Bounds of Optimal Learning

Roman V. Belavkin

***Abstract*—Learning is considered as a dynamic process described by a trajectory on a statistical manifold, and a topology is introduced defining trajectories continuous in information. The analysis generalises the application of Orlicz spaces in non-parametric information geometry to topological function spaces with asymmetric gauge functions (e.g. quasi-metric spaces defined in terms of KL divergence). Optimality conditions are formulated for dynamical constraints, and two main results are outlined: 1) Parametrisation of optimal learning trajectories from empirical constraints using generalised characteristic potentials; 2) A gradient theorem for the potentials defining optimal utility and information bounds of a learning system. These results not only generalise some known relations of statistical mechanics and variational methods in information theory, but also can be used for optimisation of the exploration-exploitation balance in online learning systems.**

I. Introduction

OPTIMAL control theory [1], [2] and methods of conditional Markov processes in stochastic optimal control [3] have had a great impact on the development of adaptive and learning systems [4], [5]. Under certain conditions required for the convergence of empirical distributions, these systems can satisfy some optimality criteria asymptotically, even if the information they use is incomplete. One of the greatest challenges, however, is the development of a non-asymptotic theory of learning and optimisation that could be applied to a wider class of problems, such as optimisation of non-stationary processes and learning from non-independent and not identically distributed sequences. Such processes have been studied in information theory and non-stationary non-equilibrium thermodynamics. This paper presents a non-asymptotic approach to optimisation of learning based on information value theory [6], where variational problems were first considered in the context of Bayesian learning. It presents further development of the duality theory of utility and information [7] by considering topologies on function spaces suitable to describe learning systems and their optimisation. The geometric approach has been inspired by works in information geometry [8]–[10].

The next section overviews briefly some basic concepts of utility theory and its relation to the theories of decisions under uncertainty and stochastic optimal control. It will also discuss the main problem of their application to learning systems. Section III will be concerned with a representation of learning processes in topological functional spaces. Their relation to normed spaces, such as Orlicz spaces in non-parametric information geometry [10], will be discussed. Section IV will connect these representations with optimisation theory. Generalised characteristic potentials will be introduced that allow for prametrisation of optimal trajectories from empirical constraints. The optimal trajectories will define bounds on expected utility and information gains in learning systems, which are computed as path integrals in conservative vector fields of the gradients of the potentials. The method will be illustrated on two examples for optimisation systems with a binary and uncountable utilities, and when information is represented by negative entropy. In the end, we discuss how this work is related to the maximum entropy principle of statistical mechanics, information value theory and how it generalises the optimal control.

Roman V. Belavkin is with School of Engineering and Information Sciences, Middlesex University, London, UK (email: R.Belavkin@mdx.ac.uk).
This work was supported in part by EPSRC grant EP/DO59720

II. Optimisation and Learning

Let A be a set of arbitrary nature, and let $\lesssim \subseteq A^2$ be a complete transitive binary relation (total pre-order or the *preference* relation). Then $(A, \lesssim)$ is an abstract choice set. A subset of symmetric pairs $\sim \subseteq \lesssim$ is the equivalence relation, and antisymmetric $< \subseteq \lesssim$ is a partial order. The quotient set $A\backslash\sim$ is totally ordered. We assume that the quotient set $A\backslash\sim$ can be embedded into the extended real line $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{\pm\infty\}$, and therefore the preference relation has a *utility* representation $x : A \to \overline{\mathbb{R}}$:

$$a_1 \lesssim a_2 \quad \iff \quad x(a_1) \leq x(a_2)$$

The rational choice and optimisation problems then can be solved by maximisation of the utility.

Under uncertainty, one considers probability measures $\mu : \mathfrak{R} \to [0,1] \subset \mathbb{R}$ on some σ-ring $\mathfrak{R}(A) \subseteq 2^A$. These measures are sometimes interpreted as lotteries over the choice set $(A, \lesssim)$. For example, the Dirac δ-measures ($\delta_a(da) = 1$ if $a \in da$; 0 otherwise) correspond to the elements $a \in A$ observed with certainty. Other lotteries are convex combinations of the δ-measures. The problem of optimal choice under uncertainty is solved by extending the preference relation $(A, \lesssim)$ onto the set of *all* lotteries. This extension should be compatible with $(A, \lesssim)$ in the following sense $(\Delta, \lesssim) = (A, \lesssim)$, where Δ is the set of all Dirac δ-measures. One such extension is given by the *expected utility*:

$$E_\mu\{x\} = \int_A x(a)\,\mu(da)$$

Thus, measure μ is preferred to ν if $E_\mu\{x\} \geq E_\nu\{x\}$. This method is adopted in the classical Bayesian estimation procedures [11] and stochastic optimal control [1], [3]. Moreover, it is well-known that the expected utility is the only representation satisfying the continuity and substitution independence axioms [12], typical for ordered linear spaces.

