$$\left| D^2 F - \frac{|\phi|}{|\nu|} D^2 V \right| \langle \nu^i, \nu^j \rangle - D_{ij}F - \frac{|\phi|}{|\nu|} D_{ij}V$$

from Prop. 2 as the **second order mismatch** between $V$ and $f$ at $\theta$.

One might think that so long as the Hessian of $V$ at $\theta$ is not a constant times the Hessian of $f$ at $\theta$, then the level curves of $V$ and $f$ in the vicinity of $\theta$ will not match up, and so there will be points infinitesimally close to $\theta$ where the gradients of $V$ and of $f$ are not parallel. If this were the case, it would suggest that the condition in Prop. 2 that the second-order mismatch is non-zero could be tightened (by replacing that condition with the requirement that $\hat{F}_{ij} \neq \frac{|\nu|}{|\nu|} \hat{V}_{ij}$) and yet the implication given by that Lemma — that there are points where the gradients are not parallel — would still follow.
This is not the case though: Even if the Hessian of $V$ is not a constant times the Hessian of $f$ at $\theta$, it is possible that the gradients of $f$ and $V$ are parallel throughout a neighborhood of $\theta$. As a simple example, take $d = 2, V(\theta) = \theta^1, f(\theta) = (1 + |\theta^1|^2)^2$. Even though the Hessian of this $V$ is not a constant times the Hessian of this $f$ anywhere near the origin, the gradients of $f$ and $V$ are parallel everywhere near the origin.

Prop. 2 allows the possibility that the second-order mismatch is nonzero but there is an $d - 1$-dimensional hyperplane $T$ going through $\theta$ such that $\nabla f \propto \nabla V$ throughout $T$. One might wonder if non-zero second-order mismatch in fact rules out such hyperplanes. This is not the case though. As an example, choose $d = 2, V(\theta) = \theta^1, f(\theta) = (1 + \theta^1)^2(1 + [\theta^2]^2)$. The second order mismatch is nonzero for this $V$ and $f$ at $\theta = (0, 0)$. (The Hessian of $f$ at the origin is twice the identity matrix, the Hessian of $V$ there is the zero matrix, but $\nu_i \nu_j$ only has one non-zero entry, for $i = j = 1$.) However everywhere on the line $\{\theta' = t(1, 0) : t \in \mathbb{R}\}$ going through the origin, $\nabla f(\theta')$ and $\nabla V(\theta')$ are parallel.

It is typical that second-order mismatch between $V$ and $f$ is nonzero at almost all $\theta \in \Theta$. In particular, this is usually the case for information-theoretic quantities $f$ (e.g., for entropy, information capacity, mutual information, etc.). In such cases where the second-order mismatch is nonzero, for almost all $\theta$, there is an $\epsilon > 0$ such that there are directions $\delta \theta$ with negative value of $f$ at all $\theta' \in B_{\epsilon}(\theta)$ except (possibly) for the $\theta'$ in a subset of dimension less than $d$ (i.e., in a subset of measure zero). So for example, for almost all game parameter vectors $\theta$, any player $i$, and any information-theoretic quantity of interest, there will be a direction $\delta \theta$ such that taking an infinitesimal step in that direction away from $\theta$ will increase that information-theoretic quantity but hurt player $i$.

In light of this, arguably the important issue is not whether at some particular $\theta$ there is a direction in which information increase but expected utility decreases. That is almost always the case. Rather our results redirect attention to different issues. One such issue is what fraction of all $\theta$ and all possible infinitesimal changes to $\theta$ have negative value of information.

Another important issue arises whenever there are constraints on the allowed directions in $\Theta$. Such constraints are quite common in the real world. As an example, there might be constraints on the information capacity of a channel, which would manifest themselves in constraints on the allowed change in parameter space. Another example on the allowed change is to the parameter vector is the stipulation that it must be an insertion (or removal) of a garbling node, in the sense of Blackwell’s theorem [17].
With such constraints, in general Prop. 1 does not hold. Geometrically, such violations of Prop. 1 occur if the set of infinitesimal changes to \( \theta \) that respect the constraints to first order has zero intersection with \( C_\pm(\theta) \). This is why, for example, the result of Blackwell that there are no garbling nodes that one can insert just before a decision node in an ID that increase maximal expected utility is consistent with our results — the insertion of such a garbling node corresponds to a constrained change in \( \theta \), a change that lies in a cone that does not intersect \( C_\pm(\theta) \).

We established above that so long as \( C_{i,\pm}(\theta) \) for a particular player \( i \) is nonempty, there are directions in which the value of \( f \) is negative for \( i \). In other words, such directions exist if \( \nabla f \) and \( \nabla V \) are not positively proportional. We can generalize this to give a condition for there to be directions in which the value of \( f \) is negative for all players.

To do this define \( V_{i,\delta\theta}(\theta) \) as the value of \( f \) in direction \( \delta\theta \) to player \( i \) in a MAID. Also write the conic hull \[ \text{(7)} \] of a set of vectors \( \{a(i)\} \) all living in \( \mathbb{R}^d \) as

\[
\text{Con}(\{a(i)\}) \equiv \left\{ \sum_i \alpha_i a(i) : \alpha_i \geq 0 \forall i \right\}
\]

**Proposition 3.** Fix a set of MAIDs indexed by an \( d \)-dimensional differentiable manifold \( \Theta \) of game parameter vectors. Let \( \theta \) be a point in the interior of \( \Theta \), and fix a function \( f : \Theta \to \mathbb{R} \). Then

\[
\exists \delta\theta : V_{f,\delta\theta}(\theta) < 0 \\forall i
\]

\[
\Leftrightarrow
\]

\[
\nabla f \notin \text{Con}(\{\nabla V_i\})
\]

where all \( N + 1 \) gradients are evaluated at \( \theta \). In particular, in this case the cone \( \text{Con}(\{\nabla V_i\}) \) is pointed.

**Proof.** Let \( C_{i,\pm}(\theta) \) indicate the set \( C_{i,\pm}(\theta) \) for player \( i \). There is a direction \( \delta\theta \) such that \( V_{f,\delta\theta}(\theta) < 0 \ \forall i \) if and only if \( \cap_i C_{i,\pm}(\theta) \) is nonempty. \( \text{Con}(\{\nabla V_i\}) \), however, is characterized by the property that for each \( \delta\theta \in \text{Con}(\{\nabla V_i\}) \), there is at least one \( i \) such that \( \langle \delta\theta, \nabla V_i \rangle > 0 \). In fact, \( \delta\theta \in \text{Con}(\{\nabla V_i\}) \) means that \( \delta\theta = \sum_i \lambda_i \nabla V_i \) with \( 0 \leq \lambda_i \leq 1 \) for all \( i \), and hence

\[
0 < \langle \delta\theta, \delta\theta \rangle = \langle \delta\theta, \sum_i \lambda_i \nabla V_i \rangle = \sum_i \lambda_i \langle \delta\theta, \nabla V_i \rangle,
\]

and so, there has to be at least one \( i \) with \( \langle \delta\theta, \nabla V_i \rangle > 0 \).

\[ \square \]

When the second condition in Prop. 3 holds, there is a direction from \( \theta \) that is Pareto-improving, yet reduces \( f \). So for example, if that condition holds for a particular set of MAIDs when \( f \) is the
information capacity of a channel in the MAID, then all players would benefit from reducing the capacity of that channel.

