Optimal measures and transition kernels*

Roman V. Belavkin

School of Engineering and Information Sciences
Middlesex University, London NW4 4BT, UK
Tel: +442084116263
R.Belavkin@mdx.ac.uk

Abstract. We study positive measures that are solutions to an abstract optimisation problem, which is a generalisation of a classical variational problem with a constraint on information of a Kullback-Leibler type. The latter leads to solutions that belong to a one parameter exponential family, and such measures have the property of mutual absolutely continuity. Here we show that this property is related to strict convexity of a functional that is dual to the functional representing information, and therefore mutual absolute continuity characterises other families of optimal measures. This result plays an important role in problems of optimal transitions between two sets: Mutual absolute continuity implies that optimal transition kernels cannot be deterministic, unless information is unbounded. For illustration, we construct an example where, unlike non-deterministic, any deterministic kernel either has negatively infinite expected utility (unbounded expected error) or communicates infinite information.

1 Introduction

Let $X := \cup C_c(\Omega)$ be the union of spaces of continuous functions $x : \Omega \to \mathbb{R}$ with compact support in a locally compact topological space Ω. Thus, X is a normed space with the Chebyshev norm $\|x\|_\infty := \sup_\omega |x(\omega)|$, and in fact it is an ordered commutative C^*-algebra with pointwise multiplication and ordering. The dual of X is the space $Y := \mathscr{M}(\Omega)$ of Radon measures on Ω [7], which includes σ-additive and regular Borel measures. Thus, Y is a Banach space with the norm $\|\cdot\|_1$, and in fact it is a module over algebra X with pointwise multiplication. Given a fixed element $x \in X$, let $\{y_\beta\}_x \subset Y$ be a family indexed by $\beta \geq 0$, where each y_β is defined as

$$y_\beta := e^{\beta x} y_0 , \quad y_0 > 0 \tag{1}$$

The elements y_β correspond to positive one-parameter exponential Radon measures, and normalised elements $p_\beta := y_\beta / \|y_\beta\|_1$ are the corresponding exponential probability measures. A similar construction can be made in the case when X is a non-commutative C^*-algebra, such as the algebra of compact Hermitian operators on a Hilbert space. However, the exponential family can be defined in different ways, such as $y_\beta := e^{\beta x + \ln y_0}$ or $y_\beta := e^{\beta x/2} y_0 e^{\beta x/2}$, which are not equal in the non-commutative case.

* This work was supported by EPSRC grant EP/H031936/1.

The exponential family plays an important role in mathematical statistics, physics and information theory. Many important probability distributions are members of this family. In fact, the lower bound for the variance of the unbiased estimator of an unknown parameter, defined by the Rao-Cramer inequality, is attained if and only if the probability distribution is a member of the exponential family [9,19]. The Boltzmann (or Gibbs) distribution is a member of this family, and it is known to maximise entropy of a thermodynamical system under the constraint on energy [10]. A closely related variational problem is minimisation of Kullback-Leibler distance [14] (negative relative entropy) of one probability measure from another subject to a constraint on the expected value. These problems were studied in information theory [21,22,23], and it was established that exponential distributions maximise the capacity of an information channel. More recently, the exponential family has been studied in information geometry, and it was shown that the family is a Banach space with an Orlicz norm [18]. These result have been generalised to quantum systems [6,24].

As will be shown later in this paper, most of the above properties are related to the fact that exponential measures are optimal solutions to one specific variational problem. In this paper, we shall study a generalisation of this problem, which we shall refer to as *optimisation with an information constraint*. The abstract information constraint will be defined using a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$, such that its values $F(y)$ are associated with the values $I(y, y_0)$ of some information resource (or distance) of measure y relative to y_0. A specific form of this functional will lead to a specific family $\{y_\beta\}_x$ of optimal solutions, such as the exponential family (1) if $F(y)$ is associated with the Kullback-Leibler information distance.

The main motivation to study this generalisation was the observation that measures in the exponential family have a remarkable property of being mutually absolutely continuous. We remind that measure y_1 is absolutely continuous with respect to measure y_2 if $y_2(E) = 0$ implies $y_1(E) = 0$ for all $E \in \mathscr{R}(\Omega)$ (here and elsewhere $\mathscr{R}$ denotes a σ-algebra of subsets of Ω). Mutual absolute continuity is the case when the implication holds in both directions. The main question we investigate in this paper is what other families of optimal positive measures have the mutual absolute continuity property.

The answer to this question is related to the properties of the information functional F, and in fact to the properties of its dual functional F^*. In this paper, we prove that it is strict convexity of F^* that makes all optimal positive measures mutually absolutely continuous. We argue also that strict convexity of F^*, the dual of an information functional, is a property that is natural in the context of optimisation problems. Mutual absolute continuity becomes particularly interesting property for optimal measures, defined on set $\Omega = A \times B$ representing a composite system. In this case, the optimal family defines Markov transition kernels between elements of A and B that realise only non-deterministic transitions; deterministic transitions are suboptimal if information, understood broadly here, is bounded. We illustrate this result by constructing an example, where any deterministic kernel either has a negatively infinite expected utility (unbounded error) or communicates infinite information, but a non-deterministic kernel can have both finite expected utility and finite information.

In the next section, we introduce the notation, define the generalised optimisation problem and recall some basic relevant facts. Then we establish several properties of optimal solutions to the problem and use them to prove the main result on mutual absolute continuity of optimal positive measures. The proof is based on standard techniques of convex analysis, and it does not depend on commutativity. Therefore, the result applies to a general, non-commutative setting. The last two sections of the paper are devoted to optimal probability measures and optimal Markov transition kernels. For simplicity, we study them in the classical (commutative) setting. The paper concludes by a discussion of these results.

2 Preliminaries

Let X and Y be complex linear spaces put in duality via bilinear form $\langle\cdot,\cdot\rangle : X\times Y\to\mathbb{C}$:

$$\langle x,y\rangle=0\,,\ \forall x\in X\ \Rightarrow y=0\,,\qquad \langle x,y\rangle=0\,,\ \forall y\in Y\ \Rightarrow x=0$$

The dual space of a locally convex space X will also be denoted by X', and the dual of a normed space $(X,\|\cdot\|)$ will be denoted by X^*. We denote by $X^\sharp$ the algebraic dual space of X. The same notation applies to the dual spaces of Y.

The main results of this paper are derived using only the fact that X and Y are ordered linear spaces in duality. However, in applications, these spaces can have richer algebraic structures [5]. In particular, space X is usually closed with respect to an associative, but generally non-commutative binary operation $\cdot : X\times X\to X$ (e.g. pointwise multiplication or matrix multiplication) and involution as a selfinverse antilinear map $*: X\to X$ reversing the multiplication order, $(x^*z)^*=z^*x$, so that X is a $*$-algebra with the positive cone X_+ of x^*x generating X. The dual space Y is closed under the transposed involution $*: Y\to Y$, defined as $\langle x,y^*\rangle=\langle x^*,y\rangle^*$, has a positive cone Y_+, dual of X_+, and it has identity $y_0\in Y_+$ (also called the reference measure), which is a strictly positive linear functional such that $\langle x^*x,y_0\rangle>0$ for all $x\neq 0\in X$. The ordering $\langle x,y\rangle\geq 0$ is understood as $\mathrm{Re}\langle x,y\rangle\geq 0$ for $\langle x,y\rangle\in\mathbb{C}$. However, we shall mostly deal with Hermitian elements $x=x^*$ and $y=y^*$ such that $\langle x,y\rangle\in\mathbb{R}$. If the pairing $\langle\cdot,\cdot\rangle$ has the property that for each $z\in X$ there exists a transposed element $z'\in Y$ such that $\langle zx,y\rangle=\langle x,z'y\rangle$, then $Y\supset X$ is a left (right) module with respect to the transposed left (right) action $y\mapsto z'y$ ($y\mapsto yz^{*\prime *}$) of X on Y such that $(xz)'=z'x'$ and $\langle x,yz^{*\prime *}\rangle=\langle x^*,z^{*\prime}y^*\rangle^*=\langle z^*x^*,y^*\rangle^*=\langle xz,y\rangle$. In many practical cases, the pairing $\langle\cdot,\cdot\rangle$ is *central* (or *tracial*) so that the left and right transpositions act identically on y_0: $z^{*\prime}y_0=y_0z^{\prime *}$ for all $z\in X$. In this case, the element $y=z^{*\prime}y_0=y_0z^{\prime *}$ can be identified with a complex conjugation of $z\in X$.

Below are three main examples of pairing of X and Y by a sum, an integral or trace:

$$\langle x,y\rangle:=\sum_{\Omega}x(\omega)y(\omega)\,,\qquad \langle x,y\rangle:=\int_{\Omega}x(\omega)\,dy(\omega)\,,\qquad \langle x,y\rangle:=\mathrm{tr}\,\{xy\}\qquad (2)$$

The main examples of X are the commutative C^*-algebra $(\cup C_c(\Omega),\|\cdot\|_\infty)$ of continuous functions with compact support in a locally compact topological space Ω or the non-commutative C^*-algebra $(C_c(\mathscr{H}),\|\cdot\|_\infty)$ of compact Hermitian operators on a separable Hilbert space $\mathscr{H}$. The main examples of $Y=X^*$ are the Banach space $(\mathscr{M}(\Omega),\|\cdot\|_1)$ of Radon measures on Ω or its non-commutative generalisation $(\mathscr{M}(\mathscr{H}),\|\cdot\|_1)$.

Let $F : Y \to \mathbb{R} \cup \{\infty\}$ be a closed functional — sublevel sets $\{y : F(y) \leq \lambda\}$ are non-empty for some and closed in the weak topology $\sigma(Y,X)$ for each $\lambda < \sup F$ (defined in this way, F is also lower-semicontinuous). In this paper, we shall study solutions $y_\beta \in Y$ to an optimisation problem defining the following *optimal value function*:

$$\overline{x}(\lambda) := \sup\{\langle x,y\rangle : F(y) \leq \lambda\} \tag{3}$$

We define $\overline{x}(\lambda) = -\infty$ if $\lambda < \inf F$. It is clear from the definition that $\overline{x}(\lambda)$ is isotone. Set $C := \{y : F(y) \leq \lambda\}$ is the set of feasible solutions, and if it is non-empty, then $\overline{x}(\lambda)$ coincides with its support function $s_C(x) := \sup\{\langle x,y\rangle : y \in C\}$.

Function $\overline{x}(\lambda)$ has the following inverse

$$\overline{x}^{-1}(\upsilon) := \inf\{F(y) : \langle x,y\rangle \geq \upsilon\} \tag{4}$$

In addition to $\overline{x}(\lambda)$ and its inverse $\overline{x}^{-1}(\lambda)$, we shall consider also the following functions:

$$\underline{x}(\lambda) := \inf\{\langle x,y\rangle : F(y) \leq \lambda\} \tag{5}$$

$$\underline{x}^{-1}(\upsilon) := \inf\{F(y) : \langle x,y\rangle \leq \upsilon\} \tag{6}$$

Observe that $\underline{x}(\lambda) = -\overline{(-x)}(\lambda) = -\sup\{-\langle x,y\rangle : F(y) \leq \lambda\}$ is an antitone function, and $\underline{x}(\lambda) = \infty$ if $\lambda < \inf F$.