Although in many cases the methods of optimal control can be applied under certain assumptions to the learning systems, one has to take into account that probability measures in learning problems are not known. Instead, the learner has to propose a hypothesis (a prior) about the measure before making a decision, and then update the hypothesis using posterior information. This important difference means that the described above theory is optimal only approximately (or asymptotically).

In this work, we consider an evolution of the learning system as a trajectory $\mu = \mu(t)$ in the space of all probability measures (i.e. the space of all hypotheses). A topology in this space will allow us to consider continuous trajectories. This topology will be related to empirical information that the learning system receives, and our main goal will be to define optimality conditions and parametrisation of the optimal trajectories. The theory has close relation to the use of Orlicz spaces in non-parametric information geometry [10] and conjugate duality in optimisation theory [13].

III. Topological Spaces of Learning Systems

First, we point out that all measures we deal with are Radon (tight, inner-regular) measures. This is because a preference relation $(A, \lesssim)$ with a utility representation is a separable, complete metric space. Radon measures are finite on compact subsets: $\mu(A_c) < \infty$, if $A_c \subset A$ is compact. They can be defined as non-negative, continuous linear functionals $\mu(f)$ on the space $\mathcal{C}_c^\infty(A)$ of continuous functions with compact support.

All measures μ that are absolutely continuous with respect to some dominating measure ν have coordinatisations in Lebesgue space $L_1(A, \nu)$ given by the Radon-Nikodym derivatives $y(a) = d\mu/d\nu$. For example, if A is compact, then one can take some constant dominating measure da (e.g. the Lebesgue measure), and then measures can be identified with the density functions $y \in Y$, where $Y = \mathcal{C}_c^*(A)$ is the dual of space $\mathcal{C}_c^\infty(A)$ with respect to a bilinear form $\langle\cdot,\cdot\rangle : Y \times \mathcal{C}_c^\infty \to \mathbb{R}$, represented by the inner product

$$\langle y, f\rangle = \int_A f(a)\, y(a)\, da\,, \quad f \in \mathcal{C}_c^\infty(A) \qquad (1)$$

Thus, probability measures correspond to non-negative linear functionals $y \in Y$ with the L_1 norm equal to one: $\|y\|_1 = \int_A |y(a)|\, da = 1$. The set of all such measures is sometimes referred to as the *statistical manifold*, and it represents the space of hypotheses of the learning system. The statistical manifold is a Choquet simplex — a convex hull in L_1 with extreme points corresponding to the Dirac δ-measures. Note that if set A is infinite, then δ-measures are singular with respect to da, and therefore do not belong to $L_1(da)$. However, they can be considered as generalised limits in the extended optimisation problems (see [13]).

The expected utility representation requires not only that the utility functions are measurable, but also summable with respect to the measures being compared. Thus, suitable utility functions are points in space $X = Y^*$, the dual of Y with respect to the transposed bilinear form (1). This means that $X = \mathcal{C}_c^{**}(A)$ is the second dual of $\mathcal{C}_c^\infty(A)$, and space Y is the pre-dual of X. If set A is finite, then these spaces are reflexive (because they are finite-dimensional). In this case, space X of utility functions is isomorphic to $\mathcal{C}_c^\infty(A)$, and Y is also the dual of X. Here, we consider the general case when set A can be infinite, and therefore the corresponding spaces can be infinite-dimensional.

We now introduce a topology on Y (and the statistical manifold) that is stronger than the topology of L_1. This topology will also induce polar topology on the space X of utility functions. Let $y_0 \in Y$ be some initial point in Y (e.g. the prior hypothesis), and let $K_Y \subset Y$ be some closed neighbourhood of y_0 containing all measures consistent with posterior information. This neighbourhood can be defined using the fact that learning systems are characterised by incomplete information. In particular, the Dirac δ-measures cannot be included into K_Y because they correspond to complete certainty and maximum (possibly infinite) information. Thus, closed neighbourhoods of y_0 in the topology of systems with information constraints should be defined by absorbing sets K_Y that exclude the δ-measures.

