# Chapter 14
# Adopting Convex Analysis in Statistical Applications




**Christos P. Kitsos and Filipe Amaral**

**Abstract** This paper provides a number of Mathematical areas where Convex Analysis is essential. It is mainly concentrated on Statistical Applications from different fields of Statistics: Martingales, Decision Theory, or Optimal Experimental Design. The appropriate theoretical insight is considered in a compact and critical view.

**Keywords** Convex analysis · Martingales · Decision theory · Optimal experimental design

## 14.1 Introduction

In principle, Convex Analysis is devoted on the study of convex function and sets. Different mathematical fields share such properties: Topology, Schubert [35], Numerical Analysis, Ortega and Rheinboldt [32], Optimal Designs, Federov [15], Probability Theory, Levy [31], with the Theory of Optimization to be the fundamental area, Du et al. (2001).

The Convex Optimization involves either a convex function or a feasible convex set, when the imposed constrained equality restrictions it is usually required to be affine, in the sense that remains invariant to linear transformations.

The main scientific area based on convex optimization, developed relatively recently by G. Dantzig we would say, despite the theoretical work of L. Kantorovich, the application attempts with his "tables" of W. Leontief, and latter by T. Koopmans, is Linear Programming (LP). Needless to say that John Von Newman immediately realized the importance of the "simplex method" and contributed for the development of LP. The Least Square Method (LSM) is another field using widely in the Convex Analysis.

C. P. Kitsos (✉)
University of West Attica, Athens, Greece
e-mail: xkitsos@uniwa.gr

F. Amaral
Escola Superior de Educação Paula Frassinetti, Porto, Portugal
e-mail: fma@esepf.pt

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

T. M. Rassias and P. M. Pardalos (eds.), *Convex and Variational Analysis with Applications*, Springer Optimization and Its Applications 230,
https://doi.org/10.1007/978-3-032-07860-5_14

Despite the fact that Legendre published in 1805 for the method, who also named it, it was Gauss who first worked on LSM. It is one of the biggest scientific fights between Gauss and Legendre who was the first that worked on the method.

Based on these two fundamental areas of Applied Mathematics, LSM and LP, more methods were developed, rather to solve real life problem with a strong mathematical background and a stronger computational support, as the calculation needed are really tedious. Recall that the number of calculations need to find the inverse of an $n \times n$ matrix is of order $n!$.

In this paper we shall try to avoid exploring the widely used areas of applications of Convex Analysis and we shall try to approach and discuss areas where the Convex Analysis is essential, but it is not clear, yet.

## 14.2 Background

A set $K$ is convex iff every convex combination of points of $K$ belongs to $K$, i.e., if we let $\{e_1, ..., e_n\}$ to form a basis in $\mathbb{R}^n$ and $e_i \in K$ then the point $x = \sum_{i=1}^{n} \lambda_i e_i \in K$ with $\lambda_i \in \mathbb{R}$ and $\sum_{i=1}^{n} \lambda_i = 1$.

The set $K_0$ of the real space $\mathbb{R}^n$ is convex hull, $conv\,(K_0)$, say iff it is the minimal convex set that contains $K_0$, i.e., is the intersection of all convex sets that cover $K_0$.

Recall that a differentiable function is convex iff $\bigtriangledown f$ is monotone, Rockafellar [34].

i. Two convex sets $K_1, K_2 \subseteq \mathbb{R}^n$, the $K_1 + K_2$ is convex $\lambda K_1$ or $(\lambda + \eta)K_1$

**Theorem 14.2.1** (Caratheodory) *Consider a non-empty set $K \subseteq \mathbb{R}^n$. Then any point of the convex hull of $K$ is representable as a convex combination of, at most, $n + 1$ points of $K$.*

The convex conjugate of a closed convex function is again a closed function. Thus the convex conjugate of a polyhedral convex function is again a polyhedral convex function.