We use function (3) to represent generally optimisation problems with a constraint on information, and its inverse function (4) to represent generally optimisation problems with a utility constraint. Indeed, consider the case $X = (\cup C_c(\Omega), \|\cdot\|_\infty)$ and $Y = (\mathscr{M}(\Omega), \|\cdot\|_1)$. Then probability measures on Borel σ-algebra $\mathscr{R}(\Omega)$ are positive elements $p \in \mathscr{M}(\Omega)$ with $\|p\|_1 = 1$. We shall refer to the set of all probability measures

$$\mathscr{P}(\Omega) := \{y \in \mathscr{M}(\Omega) : y > 0\,,\ \|y\|_1 = 1\}$$

as *statistical manifold* by analogy with information geometry [1,8,18]. In the classical probability theory (X is a commutative algebra), set $\mathscr{P}$ is a Choquet simplex — a compact convex set such that every $p \in \mathscr{P}$ is uniquely represented by the extreme points $\delta \in \partial\mathscr{P}$ [17]. Here, $\partial\mathscr{P}$ denotes the boundary of $\mathscr{P}$, and we shall denote by $\operatorname{ext}\mathscr{P} \subseteq \partial\mathscr{P}$ the set of all extreme points of $\mathscr{P}$. Thus, in the classical probability theory, we can identify $\operatorname{ext}\mathscr{P}$ with Ω.

In the non-classical probability (X is non-commutative), a similar construction can be made. For example, if $X = (C_c(\mathscr{H}), \|\cdot\|_\infty)$ and $Y = (\mathscr{M}(\mathscr{H}), \|\cdot\|_1)$, then quantum probability measures, representing states, are positive elements $p \in \mathscr{M}(\mathscr{H})$ with $\|p\|_1 = 1$, and the *quantum statistical manifold* is

$$\mathscr{P}(\mathscr{H}) := \{y \in \mathscr{M}(\mathscr{H}) : y > 0\,,\ \|y\|_1 = 1\}$$

The quantum statistical manifold is also compact and convex, but it is not a simplex. This is because representations by $\delta \in \operatorname{ext}\mathscr{P}$ are not unique. This fundamental difference of quantum probability is the consequence of non-commutativity. However, our results will apply to both classical and non-classical case, which have many similarities.

Observe that the expected value $\mathbb{E}_p\{x\}$ of a classical random variable $x : \Omega \to \mathbb{R}$ is a linear functional $p(x) = \langle x, p\rangle$, $p \in \mathscr{P}$. Note that often we can consider x as an element of space Y', dual of Y, and $\mathbb{E}_p\{x\}$ as a linear functional $x(p) = \langle x, p\rangle$, where p is varied over some subset of $\mathscr{P}$, as in function (3). In quantum physics, operator $x \in Y'$ is often called an *observable*, and the linear functional $\mathbb{E}_p\{x\} = \langle x, p\rangle$ takes values in the spectrum $\sigma(x)$, which is real, if x is Hermitian.

If $x \in X$ is a classical utility function $x : \Omega \to \mathbb{R}$, then maximisation of linear functional $x(p) = \langle x, p\rangle$ is the problem of maximisation of the expected utility. We remind that given a *preference relation* $\lesssim$ on Ω (a total pre-order), a *utility* function is a pre-order embedding $x : (\Omega \lesssim) \to (\mathbb{R}, \leq)$: $\omega_1 \lesssim \omega_2$ if and only if $x(\omega_1) \leq x(\omega_2)$. A non-classical utility operator is defined similarly using pre-order on its eigenstates (see [4]). It is well-known that the expected utility $\mathbb{E}_p\{x\}$ (linear functional $\langle x, y\rangle$) is the only functional that makes statistical manifold $\mathscr{P}$ (linear space Y) totally pre-ordered, and such that $(\mathscr{P}, \lesssim) \subset (Y, \lesssim)$ is compatible with the linear structure of Y and is an Archimedian pre-order [16].

It is clear from the above that the optimal value function (3) corresponds to optimisation under uncertainty over the set of probability measures $\mathscr{P} \cap C$, where C is defined by the constraint $F(y) \leq \lambda$. In particular, if $F(y) := I(y, y_0)$, where $I : Y \times Y \to \mathbb{R}_+ \cup \{\infty\}$ is some information distance, then $F(y) \leq \lambda$ is the constraint $I(y, y_0) \leq \lambda$ on information distance. Note that without the constraint, $x(p)$ is always maximised at least in one of the extreme points $\delta_x \in \operatorname{ext} \mathscr{P}$.

Proposition 1. *Let $\mathscr{P}$ be a non-empty compact convex subset of a locally convex space Y. Then for any $x \in X \subseteq Y^\sharp$ there exists $\delta_x \in \operatorname{ext} \mathscr{P}$ such that*

$$\langle x, \delta_x\rangle := \sup\{\langle x, p\rangle : p \in \mathscr{P}\} \in \mathbb{R} \cup \{\infty\} \tag{7}$$

Proof. If there exist a non-empty subset $\Delta \subseteq \operatorname{ext} \mathscr{P}$ of extreme points δ_x such that $\langle x, \delta_x\rangle = \sup\{\langle x, p\rangle : p \in \mathscr{P}\}$ for any $\delta_x \in \Delta$, then by linearity $\langle x, \delta_x\rangle = \langle x, \bar{p}\rangle$ for any $\bar{p} \in \operatorname{cl} \operatorname{co} \Delta$ (here $\operatorname{cl} \operatorname{co}$ denotes convex closure of a set). Dually, the set Δ is empty only if there exist $\bar{p}$ such that $\langle x, \bar{p}\rangle > \langle x, p\rangle$ for all $p \in \operatorname{cl} \operatorname{co} \operatorname{ext} \mathscr{P}$. But $\operatorname{cl} \operatorname{co} \operatorname{ext} \mathscr{P} = \mathscr{P}$ (Krein-Milman theorem), and therefore $\bar{p} \notin \mathscr{P}$. □

In many practical applications, finding the optimal extreme points δ_x solving optimisation problems may not be feasible as it may require a large (possibly infinite) amount of information. A feasible solution can be found by taking into account the information constraint.

Definition 1 (Information (feasibility) constraint). *A value λ of a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$ is called an* information *or* feasibility constraint *in problem (3), if $\bar{x}(\lambda) < \langle x, \delta_x\rangle$, where δ_x is defined in Proposition 1. Generally,* $\inf F \leq \lambda < F(\delta_x) \leq \sup F$.

It will be shown later that if feasible solutions to problem (3) exist, then they are also solutions to problem (4), defined by the inverse function $\bar{x}^{-1}(\upsilon)$. One often seeks non-trivial solutions $p \in \mathscr{P}$ to optimisation problems such that the expected utility is greater than that of a solution requiring no information (i.e. a trivial solution). Non-trivial solutions can be found by taking into account a utility constraint.

Definition 2 (Utility (non-triviality) constraint). *A value υ of a linear functional $x : Y \to \mathbb{R}$ is called an expected* utility *or* non-triviality constraint *in problem (4), if $\overline{x}^{-1}(\upsilon) > \inf F$. Generally, $\upsilon > \overline{\upsilon}_0$, where*

$$\overline{\upsilon}_0 := \lim_{\lambda \downarrow \inf F} \sup\{\langle x, y\rangle : F(y) \leq \lambda\} \in \mathbb{R} \cup \{-\infty\} \tag{8}$$

Remark 1. One can show in a way similar to Proposition 1 and using the equality $\underline{x}(\lambda) = -\overline{(-x)}(\lambda)$ that there exists $\delta_{-x} \in \operatorname{ext} \mathscr{P}$ such that

$$\langle x, \delta_{-x}\rangle := \inf\{\langle x, y\rangle : y \in \mathscr{P}\} \in \mathbb{R} \cup \{-\infty\} \tag{9}$$

An information constraint in problem (5) is such that $\underline{x}(\lambda) > \langle x, \delta_{-x}\rangle$, and generally $\inf F \leq \lambda < F(\delta_{-x}) \leq \sup F$. A utility constraint in problem (6) is such that $\underline{x}^{-1}(\upsilon) > \inf F$, and generally $\upsilon < \underline{\upsilon}_0$, where

$$\underline{\upsilon}_0 := \lim_{\lambda \downarrow \inf F} \inf\{\langle x, y\rangle : F(y) \leq \lambda\} \in \mathbb{R} \cup \{\infty\} \tag{10}$$

Note that often $\langle x, \delta_x\rangle \neq -\langle x, \delta_{-x}\rangle$. Indeed, if x is a real function on Ω, then $\langle x, \delta_x\rangle = \sup x(\omega)$ and $\langle x, \delta_{-x}\rangle = \inf x(\omega)$. Furthermore, generally $F(\delta_x) \neq F(\delta_{-x})$ and $\overline{\upsilon}_0 \neq -\underline{\upsilon}_0$.

Problems (3) and (4), considered on the statistical manifold $\mathscr{P}$, generalise several related variational problems in information theory and physics, in which $F(p)$ corresponds to the Kullback-Leibler information distance $I_{KL}(p,q) := \mathbb{E}_p\{\ln p - \ln q\}$ of probability measure p from a reference measure q. An important example is when $I_{KL}(p,q)$ is Shannon information between random variables $a \in A$ and $b \in B$, which is defined as $I_{KL}(p,q)$, where $p = p(A \mid b)$ is the conditional and $q = p(A)$ is the marginal probability. Function (3) in this case defines the *value of Shannon information*, which was introduced and studied by Stratonovich [22,23]. The general form of problems (3) or (4) allows us to study families of optimal solutions independent of the way information distance $I(y, y_0)$ or functional $F(y)$ is defined.

We shall study the question of existence of feasible and non-trivial solutions to problem $\overline{x}(\lambda)$, but not necessarily to $\overline{(-x)}(\lambda)$. Because solutions may exist even for unbounded linear functionals $x \in Y^\sharp$, we shall refer to such $x \in Y^\sharp$ as *information bounded* or simply as *F-bounded* elements.

Definition 3 (F-bounded linear functional). *An element $x \in Y^\sharp$ is bounded relative to a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$ or F-*bounded *if $\overline{x}(\lambda) \in \mathbb{R}$ for each $\lambda \in (\inf F, F(\delta_x))$.*

Topological questions will not be addressed in this paper. However, the following should be noted about the space of all F-bounded elements. Information distance $I(y, y_0)$ or functional $F(y) = I(y, y_0)$ can be used to define a topology on Y (and $\mathscr{P} \subset Y$), in which the collection of sublevel sets $C := \{y : F(y) \leq \lambda\}$ of feasible and non-trivial solutions is the base of closed neighbourhoods of y_0, $\inf F = F(y_0)$ (e.g. see [3]). The support function $s_C(x) := \sup\{\langle x, y\rangle : y \in C\}$, which is a generalisation of a seminorm, defines a topology on $X \subseteq Y^\sharp$ that is compatible with duality. Because $s_C(x) = \overline{x}(\lambda)$, this topological space is precisely the space of all F-bounded elements. Such topological spaces, however, are generally not topological vector spaces, because sets C can be 'unbalanced' if $I(y, y_0) \neq I(y_0, y)$ (or $F(y - y_0) \neq F(y_0 - y)$), and therefore $s_C(x) \neq s_C(-x)$. Thus, the topologies on Y and X, generated respectively by the information distance $I(y, y_0)$ and support function $s_C(x)$, are different from the norm topologies $(Y, \|\cdot\|)$ and $(Y^*, \|\cdot\|^*)$. In particular, Y^* may contain elements x or $-x$ that are not F-bounded, so that solutions to problem $\overline{x}(\lambda)$ or $\overline{(-x)}(\lambda)$ may not exist, and sets C can be unbounded in $(Y, \|\cdot\|)$. On the other hand, there can be F-bounded elements outside Y^*. These facts will be illustrated on an example later.