The following lemma may help intuition concerning that second condition:

**Lemma 4.** If a set of vectors \( \{v(i)\} \) in \( \mathbb{R}^d \) is linearly independent then \( \text{Con}(\{v(i)\}) \) is pointed.

**Proof.** Otherwise, there exists some \( v \neq 0 \) with both \( v \) and \( -v \) in \( \text{Con}(\{v(i)\}) \). Thus, we have both \( v = \sum_i \lambda_i v(i) \) and \( -v = \sum_i \mu_i v(i) \) with \( \lambda_i \geq 0, \mu_i \geq 0 \) for all \( i \), and not all of them simultaneously 0. But then \( \sum_i (\lambda_i + \mu_i)v(i) = 0 \) which is a nontrivial linear relation between the \( v(i) \) violating our assumption that they be linearly independent. \( \square \)

The converse of Lemma 4 does not hold in general. For example, in \( \mathbb{R}^2 \), if \( v_1 = (1, 0) \), \( v_2 = (0, 1) \) and \( v_3 = v_1 + v_2 \), then the set \( \{v^i\} \) is both linearly dependent and pointed.

Instead of investigating the question of linear dependence between \( f \) and some \( V^i \), we could also use the preceding results for characterizing linear dependencies between various \( V^i \)s. For instance, as in Prop. 1, \( V^1 \) and \( V^2 \) are not positively collinear if

\[
\langle \text{grad}(V^1), \text{grad}(V^2) \rangle < \|\text{grad}(V^1)\|\|\text{grad}(V^2)\|.
\]

(27)

And as in Prop. 2, they are not positively aligned in an entire neighborhood of \( \theta \) if

\[
|\text{grad}(V^1)|D^2 V^2 - |\text{grad}(V^2)|D^2 V^1| \frac{\text{grad}(V^1)^i \text{grad}(V^1)^j}{|\text{grad}(V^1)|^2} - |\text{grad}(V^1)|D_{ij} V^2 - |\text{grad}(V^1)|D_{ij} V^1
\]

is nonzero. Thus, when the corresponding quantities are nonzero for any two \( V^k \) and \( V^\ell \), then these gradients must be linearly independent generically in a neighborhood of \( \theta \). Of course, this is possible only when the number \( N \) of players does not exceed the dimension \( d \) of our parameter space \( \Theta \). Similarly, when \( N < d \), we can also check such conditions not only between the various \( V^k \)s, but also between \( f \) and any \( V^k \).

\( \text{Con}(\{\text{grad}(V^i)\}) \) is pointed for many sets of MAIDs. But there are some important exceptions. As an illustration, consider a set of \( N \)-player MAIDs indexed by \( \Theta \) and a \( \theta \in \Theta \) where there is some ball \( B(\theta) \) centered on \( \theta \) such that the MAIDs of all \( \theta' \in B(\theta) \) have the same joint-move space \( X \). Suppose further that the utility functions for all those \( \theta' \) specify that the associated game is constant-sum, with the sum of utilities having the same value \( U \) for all those \( \theta' \in B(\theta) \). (For example, it may be that the \( \theta' \in B(\theta) \) vary only in the conditional distributions they assign to the chance nodes of the MAID — so that the utility functions of the players are the same throughout
$B(\theta)$ — and that the utility functions at $\theta$ form a constant-sum game with utility sum $\mathcal{U}$.) Then no matter what equilibrium concept we use, the sum of the expected utilities of the players will equal $\mathcal{U}$ for all $\theta' \in B(\theta)$. Therefore the sum of the vectors $\text{grad}(V^i)$ evaluated at $\theta$ equals zero. So $\text{Con}(\{\text{grad}(V^i)\})$ is not pointed in such sets of MAIDs.

However it is important to remember that Pareto-improving changes to $\theta$ are not the same thing as Pareto-improving changes to a strategy profile in a game specified by a single $\theta$. We can have Pareto-improving changes to $\theta$ even if $\theta$ specifies a game in which there are no Pareto-improving changes to the equilibrium strategy profile. In particular, $\theta$ can specify a constant-sum game — which allows no Pareto-improving changes to strategy profiles — while there are changes to $\theta$ that (result in changes to the equilibrium profile that) are Pareto-improving.

This is reflected in the fact that even if $\theta$ specifies a constant-sum game, it may be that $\text{Con}(\{\text{grad}(V^i)\})$ is non-pointed. Due to this fact, Prop. 3 does not mean that if $\theta$ specifies a constant-sum game, then there is no direction in $\Theta$ that is Pareto-improving while reducing $f$. The character of games infinitesimally close (in $\Theta$) to $\theta$ is what is important.

Note that the bit of whether or not $\text{grad}(f) \in \text{Con}(\{\text{grad}(V^i)\})$ is a covariant quantity. So the implications of Prop. 3 do not change if we change the coordinate system. In fact, the value of that bit is independent of our choice of the metric, so long as no $\text{grad}(V^i)$ is in the kernel of the metric. This means in particular that the implications of Prop. 3 do not vary with the choice of metric, so long as we stick to Riemannian metrics.

V. ILLUSTRATING EXAMPLE

To illustrate the foregoing, in this section we work through the formulas given in Sec. IV for the case of a game against Nature (the Blackwell ID). Then we do the same for a two-player, two-stage leader-follower game. In the first example, we show explicitly how to calculate all partial derivatives, while focusing on the results about differential value of information in the second example.

7 For example it may be that all games infinitesimally close to $\theta$ are also constant-sum — but with a different sum of utilities from the sum at $\theta$. In this case changing the game from $\theta$ to some infinitesimally close $\theta'$ will change the sum of expected utilities of the players, and so the sum of the $\text{grad}(V^i)$ is non-zero.
A. Decision problem

In a decision problem one agent plays against nature. (So this MAID is the special case of an influence diagram.) A simple example, which we will use to illustrate our notion of differential value of information is shown in Fig. 1 as the DAG of a ID.

In this MAID there is a state of nature random variable \( X \) taking on values \( x \) according to a distribution \( p(x) \). The agent observes \( x \) indirectly through a noisy channel that produces a value \( s \) of a signal random variable \( S \) according to a probability distribution \( p(s \mid x; \theta) \) parametrized by \( \theta \). The agent then takes an action, which we write as the value \( a \) of the random variable \( A \), according to the distribution \( \sigma(a \mid s) \). Finally, the utility \( u(x, a) \) is a function that depends only on \( x \) and \( a \).