For a given function $f : \mathbb{R}^n \to \mathbb{R}$ the gradient $\nabla f(x)$ is defined as

$$\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n})$$

The Gâteaux derivative $G_f(x, y)$, say, of $f$ at $x$ in the direction $y$ is defined

$$G_f(x, y) := \lim_{h \to 0^+} \frac{1}{h} \{f(x + hy) - f(x)\} \tag{14.1}$$

For $f$ is differentiable at $x \in \mathbb{R}^n$, then it holds that

$$G_f(x, y) = \langle \nabla f(x), y \rangle = \sum_{i=1}^{n} y_i \frac{\partial f(x)}{\partial x_i}$$

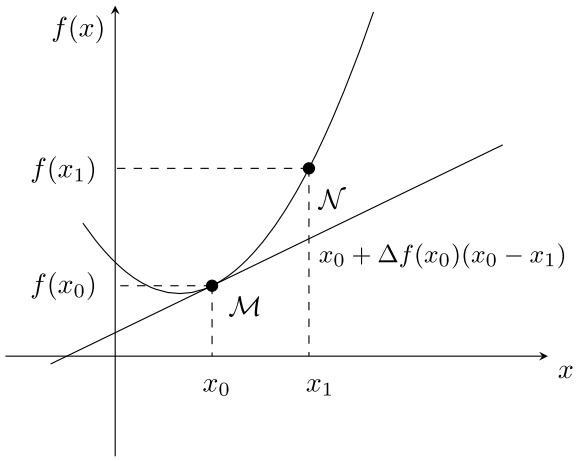



**Fig. 14.1** $\mathcal{M}(x_0, f(x_0)) = \min_{x+k} f(x)$, $K \subseteq \mathbb{R}^n$ closed and convex

We note in the sequence the main results, relative to the above definitions in a compact form, see also Fig. 14.1.

1. $G_f(x, y - x) = \langle \nabla f(x), y - x \rangle$
2. $f(y) - f(x) = G_f(x, y - x) + \circ(\|y - x\|)$

Recall the following theoretical results,

**Proposition 1** *Let $f : \Delta \subseteq \mathbb{R}^n \to \mathbb{R}$ be a Gâteaux differentiable in a convex subset $E$ of $\Delta$. Then for $\forall x, y \in E$*

*(i)* *$f$ is convex in $E \Leftrightarrow f'(x)(x - y) \leq f(y) - f(x)$*
*(ii)* *$f$ is strictly convex in $E \Leftrightarrow f'(x)(x - y) < f(y) - f(x), x \neq y$*
*(iii)* *$f$ convex on $E \Leftrightarrow \left[f'(y) - f'(x)\right](y - x) \geq 0$*

*With $\Delta \subseteq \mathbb{R}^n$ assumed to be a Hilbert space.*

Let $f$ be a function as in Proposition 1, then it holds

**Proposition 2** *For the function $f$ the following three statements are equivalent*

*(i)* *$f$ is uniformly convex in $E$ with constant $k$*
*(ii)* *$f(y) - f(x) \geq f'(x)(x - y) + k\|y - x\|^2, \forall x, y \in \Delta$*
*(iii)* *$f'(y) - f(x)(y - x) \geq 2k\|y - x\|^2, \forall x, y \in E$.*

Still $\Delta$ is a Hilbert space, $E \subset \Delta$ and the constrained optimization problem is

$$\min f(x), x \in \Delta \tag{14.2}$$

For the optimization problem as in (14.2) holds the following theorem,

**Theorem 14.2.2** *If $f$ is strictly convex and $E$ is convex then the problem admits a unique solution (minimizer).*

We define as Frechet derivative of $f$ at "$x$" in the direction of "$y$"

$$F_f(x, y) = G_f(x, y - x) = \langle \nabla f(x), y - x \rangle$$

Therefore for $f$ differentiable at $x$ it holds that

$$f(y) - f(x) \cong F_f(x, y) + \circ(\|y - x\|) \tag{14.3}$$

The above, we believe, provide the link with various areas of applications.