In the next section, we show that solutions y_β, if exist, are the elements of subdifferential of functional F^*, dual of F. We remind that $F^* : X \to \mathbb{R} \cup \{\infty\}$ is the Legendre-Fenchel transform of F:

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

and it is aways closed and convex (e.g. see [20,25]). Condition $F^{**} = F$ implies F is closed and convex. Otherwise, the epigraph of F^{**} is a convex closure of the epigraph of F in $Y \times \mathbb{R}$. Closed and convex functionals are continuous on the (algebraic) interior of the effective domain $\operatorname{dom} F := \{y : F(y) < \infty\}$, and they have the property

$$x \in \partial F(y) \quad \iff \quad \partial F^*(x) \ni y \tag{11}$$

where set $\partial F(y_0) := \{x : F(y) \geq F(y_0) + \langle x, y - y_0\rangle,\ \forall y \in Y\}$ is *subdifferential* of F at y_0, and its elements are called *subgradients*. In particular, $0 \in \partial F(y_0)$ implies $F(y_0) \leq F(y)$ for all y (i.e. $\inf F = F(y_0)$). If F is weakly (Gâteaux) differentiable (or if F^* is strictly convex), then $\partial F(y) = \{x\}$, and the correspondence $y \mapsto x \in \partial F(y)$ is a function.

Recall also that subgradients satisfy the following monotonicity condition [11]:

$$\langle x_1 - x_2, y_1 - y_2\rangle \geq 0\,, \quad \forall y_i \in \partial F^*(x_i) \tag{12}$$

If the inequality is strict for all $x_1 \neq x_2$, then ∂F^* is strictly monotone, and F^* is strictly convex.

We remind also that $H : Y \to \mathbb{R} \cup \{-\infty\}$ is *concave* if $F(y) = -H(y)$ is convex. By analogy, one defines *supdifferential* of concave function [20], and the correspondence $x \mapsto y \in \partial H^*$ is antitone. The dual of H in concave sense is $H^*(x) := \inf\{\langle x, y\rangle - H(y)\}$.

3 General properties of optimal solutions and the optimal value function

In this section, we study general properties of the optimal value function (3) and optimal feasible solutions — elements y_β such that $\overline{x}(\lambda) = \langle x, y_\beta\rangle = \upsilon < \infty$. First, we apply the standard method of Lagrange multipliers to derive solutions y_β to problem (3).

Proposition 2 (Necessary and sufficient optimality conditions). *Element $y_\beta \in Y$ solves problem (3) with closed $F : Y \to \mathbb{R} \cup \{\infty\}$ if and only if the following conditions hold*

$$y_\beta \in \partial F^*(\beta x)\,, \quad F(y_\beta) = \lambda\,, \quad \beta^{-1} \in \partial \overline{x}(\lambda)\,, \quad \beta^{-1} > 0$$

Proof. If the solution y_β to problem (3) exists, then it belongs to the boundary of the sublevel set $C = \{y : F(y) \leq \lambda\}$, because $\langle x, \cdot\rangle$ is linear and the sublevel set is closed. Moreover, y_β belongs to the boundary of a convex closure of set C in Y, because it is the intersection of all closed half-spaces $\{y : \langle x,y\rangle \leq \langle x,y_\beta\rangle\}$ containing C. Observe also that

$$\operatorname{cl}\operatorname{co}\{y : F(y) \leq \lambda\} = \{y : F^{**}(y) \leq \lambda\}$$

and therefore solutions satisfy condition $F(y_\beta) = F^{**}(y_\beta)$. The latter implies also $\partial F(y_\beta) = \partial F^{**}(y_\beta)$ (e.g. see [20], Theorem 12). Thus, the Lagrange function for the conditional extremum in (3) can be written in terms of F^{**} as follows

$$K(y,\beta^{-1}) = \langle x,y\rangle + \beta^{-1}[\lambda - F^{**}(y)]\,,$$

where β^{-1} is the Lagrange multiplier for the constraint $\lambda \in (\inf F, F(\delta_x))$. Because $x(y) = \langle x,y\rangle$ is linear and F^{**} is convex, the Lagrange function is concave for $\beta^{-1} > 0$. In this case, condition $\partial K(y_\beta,\beta^{-1}) \ni 0$ is both necessary and sufficient for y_β and β^{-1} to define its least upper bound, which gives

$$\begin{aligned} \partial_y K(y_\beta,\beta^{-1}) &= x - \beta^{-1}\partial F^{**}(y_\beta) \ni 0\,, \quad &\Rightarrow \quad & y_\beta \in \partial F^*(\beta x) \\ \partial_{\beta^{-1}} K(y_\beta,\beta^{-1}) &= \lambda - F^{**}(y_\beta) \ni 0\,, \quad &\Rightarrow \quad & F^{**}(y_\beta) = \lambda \end{aligned}$$

Note that if $F \neq F^{**}$, then generally $F^{**}(y) \leq F(y)$, and condition $F^{**}(y_\beta) = \lambda$ must be replaced by the stronger condition $F(y_\beta) = \lambda$. Noting that $\overline{x}(\lambda) = \langle x,y_\beta\rangle + \beta^{-1}[\lambda - F(y_\beta)]$, the Lagrange multiplier is defined by $\partial\overline{x}(\lambda) \ni \beta^{-1}$. Note that $\partial\overline{x}(\lambda) \geq 0$, because $\overline{x}(\lambda)$ is isotone (non-decreasing), and $\beta^{-1} = 0$ if and only if $\lambda = F(\delta_x)$. □

Remark 2. Solutions to problem (4), defining the inverse function $\overline{x}^{-1}(\upsilon)$, are given by similar conditions. Indeed, the corresponding Lagrange function is

$$K(y,\beta) = F^{**}(y) + \beta[\upsilon - \langle x,y\rangle]$$

and the corresponding necessary and sufficient conditions are

$$y_\beta \in \partial F^*(\beta x)\,, \quad \langle x,y_\beta\rangle = \upsilon\,, \quad \beta \in \partial\overline{x}^{-1}(\upsilon)\,, \quad \beta > 0$$

Function $\underline{x}(\lambda)$, defined by equation (5), is antitone, because $\underline{x}(\lambda) = -\overline{(-x)}(\lambda)$. The necessary and sufficient conditions for the infimum in $\underline{x}(\lambda)$ are identical to those in Proposition 2 with the only exception that $\beta^{-1} < 0$. Similarly, conditions defining the infimum in $\underline{x}^{-1}(\upsilon)$ are identical those of $\overline{x}^{-1}(\upsilon)$, given above, but with $\beta < 0$.

Remark 3. If there exist $y_0 \in \operatorname{dom} F$ such that $\inf F = F(y_0)$ (i.e. $0 \in \partial F(y_0)$), then $\overline{\upsilon}_0 = \sup\{\langle x,y\rangle : y \in \partial F^*(0)\}$ and $\underline{\upsilon}_0 = \inf\{\langle x,y\rangle : y \in \partial F^*(0)\}$. If y_0 is unique (i.e. $\partial F^*(0) = \{y_0\}$), then $\overline{\upsilon}_0 = \underline{\upsilon}_0$; otherwise, $\overline{\upsilon}_0 \geq \underline{\upsilon}_0$.

In previous section, we defined an F-bounded linear functional x that admits solutions to problem (3) or (4) for each constraint $\lambda \in (\inf F, F(\delta_x))$. It was mentioned also that solutions may not exist even for some $x \in Y^*$, so that the sets $\{y : F(y) \leq \lambda\}$ are unbounded in the norm topology $(Y, \|\cdot\|)$. If they are bounded, then $0 \in \operatorname{Int}(\operatorname{dom} F^*)$ (see [2,15]). Thus, the concept of an F-bounded element allows us to consider such $x \in Y^\sharp$ that the origin of a one-dimensional subspace $\mathbb{R}x := \{\beta x : \beta \in \mathbb{R}\}$ is not on the interior of $\operatorname{dom} F^*$. Also, condition $\overline{x}(\lambda) \in \mathbb{R}$ does not imply $\overline{(-x)}(\lambda) \in \mathbb{R}$, so that x is F-bounded, but $-x$ is not. Furthermore, an F-bounded x can be unbounded relative to a norm $\|\cdot\|$ on Y, and therefore it can be outside Banach space Y^* (in fact, Y^* is precisely the space of $\|\cdot\|$-bounded elements). For an illustration, consider the following example.

Example 1. Let $\Omega = \mathbb{N}$ and let X, Y be the spaces of real sequences $\{x(n)\}$ and $\{y(n)\}$ with pairing $\langle\cdot,\cdot\rangle$ defined by the sum (2). Let $F(y) = \langle \ln y - 1, y\rangle$ for $y > 0$, so that the gradient $\nabla F(y) = \ln y$, and F is minimised at the counting measure $y_0(n) = 1$. The optimal solutions have the form $y_\beta = e^{\beta x}$, and the optimal value functions $\overline{x}(\lambda)$ and $\overline{(-x)}(\lambda)$ are respectively

$$\langle x, y_\beta\rangle = \sum_{n=1}^{\infty} x(n)\, e^{\beta x(n)} \quad \text{and} \quad -\langle x, y_\beta\rangle = -\sum_{n=1}^{\infty} x(n)\, e^{-\beta x(n)}\,, \qquad \beta^{-1} > 0$$

In particular, for $x(n) = -n$, the first series converges to $-e^{\beta}(e^{\beta} - 1)^{-2}$, but the second diverges for any $\beta^{-1} > 0$. Thus, x is F-bounded, but $-x$ is not. Observe also that both x and $-x$ are unbounded relative to the norm $\|\cdot\|_1$ on Y, because there is no real number $\|x\|_\infty := \sup\{|\langle x, y\rangle| : \|y\|_1 \leq 1\} = \sup_n\{x(n), -x(n)\}$ for such x. On the other hand, any constant sequence $x(n) = \alpha$, where $\alpha \in (0, \infty)$, is bounded, but it is not F-bounded.

The criteria for an F-bounded element $x \in X$ follow from the optimality conditions, obtained in Proposition 2.

Proposition 3 (Existence of solutions). *Solutions $y_\beta \in Y$ maximising $x(y) = \langle x, y\rangle$ on closed sets $\{y : F(y) \leq \lambda\}$ exist for each $\lambda \in (\inf F, F(\delta_x))$, where $F : Y \to \mathbb{R} \cup \{\infty\}$ is a closed functional, if and only if there exists at least one number $\beta^{-1} > 0$ such that $F^*(\beta x) < \sup F^*$. In other words, $x \in X$ is F-bounded if and only if it is absorbed by the set $\{w : F^*(w) \leq \lambda^*\}$ for some $\lambda^* \in (\inf F^*, \sup F^*)$.*

Proof. ($\Rightarrow$) Assume there exists number $\beta^{-1} > 0$ such that $F^*(\beta x) \in (\inf F^*, \sup F^*)$. Then there exists $y_\beta \in \operatorname{dom} F^{**}$ such that $F^*(\beta x) = \beta\langle x, y_\beta\rangle - F^{**}(y_\beta) \geq \beta\langle x, y_\beta\rangle - F(y_\beta)$. In fact, solutions to problem (3) are y_β such that $F(y_\beta) = F^{**}(y_\beta)$ and $y_\beta \in \partial F^*(\beta x)$ (Property (11)), and therefore

$$\langle x, y_\beta\rangle = \beta^{-1}\left[F^*(\beta x) + F^{**}(y_\beta)\right] \in \mathbb{R}$$

Therefore, $y_\beta \in \partial F^*(\beta x)$ solve problem (3) for $\lambda = \beta\langle x, y_\beta\rangle - F^*(\beta x) \in (\inf F, F(\delta_x))$.