A neighbourhood K_Y of y_0 in a linear space can be defined with the distance of $y \in K_Y$ from y_0. In particular, the distance from $0 \in \mathrm{Int}(K_Y)$ is given by the *gauge*

$$D(y) = \inf\{\beta > 0 : y \in \beta K_Y\}$$

with $D(y_0) = 0$. It is well-known also that if K_Y is convex, then the gauge can be computed as the *support* function of the polar set $K_Y^* = \{x : \langle x, y\rangle \le 1, y \in K_Y\} \subset X = Y^*$

$$D(y) = \sup\{\langle x, y\rangle : x \in K_Y^*\}$$

Note that the polar set is always convex, and it is absorbing if and only if set K_Y is bounded. Clearly, it is also true for the bipolar set K_Y^{**}. Here, we shall only consider the case, when $K_Y = K_Y^{**}$ (i.e. $K_X^* = K_Y$), and therefore closed neighbourhoods K_Y and K_X in spaces Y and X are bounded and absorbing closed convex sets (convex sets with interior points are also called convex bodies). However, we shall not require these sets to be balanced (symmetric). The gauge (also called the *Minkowski functional*) and the support function are positively homogeneous of the first degree and sub-additive on convex sets (and therefore convex). They generalise norms and semi-norms, and many important results of functional analysis still hold for these functions. For example, the Hahn-Banach theorem holds for Minkowski functionals in linear spaces over $\mathbb{R}$. Other results, which will not be reported here, include the completeness of the dual and separability (under certain conditions) of the pre-dual space, boundedness of continuous linear functionals on bounded sets, and so on. Because information is non-symmetric, it is attractive to develop the theory using a non-symmetric topology. This is a generalisation of the normed spaces, such as the Orlicz spaces, used in information geometry [10]. Another reason for this generalisation is monotonicity, which will be pointed out later.

If the amount of empirical information in the learning system can be measured by some convex functional $F : Y \to \overline{\mathbb{R}}$, then convex body K_Y can be defined by the restriction $F(y) \le I \in \mathrm{im}(F) \setminus \{+\infty\}$. The polar set K_X is defined by the corresponding restriction of the dual convex functional $F^*(x) \leq I^* \in \mathrm{im}(F^*) \setminus \{+\infty\}$. For the theory of convex functions see [13], [14]. Here we recall some basic concepts.

Let $F : Y \to \overline{\mathbb{R}}$ be a proper, closed (lower semi-continuous), convex functional. Note that $F : Y \to \overline{\mathbb{R}}$ is *proper* if its *effective domain* $\mathrm{dom}\, F \equiv \{y : F(y) < \infty\}$ is non-empty and $F(y) > -\infty$. Such functionals are continuous on the interior of $\mathrm{dom}\, F$. Gâteaux differential (directional derivative) is generalised by the *sub-differential*:

$$\partial F(y) \equiv \{x \in X : \langle x, y' - y\rangle \leq F(y') - F(y)\}\,, \quad \forall\, y' \in Y$$

where $\partial F \subset X$ is a non-empty, bounded and closed convex set. If $F(y)$ is differentiable at y, then $\partial F(y) = \{x\}$ (i.e. a singleton set); otherwise, $x \in \partial F(y)$. Thus, the minimum of F (if exists) is $y_0 \in \mathrm{dom}\, F$ such that $0 \in \partial F(y_0)$.

The dual functional $F^* : X \to \overline{\mathbb{R}}$ is defined by the Legendre-Fenchel transform:

$$F^*(x) = \sup\{\langle x, y\rangle - F(y)\}\,, \qquad x \in \partial F(y)$$

It is always convex, closed and proper. An important property of the dual convex functionals F and F^* is that they are both differentiable if and only if they are both strictly convex.

The gauge for set K_X and the support of its polar set K_Y for $x_0 = 0$ and $y_0 = 0$ can be defined as:

$$D^*(x) = \inf\{\beta^{-1} > 0 : F^*(\beta\, x) \leq I^*\} \tag{2}$$

$$D^*(x) = \sup\{\langle x, y\rangle : F(y) \leq I\} \tag{3}$$

In the same manner, one defines function $D(y)$ that is the gauge for set K_Y and the support of K_X. Topological spaces L_F and L_F^*, associated with dual functionals F and F^*, are the totalities of all elements $y \in Y$ and $x \in X$, such that $D(y) < \infty$ and $D^*(x) < \infty$. Thus, closed neighbourhoods of $y_0 \in L_F$ are sets $K_Y = \{y : F(y) \leq I < \infty\} \subset Y$. This topology can be considered along with the *upper* topology on $(\mathbb{R}, \leq)$, in which closed sets are segments $(-\infty, I] \subset \mathbb{R}$. In this case, information constraints $I_1 \leq I_2$ correspond to sets $K_1 \subseteq K_2 \subset L_F$. The gauge $D(y)$ with respect to a 'unit ball' $\{y : D(y) \leq 1\}$ defines a complete pre-order on L_F, and any function $y = f(I)$ that is monotonic with respect to pre-orders $(L_F, \lesssim)$ and $(\mathbb{R}, \leq)$ is also continuous in these topologies.