![Fig. 1. A simple decision situation.](image)

We will refer to this MAID as the Blackwell ID.

In the Blackwell ID, for a given signal \( s \) and action \( a \), the conditional expected utility is

\[
\mathbb{E}(u \mid s, a; \theta) = \sum_x p(x \mid s; \theta) u(x, a) \quad (29)
\]

\[
p(x \mid s; \theta) = \frac{p(s \mid x; \theta) p(x)}{\sum_{x'} p(s \mid x'; \theta) p(x')} \quad (30)
\]

and given any distribution \( p(a \mid s) \) by the decision-maker, the associated unconditioned expected utility is \( \sum_{s,a} p(a \mid s) p(s; \theta) \mathbb{E}(u \mid s, a; \theta) \) where \( p(s; \theta) = \sum_x p(x; \theta) p(s \mid x; \theta) \). A fully rational decision maker will set \( p(a \mid s) \) to maximize this unconditioned expected utility. That would be equivalent to setting \( p(a \mid s) \) for each possible \( s \) to maximize \( \mathbb{E}(u \mid s, a; \theta) \).

Here, to have a differentiable strategy, we assume that the agent is not fully rational but plays a quantal best response with a finite \( \beta \):

\[
\sigma(a \mid s) = \frac{1}{Z(s)} e^{\beta \mathbb{E}(u \mid s, a; \theta)} \]

where \( Z(s) = \sum_a e^{\beta \mathbb{E}(u \mid s, a; \theta)} \) denotes the normalization constant.
1. **Calculating the gradient of the expected utility**

In order to identify the direction in the parameter space of the channel that is relevant for utility changes we have to calculate the gradient of $V(\beta, \theta)$ with respect to $\theta$. The expected utility depends on $\theta$ in two ways: (1) Directly via the change of the channel distribution

$$\frac{\partial}{\partial \theta} p(s \mid x; \theta) \equiv c_k(s \mid x)$$

and (2) indirectly via the induced change in the strategy

$$\frac{\partial}{\partial \theta^k} \sigma(a \mid s) .$$

The last term has two summands:

$$\frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, a) = \sigma(a \mid s) - \frac{1}{Z(\beta, \theta)} \frac{\partial Z}{\partial \theta^k} + \beta \sigma(a \mid s) \frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, a)$$

with

$$\frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, a) = \frac{\sum_s u(x, a) c_k(s \mid x) p(x)}{\sum_{x'} p(s \mid x'; \theta) p(x')} - \frac{\sum_s p(s \mid x; \theta) p(x) u(x, a)}{(\sum_{x'} p(s \mid x'; \theta) p(x'))^2} \sum_{x'} c_k(s \mid x') p(x')$$

and

$$\frac{\partial Z}{\partial \theta^k} = \beta \sum_a \exp \beta \mathbb{E}(u \mid s, a) \frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, a)$$

Putting all terms together leads to

$$\frac{\partial}{\partial \theta^k} V(\beta, \theta) = \sum_{a,s} \sigma(a \mid s) \mathbb{E}_{p_X}(U(X, a) C_k(s \mid X))$$

$$+ \sum_{a,s} \beta \sigma(a \mid s) p(s \mid x, a) \left( \frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, a) - \mathbb{E}_{\sigma(A \mid s)} \left( \frac{\partial}{\partial \theta^k} \mathbb{E}(u \mid s, A) \right) \right)$$

$$= \sum_{a,s} \sigma(a \mid s) \left\{ \mathbb{E}_{p_X}(U(X, a) C_k(s \mid X))$$

$$+ \beta \mathbb{E}(u \mid s, a) \left( \mathbb{E}_{p_X}(U(X, a) C_k(s \mid X)) - \mathbb{E}_{p_X(X; \theta)}(U(X, a) \mathbb{E}_{p_X}(C_k(s \mid X))) \right)$$

$$- \beta \mathbb{E}(u \mid s, a) \sum_{a'} \sigma(a' \mid s) \left( \mathbb{E}_{p_X}(U(X, a') C_k(s \mid X)) - \mathbb{E}_{p_X(X; \theta)}(U(X, a') \mathbb{E}_{p_X}(C_k(s \mid X))) \right) \right\}$$
As explained above, in order to obtain a proper contra-variant gradient vector we need a metric. We start from the observation that the joint distribution on the MAID, i.e.

\[ p(x, s, a; \theta) = p(x)p(s|x; \theta)\sigma(a|s) \]

depends on the channel parameters \( \theta \) via the channel and implicitly the strategy of the agent. According to eq. (3), the Fisher metric is given by

\[ g_{kl}(\theta) = \sum_{x, s, a} p(x, s, a; \theta) \frac{\partial \log p(x, s, a; \theta)}{\partial \theta^k} \frac{\partial \log p(x, s, a; \theta)}{\partial \theta^l} \]

Thus, even when using a binary channel with parameters \( \theta = (\epsilon^1, \epsilon^2) \), the metric is not the one that we calculated in Sec. II.C but includes additional terms which reflect that the strategy of the decision maker adapts to changes of the channel parameters as well. Using \( p(x, s, a; \theta) = \log p(x) + \log p(s | x; \epsilon^1, \epsilon^2) + \log \sigma(a|s) \) and that \( p(x) \) does not depend on \( \epsilon^1, \epsilon^2 \), we obtain:

\[ g_{kl}(\epsilon^1, \epsilon^2) = \sum_{x, s, a} p(x)p(s | x)\sigma(a | s)\left( \frac{\partial \log p(s | x)}{\partial \epsilon^k} \frac{\partial \log p(s | x)}{\partial \epsilon^l} + \frac{\partial \log \sigma(a | s)}{\partial \epsilon^k} \frac{\partial \log \sigma(a | s)}{\partial \epsilon^l} \right) \]

\[ = \sum_{x, s} p(x) \frac{\partial p(s | x)}{\partial \epsilon^k} \frac{\partial p(s | x)}{\partial \epsilon^l} + \sum_{x, s, a} p(x)p(s | x)\sigma(a | s) \frac{\partial \log \sigma(a | s)}{\partial \epsilon^k} \frac{\partial \log \sigma(a | s)}{\partial \epsilon^l} \]

Thus, in addition to the first term which corresponds to the result calculated above, we have three additional terms which take the dependence of the players decision into account. The gradient of the expected utility is then obtained by \( (\text{grad}V)^j = \frac{\partial}{\partial \theta} V(\theta)g^{ij}(\theta) \) where \( g^{ij} \) is the inverse of the Fisher metric \( g_{ij}(\theta) \).
In general, we could also include the rationality parameter $\beta$ of the player into the metric, i.e. $g(\epsilon^1, \epsilon^2, \beta)$ would then be a $3 \times 3$ matrix. This would take into account how the strategy of the player changes when $\beta$ is varied. Here, since we want to study the effect of information changes, we restrict changes to the channel parameters $(\epsilon^1, \epsilon^2)$ and consider $\beta$ as fixed. The general framework, can also be used to study the effect of changes to $\beta$, in particular investigating trade-offs between information and rationality. In any case, depending on which parameter changes one wants to study, the metric should be chosen accordingly.