## 14.3 Statistical Applications

In this section we refer to the main points that convex functions influence the theoretical background of a number of different fields in Statistics. It is essential to point out that a number of theoretical values are valid only for convex functions.

### *14.3.1 Convex Functions and Martingales*

For the convex function $\varphi$, twice differentiable and the real random variable $X$ and $E$ being the Expectation operator, Jensen's inequality states,

$$\varphi(E(X)) \leq E(\varphi(X) \tag{14.4}$$

Now, let $X_n$ and $Y_n$ for $n = 0, 1, ...$ be stochastic process, Karlin and Tylor [21], then recall that $(X_n)$, $n = 0, 1, ...$ is a supermantingale with the respect to $(Y_n)$ if for all $n$ and with the notation $x^- = \min\{x, 0\}$ , $x^+ = \max\{x, 0\}$, it holds:

**Proposition 1** *For the above stated notation the following is true.*

*(i)* $-\infty < E(X_n)$
*(ii)* $E(X_{n+1}|Y_0, ..., Y_n) \leq X_n$
*(iii)* $X_n$ *is a function of* $(Y_0, Y_1, ..., Y_n)$

By duality, we call, say that $(X_n)$ is a submartingale with respect to $(Y_n)$ if, for all $n$:

(a) $E(X_n^+) < +\infty$
(b) $X_n \leq E(X_{n+1}|Y_0, ..., Y_n)$
(c) As in (iii.)

The equality of (b) above or (ii) states that $X_n$ is determined by the history up to the true $n$. This is in principle the idea behind the definition of a martingale, the current situation is depending only on the previous situation.

$$X_n = E(X_{n+1}|Y_0, ..., Yn) \tag{14.5}$$

is valid, by the definition of a martingale, which if does not hold, we obtain either a sub or super martingale, i.e., inequality (ii) or (b) holds.

Due to Jansen's inequality it can be proved that if $(X_n)$ is a martingale with respect to $(Y_n)$, and $\varphi$ convex, i.e., $\varphi$ is twice differentiable with $\frac{d^2\varphi}{dx^2} \geq 0$ for all $x$, such that

$$E\left[\varphi(X_n)^+\right] < \infty \text{ for all } n \tag{14.6}$$

then $\varphi(X_n)$ is a submartingale with respect to $Y_n$.

Moreover if $(X_n)$ is a submartingale with respect to $(Y_n)$ and $\varphi$ is convex and increasing function with $E\left[\varphi(X_n)^+\right] < \infty$ then $\varphi(X_n)$ is a submartingale.

The result can be useful if we truncate "the history" of the sequence below by a constant value, say, $c$.

That is if $(X_n)$ is submartingale then, for a given constant, we can define a new "sequence" truncated (the values less to $-c$ are not considered) such as

$$X_n^* = \begin{cases} X_n & \text{if} \quad X_n > -c \\ -c & \text{if} \quad X_n \leq -c \end{cases} \tag{14.7}$$

is a submartingale for which $E(\left|X_n^*\right|) < \infty$ for all $n$.

Jansen's inequality (14.4) is applied to many cases in Statistics. One more is the following: by Jensen's inequality it holds that

$$E(e^{\xi X}) \geq e^{\xi E(X)} \tag{14.8}$$

with $X$ given a random variable and $\xi$ a scalar, based on the fact that the exponential function is convex. Recall that the moment generating function of the random variable, identical to Laplace transform, is

$$M(\xi) = E(e^{\xi X}) \tag{14.9}$$

The Chernoff bound, Chernoff [8], for random variable $X$ and for $\xi > 0$ is then defined as

$$P(X = \tau) = P(e^{\xi X} \geq e^{\xi \tau}) \leq M(\xi)e^{-\xi \tau}$$

Therefore it holds

$$P(X \geq \tau) \leq \inf_{\xi > o} \left[M(\xi)e^{-\xi \tau}\right] \tag{14.10}$$

due to (14.9),

$$P(X \geq \tau) \leq \inf_{\xi > o} \left[ E(e^{\xi(X-\tau)}) \right] \tag{14.11}$$