($\Leftarrow$) Assume there exists a solution y_β to problem (3) for $\lambda \in (\inf F, F(\delta_x))$. Then $y_\beta \in \partial F^*(\beta x)$ (Proposition 2), where $0 < \beta^{-1} < \infty$ (otherwise, $\lambda = F(\delta_x) \leq \sup F$ or $\lambda = \inf F$).

The existence of solution y_β implies that sublevel set is bounded by closed half-space $\{y : \langle x, y\rangle \le \langle x, y_\beta\rangle\}$. Because F is closed, its sublevel sets $\{y : F(y) \le \lambda\}$ are closed for all λ, and therefore the existence of a solution for one such λ implies that solutions exist for all $\lambda \in (\inf F, F(\delta_x))$. The converse is true and trivial.

Observe also that βx is on the boundary of the closed convex set $\{w : F^*(w) \le F^*(\beta x)\}$, which is bounded by the closed half-space $\{w : \langle w, y_\beta\rangle \le \beta\langle x, y_\beta\rangle\}$. In particular, elements $\beta_1 x$ and $\beta_2 x$ for $\beta_1 < \beta < \beta_2$ are respectively on the interior and on the exterior of this set, which is equivalent to x being absorbed by the set. □

Proposition 4 (Monotonicity). *Functions $\overline{x}(\lambda)$, $\underline{x}(\lambda)$, $\overline{x}^{-1}(\upsilon)$ and $\underline{x}^{-1}(\upsilon)$, defined by equations (3), (5), (4) and (6) for a closed $F : Y \to \mathbb{R} \cup \{\infty\}$ and $x \neq 0$, have the following properties:*

1. *The mappings $\lambda \mapsto \beta$, $\beta^{-1} \in \partial\overline{x}(\lambda)$, and $\upsilon \mapsto \beta \in \partial\overline{x}^{-1}(\upsilon)$ are isotone.*
2. *If in addition F^* is strictly convex on $\operatorname{dom} F^*$, then these mappings are continuous.*
3. *$\overline{x}(\lambda)$ is concave and strictly increasing for $\lambda \le F(\delta_x)$.*
4. *$\underline{x}(\lambda)$ is convex and strictly decreasing for $\lambda \le F(\delta_{-x})$.*
5. *$\overline{x}^{-1}(\upsilon)$ is convex and strictly increasing for $\upsilon \ge \overline{\upsilon}_0$.*
6. *$\underline{x}^{-1}(\upsilon)$ is convex and strictly decreasing for $\upsilon \le \underline{\upsilon}_0$.*

where δ_x, δ_{-x}, $\overline{\upsilon}_0$ and $\underline{\upsilon}_0$ are defined by equations (7), (9), (8) and (10) respectively.

Proof.
1. Let y_{β_1}, y_{β_2} be two solutions to problem (3) with constraints $\lambda_1 \le \lambda_2$ respectively, and let $\upsilon_1 = \langle x, y_{\beta_1}\rangle$ and $\upsilon_2 = \langle x, y_{\beta_2}\rangle$. Using condition $y_\beta \in \partial F^*(\beta x)$ of Proposition 2 and monotonicity condition (12) for convex F^*, we have
$$\langle \beta_2 x - \beta_1 x, y_{\beta_2} - y_{\beta_1}\rangle = (\beta_2 - \beta_1)\langle x, y_{\beta_2} - y_{\beta_1}\rangle \ge 0$$
Function $\overline{x}(\lambda)$ is isotone (by the inclusion $\{y : F(y) \le \lambda_1\} \subseteq \{y : F(y) \le \lambda_2\}$), and therefore $\lambda_1 \le \lambda_2$ implies $\langle x, y_{\beta_2} - y_{\beta_1}\rangle = \upsilon_2 - \upsilon_1 \ge 0$. It follows from the inequality above that $\lambda_1 \le \lambda_2$ (or $\upsilon_1 \le \upsilon_2$) implies $\beta_1 \le \beta_2$, which proves that $\lambda \mapsto \beta$ and $\upsilon \mapsto \beta$ are isotone.
2. Optimality condition $y_\beta \in \partial F^*(\beta x)$ is equivalent to $\beta x \in \partial F^{**}(y_\beta)$ by property (11), and together with condition $F(y_\beta) = \lambda$ (or $\langle x, y_\beta\rangle = \upsilon$) it implies that different $\beta_1 < \beta_2$ can correspond to the same λ (υ) if and only if $\partial F^{**}(y_\beta)$ includes both $\beta_1 x$ and $\beta_2 x$. This implies that F^* is not strictly convex on $[\beta_1 x, \beta_2 x] \subseteq \partial F^{**}(y_\beta)$. Conversely, if F^* is strictly convex on $\operatorname{dom} F^*$, then $\beta_1 \neq \beta_2$ implies $\lambda_1 \neq \lambda_2$ ($\upsilon_1 \neq \upsilon_2$). Therefore, $\overline{x}(\lambda)$ ($\overline{x}^{-1}(\upsilon)$) is a differentiable real function, and its derivative is continuous.
3. Function $\overline{x}(\lambda)$ is strictly increasing, because $\partial\overline{x}(\lambda) \ni \beta^{-1} > 0$ if $\lambda < F(\delta_x)$, and $\beta^{-1} = 0$ if and only if $\lambda \ge F(\delta_x)$ (Proposition 2). Moreover, the mapping $\lambda \mapsto \beta^{-1} \in \partial\overline{x}(\lambda)$ is antitone (because $\lambda \mapsto \beta$ is isotone), and therefore $\overline{x}(\lambda)$ is concave.
4. By the same reasoning as above, function $\overline{(-x)}(\lambda)$ is concave and strictly increasing for $\lambda \le F(\delta_{-x})$. Thus, $\underline{x}(\lambda) = -\overline{(-x)}(\lambda)$ is convex and strictly decreasing.
5. Function $\overline{x}^{-1}(\upsilon)$ is strictly increasing for $\overline{\upsilon}_0 \le \upsilon$, because $\partial\overline{x}^{-1}(\upsilon) \ni \beta > 0$, and $\beta = 0$ if and only if $\upsilon = \langle x, y_0\rangle \le \overline{\upsilon}_0$ for any $y_0 \in \partial F^*(0)$ ($\inf F = F(y_0)$). Moreover, the mapping $\upsilon \mapsto \beta \in \partial\overline{x}^{-1}(\upsilon)$ is isotone, and therefore $\overline{x}^{-1}(\upsilon)$ is convex.

6. Function $\underline{x}^{-1}(\upsilon)$ is the inverse of convex and strictly decreasing function $\underline{x}(\lambda)$. Thus, $\underline{x}^{-1}(\upsilon)$ is also convex and strictly decreasing for $\upsilon \leq \underline{\upsilon}_0$.

□

Remark 4 (Strict convexity). Classical information distances between probability measures are often required to satisfy the additivity axiom: $I(yz, y_0) = I(y, y_0) + I(z, y_0)$ [8]. This is why such information distances are represented using a logarithmic function, and functional F^*, dual of $F(y) = I(y, y_0)$, is represented using an exponential function, and it is strictly convex. If F^* is not strictly convex, then there may exist different quantities $\beta_1 x \neq \beta_2 x$ corresponding to the same value $\lambda = F(y_\beta)$ (or $\upsilon = \langle x, y_\beta \rangle$). If x is understood as the objective function of an optimisation problem (e.g. a utility), then without strict convexity of F^*, the information functional cannot 'distinguish' between some quantities of x. Thus, the requirement for F^* to be strictly convex is natural in the context of optimisation problems.

To distinguish between positive and negative elements, we equip spaces X and Y with order relations $\leq$ in a usual way. Let $X_+ \subset X$ be a pointed convex cone of non-negative elements in X so that $w \leq x$ if and only if $x - w \in X_+$. We also demand that X_+ is reproducing: $X_+ - X_+ = X$ or

$$x = x_+ - x_- \,, \quad x_+, x_- \in X_+ \,,\ \forall x \in X$$

For example, if X is a function space, then X_+ is the set of positive functions with respect to the natural pointwise order. If X is the space of operators on a Hilbert space, then X_+ is the cone of elements $x^* x \in X$. The order on Y is induced by the dual cone:

$$Y_+ := \{y \in Y : \langle x, y \rangle \geq 0 \,,\ \forall x \geq 0\}$$

Proposition 5 (Zero solution). *If solutions y_β to problem (3) for all values λ of a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$ are non-negative (i.e. $y_\beta \in Y_+$ for all $\lambda = F(y)$) and $y_\beta = 0$ for some λ, then*

$$x = 0 \quad \textit{or} \quad \inf F = F(0) \quad \textit{or} \quad F(\delta_x) = F(0)$$

Proof. Assume the opposite: $x \neq 0$ and $\inf F < F(0) < F(\delta_x)$. Then function $\overline{x}(\lambda) = \langle x, y_\beta \rangle$ is strictly increasing (Proposition 4), and sets $\{y : F(y) < F(0)\}$ and $\{y : F(0) < F(y)\}$ are non-empty (F is closed). Thus, there exist solutions y_1 and y_2 such that

$$F(y_1) < F(0) < F(y_2) \quad \text{and} \quad \langle x, y_1 \rangle < 0 < \langle x, y_2 \rangle$$

Using decomposition $x = x_+ - x_-$, x_+, $x_- \in X_+$ and y_1, $y_2 \in Y_+$, we conclude that

$$\begin{aligned} \langle x_+ - x_-, y_1 \rangle < 0 \quad &\Rightarrow \quad x_+ < x_- \quad (x_- - x_+ \in X_+) \\ \langle x_+ - x_-, y_2 \rangle > 0 \quad &\Rightarrow \quad x_+ > x_- \quad (x_+ - x_- \in X_+) \end{aligned}$$

This implies $x = 0$, which is a contradiction. □

4 Main result: Mutual absolute continuity

Our interest is in the support set of optimal positive measures solving problem (3). We remind that X is a linear algebra, which can be associated with the algebra $\mathscr{R}(\Omega)$ of subsets of Ω in the classical (commutative) setting, or with the algebra $\mathscr{R}(\mathscr{H})$ of subspaces of $\mathscr{H}$ in non-classical (non-commutative) setting. A subalgebra $\mathscr{R}(E)$ of subset $E \subset \Omega$ or subspace $E \subset \mathscr{H}$ corresponds in each case to a subspace $M \subset X$. Thus, we shall treat these cases generally by defining a continuous linear projection $P_M : X \to M \subset X$ and using notation $y(M) = 0$ to denote measures that are zero on subset or subspace E.

We remind also that if Y is the dual of X, then the dual of subspace $M \subset X$ is the factor space $Y/M^{\perp}$ of equivalence classes $[y] := \{z \in Y : y - z \in M^{\perp}\}$ generated by the annihilator $M^{\perp} := \{y \in Y : \langle x, y\rangle = 0\,, \forall x \in M\}$. Thus, the elements of $Y/M^{\perp}$ correspond to measures that are equivalent on M. In particular, $[0] \in Y/M^{\perp}$ is the annihilator $M^{\perp}$, and it is a subspace of Y corresponding to measures such that $y(M) = 0$. The restriction of F^* to M is given by $F^*(P_M x)$, and the dual of $F^*(P_M x)$ is defined on $Y/M^{\perp}$ as $F^{**}([y]) := \inf\{F^{**}(y) : y \in [y]\}$.