**FIG. 2.** “Unit balls” of the Fisher metric depending on the rationality $\beta$ of the decision maker (see text for details).

Fig. 2 illustrates the effect of the players rationality on the metric $g(\epsilon^1, \epsilon^2)$. For different channel parameters, the shape of the metric is represented by plotting an ellipse that illustrates how the “unit ball” wrt the Fisher metric appears when plotted wrt the Euclidean coordinates $\epsilon^1, \epsilon^2$. $\beta = 0$ corresponds to a player that always plays uniformly at random and in this case all partial derivatives $\frac{\partial \sigma(a|x)}{\partial \epsilon^i}$ vanish. The figure shows that a more rational player reacts more strongly to changes of the channel noise and does so at successively higher noise levels.

As explained above, the Fisher metric quantifies the sensitivity of a distribution $p(x; \theta)$ to changes of the parameters $\theta$. Thus, when the distribution is insensitive to changes of $\theta$, a unit change to the distribution requires a large parameter change and the unit ball appears stretched in the $i$th coordinate direction.
B. Gradient of the mutual information

In order to understand better the role of information we compare the gradient of the expected utility with the gradient of the mutual information. The mutual information between $S$ and $X$ is

$$I(S;X) = \sum_{s,x} p(s | x, \theta) p(x) \log \frac{p(s | x, \theta)}{\sum_{x'} p(s | x', \theta) p(x)}.$$  \hfill (37)

$$\frac{\partial}{\partial \theta^k} I(S;X) = \sum_{s,x} c_k(s | x) p(x) \log \frac{p(s | x, \theta)}{\sum_{x'} p(s | x, \theta) p(x)}.$$  \hfill (38)

Again, the gradient $\text{grad} I(S; X)$ is obtained by multiplying the vector of partial derivatives with the inverse of the Fisher metric.

It seems to be possible that more mutual information does not correspond to more expected utility: If the agent gains a lot of irrelevant (less valuable) information, but loses valuable information an increase of mutual information can correspond to a smaller expected utility. That means, however, that even in a decision theoretic problem there is a possibility for a negative value of information. In particular, Prop. [1] states that there is a change to the channel parameters such that the player gains information while losing utility whenever $\text{grad} V$ and $\text{grad} I(X; S)$ are not collinear, i.e. parallel to each other.

C. Differential value of information

To illustrate the above concepts, we consider a simple setup using a binary state of nature $p(X = 0) = p(X = 1) = \frac{1}{2}$ and a binary channel $p(s | x; \theta)$ as in eq. (5). The player has two moves $A = 0, A = 1$ and obtains utility

$$u(x, a) = \begin{cases} 0 & x = 0, a = 0 \\ -2 & x = 0, a = 1 \\ 0 & x = 1, a = 0 \\ 1 & x = 1, a = 1 \end{cases}$$

Thus, the player has a save action $A = 0$, but can obtain a higher utility by playing 1 when he is certain enough that the state of nature is 1. Fig. [3] show the iso-clines of the mutual information $I(X; S)$ and the expected utility $V(\beta, \epsilon^1, \epsilon^2)$ for $\beta = 5$. Both, mutual information and expected utility improve with decreasing channel noise, i.e. small $\epsilon$. Nevertheless, the iso-clines do not agree.
and thus it is possible to change the channel parameters such that the expected utility increases while the mutual information decreases.

This can also be seen from the gradient vectors $\text{grad} I(X; S)$ and $\text{grad} V$. Only if those are collinear, any change increasing the expected utility would necessarily increase the mutual information as well. Here, we clearly see that the directions of steepest ascent of $I(X; S)$ and $V$ are different and according to Prop. 1 there exist directions such that $V$ increases while $I(X; S)$ decreases (and vice-versa).

![Isoclines: Beta = 5](image)

**FIG. 3.** Iso-clines of expected utility $V$ and mutual information $I(X; S)$ with corresponding gradient vectors showing the directions of steepest ascent (wrt the Fisher metric).

Fig. 4 shows the differential value of information (Def. 4). Recall that the differential value of information quantifies how much expected utility $V$ changes per unit change of the mutual information $(X; S)$. This again shows where the gradients $\text{grad} V$ and $\text{grad} I(X; S)$ are aligned and where they are rather different. The value of information is large when the gradients are almost collinear and can even become negative when their angle is above 90 degrees.

Next, we extend the above setting to a game involving two players.
VI. TWO-PLAYER GAMES

Instead of \( X \) being a nature node, it is now also a player (see Fig. 5). Therefore we have two players, the leader \( A^1 \) and the follower \( A^2 \).

The main difference is now that we have to replace the distribution \( \rho(X) \) of the state of nature by the equilibrium strategy \( \sigma(a_1) \) of player 1 that will also depend on the parameters \( \theta \). The channel
FIG. 5. A 2-player game where player $A^1$ (leader) can move before player $A^2$ (follower).

is characterized by the conditional probability distribution $p(s|a_1; \theta)$. Its gradient with respect to the parameter will be denoted again as

$$\frac{\partial}{\partial \theta} p(s|a_1; \theta) \equiv c_k(s|a_1)$$

The equilibrium strategies are fixed points of the QRE equations

$$\sigma_1(a_1) = \frac{1}{Z_1(\beta, \theta)} \exp \beta \mathbb{E}(u_1|a_1) \quad Z_1(\beta, \theta) = \sum_{a_1} \exp \beta \mathbb{E}(u_1|a_1)$$

$$\sigma_2(a_2|s) = \frac{1}{Z_2(\beta, \theta)} \exp \beta \mathbb{E}(u_2|s, a_2) \quad Z_2(\beta, \theta) = \sum_{a_2} \exp \beta \mathbb{E}(u_2|s, a_2)$$

and

$$\mathbb{E}(u_1|a_1) = \sum_{s, a_2} \sigma_2(a_2|s)p(s|a_1; \theta)u_1(a_1, a_2)$$

$$\mathbb{E}(u_2|s, a_2) = \sum_{a_1} p(a_1|s; \theta)u_2(a_1, a_2)$$

$$= \frac{\sum_{a_1} p(s|a_1; \theta)\sigma_1(a_1)u_2(a_1, a_2)}{\sum_{a_1'} p(s|a_1'; \theta)\sigma_1(a_1')}$$

The expected utilities in equilibrium are therefore

$$\hat{V}_1(\beta, \theta) = \sum_{a_2, s} \sigma_2(a_2|s) \sum_{a_1} \sigma_1(a_1)p(s|a_1; \theta)u_1(a_1, a_2)$$

$$\hat{V}_2(\beta, \theta) = \sum_{a_2, s} \sigma_2(a_2|s) \sum_{a_1} \sigma_1(a_1)p(s|a_1; \theta)u_2(a_1, a_2)$$

Now the $\theta$ dependence is even more complicated than in the decision theoretic case, because now we have to take into account that both players adapt their strategy if the channel parameters are changed. Thus, the equilibrium of the game moves and we need to use the implicit differentiation formula eq. [11].
Similarly, for the same reason, the Fisher metric becomes more involved and has additional terms taking the change in equilibrium into account. Here, we will not detail the required calculations, instead we directly present an example.