The Chernoff bound as above plays a fundamental role to define in Economics the entropic value of risk, see also Sect. 14.4, usually denoted by EVaR, with coefficient level $1 - \alpha$ defined as,

$$\text{EVaR}_{1-\alpha} = \inf \left\{ \xi^{-1} \frac{\mathcal{M}(\xi)}{\alpha} \right\} \tag{14.12}$$

The notation is as in (14.12) for the random variable $X$ and level $\alpha$ replaces $\tau$ as in (14.7), to follow the Economic Theory notation. Indeed, from,

$$P(X \geq \alpha) \leq e^{-\xi\alpha} \mathcal{M}(\xi) = \alpha$$

one obtains

$$\alpha = \alpha(X; \xi) = \xi^{-1} \ln \left[ \frac{\mathcal{M}(\xi)}{\alpha} \right] \tag{14.13}$$

i.e., (14.12) can be evaluated, Ahmadi-Javid [1]

As an example, we let the random variable $X$ follow a Normal distribution, with mean $\mu$ and variance $\sigma^2$, i.e., $X \sim N(\mu, \sigma^2)$, then, it can be evaluated that

$$\text{EVaR}_{1-\alpha} = \mu + \sigma \left[ -2 \ln \alpha \right]^2 \tag{14.14}$$

Relation (14.14) deserves a particular investigation for the Economical applications.

### *14.3.2 Convex Functions in Decision Theory*

Now, let $\tau(\cdot)$ be any Statistical Test, Dudewicz [12], among others, for this test the Type I error $\alpha(\tau)$, is assigned to it and the Type II error, usually devoted as $1 - \beta(\tau)$. Practically the Type I error is the probability of rejecting $H_0$ hypothesis, Fergunson [16], while Type II error is the probability of accepting a false $H_0$. Thus the point $r(\tau) = (\alpha(\tau), 1 - \beta(\tau))$ known as the "risk point" is important in Decision Theory, as it "measures" how "good" or how "bad" is the test $\tau(\cdot)$, or the smaller the coordinates the better the test is.

The set of the risk points is convex in the sense that the following theorem holds, Dudewicz [12], Ferguson [16], among others. The Risk Analysis can be extended to a number of applications, Kitsos [27].

**Theorem 14.3.1** *For any hypothesis-testing problem of the form $H_0$ vs $H_1$, with both $H_i$ , $i = 0, 1$, to be simple the set $K_r = \{r(\tau) : \tau(\cdot)$ is a test$\}$ of risk points, of all possible test is a convex subset of the unit square $I_2 = [0, 1]^2$*

Notice that the above mentioned $K_r$ set includes the points $(1, 0)$ and $(0, 1)$ that is includes the extreme case for type I or type II error, as trivially this is true for $\tau_1 \equiv 1$ and $\tau_2 \equiv 0$ the $(1, 0)$ and $(0, 1)$ values are evaluated. Practically this means that the researcher can reach a "certain" test (with a very "big" sample size and well collected sample, for example, or for obvious "social/medical" problems).

Notice that the Risk Analysis problem as it has been expressed by Kitsos [27] includes a wide class of problems and not only the Decision Theory approach, the pioneering field of Risk Analysis with Game Theory, to be one of the first areas that devoted interest on Risk Theory.

In principle a statistical decision problem (sdp) is a triplet $(\Theta, D, l)$ and a random variable (rv) $X$.

The rv $X$, known as the data, follows a distribution function (d.f) $F(x, \theta)$ where $\theta$ is unknown, but it is known that $\theta \in \Theta$ with $\Theta$ known as the parameter space. Let $\mathcal{X}$ be a the set of possible values of the rv $X$, and $\theta$ is called the state of nature (here is a difference with the known parameter and parameter space in classical situation) while $\Theta \neq \varnothing$ is the parameter space and $D \neq \varnothing$ is called the decision space with "$l$" known as the loss function.