Theorem 1 (Mutual absolute continuity). *Let $\{y_\beta\}_x \subset Y_+$ be a family of non-negative linear functionals on X that are solutions to problem (3) for all values λ of a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$. If F^*, the dual of F, is strictly convex for all $x \in \operatorname{dom} F^*$, then:*

1. *There is a subfamily $\{y^\circ_\beta\}_x \subseteq \{y_\beta\}_x$ containing y°_β for each $\lambda \in (\inf F, F(\delta_x))$, and y°_β correspond to mutually absolutely continuous positive measures.*
2. *If $0 \in \operatorname{dom} F^*$ ($\operatorname{dom} F^{**}$ is closed), then there exist y_0 (δ_x) in $\{y_\beta\}_x$ such that $\inf F = F(y_0)$ ($\sup\{\langle x, y\rangle : y \in \operatorname{dom} F\} = \langle x, \delta_x\rangle$), and it is absolutely continuous w.r.t. all y°_β.*
3. *If in addition F^{**} is strictly convex for all $y \in \operatorname{dom} F^{**}$, then $\{y^\circ_\beta\}_x = \{y_\beta\}_x \setminus \{y_0, \delta_x\}$.*

Proof. Let y_β be a solution for some $\lambda \in (\inf F, F(\delta_x))$. Then $y_\beta \in \partial F^*(\beta x)$, $0 < \beta^{-1} < \infty$ (Proposition 2). Let $P_M : X \to M$ be a continuous linear projection onto subspace $M \subset X$. Then $[y_\beta] \in \partial F^*(\beta P_M x)$, $[y_\beta] \in Y/M^{\perp}$. Assume that the corresponding measure $y_\beta(M) = 0$. Then $y_\beta \in [0] \in Y/M^{\perp}$, where $[0] = M^{\perp}$, and because $[y_\beta] \geq 0$ (P_M is a positive operator), $[y_\beta] = [0]$ implies by Proposition 5

$$P_M x = 0 \quad \text{or} \quad \inf F^{**} = F^{**}([0]) \quad \text{or} \quad F^{**}([\delta_x]) = F^{**}([0])$$

Observe that $\partial F^{**}([0])$ is a singleton set, because F^* (and hence $F^*(P_M x)$) is strictly convex on $\operatorname{dom} F^*$. Therefore, the latter two cases above are false, because otherwise $\partial F^{**}([0])$ would contain the intervals $[0, \beta P_M x]$ or $[\beta P_M x, \infty)$, $0 < \beta < \infty$. Thus, the only true case is $P_M x = 0$. But then $\beta P_M x = 0$ for all β, and therefore

$$[0] \in \partial F^*(\beta P_M x)\,, \quad \forall \beta \in \mathbb{R}$$

In other words, for each $\lambda \in (\inf F, F(\delta_x))$, there is a solution y_β, such that the corresponding measure $y_\beta(M) = 0$.

These measures are not mutually absolutely continuous only if there exists solution y_β° for some $\lambda \in (\inf F, F(\delta_x))$ such that the corresponding measure $y_\beta^\circ(M') = 0$ on some larger subspace $M' \supset M$. The subfamily $\{y_\beta^\circ\}_x \subseteq \{y_\beta\}_x$ corresponding to mutually absolutely continuous measures for all $\lambda \in (\inf F, F(\delta_x))$ is constructed by taking

$$M = \sup\{M' \subset X : \exists y_\beta^\circ \in \{y_\beta\}_x,\ y_\beta^\circ(M') = 0\}$$

where supremum is with respect to ordering by inclusion.

If $0 \in \operatorname{dom} F^*$ ($\operatorname{dom} F$ is closed), then $\inf F$ ($\sup\{\langle x, y\rangle : y \in \operatorname{dom} F\}$) is attained at some y_0 (δ_x) corresponding to $\beta = 0$ ($\beta^{-1} = 0$). Because $P_M x = 0$ implies that $\beta P_M x = 0$ for $\beta = 0$ ($\beta^{-1} = 0$), the measure corresponding to y_0 (δ_x) is absolutely continuous with respect to all measures in $\{y_\beta^\circ\}$.

If F^{**} is strictly convex on $\operatorname{dom} F^{**}$, then $\partial F^*(\beta x)$ contains unique element y_β° for each $\beta^{-1} > 0$, and $\{y_\beta^\circ\}_x = \{y_\beta\}_x \setminus \{y_0, \delta_x\}$. □

Remark 5. If F^{**} is continuous at $y_\beta \in \operatorname{Int}(\operatorname{dom} F^{**})$, then it is G-differentiable at y_β if and only if $\partial F^{**}(y_\beta)$ is a singleton set (e.g. see [25], Chapter 2, Section 4.1). Our interest, however, is in solutions $y_\beta \in \partial F^*(\beta x)$ that can be on the boundary of $\operatorname{dom} F^{**}$, such as in the case when $\operatorname{dom} F^{**}$ is the positive cone Y_+ of $(Y, \leq)$, and all solutions $y_\beta \in Y_+$ correspond to positive measures. In this case, $y_\beta(M) = 0$ for some $M \subset X$ implies y_β is on the boundary of Y_+. The condition of strict convexity of F^* on $\operatorname{dom} F^*$ in Theorem 1 implies that $\partial F^*(x_1) \neq \partial F^*(x_2)$ for all $x_1 \neq x_2$ in $\operatorname{dom} F^*$, even if $\partial F^*(x_i)$ are on the boundary of $\operatorname{dom} F^{**}$.

Corollary 1 (Support). *Under the assumptions of Theorem 1, the support of element $x \in X$ is a subset of the support of optimal measures y_β for all $\lambda \in (\inf F, F(\delta_x))$.*

Proof. During the proof of Theorem 1, we established under its assumptions, that if $\lambda \in (\inf F, F(\delta_x))$, then condition $y_\beta(M) = 0$ implies $P_M x = 0 \in M$. Dually, if $P_M x \neq 0$ for some $M \subset X$, then $y_\beta(M) \neq 0$ for all $y_\beta \in \{y_\beta\}_x$. □

Example 2 (Relative Information). Let us define $F : Y \to \mathbb{R} \cup \{\infty\}$ as

$$F(y) := \begin{cases} \left\langle \ln \frac{y}{y_0}, y \right\rangle - \langle 1, y - y_0\rangle & \text{if } y > 0 \text{ and } y_0 > 0 \\ \langle 1, y_0\rangle & \text{if } y = 0 \text{ and } y_0 > 0 \\ \infty & \text{otherwise} \end{cases} \tag{13}$$

This functional is closed, strictly convex and weakly differentiable on the interior of $\operatorname{dom} F$:

$$\nabla F(y) = \ln \frac{y}{y_0} \quad \Longleftrightarrow \quad e^x y_0 = \nabla F^*(x)$$

One can define $F^* : X \to \mathbb{R} \cup \{\infty\}$ as

$$F^*(x) := \langle 1, e^x y_0\rangle$$

which is also closed, strictly convex and weakly differentiable for all $x \in X$, where it is finite (i.e. on $\mathrm{dom} F^*$). Solutions to problem (3) with F defined above belong to the exponential family (1), and they correspond to exponential measures that are mutually absolutely continuous.

Note that generally $\mathrm{dom} F^* \subset X$. For example, if X is the space of sequences $x : \mathbb{N} \to \mathbb{R}$, then there are unbounded sequences in X. However, some unbounded sequences are F-bounded, if there exists $\beta^{-1} > 0$ such that $F^*(\beta x) < \infty$ (e.g. see Example 1). Observe that this property depends on the choice of element $y_0 = \nabla F^*(0)$ minimising F.

The relative information functional (13) is a generalisation of the classical Kullback-Leibler information distance $I_{KL}(p,q) := \mathbb{E}_p\{\ln p - \ln q\}$ on $\mathscr{P}(\Omega)$ [14]. Indeed, for positive measures with equal norm $\|\cdot\|_1 = \langle 1, \cdot \rangle$, we have $\langle 1, y - y_0 \rangle = 0$. Functional (13), however, is non-negative for all elements y and y_0 (i.e. not necessarily with equal norms), and the gradient of F has a convenient form. If X and Y are commutative algebras, such as algebras of real functions on Ω, then the pairing $\langle \cdot, \cdot \rangle$ is defined by the sum or the integral (2), and (13) reduces to the classical measures of relative information. For non-commutative algebras, such as the algebra of compact Hermitian operators on a separable Hilbert space and the trace pairing (2), functional (13) is a generalisation of some quantum information distances, which depend on the way yy_0^{-1} is defined (e.g. as $\exp\{\ln y - \ln y_0\}$ or $y^{1/2} y_0^{-1} y^{1/2}$).

Example 3 (Counter-example). This example is based on a counter-example, proposed by one of the reviewers of an earlier version of the paper. Let $w \in X$ be a fixed 'weight' vector, and let $F : Y \to \mathbb{R} \cup \{\infty\}$ be defined as follows

$$F(y) := \langle w, |y| \rangle^2 - \langle w, |y| \rangle$$

where $|y| = \sup\{-y, y\}$. Its subdifferential is

$$\partial F(y) = \begin{cases} (2\langle w, y \rangle - 1)w & \text{if } y > 0 \\ [-w, w] & \text{if } y = 0 \\ (1 - 2\langle w, y \rangle)w & \text{if } y < 0 \end{cases}$$

It is clear from the above that $\inf F = F(0)$, because $0 \in \partial F^*(x)$. For any $\lambda > \inf F$, there is a unique solution $y_\beta \in \partial F^*(\beta x)$ to problem (3) such that $\beta x \notin [-w, w]$. However, if $\beta_1 P_M x \in [-P_M w, P_M w]$, then all solutions $y_{\beta_1} \in [0]$ (and $y_{\beta_1}(M) = 0$) for the corresponding $\lambda > \inf F$. If $\beta_2 P_M x \notin [-P_M w, P_M w]$, then $y_{\beta_2} \notin [0]$ (and $y_{\beta_2}(M) \neq 0$). Therefore, solutions y_β do not correspond to a family of mutually absolutely continuous measures. It is quite clear, however, that functional F is defined above in such a way that its dual F^* is not strictly convex, and therefore it does not satisfy the conditions of Theorem 1. As mentioned in Remark 4, such functionals are not very good for measuring information in optimisation problems, because their values cannot distinguish between some quantities βx of utility.

5 Optimal probability measures

Let us now consider the case, when the optimisation problem (3) is restricted to statistical manifold $\mathscr{P} \subset Y$. In this case, solutions $p_\beta = y_\beta / \|y_\beta\|_1$ are optimal probability measures maximising expected value $\mathbb{E}_p\{x\} := \langle x, p \rangle$ subject to $F(p) \leq \lambda$ and $\|y\|_1 = 1$. In this context, we interpret $F(p) \leq \lambda$ as an information constraint. Theorem 1 and Corollary 1 establish general properties of optimal measures in a broad class of functionals F. However, a little bit more can be said about optimal probability measures.

As was mentioned earlier, measures such that $y(M) = 0$ for some $M \subset X$ belong to the same subspace $M^\perp \subset Y$. Therefore, all mutually absolutely continuous measures, such as the family $\{y^\circ_\beta\}$ of optimal solutions in Theorem 1, belong to the same subspace in $M^\perp \subset Y$. Recall that statistical manifold $\mathscr{P}$ is a compact convex set that is the base of projective positive cone Y_+. Thus, all mutually absolutely continuous optimal probability measures p°_β corresponding to $y^\circ_\beta \in \{y^\circ_\beta\}_x$ belong to the interior of the base of the projective subcone $M^\perp_+$, or in other words to the interior of a statistical sub-manifold. In the classical case, this sub-manifold is a simplex $\mathscr{P}(\Omega \setminus E)$, and it is a facet of the simplex $\mathscr{P}(\Omega)$.