As in the decision problem, we consider binary state spaces and an asymmetric binary channel with parameters \( \theta = (\epsilon^1, \epsilon^2) \). We take the utility function of the players from [4] where Bagwell gave an example of a leader-follower game with negative value of information for the leader:

|    | follower | leader |
|----|----------|--------|
|    | \( u_1 / u_2 \) | \( L \) | \( R \) |
| \( L \) | 5 / 2 | 3 / 1 |
| \( R \) | 6 / 3 | 4 / 4 |

In particular, in pure strategy Nash equilibria the leader can only take advantage of moving first (by playing \( L \)), when the follower can observe his move perfectly (Stackelberg solution). As soon as the slightest amount on noise is added only the equilibrium of the simultaneous move game (both playing \( R \), the Cournot solution) remains.

Here, we use our refined, differential analysis to demonstrate a much richer structure. In particular, we show there exist a QRE branch and parameters for the noise of the channel such that both players prefer more noise.

In the decision case, we used \( I(X; S) \) to quantify the amount of information that is available to the player. In the game setting, the corresponding information \( I(A_1; S) \) strongly depends on the move of the leader. To demonstrate that this is actually the case, we first consider a symmetric channel \( p(s|a_1) \) parametrized by \( \epsilon = \epsilon^{1,2} \). Fig. 6 shows the strategy of the leader, i.e. \( \sigma(A_1 = L) \), depending on the channel noise and the rationality \( \beta = \beta^{1,2} \) of the players.

For sufficiently rational players, i.e. \( \beta > 5 \), there exist multiple QRE solutions. For \( \beta \to \infty \) the three QRE equilibria converge to the pure strategy Nash equilibrium where both play \( R \) (Cournot outcome, lower branch in red/green) and the two mixed strategy Nash equilibria of the original game. For \( \epsilon = 0 \), the upper mixed strategy equilibrium coincides with the leader advantage equilibrium mentioned above. In the following, we focus on the branch that smoothly connects to the origin \( \beta = 0 \), also called “principal branch”, and includes the upper equilibrium.

Fig 7 shows the channel capacity as well as the mutual information \( I(A_1; S) \) that is actually transferred across the channel. As soon as the leader is rational enough, he starts to prefer the move

---

9 The game is a discretization, with two moves for each player, of the Stackelberg duopoly.
10 There are additional mixed equilibria, which change smoothly with the noise. These are mentioned, but not discussed further in Bagwell [4].
FIG. 6. Surface of QRE equilibria for the leader-follower game.

$L$ and the mutual information $I(A_1; S)$ decreases. Remember that $I(A_1; S) = H(A_1) - H(A_1|S)$ is upper bounded by the entropy $H(A_1)$ and thus vanishes if the leader plays a pure strategy. Nevertheless, the potential information that could be transferred, i.e. the channel capacity, is still high. Thus, we argue that channel capacity is the better measure of information in the leader-follower game, since it quantifies how informative the signal could be to the follower independent of the strategy of the leader. In this sense, it is a property of the channel as such and does not depend on how the channel is actually used. Since a noisy channel can be considered as (a generalization of) an information partition, studying the capacity is in line with standard game theory where the information partition is considered as part of the game independent of which equilibrium is subsequently played.

To further investigate the role of information in the leader-follower game, we consider an asymmetric channel, i.e. $p(s|a_1)$ is parametrized by two noise parameters, $\epsilon_1, \epsilon_2$, giving the probability of error for the two inputs $L$ and $R$ respectively, and fix $\beta = 10$ for both players. In this case, we again find multiple QRE solutions when the channel noise is small enough. Fig. 8 shows the strategy of the first player ($\sigma(A_1 = L)$ in panel A) and the resulting expected utilities for both players (panels B and C).
FIG. 7. Channel capacity and mutual information $I(A_1; S)$ in the leader-follower game.

A) B) C)

FIG. 8. Surface of QRE equilibria ($\beta^{1,2} = 10$) for the leader-follower game. Panel A shows $\sigma(A_1 = L)$, while panels B and C show the corresponding expected utilities for the leader and the follower. The advantage of the leader is biggest when choosing $L$ with high probability.

Since, the strategy of the players is basically indifferent wrt the channel parameters for the lower branch, we focus on the upper branches. In particular, we investigate the top branch (pink) where the leader has the biggest advantage and can achieve the highest utility while the utility of the follower is lowest. In the following, we refer to this QRE solution as the “Stackelberg branch”.

Fig. 9 shows the differential value of channel capacity as well as mutual information $I(A_1; S)$ for the leader and the follower on the Stackelberg branch. Again, we observe a large difference
FIG. 9. Top: Differential value of channel capacity for the leader (panel A) and the follower (panel B). Bottom: Differential value of mutual information $I(A_1; S)$ for the leader (panel C) and the follower (panel D).
between channel capacity and mutual information. From Fig. 8 (panel B) we can see that the expected utility of the leader is high if the channel noise is low, thus we would expect a positive value of channel capacity. For the follower, the situation is reversed (panel C, same figure) and accordingly his differential value of information should be negative. This is mostly confirmed, Fig. 9 (panel A and B), but there are regions of the parameter space where the differential value of channel capacity is negative for both players. This occurs, because \( \epsilon^1 \) is more important, in terms of expected utility, to the leader than \( \epsilon^2 \), which is not reflected by the channel capacity.

Understanding the differential value of information \( I(A_1;S) \) is much more difficult. Here, we have to take into account that the strategy of the leader becomes more deterministic (thus its entropy is reduced) when the channel noise is low and accordingly the mutual information can be reduced even though the channel capacity is increased. In fact, the mutual information has a saddle point around \((\epsilon^1 = 0.15, \epsilon^2 = 0.3)\) which leads to a singularity of the differential value of information at this point. For this reason, and the ones given above, we focus on the more significant channel capacity and investigate whether there are directions such that both players prefer less capacity.