Moreover for every $\theta \in \Theta$ and $d \in D$ a real number $l(\theta, d)$ is assigned, while a decision rule is defined as

$$s : \mathcal{X} \to D : x \to s(x) = d$$

and the risk function,

$$r_\theta = E_\theta \left[ l(\theta, s(x)) \right] \tag{14.15}$$

Now if we consider $(\Theta, D, l)$, $X$ to be a statistical decision problem with $\Theta$ finite that is $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$, then the set of all risk points of decision rules is

$$S = \left\{ (y_1, \cdots, y_k) : y_i = r_{\theta_i}(s) \ , \ i = 1, 2, \cdots, k \right\}$$

is a convex subset of $\mathbb{R}^k$ , $S \subseteq \mathbb{R}^k$.

A converse of the above thought is: Any convex set $S \subseteq \mathbb{R}^k$ is the set of risk points of at least one decision problem $(\Theta, D, l)$, $X$. Note that convex sets and convex analysis are widely used in a number of numerical examples in Dudewicz [12].

### *14.3.3 Convex Functions for ML*

The main problem in Statistical Inference is related with the choice of the appropriate method to Maximize the Likelihood (ML) and obtain the Maximum Likelihood Estimation (MLE). Usually the Scoring Method, Hald [19], provides a solution. The method might be impractical when many parameters are considered. In the particular case of GLM (Generalized Linear Models) the log-Likelihood can be proved helpful to the researcher, Green (1984).

From Numerical Analysis, point of view, associated with the statistical inference, the modified Newton iteration can be adapted, or even more complicated algorithms the Davidon-Flatcher-Power being an early one, which is a globally convergent from convex functions algorithm, Powell (1971), or solving the "canonical equations" to evaluated the MLE numerically, Ortega and Rheinboldt [32].

In principle the exponentially family of distributions $f(x;\theta)$ , $\theta \in \Theta \subseteq \mathbb{R}^p$ is generated by a non-negative measure $\mu$ say, on the sample space $\Omega$, when a sufficient statistic, $S(x)$ say, with $S(x) \in \mathbb{R}^p$ such that $\mu \circ S^{-1}$ it is not concentrated on any affine hyperplane, is defined as

$$f(x;\theta) = \frac{1}{\phi(\theta)} \exp\left[S(x)^T\theta\right] \ , \ \theta \in \Theta \subseteq \mathbb{R}^p, p = \dim \Theta \tag{14.16}$$

with $\Theta = \left\{\theta | \phi(\theta) = \int \exp\left\{\left[S(x)^T\theta\right]\right\} \mu(dx) < \infty\right\}$ where $S^T$ is the transposed $S$.

The family is regular when $\Theta$ is open and moreover when the observations $S(x)$ belong in the interior of the convex support then MLE exists, is unique within $\Theta$ and satisfies

$$E_\theta\left[S(x)\right] = S(x) \tag{14.17}$$

see Barndorff-Nielsen [4] for details, on this fundamental result on the regular exponential family.

In particular when $p = 1$, with the observations of the sufficient statistic contained in the interior of its convex support, then, the modified Newton-Raphson algorithm, Ortega and Rheinboldt [32], converges to MLE for any initial value $\theta_0 \in \Theta$. The choice of $\theta_0$ is crucial, as if $\theta_0$ is "far" from the true parameter value, i.e., does not lie in the neighborhood of the "true" parameters, the Newton-Raphson diverges, Ford et al. [17]. The evaluation of the MLE it is clear it is fundamental in Statistic, Hald [19]. The iterative scheme of Newton-Raphson supports the calculations as well as a number of Numerical Analysis, and not only, techniques to calculate the optimal points. The MLE evaluated is a typical example where three different fields are facing the problem: Mathematics, Numerical Analysis and Statistics. The Mathematical framework suffers in the discrete case but still is applied.