The restriction of $y \in Y_+$ to the statistical manifold $\mathscr{P}$ factorises the dual space $X \subseteq Y^\sharp$. Observe that $\mathscr{P}$ is a subset of the affine set N:

$$N := \{y \in Y : \langle 1, y \rangle = 1\} = \{1\}_\perp + q\,, \qquad q \in \mathscr{P}$$

where subspace $\{1\}_\perp$ is the annihilator of linear functional $1 \in X$. This unit functional is the extension of the norm $\|\cdot\|_1$, which is additive on the cone Y_+ of positive elements, from Y_+ to the whole space Y: $\langle 1, y \rangle = \|y\|_1$ if $y \geq 0$. Thus, every probability measure $p \in \mathscr{P}$ is equivalently represented by elements $y \in \{1\}_\perp$ as $p = y + q$, $q \in \mathscr{P}$.

The space of random variables (observables) is the dual of subspace $\{1\}_\perp$, and it is the factor space $X/\mathbb{R}1$, generated by the subspace $\mathbb{R}1 := \{\beta 1 : \beta \in \mathbb{R},\ 1 \in X\}$ of constant vectors. Random variables are shifts $[x] = \mathbb{R}1 + x$, and they are equivalence classes: x is equivalent to x' if and only if $x - x' \in \mathbb{R}1$ or equivalently $\langle x - x', p - q \rangle = 0$ for any p, $q \in \mathscr{P}$. Thus, different random variables $[x]$ and $[w]$ correspond to elements $x, w \in X$ such that $\langle x - w, p - q \rangle \neq 0$ or $x - w \notin \mathbb{R}1$.

In Corollary 1, it was established that for strictly convex F^*, the support of $x \in X$ is a subset of the support of optimal measures y_β for all $\lambda \in (\inf F, F(\delta_x))$. Observe now that zero in the space of random variables $X/\mathbb{R}1$ is subspace $\mathbb{R}1$ of constant vectors. Therefore, if $P_M x \notin \mathbb{R}1$, then $p_\beta(M) > 0$. Conversely, $p_\beta(M) = 0$ implies that $P_M x \in \mathbb{R}1$. In the language of classical probability this result can be stated as follows: if $x(\omega_1) \neq x(\omega_2)$ for some ω_1, $\omega_2 \in E \subset \Omega$, then $p_\beta(E) > 0$ for all optimal probability measures with $\lambda \in (\inf F, F(\delta_x))$. Conversely, $p_\beta(E) = 0$ implies that $x(\omega) = \text{const}$ for all $\omega \in E$.

6 Optimal transition kernels

In this section, we consider a composite system $\Omega = A \times B$ and the problem of optimisation of transitions between the elements of A and B. For simplicity, our exposition will be in the classical setting of commutative algebra X. This is because joint and conditional probability measures are well-defined and understood in this setting. In the non-classical case, the analogue of a conditional probability operator can also be defined, and the results of this section can then be transferred to this setting. However, this leads to unnecessary complications, which we shall avoid.

Optimisation problems for composite systems appear in theories of optimal decisions and control, where optimality is defined relative to a utility function $x : A \times B \to \mathbb{R}$, and the main objective is optimisation of transitions between the elements of sets A and B. In some cases, optimal transitions are deterministic corresponding to some functions $a = f(b)$ or $b \in f^{-1}(a)$. Non-deterministic transitions are represented by Markov transition kernels.

Let $\mathscr{P}(A)$ and $\mathscr{P}(B)$ be classical statistical manifolds associated with measurable sets $(A, \mathscr{A})$ and $(B, \mathscr{B})$ respectively. Recall that a *transition kernel* $\tau : B \to \mathscr{P}(A)$ (e.g. [8]) is a conditional probability measure $\tau(b) = p(A_i \mid b)$ that is measurable with respect to $\mathscr{B}(B)$ for each $A_i \in \mathscr{A}(A)$. Transition kernel defines linear operator $T : \mathscr{P}(B) \to \mathscr{P}(A)$ as follows:

$$T\,p(B_j) := \int_{B_j} p(A_i \mid b)\, dp(b) = p(A_i)$$

Elements $p \in \mathscr{P}(A \times B)$ are joint probability measures $p(A_i \cap B_j) = p(A_i \mid B_j)\, p(B_j)$, and for $p(B_j) > 0$, the conditional probability is defined by the Bayes formula:

$$p(A_i \mid B_j) = \frac{p(A_i \cap B_j)}{p(B_j)}\,,$$

A random variable a is statistically independent of b if and only if $p(A_i \mid b) = p(A_i)$ for each $b \in B$, $A_i \in \mathscr{A}(A)$. In this case, $p(A_i \cap B_j) = p(A_i)p(B_j)$. On the other hand, deterministic dependency $a = f(b)$ corresponds to transition kernel

$$p(A_i \mid b) = \delta_{f(b)}(A_i) = \begin{cases} 1 & \text{if } f(b) \in A_i \\ 0 & \text{otherwise} \end{cases}$$

In this case, $p(A_i \cap B_j) = \delta_{f(b)}(A_i)\, p(B_j) = 0$ for all $f(b) \notin A_i$. If $a = f(b)$ is an injective function, then $p(A_i) = p(B_j)$ for each $A_i = f(B_j)$, and $p(B_j \mid a) = \delta_{f^{-1}(a)}(B_j)$. Thus, we can classify $p \in \mathscr{P}(A \times B)$ into deterministic or non-deterministic.

Definition 4 (Deterministic composite state). *A joint probability measure $p \in \mathscr{P}(A \times B)$ is* deterministic, *if and only if it defines a deterministic transition kernel $\tau(b) = \delta_{f(b)}(A_i)$ for a measurable function $f : B \to A$ or $f^{-1} : A \to B$. Otherwise, p is* non-deterministic.

Example 4 (Exponential kernels). Let $\Omega = A \times B$, where $A = (A, da)$ and $B = (B, db)$ are equivalent Lebesgue spaces. Let $X = X(A \times B)$ be a commutative algebra, and Y be in duality with X via $\langle \cdot, \cdot \rangle$, defined by the integral (2). Let $F : Y \to \mathbb{R} \cup \{\infty\}$ be the relative information functional (13), such that for each $y \in Y_+$, F is minimised at $y_0 = y(A)y(B)$, where $y(A) = \langle 1, y \rangle|_B$, $y(B) = \langle 1, y \rangle|_A$ are the corresponding marginal measures. Then the restriction of F to $\mathscr{P}(\Omega)$ is the Shannon mutual information [21]:

$$F_S(p) := \int_{A \times B} \ln \left[\frac{dp(a,b)}{dp(a)\, dp(b)} \right] dp(a,b) = \int_B dp(b) \int_A \ln \left[\frac{dp(a \mid b)}{dp(a)} \right] dp(a \mid b) \quad (14)$$

Solutions $p_\beta \in \mathscr{P}(A \times B)$ to problem (3) with constraints on Shannon information belong to the exponential family:

$$dp_\beta(a,b) = e^{\beta\,[x(a,b) + \Phi(\beta^{-1})]}\, dp(a)\, dp(b)\,,$$

where $\Phi(\beta^{-1})$ is determined from the normalisation condition as

$$\Phi(\beta^{-1}) = -\beta^{-1}\ln F^*(\beta x) = -\beta^{-1}\ln\int_{A\times B} e^{\beta\, x(a,b)}\, dp(a)\, dp(b)$$

The solutions define exponential transition kernels:

$$dp_\beta(a\mid b) = e^{\beta\,[x(a,b)+\Phi(\beta^{-1},b)]}\, dp(a)\,, \qquad dp_\beta(b\mid a) = e^{\beta\,[x(a,b)+\Phi(\beta^{-1},a)]}\, dp(b)$$

where $\Phi(\beta^{-1},b)$ and $\Phi(\beta^{-1},a)$ now depend on b and a, as they are computed using partial integrals in A and B respectively. Observe also that because $dp(a) = \int_B dp(a,b)$ and $dp(b) = \int_A dp(a,b)$, the following conditions hold

$$\int_B e^{\beta[x(a,b)+\Phi(\beta^{-1},b)]}\, dp(b) = 1\,, \qquad \int_A e^{\beta[x(a,b)+\Phi(\beta^{-1},a)]}\, dp(a) = 1$$

If $A = (A,+)$ and $B = (B,+)$ are groups, and the utility function is translation invariant $x(a+c, b+c) = x(a,b)$, then it follows from the conditions above that

$$e^{\beta\,\Phi(\beta^{-1},b)}\, dp(b) = \left(\int_B e^{\beta\, x(a,b)}\, db\right)^{-1}, \qquad e^{\beta\,\Phi(\beta^{-1},a)}\, dp(a) = \left(\int_A e^{\beta\, x(a,b)}\, da\right)^{-1}$$

and the exponential transition kernels take the following simple form

$$dp_\beta(a\mid b) = \frac{e^{\beta\, x(a,b)}\, da}{\int_A e^{\beta\, x(a,b)}\, da}\,, \qquad dp_\beta(b\mid a) = \frac{e^{\beta\, x(a,b)}\, db}{\int_B e^{\beta\, x(a,b)}\, db}$$

The normalising integrals above are constant as they do not depend on a or b. In this important case, one can introduce the *free energy* function $\Phi_0(\beta^{-1}) := -\beta^{-1}\ln\int_B e^{\beta\, x(a,b)}\, db$ or the *free cumulant generating function* $\Psi_0(\beta) = -\beta\,\Phi_0(\beta^{-1})$. If one of the marginal distributions, say $p(B)$, is fixed, then Shannon information has the following expression:

$$\begin{aligned}
F_S(p_\beta) &= \int_A dp(a)\int_B \ln\frac{dp(b\mid a)}{dp(b)}\, dp(b\mid a)\\
&= \int_A dp(a)\int_B \left\{\ln(e^{\beta\, x(a,b)}) - \ln\int_B e^{\beta\, x(a,b)}\, db - \ln[dp(b)/db]\right\} dp(b\mid a)\\
&= \beta\,\mathbb{E}_{p_\beta}\{x\} - \Psi_0(\beta) + H\{p(B)\}\,, \qquad (15)
\end{aligned}$$

Observe also that the expected utility is the derivative of $\Psi_0(\beta) = \ln\int_B e^{\beta\, x(a,b)}\, db$:

$$\mathbb{E}_{p_\beta}\{x\} = \int_A dp(a)\int_B \frac{x(a,b)\, e^{\beta\, x(a,b)}}{\int_B e^{\beta\, x(a,b)}\, db}\, db = \frac{d\Psi_0(\beta)}{d\beta}\int_A dp(a) = \Psi_0'(\beta) \qquad (16)$$

Here, $H\{p(B)\} = -\int_B \ln[dp(b)/db]\, dp(b)$ is the differential entropy of $p(B)$ (assuming that the density $dp(b)/db$ exists). Also, because Shannon information can be represented as the difference $F_S(p) = H\{p(B)\} - H\{p(B\mid a)\}$, the quantity $\Psi_0(\beta) - \beta\,\Psi_0'(\beta)$ is clearly the conditional differential entropy $H\{p(B\mid a)\}$.