Fig. 10 shows the iso-clines of expected utility, for both leader and follower, and of the channel capacity together with the corresponding gradient fields. We immediately see that the gradients are nowhere collinear, thus from Prop. 1 we know that the channel noises can be changed such that the capacity increases whereas the expected utility (of either the leader or follower) decreases\(^\text{11}\). More Interestingly, the gradient field shows that everywhere \( \nabla \text{Capacity} \notin \text{Con}(\{ \nabla V^1, \nabla V^2 \}) \). Thus by Prop. 3 there must be directions such that both players are better off when the channel capacity is decreased. Simply considering the negative of the capacity as the function of interest, and thus inverting the corresponding gradient vectors, the same condition can be used to identify regions of channel parameters such that both players prefer more capacity. Since \( \nabla (-\text{Capacity}) \notin Con((\nabla V^1, \nabla V^2)) \), except for a small region in the upper left corner, e.g. \((\epsilon^1 = 0.05, \epsilon^2 = 0.35)\), we can immediately conclude that there are no such directions, for all parameter values outside this region.

Overall, using our framework, we find a very rich structure of differential value of information already in the rather simple leader-follower game. Nevertheless, as in the decision situation, general conditions on the gradient vectors can easily identify regions of channel parameters with negative value of information for both players. Furthermore, instead of comparing different infor-

\(^{11}\) By Prop. 2 this behavior is rather generic and thus expected.
FIG. 10. Iso-clines of the expected utilities for the leader and the follower as well as the channel capacity on the Stackelberg branch. The expected utility levels are color coded and the channel capacity (dotted) increases towards the origin (lower-left corner). The gradient vectors show the corresponding directions of steepest ascent (wrt the Fisher metric).

In this context, we quantify how much the players value changes to the channel parameters in units of utility per bits of information.

VII. CONCLUSIONS AND OUTLOOK

Here, we have proposed a framework to study the value of information in terms of information geometry. Our starting point is Shannon’s information theory which allows to quantify the amount of information that is transferred across a noisy channel. We then relate (quantitative) changes of information to changes of the expected utility of the players. From the concept of marginal utility in decision scenarios, we are naturally lead to study infinitesimal changes to the information structures in non-cooperative games.

To do so, we use Multi-agent influence diagrams, which represent the structure of a game in
terms of random variables as required for an information theoretic analysis. Furthermore, information geometry suggests to use the Fisher metric to calculate gradients and scalar products. Whereas the framework is in principle agnostic to the choice of the metric, some metric is needed. Otherwise, the results would depend on how the information channels in the game are parametrized.

Based on scalar products between gradients between expected utility and some information quantity of interest, we propose a measure of “differential value of information”. We then derive general conditions for the existence of negative value of information. In particular, we show that, generically, there are directions of changes to the channel parameters, i.e. changing the information structure of the game, such that more information hurts a player. Furthermore, we characterize the set of changes where this holds simultaneously for all players, i.e. every player’s expected utility decreases while information increases.

We illustrate our method using a decision situation as well as a two player leader-follower game. First, we show that negative value of information already exists in a decision context. This occurs, despite Blackwell’s theorem, since we allow arbitrary changes to the channel parameters which can not necessarily be represented as garbling in the sense of Blackwell. Then, we demonstrate an even richer structure of the differential value of information in a leader-follower game. In particular, we discuss the difference between information that is actually transferred and the information that could possibly be transferred, the channel capacity, across an information channel. We argue that, in a game context, channel capacity behaves more reasonably since it is a property of the (static) information structure of the game and does not depend on how the players behave in a certain equilibrium.

Overall, we believe that the value of information should be studied using information theory. Furthermore, allowing for infinitesimal changes to information channels supports a more fine-grained analysis of the value of information than by merely refining the information partitions. In particular, similar to marginal utility, our differential value of information has meaningful units of utility per bits. In the future, we want to use our framework, to study how certain changes to the structure of a game, as represented in a MAID, effects the value of information. Thus, answering questions such as under what conditions does adding an information channel, with infinitesimal capacity, before or after a decision node increase the expected utility of any player.

12 Having real valued parameters $\theta \in \mathbb{R}^d$ just taking partial derivatives corresponds to an implicit choice of the Euclidean metric on the parameter space.
Acknowledgments

NB acknowledges support by the Klaus Tschira Stiftung. The research of JJ was supported by the ERC Advanced Investigator Grant FP7-267087.

APPENDIX: BASIC CONCEPTS OF DIFFERENTIAL GEOMETRY

In this appendix, we provide an introduction to the basic concepts of differential geometry as needed and utilized in the main text. References include the monographs [3, 13].

Classical differential geometry works with coordinate representations of geometric objects and the transformations of those representations under coordinate changes. The geometric objects themselves are invariantly defined, but their coordinate representations are not, and in order to resolve this tension, the tensor calculus has been developed.

We start with some conventions:

1. Einstein summation convention

\[ a^i b_i := \sum_{i=1}^{d} a^i b_i \]  (46)

The content of this convention is that a summation sign is omitted when the same index occurs twice in a product, once as an upper and once as a lower index. The conventions about when to place an index in an upper or lower position will be given subsequently. One aspect of this, however, is

2. When \( G = (g_{ij}) \) is a metric tensor (a notion to be explained below), with with indices \( i, j \) ranging from 1 to \( d \), the inverse metric tensor is written as \( G^{-1} = (g^{ij}) \), that is, by raising the indices. In particular, the fact that the product of a matrix and its inverse is the identity matrix turns into

\[ g^{ij} g_{jk} = \delta^i_k := \begin{cases} 1 & \text{when } i = k \\ 0 & \text{when } i \neq k, \end{cases} \]  (47)

the so-called Kronecker symbol.

3. Combining the previous rules, we obtain more generally

\[ v^i = g^{ij} v_j \text{ and } v_i = g_{ij} v^j. \]  (48)
A (finite dimensional) manifold $M$ is locally modeled after $\mathbb{R}^d$. Thus, locally, it can be represented by coordinates $x = (x^1, \ldots, x^d)$ taken from some open subset of $\mathbb{R}^d$. That is, while their global topology may be intricate, local patches of a manifold can be represented by coordinates taken from an open set in $\mathbb{R}^d$. These coordinates, however, are not canonical, and we may as well choose other ones, $y = (y^1, \ldots, y^d)$, with $x = F(y)$ for some homeomorphism $F$. When the manifold $M$ is differentiable, we can cover it by local coordinates in such a manner that all such coordinate transitions are diffeomorphisms, that is, bijective maps that are differentiable and whose inverses are differentiable as well. For simplicity, by “differentiable” we shall mean “infinitely often differentiable” in the sequel. Again, the choice of coordinates is non-canonical. The basic content of classical differential geometry then is to investigate how various expressions representing objects on $M$ like tangent vectors transform under coordinate changes. Here and in the sequel, all objects defined on a differentiable manifold will be assumed to be differentiable themselves. This is checked in local coordinates, but since coordinate transitions are diffeomorphic, the differentiability property does not depend on the choice of coordinates.