If convex functionals F and F^* can be represented in the integral form $F(y) = \int f(y(a))\, da$, where $f : \mathbb{R} \to \overline{\mathbb{R}}$ is an even (symmetric), convex function with $f(0) = 0$ and $\lim f(t) = \infty$ for $t \to \infty$, then spaces L_f and L_f^* are the dual Orlicz spaces. These spaces are complete normed spaces (Banach spaces), in which the gauge is a norm, and they are a generalisation of the classical L_p spaces (sequence spaces can also be considered). The theory of such spaces is well-developed, and Orlicz spaces have been used in non-parametric information geometry [10]. Here, however, we shall not require dual functionals F and F^* to be symmetric, and therefore the gauge will not be a norm. An important reason for this is to avoid non-monotonic operations, such as $x \mapsto |x|$ on the utility functions, because they do not preserve the preference relation $(A, \lesssim)$ on the domain. In optimisation problems, monotonicity is more important than symmetry.

Note also that differentials of convex functionals are monotone operators between dual spaces [17]. In fact, if dual functionals F and F^* are both strictly convex, then they are both differentiable, and the pair of strictly monotone operators ∂F and ∂F^* ($\partial F = (\partial F^*)^{-1}$) set up a Galois connection between pre-ordered topological spaces L_F and L_F^*. Below are precisely such asymmetric functionals that appear often in problems with information constraints:

$$F(y) = \int_A \Big(\ln \frac{y(a)}{y_0(a)} - 1\Big)\, y(a)\, da \tag{4}$$

$$F^*(x) = \int_A e^{x(a)}\, y_0(a)\, da \tag{5}$$

Functional (4) is information (KL) divergence [15], [16] also known as the relative entropy. Its effective domain contains $y \geq 0$ absolutely continuous w.r.t. $y_0 \geq 0$. If set A is infinite, then we assume $F(\delta) = +\infty$, because $\lim F(y) = +\infty$ as $y(a) \to \delta_a(da)/da$. Its unique minimum is achieved at $y = y_0$. Information divergence (4) generalises several functionals used in statistical mechanics and information theory (e.g. negative Boltzmann entropy and Shannon mutual information). Its dual is functional (5). It is positive for $y_0(a) > 0$, and $\inf F^*(x) = \lim F^*(x) = 0$ as $x(a) \to x_0(a) = -\infty$ (in the sense of pointwise convergence). Normalisation of functions y corresponds to transformation $F^*(x) \mapsto \ln F^*(x)$.

Learning systems are characterised by increasing information dynamics. $I = I(t)$, where $t \in \mathbb{R}$ is time (discrete or continuous). In this case, continuous functions $y = f(I)$, defined in terms of topology L_F and upper topology on $\mathrm{im}\,(F) \subset (\mathbb{R}, \leq)$, can also be considered as $y = f \circ I(t)$ continuous with respect to the upper topology on $t \in \mathbb{R}$. This represents that learning systems use new information to update the hypothesis $y \in L_F$. With some abuse of notation, we can represent the evolution of a learning system by a continuous trajectory $y(t)$ on the statistical manifold. In the next section, we shall consider the problem of optimisation of the evolution of a learning system. The analysis is closely related to the information value theory [6] and variational methods in approximate inference.

IV. Optimality, Parametrisation and Bounds

If $x \in X$ is the utility representation of $(A, \lesssim)$, then maximisation of the expected utility is just maximisation of the linear functional $x(y) = \langle x, y\rangle$ on the statistical manifold. Incomplete information in a learning system imposes additional constraints, restricting the set of feasible solutions to some subset K_Y. This is a standard optimisation problem with constraints, and it coincides with the support function (3) of set K_Y or the gauge (2) of the polar set K_Y^*. New information may relax the constraints, but unless the system can receive the maximum (possibly infinite) information amount, the constraints remain active. Observe also that the statistical manifold is a simplex, and without the constraints, a continuous linear functional always achieves the extremum in one of the extreme points (i.e. one of the δ-measures).