Remark 6 (Information of deterministic kernels). Transition kernels define information in a more traditional sense as communication between the elements of sets A and B. The maximum amount of information in this sense can be communicated by an injective function $a = f(b)$, because the preimage $f^{-1}(a)$ uniquely determines b. If a function is not injective, then $b \in f^{-1}(a)$ is determined up to the probability $1/|f^{-1}(a)|$. Note that $\sum_{a\in f(B)} |f^{-1}(a)| = |B|$, and the expected value of $|f^{-1}(a)|$ with respect to a uniform distribution of $a \in f(B)$ is $|B|/|f(B)|$. Therefore, one can put $p(a) = p(b)|f(B)|/|B|$. This reasoning demonstrates that for deterministic $p_f \in \mathscr{P}(A\times B)$, information constraints $F(p_f) \le \lambda < \sup F$, understood in the sense of communication, impose constraints $|f(B)| < |B|$ on the cardinality of the image of the function. If B is infinite, then there can be an infinite number of constraints $\lambda < \sup F$ such that the image $f(B)$ is finite. Moreover, if B is countable, then $f(B)$ is finite for all $\lambda < \sup F$. The infimum of information corresponds to constant functions (including the empty function). These facts can be well illustrated using Shannon information:

$$\begin{aligned} F_S(p_f) &= \int_B dp(b) \int_A \ln \frac{\delta(f(b)-b)}{dp(a)}\, \delta(f(b)-b) \\ &= -\int_B dp(b) \ln\Big(dp(f(b))\Big) = -\int_B dp(b) \ln\left(\frac{|B|}{|f(B)|}\, dp(b)\right) \\ &= \ln|f(B)| - \ln|B| + H\{p(B)\} \end{aligned}$$

where $p_f = \delta(f(b)-b)\,dp(b)$, and $H\{p(B)\}$ is the entropy of $p(B)$. As is well-known, $H\{p(B)\} \le \ln|B|$, and therefore $F_S(p_f) \le \ln|f(B)|$. Moreover, because $H\{p(A\mid b)\} = 0$ for $p(A\mid b) = \delta_{f(b)}(A)$, we have $F_S(p_f) = H\{p(A)\} \le H\{p(B)\}$. The maximum amount of information $F_S(p_f) = H\{p(B)\}$ is communicated if and only if $f: B \to A$ is injective on the support of $p(B)$.

The application of Theorem 1 to the case $\Omega = A\times B$ yields the following result.

Corollary 2 (Optimal transition kernels). *Let $\{p_\beta\}_x \subset \mathscr{P}(A\times B)$ be a family of joint probability measures that are optimal solutions to problem (3) for all values λ of a closed functional $F : Y \to \mathbb{R}\cup\{\infty\}$. If F^*, the dual of F, is strictly convex on* $\operatorname{dom} F^*$ *and F is minimised at $p_0 \in \partial F^*(0) \subset \operatorname{Int}(\mathscr{P}(A\times B))$, then p_β is deterministic if and only if $\lambda \ge F(\delta_x)$ or $\langle x, p_\beta\rangle = \langle x, \delta_x\rangle$.*

Proof. Assume there exists $p_\beta \in \{p_\beta\}_x$ for $\lambda < F(\delta_x)$ (and hence $\langle x, p_\beta\rangle < \langle x, \delta_x\rangle$) and such that the corresponding transition kernel is deterministic: $p_\beta(A_i \mid B_j) = 1$ if $A_i = f(B_j)$ and $p_\beta(A\setminus A_i \mid B_j) = 0$. In this case, $p_\beta(A\setminus A_i, B_j) = 0$, and therefore it is not on the interior of $\mathscr{P}(A\times B)$, $p_\beta \notin \partial F^*(0)$ and $F(p_\beta) = \lambda \in (\inf F, F(\delta_x))$. But then $p_\beta(A\setminus A_i, B_j) = 0$ for all $\lambda \in [\inf F, \sup F]$ by Theorem 1. In particular, there exists $p_0 \in \partial F^*(0)$ such that $p_0 = 0$ if $f(b) \notin A_i$, and therefore p_0 is not on the interior of $\mathscr{P}(A\times B)$. Thus, by contradiction we have proven $\lambda \ge F(\delta_x)$ (and hence $\langle x, p_\beta\rangle = \langle x, \delta_x\rangle$). Conversely, if $\lambda \ge F(\delta_x)$, then by there exist solution $\delta_x \in \operatorname{ext}\mathscr{P}(A\times B)$ (Proposition 1) corresponding to some function $f(b) = a$. $\square$

Remark 7. The assumptions of Corollary 2 are quite general. Strict convexity of F^* was justified in Remark 4, and condition $p_0 \in \operatorname{Int}(\mathscr{P}(A\times B))$ is very natural. Indeed, each facet of the simplex $\mathscr{P}(A\times B)$ is also a simplex of some subset of $A\times B$. Therefore, the element p_0 is always in the interior of some simplex $\mathscr{P}(A_i\times B_j)$, unless $p_0=\delta\in\operatorname{ext}\mathscr{P}(A\times B)$. In all practical cases, information is minimised at $p_0\neq\delta$. In particular, one often chooses $p_0=p(A_i)p(B_j)$, so that a and b are independent, and the support of $p(A_i)$ and $p(B_j)$ includes more than one element.

Corollary 3 (Strict inequalities). *Let $\{p_\beta\}_x\subset\mathscr{P}(A\times B)$ and $F:Y\to\mathbb{R}\cup\{\infty\}$ be defined as in Corollary 2. Then, for any F-bounded element $x\in X$*

$$\langle x,p_f\rangle<\langle x,p_\beta\rangle$$

for all deterministic $p_f\in\mathscr{P}(A\times B)$ such that $F(p_f)=F(p_\beta)\in(\inf F,F(\delta_x))$. Similarly,

$$F(p_f)>F(p_\beta)$$

for all deterministic $p_f\in\mathscr{P}(A\times B)$ such that $\langle x,p_f\rangle=\langle x,p_\beta\rangle\in(\overline{\upsilon}_0,\langle x,\delta_x\rangle)$.

Proof. For all $x\in X$ and $y\in Y$, the Young-Fenchel inequality holds:

$$\langle x,y\rangle\leq F^*(x)+F(y)$$

Moreover, the above holds with equality if and only if $y\in\partial F^*(x)$ (e.g. see [25]). If $x\in X$ is F-bounded and $F(y)=\lambda\in(\inf F,F(\delta_x))$, then $\partial F^*(\beta x)$ is non-empty by Proposition 3. Assume $p_\beta\in\partial F^*(\beta x)$. Then $\langle x,p_\beta\rangle=\beta^{-1}[F^*(\beta x)+F(p_\beta)]$. On the other hand, if p_f is deterministic, then $p_f\notin\partial F^*(\beta x)$, unless $F(p_f)\geq F(\delta_x)$ (Corollary 2). Thus, for any p_f such that $F(p_f)=F(p_\beta)$, we have $\langle x,p_f\rangle<\beta^{-1}[F^*(\beta x)+F(p_\beta)]$, which proves the first inequality.

By definition of the Legendre-Fenchel transform, $F^{**}(y)\geq\langle x,y\rangle-F^*(x)$, and the equality holds if and only if $y\in\partial F^*(x)$. Thus, if p_f is deterministic, then $p_f\notin\partial F^*(\beta x)$, unless $\langle x,p_f\rangle=\langle x,\delta_x\rangle$ (Corollary 2). Moreover, $\beta>0$ for any $\lambda>\inf F$ corresponding to $\langle x,p_\beta\rangle>\overline{\upsilon}_0$. In this case, $F^{**}(p_f)>\beta\langle x,p_f\rangle-F^*(\beta x)=\beta\langle x,p_\beta\rangle-F^*(\beta x)=F^{**}(p_\beta)$. Using the facts that $F(p_f)\geq F^{**}(p_f)$ and $F^{**}(p_\beta)=F(p_\beta)$ for solutions $p_\beta\in\partial F^*(\beta x)$, we obtain the second inequality. □

Strict inequalities of Corollary 3 present an interesting opportunity for constructing an example such that $\langle x,p_f\rangle=-\infty$ or $F(p_f)=\infty$ for any deterministic transition kernel satisfying a given information or utility constraint. The inequalities $\langle x,p_\beta\rangle>-\infty$ or $F(p_\beta)<\infty$ would imply the existence of a non-deterministic transition kernel satisfying the same information or utility constraints and having a finite expected utility and information. Such an example can be relevant in the context of the computational complexity theory. Let us consider one prototypical example.

Example 5 (Optimal communication). Let $a\in A$ and $b\in B$ be real variables, and let us consider the problem of information transmission between A and B that is optimal with respect to a measurable utility function $x:A\times B\to\mathbb{R}$. If $b\in(\mathbb{R},\mathscr{B},p)$ is a random variable with known distribution, then the expected utility $\mathbb{E}_p\{x\}$ is:

$$\mathbb{E}_p\{x\}=\int_A\int_B x(a,b)\,dp(a,b)=\int_B dp(b)\int_A x(a,b)\,dp(a\mid b)=\int_B\mathbb{E}_p\{x\mid b\}\,dp(b)$$

Here $\mathbb{E}_p\{x \mid b\}$ denotes conditional expected utility, and it is maximised by choosing the optimal conditional probability $dp(a \mid b)$. The maximum of information is communicated by an injective function $a = f(b)$, represented by a deterministic transition kernel. The optimal function is defined by the utility function. On the other hand, if no information about $b \in B$ can be communicated, then $dp(a \mid b) = dp(a)$, and one can only consider constant functions. Note, however, that one can still choose the optimal constant function $\bar{a} = f(b)$. Indeed, if $x(a,b)$ is differentiable and concave in a, then $\bar{a}$ is a solution to the equation $\nabla_a \int_B x(a,b)\,dp(b) = 0$. In particular, if $x(a,b) = -\frac{1}{2}(a-b)^2$, then $\nabla_a \int_B x(a,b)\,dp(b) = \int_B (b-a)\,dp(b)$, and $\bar{a} = \int_B b\,dp(b) = \mathbb{E}_p\{b\}$, which is the well-known classical method minimising mean-squared deviation. Therefore

$$\mathbb{E}_{p_f}\{x\} = -\frac{1}{2}\int_B (a-b)^2\,dp(b) \leq -\frac{1}{2}\mathrm{Var}\{b\}$$

The value on the right depends on the distribution $p(B)$, and there are many examples of distributions with unbounded variance, such as $dp(b) = [\pi(b^2+1)]^{-1}\,db$ (Cauchy distribution). Indeed, the integral $\int_B (a-b)^2(b^2+1)^{-1}\,db$ does not converge on $B = (-\infty,\infty)$. We note that $\mathbb{E}_p\{b\}$ is also undefined in this case. However, if $p(B)$ is unimodal and symmetric, then one can choose $\bar{a}$ to be the mode of $p(B)$.