First of all, we can consider differentiable functions $\phi$. Their values are, of course, independent of the choice of coordinates, that is, if $x = F(y)$, then $\phi(x) = \phi(F(y))$.

Next, there are the tangent vectors. A tangent vector for $M$ at some point represented by $x_0$ in local coordinates $x$ is an expression of the form

$$V = v^j \frac{\partial}{\partial x^j};$$

this means that it operates on a function $\phi(x)$ in our local coordinates as

$$V(\phi)(x_0) = v^j \frac{\partial \phi}{\partial x^j} |_{x=x_0}.$$  

(50)

The tangent vectors at $p \in M$ form a vector space, called the tangent space $T_pM$ of $M$ at $p$. The question then is how the same tangent vector is represented in different local coordinates $y$ with $x = F(y)$ as before. The answer comes from the requirement that the result of the operation of the tangent vector $V$ on a function $\phi$, $V(\phi)$, should be independent of the choice of coordinates. Applying here and in the sequel always the chain rule, this yields

$$V = v^j \frac{\partial y^\alpha}{\partial x^i} \frac{\partial}{\partial y^\alpha}.$$  

(51)

Thus, the coefficients of $V$ in the $y$-coordinates are $v^j \frac{\partial y^\alpha}{\partial x^i}$. This is verified by the following computation

$$v^j \frac{\partial y^\alpha}{\partial x^i} \frac{\partial}{\partial y^\alpha} \phi(F(y)) = v^j \frac{\partial y^\alpha}{\partial x^i} \frac{\partial \phi}{\partial x^j} \frac{\partial x^j}{\partial y^\alpha} = v^j \frac{\partial x^j}{\partial y^\alpha} \frac{\partial \phi}{\partial x^j} = v^j \frac{\partial \phi}{\partial x^j}.$$  

(52)
as required.

A vector field then is defined as \( V(x) = v^i(x) \frac{\partial}{\partial x^i} \), that is, by having a tangent vector at each point of \( M \). As indicated above, we assume here that the coefficients \( v^i(x) \) are differentiable.

Returning to a single tangent vector, \( V = v^j \frac{\partial}{\partial x^j} \) at some point \( x_0 \), we consider a covector \( \omega = \omega_i dx^i \) at this point as an object dual to \( V \), with the rule

\[
dx^i(\frac{\partial}{\partial x^j}) = \delta^i_j
\]

yielding

\[
\omega_j dx^j(\frac{\partial}{\partial x^i}) = \omega_i v^j \delta^i_j = \omega_i v^j.
\]

This expression depends only on the coefficients \( v^j \) and \( \omega_i \) at the point under consideration and does not require any values in a neighborhood. We can write this as \( \omega(V) \), the application of the covector \( \omega \) to the vector \( V \), or as \( V(\omega) \), the application of \( V \) to \( \omega \).

We have the transformation behavior

\[
dx^i = \frac{\partial x^i}{\partial y^\alpha} dy^\alpha
\]

required for the invariance of \( \omega(V) \). Thus, the coefficients of \( \omega \) in the \( y \)-coordinates are given by the identity

\[
\omega_j dx^j = \omega_i \frac{\partial x^i}{\partial y^\alpha} dy^\alpha.
\]

The transformation behavior of a tangent vector as in (51) is called contravariant, the opposite one of a covector as (56) covariant.

A 1-form then assigns a covector to every point in \( M \), and thus, it is locally given as \( \omega_i(x) dx^i \).

Having derived the transformation of vectors and covectors, we can then also determine the transformation rules for other tensors. A lower index always indicates covariant, an upper one contravariant transformation.

The metric tensor, written as \( g_{ij} dx^i \otimes dx^j \), with \( g_{ij} = \langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \rangle \) being the product of those two basis vectors, operates on pairs of tangent vectors. It therefore transforms doubly covariantly, that is, becomes

\[
g_{ij}(F(y)) \frac{\partial x^i}{\partial y^\alpha} \frac{\partial x^j}{\partial y^\beta} dy^\alpha \otimes dy^\beta.
\]

We require that the metric tensor be positive definite and symmetric, that is,

\[
g_{ij} = g_{ji} \text{ for all indices } i, j.
\]
The function of the metric tensor is to provide a Euclidean product of tangent vectors,
\[ \langle V, W \rangle = g_{ij}v^i w^j \] (59)
for \( V = v^i \frac{\partial}{\partial x^i}, W = w^j \frac{\partial}{\partial x^j}. \) As a check, in this formula, \( v^i \) and \( w^j \) transform contravariantly, while \( g_{ij} \) transforms doubly covariantly so that the product as a scalar quantity remains invariant under coordinate transformations.

A differentiable manifold equipped with such a metric tensor is called a Riemannian manifold.

Vectors \( V, W \in T_p M \) with \( \langle V, W \rangle = 0 \) are called orthogonal. The norm of a vector \( V \in T_p M \) is defined as
\[ |V| = \sqrt{\langle V, V \rangle}. \] (60)

For a function \( \phi \), we have its differential
\[ d\phi = \frac{\partial \phi}{\partial x^i} dx^i, \] (61)
a 1-form; this depends on the differentiable structure, but not on the metric. The gradient of \( \phi \), however, involves the metric; it is defined as
\[ \text{grad} \phi = g^{ij} \frac{\partial \phi}{\partial x^i} \frac{\partial}{\partial x^j}. \] (62)

The gradient of a function \( \phi \) is orthogonal to the level hypersurfaces \( \phi = c \), in the following sense. When \( V \in T_p M \) is tangent to such a level hypersurface, it satisfies
\[ V(\phi) = v^k \frac{\partial \phi}{\partial x^k} = 0. \] (63)

When \( V \) then satisfies (63), we have
\[ \langle \text{grad} \phi, V \rangle = g_{ik} g^{ij} \frac{\partial \phi}{\partial x^i} v^k = g_{ik} \frac{\partial \phi}{\partial x^k} v^k = 0, \] (64)
that is, \( \text{grad} \phi \) and \( V \) are orthogonal.

We also need the formula for the product of the gradients of two functions \( \phi, \psi \),
\[ \langle \text{grad} \phi, \text{grad} \psi \rangle = g_{ik} g^{ij} g^{kl} \frac{\partial \phi}{\partial x^i} \frac{\partial \psi}{\partial x^l} = g_{ij} \frac{\partial \phi}{\partial x^i} \frac{\partial \psi}{\partial x^j}. \] (65)

In differential geometry, one also needs a notion of second derivatives. The first derivative of a function is easy to compute in local coordinates; it yields a one-form, or dually, a tangent vector. When we wish to compute the second derivatives of a function, we would thus have to compute
the first derivatives of a tangent vector field. But how could this be done, as the tangent spaces at different points are not canonically identified? This is in contrast to the Euclidean situation, where any tangent space can be parallelly moved into any other one. Therefore, one also needs a notion of parallel transport in Riemann geometry upon which to build the notion of the derivative of a tangent vector field. This leads to the covariant derivative of Levi-Civitá which we now introduce.