Let us denote by $U = \langle x, \hat{y}\rangle$ the value of the linear functional for the optimal solution $\hat{y} \in K_Y$. Conditions defining this solution are obtained in a standard way by the Kuhn-Tucker theorem.

Theorem 1 (Necessary conditions of extrema)**.** *Let Y and X be a dual pair of linear spaces with respect to some non-degenerate bilinear form $\langle \cdot, \cdot\rangle : X \times Y \to \mathbb{R}$, and let $F : Y \to \overline{\mathbb{R}}$ be a proper, closed, convex functional. The solutions $\hat{y}$ of conditional extremum*

$$U(I) = \sup\{\langle x, y\rangle : F(y) \leq I \in \mathrm{im}\,(F) \setminus \{+\infty\}\}$$

satisfy the following conditions

$$\beta x \in \partial F(\hat{y})\,, \quad F(\hat{y}) = I\,, \quad \beta^{-1} \in \partial U(I)$$

Proof: The extrema are defined by zero in the subdifferential of the Lagrangian function $L(y, \beta^{-1}, I) = \langle x, y\rangle + \beta^{-1}[I - F(y)]$, where β^{-1} is the Lagrange multiplier for $F(y) \leq I$:

$$\partial_y L(y, \beta^{-1}, I) = x - \beta^{-1}\partial F(y) \ni 0\,, \quad \Rightarrow \quad \beta x \in \partial F(\hat{y})$$
$$\partial_{\beta^{-1}} L(y, \beta^{-1}, I) = I - F(y) \ni 0\,, \quad \Rightarrow \quad F(\hat{y}) = I$$

By considering the value of the Lagrangian as function of the constraint $U = U(I)$, its subdifferential gives the third condition: $\partial U(I) \ni \beta^{-1}$. □

Similar necessary conditions $\beta x \in \partial F(\hat{y})$, $\langle x, \hat{y}\rangle = U$, $\beta \in \partial I(U)$ can be obtained by solving the following (dual) minimisation problem

$$I(U) = \inf\{F(y) : \langle x, y\rangle \geq U > -\infty\}$$

Necessary conditions become also sufficient, if one considers convexity of F in the Lagrangian function: $U = \langle x, \hat{y}\rangle$ is the maximum for $\beta > 0$ and the minimum for $\beta < 0$. Convex functional $F(y)$ has only one extremum (the minimum), and $\beta = 0$ corresponds to the global minimum ($0 \in \partial F(y_0)$).

If $F(y)$ is information divergence (4), then the necessary conditions give solutions in the exponential form

$$\beta\, x(a) = \ln \frac{\hat{y}(a)}{y_0(a)} + \gamma(\beta) \;\Rightarrow\; \hat{y}(a) = y_0(a)\, e^{\beta x(a) - \gamma(\beta)}$$

where $e^{\gamma(\beta)} = \int_A y_0\, e^{\beta x}\, da$ is the normalising condition. If $y_0(a) = \mathrm{const}$, then optimal function $\hat{y}(a)$ is the density of the Gibbs distribution.

One can see that the optimal solutions $\hat{y}$ are parametrised by $\beta \in \mathbb{R}$ related to one of the constraints $F(\hat{y}) = I$ or $\langle x, \hat{y}\rangle = U$. We factorise these relations by introducing the *generalised characteristic potentials*:

$$\Phi(\beta^{-1}) \equiv \inf[\beta^{-1} I - U(I)]\,, \quad \Psi(\beta) \equiv \sup[\beta U - I(U)]$$

The extrema in the definitions above are satisfied if

$$\beta^{-1} \in \partial U(I)\,, \quad \beta \in \partial I(U)$$

These conditions correspond to the optimality condition in Theorem 1. Note, however, that the potentials are real functions $\Phi : \mathbb{R} \to \overline{\mathbb{R}}$ and $\Psi : \mathbb{R} \to \overline{\mathbb{R}}$ independent of the dimensionality of spaces Y and X. We shall assume from now that the potentials are differentiable (e.g. when the dual functionals F and F^* are strictly convex) and denote their derivatives by Ψ' and Φ'. Their applications are based on the following theorem.