Let us assume now that some limited information about the value of b can be communicated so that $dp(a \mid b) \neq dp(a)$ (and hence $dp(b \mid a) \neq dp(b)$). For example, this can be the information that b belongs to some subset of B, such as $b > 0$ or $b \leq 0$. In each case, one can choose a different optimal value $\bar{a}_1$ and $\bar{a}_2$. A more 'precise' information about b would correspond to a larger number of subsets $B_i \subset B$, and therefore one could choose a larger number of optimal values $\bar{a}_i$. The expected utility in this case is

$$\mathbb{E}_{p_f}\{x\} = -\frac{1}{2}\sum_{i=1}^{n}\int_{B_i} (\bar{a}_i - b)^2\,dp(b)$$

The cardinality $|f(B)|$ of the image of the optimal function $\bar{a}_i = f(B_i)$ is bounded by the amount of information that can be communicated. Thus, the minimum of information corresponded to the optimal constant function $\mathbb{E}_p\{b\} = f(b)$. The maximum of information would imply that the value of b can be communicated exactly, and one could select the optimal injective function $a = b$. If, however, all information cannot be communicated, then the function cannot be injective. In particular, for an infinite B, there can be an infinite number of constraints such that $|f(B)|$ is finite (see Remark 6). In this situation, one can choose only a finite number of optimal values $\bar{a}_i$, such as choosing two values $\bar{a}_1$ and $\bar{a}_2$ if the information partitions B into two subsets.

Observe now that the integral $\int (a-b)^2(b^2+1)\,db$ does not converge on the intervals $B_1 = (-\infty, 0]$ or $B_2 = [0,\infty)$. In fact, for any finite partition of the real line, there are some unbounded intervals on which the integral does not converge. Thus, in our example, b is distributed in such a way, that the expected value of utility $x(a,b) = -\frac{1}{2}(a-b)^2$ cannot be larger than $-\infty$ for any deterministic p_f satisfying constraints $\lambda < \sup F$ such that $|f(B)|$ is finite. To achieve a finite expected utility, a function must have infinite image $f(B)$. But this means that the function will transmit an infinite amount of information. Let us now demonstrate that there exist non-deterministic transition kernels for this problem achieving finite expected utility and communicating finite amount of information.

Indeed, let F be the Shannon information as in Example 4. In this case, the optimal transition kernels belong to the exponential family. Moreover, because the utility function $x(a,b) = -\frac{1}{2}(a-b)^2$ is translation invariant $x(a+c,b+c) = x(a,b)$, we can use simplified expressions from Example 4. In particular, $\Psi_0(\beta) = \ln\sqrt{2\pi\beta^{-1}}$, and optimal transition kernel $dp(a \mid b)$ is Gaussian

$$dp_\beta(a \mid b) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left\{-\frac{1}{2}\beta(a-b)^2\right\} da$$

Conditional expectation $\mathbb{E}_{p_\beta}\{x \mid b\}$ is constant for all $b \in B$:

$$\mathbb{E}_{p_\beta}\{x \mid b\} = -\frac{1}{2}\frac{1}{\sqrt{2\pi\beta^{-1}}}\int_{-\infty}^{\infty}(a-b)^2 e^{-\frac{1}{2}\beta(a-b)^2} da = -\frac{1}{2}\frac{\sqrt{2\pi\beta^{-3}}}{\sqrt{2\pi\beta^{-1}}} = -\frac{1}{2}\beta^{-1}$$

and therefore

$$\mathbb{E}_{p_\beta}\{x\} = \int_B \mathbb{E}_{p_\beta}\{x \mid b\}\, dp(b) = -\frac{1}{2}\beta^{-1}$$

The expression above can also be easily obtained from equation (16) as the derivative of $\Psi_0(\beta) = \ln\sqrt{2\pi\beta^{-1}}$. The optimal value $\beta^{-1} \geq 0$ depends on the amount λ of information, representing divergence of $dp(a \mid b)$ from $dp(a)$, and it can be found using the inverse of function $\lambda = F_S(p_\beta)$. Thus, using equation (15), we obtain

$$\beta = 2\pi e^{1-2[H\{p(B)\}-\lambda]}$$

The value of β depends on the difference $H\{p(B)\} - \lambda$, which equals to the conditional differential entropy $H\{p(B \mid a)\}$, because $\lambda = F_S(p_\beta) = H\{p(B)\} - H\{p(B \mid a)\}$. Therefore, if $H\{p(B \mid a)\}$ is finite, then $\beta > 0$ ($\beta^{-1} < \infty$), and $\mathbb{E}_{p_\beta}\{x\}$ is finite for all $\lambda > 0$. In fact, one can easily check that the following integral converges

$$\int_{-\infty}^{\infty}(a-b)^2\frac{e^{-\beta\frac{1}{2}(a-b)^2}}{b^2+1}\, db < \infty \qquad \forall\beta > 0$$

Thus, in our example, the expected utility of any deterministic p_f is $\langle x, p_f\rangle = -\infty$, unless the image $f(B)$ and the amount of information p_f communicates is infinite. A non-deterministic p_β communicating finite amount of information was shown to have finite expected utility $\langle x, p_\beta\rangle$. We point out also that the utility function $x(a,b) = -\frac{1}{2}(a-b)^2$ is unbounded, but it is F-bounded because $F^*(\beta x) = \beta\langle x, p_\beta\rangle - \lambda = -\frac{1}{2} - \lambda < \infty$ for all $\lambda \in (\inf F, F(\delta_x))$.

Remark 8. As mentioned in Remark 6, if B is countable, then any constraint $\lambda < \sup F$ implies that the image $f(B)$ is finite. Thus, one can use the same ideas of Example 5 to construct an example such that the expected utility $\langle x, p_f\rangle = -\infty$ for all $\lambda < \sup F$ and any deterministic p_f with $F(p_f) = \lambda$. For instance, if $A = B = \mathbb{N}$, then the example can be constructed by using utility $x(a,b) = -\frac{1}{2}(a-b)^2$ and taking $p(b) = [b^3\zeta(3)]^{-1}$, where $\zeta(k) = \sum_{b\in\mathbb{N}} b^{-k}$ is the Riemann zeta function. The expected utility of a deterministic p_f is negatively infinite for all $\lambda < \sup F$; the expected utility of a non-deterministic p_β is finite for all $\lambda > 0$.

7 Discussion

We studied optimal Radon measures using a generalisation of the classical variational problem of information theory. The main result is the existence of families of mutually absolutely continuous measures that are optimal solutions to problems with constraints on an abstract information resource with a strictly convex dual. Thus, we showed that this property of optimal measures, which was known for a one-parameter exponential family, is related to a geometric rather than algebraic property of information. Moreover, we argued that strict convexity of the dual functional is a natural property of information in the context of optimisation. Our method does not depend on commutativity of the algebra of observables, and for this reason the result holds both for classical and non-classical (i.e. non-commutative or quantum) measures.

In many ways, this work can be seen as a generalisation of the classical results on variational problems in information theory [21,22] and statistical physics [10]. Indeed, standard formulae of these theories relating Gibbs measures, free energy, entropy and channel capacity can be recovered simply by defining information constraints using the Kullback-Leibler divergence. However, the general approach allowed us to show that some properties of the optimal families of measures, such as mutual absolute continuity and support sets, do not depend on how the information constraints are defined.

Our results about classification of optimal transition kernels can have applications not only to optimisation problems, but also to some theoretical questions in the theory of computational complexity, where much of the effort is devoted to the question of whether non-deterministic procedures give any advantage over deterministic. It was shown here that in a broad class of optimisation problems with information constraints optimal deterministic kernels do not exist. As an illustration, an example was constructed where any deterministic kernel can only have negatively infinite expected utility (or equivalently unbounded expected error or risk), unless it communicates an infinite amount of information. On the other hand, it was shown that non-deterministic kernels can both give finite expected utility and communicate finite information in the same problem.

The results about sub-optimality of deterministic kernels do not contradict the established understanding in the classical theory of statistical decisions that asymptotically randomised policies cannot be better than deterministic (e.g. see [23] or more recently [13]). Indeed, a randomisation of the function's output can only decrease (loose) the amount of information it communicates. However, our results are about deterministic and non-deterministic kernels that communicate the same amount of information. Moreover, asymptotic results are concerned with obtaining all, possibly infinite information, in which case there are deterministic optimal kernels. A non-trivial claim that we can make here is that under information constraints deterministic kernels are not just suboptimal, but may fail to provide any meaningful solution because of an unbounded below expected utility, as was shown in Example 5. This seems to confirm common intuition in the field of applied optimisation, where numerous problems exist on which non-deterministic algorithms outperform all known deterministic methods (e.g. [12]).

References

1. Amari, S.I.: Differential-Geometrical Methods of Statistics, *Lecture Notes in Statistics*, vol. 25. Springer, Berlin, Germany (1985)
2. Asplund, E., Rockafellar, R.T.: Gradients of convex functions. Transactions of the American Mathematical Society **139**, 443–467 (1969)
3. Belavkin, R.V.: Information trajectory of optimal learning. In: M.J. Hirsch, P.M. Pardalos, R. Murphey (eds.) Dynamics of Information Systems: Theory and Applications, *Springer Optimization and Its Applications Series*, vol. 40. Springer (2010)
4. Belavkin, R.V.: Utility and value of information in cognitive science, biology and quantum theory. In: L. Accardi, W. Freudenberg, M. Ohya (eds.) Quantum Bio-Informatics III, *QP-PQ: Quantum Probability and White Noise Analysis*, vol. 26. World Scientific (2010)
5. Belavkin, V.P.: Personal communication (2010)
6. Bobkov, S.G., Zegarlinski, B.: Entropy bounds and isoperimetry, *Memoirs of the American Mathematical Society*, vol. 176. AMS (2005)
7. Bourbaki, N.: Eléments de mathématiques. Intégration. Hermann (1963)
8. Chentsov, N.N.: Statistical Decision Rules and Optimal Inference. Nauka, Moscow, U.S.S.R. (1972). In Russian, English translation: Providence, RI: AMS, 1982
9. Cramér, H.: Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ (1946)
10. Jaynes, E.T.: Information theory and statistical mechanics. Physical Review **106, 108**, 620–630, 171–190 (1957)
11. Kachurovskii, R.I.: Nonlinear monotone operators in Banach spaces. Russian Mathematical Surveys **23**(2), 117–165 (1968)
12. Kirkpatrick, S., Gelatt, C.D., Vecchi, J.M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
13. Kozen, D., Ruozzi, N.: Applications of metric coinduction. Logical Methods in Computer Science **5**(3:10), 1–19 (2009)
14. Kullback, S.: Information Theory and Statistics. John Wiley and Sons (1959)
15. Moreau, J.J.: Functionelles Convexes. Lectrue Notes, Séminaire sur les équations aux dérivées partielles. Collége de France, Paris (1967)
16. von Neumann, J., Morgenstern, O.: Theory of games and economic behavior, first edn. Princeton University Press, Princeton, NJ (1944)
17. Phelps, R.R.: Lectures on Choquet's theorem, *Lecture Notes in Mathematics*, vol. 1757, 2nd edn. Springer (2001)
18. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. The Annals of Statistics **23**(5), 1543–1561 (1995)
19. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society **37**, 81–89 (1945)
20. Rockafellar, R.T.: Conjugate Duality and Optimization, *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 16. Society for Industrial and Applied Mathematics, PA (1974)
21. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379–423 and 623–656 (1948)
22. Stratonovich, R.L.: On value of information. Izvestiya of USSR Academy of Sciences, Technical Cybernetics **5**, 3–12 (1965). In Russian
23. Stratonovich, R.L.: Information Theory. Sovetskoe Radio, Moscow, USSR (1975). In Russian

24. Streater, R.F.: Quantum Orlicz spaces in information geometry. In: The 36th Conference on Mathematical Physics, *Open Systems and Information Dynamics*, vol. 11, pp. 350–375. Torun (2004)
25. Tikhomirov, V.M.: Analysis II, *Encyclopedia of Mathematical Sciences*, vol. 14, chap. Convex Analysis, pp. 1–92. Springer-Verlag (1990)