Let \((x^1, \ldots, x^d)\) be local coordinates, as usual. The covariant derivative \(\nabla\) satisfies

\[
\nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i} = \Gamma^k_{ij} \frac{\partial}{\partial x^k} \quad \text{for all } i, j
\]

with

\[
\Gamma^j_{jk} = \frac{1}{2} g^{i\ell} (g_{j\ell,k} + g_{k\ell,j} - g_{jk,\ell}),
\]

where \((g^{ij})_{i,j=1,...,d} = (g_{ij})^{-1}\) (i.e. \(g^{ij}g_{ij} = \delta_{ij}\)) and

\[
g_{j\ell,k} = \frac{\partial}{\partial x^k} g_{j\ell}.
\]

The expressions \(\Gamma^j_{jk}\) are called the Christoffel symbols. \(\nabla\) is then extended to all vector fields \(V = v^j \frac{\partial}{\partial x^j}\) via the product rule

\[
\nabla_{\frac{\partial}{\partial x^i}} v^j \frac{\partial}{\partial x^j} = \frac{\partial v^j}{\partial x^i} \frac{\partial}{\partial x^j} + v^j \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}.
\]

Moreover,

\[
\nabla_{w^j \frac{\partial}{\partial x^i}} v^j \frac{\partial}{\partial x^j} = w^j \nabla_{\frac{\partial}{\partial x^i}} v^j \frac{\partial}{\partial x^j}.
\]

The covariant derivative \(\nabla\) is set up in such a way that it is compatible with the metric, in the sense it satisfies the product rule

\[
Z\langle V, W \rangle = \langle \nabla_Z V, W \rangle + \langle V, \nabla_Z W \rangle
\]

for all vector fields \(Z, V, W\).

The covariant derivative then allows us to introduce the Riemannian version of the Hessian of a function. The **Hessian** of a differentiable function \(f : M \to \mathbb{R}\) on a Riemannian manifold \(M\) is

\[
\nabla df.
\]

Here, we have \(df = \frac{\partial f}{\partial x^i} dx^i\) in local coordinates, hence

\[
\nabla_{\frac{\partial}{\partial x^i}} df = \frac{\partial^2 f}{\partial x^i \partial x^j} dx^j - \frac{\partial f}{\partial x^i} \Gamma^k_{jk} dx^k,
\]

50
\[ \nabla df = \left( \frac{\partial^2 f}{\partial x^i \partial x^j} - \frac{\partial f}{\partial x^k} \Gamma^k_{ij} \right) dx^i \otimes dx^j. \] (70)

We also have
\[ \nabla df(X, Y) = \langle \nabla_X \text{grad}(f), Y \rangle, \] (71)
since \( Y(f) = \langle \text{grad} f, Y \rangle \) and thus
\[ X(Y(f)) = \langle \nabla_X \text{grad}(f), Y \rangle \]
\[ = \langle \nabla_X \text{grad}(f), Y \rangle + \langle \text{grad}(f), \nabla_X Y \rangle \]
\[ = \langle \nabla_X \text{grad}(f), Y \rangle + (\nabla_X Y)(f), \]
and applying (70) to \( X \) and \( Y \) yields
\[ \nabla df(X, Y) = X(Y(f)) - (\nabla_X Y)(f). \] (72)

In local coordinates, with \( e^i = \frac{\partial}{\partial x^i} \), we then have the components of the Hessian of \( f \) given as
\[ D_{ij} f = \nabla df(e^i, e^j) = \langle \nabla_{e^i} \text{grad}(f), e^j \rangle. \] (73)

Note that the Hessian is symmetric in the sense that
\[ D_{ij} f = D_{ji} f \text{ for all } i, j. \] (74)

[1] S. I. Amari and H. Nagaoka, *Methods of information geometry*, Transl. Math. Monogr. 191, AMS and Oxford Univ. Press, 2000.

[2] Pierre-Olivier Amblard, Olivier J. J. Michel, and Saverio Morfu, *Revisiting the asymmetric binary channel: joint noise-enhanced detection and information transmission through threshold devices*, Proc. SPIE 5845 (2005), 50–60.

[3] N. Ay, J. Jost, H.V. Lê, and L. Schwachhöfer, *Information geometry*, to appear.

[4] K. Bagwell, *Commitment and observability in games*, Games and Economic Behavior 8 (1995), 271–280.

[5] Dirk Bergemann and Stephen Morris, *The comparison of information structures in games: Bayes correlated equilibrium and individual sufficiency*, Tech. report, SSRN, 2013, Economic Theory Center Working Paper No. 054-2013.
[6] David Blackwell, *Equivalent comparisons of experiments*, The Annals of Mathematical Statistics 24 (1953), no. 2, 265–272.

[7] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2003.

[8] T. Cover and J. Thomas, *Elements of information theory*, Wiley-Interscience, New York, 1991.

[9] Prashant Doshi, Yifeng Zeng, and Qiongyu Chen, *Graphical models for interactive pomdps: representations and solutions*, Autonomous Agents and Multi-Agent Systems 18 (2009), no. 3, 376–416.

[10] D. Fudenberg and J. Tirole, *Game theory*, MIT Press, Cambridge, MA, 1991.

[11] Olivier Gossner, *Ability and knowledge*, Games and Economic Behavior 69 (2010), no. 1, 95–106.

[12] Ronald A Howard and James E Matheson, *Influence diagram retrospective*, Decision Analysis 2 (2005), no. 3, 144–147.

[13] J. Jost, *Riemannian geometry and geometric analysis*, Springer, 6th edition, 2011.

[14] D. Koller and N. Friedman, *Probabilistic graphical models*, MIT Press, 2009.

[15] D. Koller and B. Milch, *Multi-agent influence diagrams for representing and solving games*, Games and Economic Behavior 45 (2003), 181–221.

[16] Ehud Lehrer, Dinah Rosenberg, and Eran Shmaya, *Garbling of signals and outcome equivalence*, Games and Economic Behavior 81 (2013), 179 – 191.

[17] Moshe Leshno and Yishay Spector, *An elementary proof of blackwell’s theorem*, Mathematical Social Sciences 25 (1992), no. 1, 95–98.

[18] P. Levine and J.P. Ponssard, *The values of information in some nonzero sum games*, International Journal of Game Theory 6 (1977), 221–229.

[19] D.J.C. Mackay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.

[20] D. H. Wolpert, M. Harre, E. Olbrich, N. Bertschinger, and J. Jost, *Hysteresis effects of changing parameters in noncooperative games*, Physical Review E 85 (2012), 036102, DOI: 10.1103/PhysRevE.85.036102.