Theorem 2 (Parametrisation)**.** *Parameter $\beta \in \mathbb{R}$, defining the solutions $\hat{y}$ of the dual extremal problems $U = \sup\{\langle x, y\rangle : F(y) \leq I\}$ and $I = \inf\{F(y) : \langle x, y\rangle \geq U\}$, is related to the constraints $I \in \mathbb{R}$ or $U \in \mathbb{R}$ by the following relations*

$$I = \Phi'(\beta^{-1})\,, \qquad U = \Psi'(\beta)$$
$$I = \beta\Psi'(\beta) - \Psi(\beta)\,, \qquad U = \beta^{-1}\Phi'(\beta^{-1}) - \Phi(\beta^{-1})$$

Proof: Consider the Legendre-Fenchel transforms of the potentials:

$$U(I) = \inf[\beta^{-1} I - \Phi(\beta^{-1})]\,, \quad I(U) = \sup[\beta U - \Psi(\beta)]$$

The necessary conditions of extrema are $I = \Phi'(\beta^{-1})$ and $U = \Psi'(\beta)$, which is the first pair of relations. Substituting them into the above transforms gives the second pair. □

It is easy to check also the relation

$$\Phi(\beta^{-1}) = -\beta^{-1}\Psi(\beta)$$

For density functions in the exponential form, characteristic potential $\Psi(\beta)$ is the *cumulant generating function* of measure $\nu(da) = y_0(a)\, da$:

$$\Psi(\beta) = \ln \int_A e^{\beta x(a)} y_0(a)\, da$$

Potential $\Phi(\beta^{-1})$ in this case is the *free energy*.

In some cases, the potentials allow for closed form parametrisation. The following are examples for two important cases of a binary and uncountable utility functions, and when $F(y)$ is information divergence with respect to a uniform measure (e.g. when $F(y)$ is negative entropy).

Example 1. *Let $A = \{a_1, a_2\}$, utility $x : A \to \{c - d, c + d\}$, and the reference measure be $da = 1/|A| = 1/2$,. Then*

$$\Psi(\beta) = \beta\, c + \ln\cosh(\beta\, d)$$
$$U(\beta) = c + d\tanh(\beta\, d)$$

Example 2. *Let A be uncountable, and let the utility function be $x : A \to [c - d, c + d] \subset \mathbb{R}$. Consider measures on equivalence classes $x(a) \in \mathbb{R}$ and a uniform reference measure $dx / \int dx$ on compact subsets of $\mathbb{R}$. Then*

$$\Psi(\beta) = \beta\, c - \ln|\beta\, d| + \ln|\sinh(\beta\, d)|$$
$$U(\beta) = c - \beta^{-1} + d\coth(\beta\, d)$$

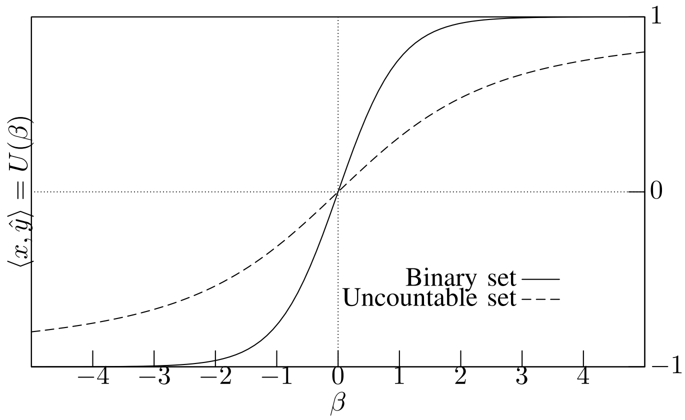

Fig. 1. Functions $U = \Psi'(\beta)$ in Examples 1 and 2.

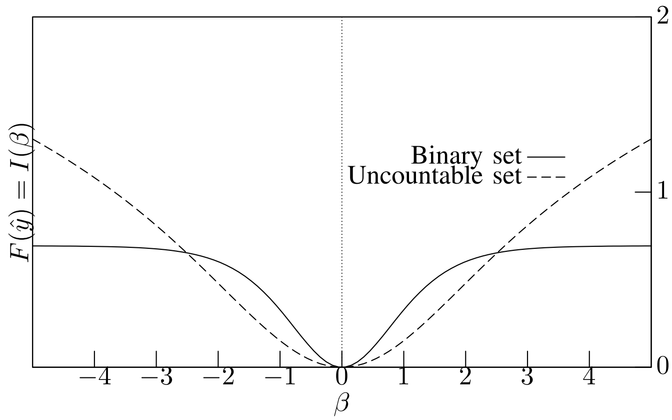

Fig. 2. Functions $I = \Phi'(\beta^{-1})$ in Examples 1 and 2.

Figures 1 and 2 show functions $U = \Psi'(\beta)$ and $I = \beta\Psi'(\beta) - \Psi(\beta)$ from Examples 1 and 2 with $c = 0$ and $d = 1$. Functions $U = \Psi'(\beta)$ are invertible, and the inverse gives parametrisation of the optimal measures from empirical constraints $\langle x, y\rangle \geq U$ on the expected utility. Alternatively, one can use the entropy of empirical distribution to obtain β^{-1} from the inverse relation of $I = \Phi'(\beta^{-1})$.

Let us now consider the evolution of a learning system, which is characterised by the dynamics of information $F(y) \leq I(t)$. The topology L_F, introduced in previous section, allows us to represent this evolution by a continuous trajectory $y = y(t)$ on the statistical manifold. The trajectory defines also the expected utility dynamics $\langle x, y(t)\rangle$. The cumulative expected utility and information gains along the trajectory on $t \in [t_1, t_2]$ are defined by the following path integrals

$$\int_{y(t_1)}^{y(t_2)} \langle x, y(t)\rangle \, dy(t) \,, \qquad \int_{y(t_1)}^{y(t_2)} F(y(t)) \, dy(t)$$

Clearly, an optimal system evolves along optimal trajectory $\hat{y}(t)$ that defines some bounds on the quantities above. These bounds should be path (trajectory) independent, which is shown by the following theorem.

Theorem 3 (Optimal bounds)**.** *Let $I = I(t)$, $U = U(t)$ be monotone functions describing constraints $F(y) \leq I(t)$ or $\langle x, y\rangle \geq U(t)$ in a learning system with a continuous trajectory $y = y(t)$, $t \in [t_1, t_2]$, on the statistical manifold. Then*

$$\begin{aligned} \int_{y(t_1)}^{y(t_2)} \langle x, y(t)\rangle \, dy(t) &\leq \Psi(\beta_2) - \Psi(\beta_1) \\ \int_{y(t_1)}^{y(t_2)} F(y(t)) \, dy(t) &\geq \Phi(\beta_1^{-1}) - \Phi(\beta_2^{-1}) \end{aligned}$$

where β_1, β_2 are determined from $I(t_1)$, $I(t_2)$ or $U(t_1)$, $U(t_2)$ using functions $\beta^{-1} = (\Phi')^{-1}(I)$ or $\beta = (\Psi')^{-1}(U)$ respectively.

Proof: The optimal trajectory is the totality of functions, satisfying optimality conditions of Theorem 1 and parametrised by $\beta(t) \in \mathbb{R}$ for each constraint $I(t)$ or $U(t)$, as shown in Theorem 2. Path integrals are independent of parametrisation, and therefore the bounds can be computed as integrals along the optimal trajectory $\hat{y}(\beta)$, $\beta \in [\beta_1, \beta_2]$. Using Theorem 2, we substitute the optimal values $\langle x, \hat{y}\rangle = U$ and $F(\hat{y}) = I$ in the integrals by the derivatives $U = \Psi'(\beta)$ and $I = \Phi'(\beta^{-1})$ of the characteristic potentials. The bounds are then given by the corresponding Riemann integrals:

$$\int_{\beta_1}^{\beta_2} \Psi'(\beta) \, d\beta \,, \qquad \int_{\beta_2^{-1}}^{\beta_1^{-1}} \Phi'(\beta^{-1}) \, d\beta^{-1}$$

The final expressions are obtained by applying the Newton-Leibniz formula. □

Theorem 3 is the analogue of the gradient theorem of vector analysis, which states that path integrals in conservative vector fields are path independent. If $F(y)$ is Shannon information, then the first expression of the theorem defines the upper bound on cumulative expected utility gain for a given change of information; the second expression is the lower bound on cumulative information required to achieve a given change of expected utility.

V. Discussion

This paper described geometric interpretation of utility and information evolution in learning systems. The most relevant results for practical applications are: 1) the parametrisation method of optimal measures from empirical constraints that is independent of dimensionality of the problem; 2) optimal bounds that can be useful for estimating system requirements. It is also interesting to point out the following uniqueness result. If information is represented by any strictly convex function $F(y)$, then the optimality conditions are satisfied by unique measures, and therefore the optimal learning trajectory $y = \hat{y}(t)$ is unique. In particular, if $F(y)$ is negative entropy, then the optimality conditions are equivalent to the maximum entropy principle [18], and the optimal trajectory describes evolution of thermodynamic equilibrium states. The optimal measures in this case are Gibbs distributions, which are often used in stochastic optimisation and some exploration algorithms. The parametrisation method defines the optimal 'temperature' (β^{-1}) for such algorithms from empirical constraints (e.g. as $\beta = (\Psi')^{-1}(U)$, where U is the empirical expected value). Because optimality conditions are formulated for local information constraints, the theory is non-asymptotic. It presents the opportunity to optimise online learning algorithms and exploration-exploitation balance.

The theory presented extends further the analogy between learning and physical systems. For example, the optimal bounds can be interpreted as the actions of an optimal learning system with respect to two conservative forces — expected utility and information. The first optimality condition in Theorem 1 can be written as eikonal equation $\partial\langle x, y\rangle = \beta^{-1}\partial F(y)$, which shows the similarity between geometric learning theory, presented here, and geometric optics (i.e. the ray method and the Fermat principle).

Optimisation of learning can be seen as a generalisation of the optimal control. Indeed, the latter optimise evolution of a system in space X (e.g. utility) and time. It can be solved, for example, by the dynamic programming method sequentially optimising the expected utility [1], which is an optimisation of a linear functional on each step. Observe that the Lagrangian function in Theorem 1 becomes linear as $\beta^{-1} \to 0$. Optimality condition $\beta^{-1} = \partial U(I)$ implies that such a control is optimal only when new information results in zero change of the expected utility. Clearly, Theorem 1 gives the required generalisation for the learning systems.

This work considered learning as an optimisation in the pre-dual space Y and information, and therefore learning can be seen as the pre-dual process of optimal control in space and time. Previously, the author applied the theory in cognitive models of human and animals [19], for optimal action selection in agents learning in non-stationary stochastic environments [20] and for stochastic reinforcement learning of procedural knowledge in neural cell-assemblies [21]. Applications of this theory to other specific problems is the subject of ongoing research.

REFERENCES

[1] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[2] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The mathematical theory of optimal processes*. John Wiley and Sons, 1962, translated from Russian.

[3] R. L. Stratonovich, "Conditional Markov processes," *Theory of Probability and its Applications (USSR)*, vol. 5, no. 2, pp. 156–178, 1960.

[4] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 1998.

[6] R. L. Stratonovich, "On value of information," *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, vol. 5, pp. 3–12, 1965, in Russian.

[7] R. V. Belavkin, "The duality of utility and information in optimally learning systems," in *7th IEEE International Conference on 'Cybernetic Intelligent Systems'*. London: IEEE, 2008.

[8] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference*. Moscow, U.S.S.R.: Nauka, 1972, in Russian, English translation: Providence, RI: AMS, 1982.

[9] S.-I. Amari, *Differential-Geometrical Methods of Statistics*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1985, vol. 25.

[10] G. Pistone and C. Sempi, "An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one," *The Annals of Statistics*, vol. 23, no. 5, pp. 1543–1561, October 1995.

[11] A. Wald, *Statistical Decision Functions*. New York: John Wiley & Sons, 1950.

[12] J. von Neumann and O. Morgenstern, *Theory of games and economic behavior*, 1st ed. Princeton, NJ: Princeton University Press, 1944.

[13] R. T. Rockafellar, *Conjugate Duality and Optimization*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. PA: Society for Industrial and Applied Mathematics, 1974, vol. 16.

[14] R. E. Showalter, *Monotone operators in Banach space and nonlinear partial differential equations*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 1997, vol. 49.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[16] S. Kullback, *Information Theory and Statistics*. John Wiley and Sons, 1959.

[17] R. I. Kachurovskii, "Nonlinear monotone operators in Banach spaces," *Russian Mathematical Surveys*, vol. 23, no. 2, pp. 117–165, 1968.

[18] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, 108, pp. 620–630, 171–190, 1957.

[19] R. V. Belavkin and F. E. Ritter, "The use of entropy for analysis and control of cognitive models," in *Proceedings of the Fifth International Conference on Cognitive Modeling*, F. Detje, D. Dörner, and H. Schaub, Eds. Bamberg, Germany: Universitäts–Verlag Bamberg, April 2003, pp. 21–26.

[20] R. V. Belavkin, "Acting irrationally to improve performance in stochastic worlds," in *Proceedings of AI–2005, the 25th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, M. Bramer, F. Coenen, and T. Allen, Eds., vol. XXII, BCS. Cambridge: Springer, 2005, pp. 305–316.

[21] R. V. Belavkin and C. Huyck, "Emergence of rules in cell assemblies of fLIF neurons," in *The 18th European Conference on Artificial Intelligence*, 2008.