## 14.4 Optimal Experimental Design

Optimal Experiment Design Theory, either Linear, Silvey [38] or Non-linear, Kitsos [24] is based on Carathéodory's Theorem, so it is clear that convex analysis is involved. Moreover both the geometrical interpretation and the optimality measures (optimal designs) are based to convex hulls and Gâteau derivatives, see (2.0.1), for the Fisher's Information matrix respectively, see Sects. 14.2 and 14.3.

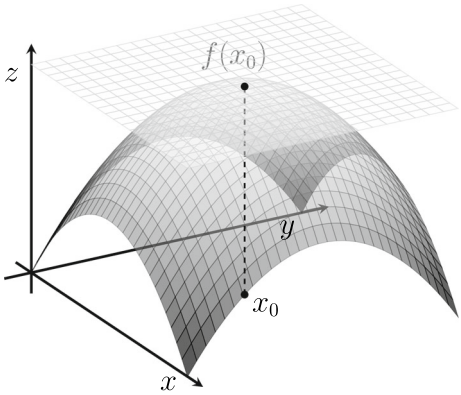



**Fig. 14.2** The Response Surface and the optimal points. The confidence interval of $f(x_0)$ lies on the tangent plane

### *14.4.1 Response Surface Method*

There is a large number of references describing the method, the most important are cited to Kitsos et al. [27]. A number of different experiment design (central composite design, Box-Behnken, etc.) are adopted to approach the second center surface, see Fig. 14.2. Most of the proper designs are D-optimal, the method uses a "hill-climbing" techniques to reach the optimum and attends many applications for the convex surface that Statistics is approaching, not only for industry. The tangent plane is useful for confidence intervals.

1. Recall that a set $T \in \mathbb{R}^n$ is a supporting hyperplane if and only if:

SH1: $T$ belongs in one of the two closed half-spaces bounded by the hyperplane.
SH2: $T$ has at least one boundary-point on the hyperplane.

2. Notice that the domain $D$ of the likelihood function is actually the parameter space, which we denoted $\Theta$ in the paper. It is considered, on principle, a subset of Euclidean space $\mathbb{R}^P$ with addiction restriction $v(\theta)$ say, i.e., $\theta = \left\{\theta \in \mathbb{R}^P : v(\theta) = 0\right\}$. This "restriction" $v(\theta)$ function is the considered $\phi(\theta)$ function in the above definition.

The Statistical line of development of the RSM, and the optimal points under investigation, in principle the maximum, has been discussed by Kitsos et al. [27], while the Response Surface problem as a general surface problem in Mathematics, has been discussed by Kitsos and Fatouros [28].

### *14.4.2 Information Matrix*

For the given design measure $\xi$ from the family of measures $\Xi$, $\xi \in \Xi$ the Fisher's Information matrix $I(\xi)$, defined through the Lebesgue-Stieltjes integral is

$$I(\xi) = \int_X f(x) f^T(x) \xi(dx) \tag{14.18}$$

The support point of $\xi$ for $x \in X$ is the one with $\xi(x) > 0$, while the set of support design points of $\xi$ is the set $S_\xi$,

$$S_\xi = \{x \in X : \xi(x) > 0\}$$

Considering the set $\mathcal{M}$ of the information matrices

$$\mathcal{M} = \{I(\xi), \xi \in \Xi\} \tag{14.19}$$

$\mathcal{M}$ is a convex set, while all the information matrices are symmetric and non-negative finite.

If the underlying model is non-linear, i.e., $f(x, \theta)$ the underlying function describing the phenomenon is non-linear, then Fisher's Information matrix depends on the unknown parameters we want to estimate.

Moreover the variance-covariance matrix $\mathrm{Cov}(\theta)$ is related (approximately in non-linear case) to Fisher's Information matrix, i.e., $\mathrm{Cov}(\theta) \cong I^{-1}(\theta, \xi)$, is depending, then, on $\theta$.

For the continuous case $I(\theta, \xi)$ is defined as

$$I^*(\theta, \xi) = \int_\chi I(\theta, x) \xi(dx) \tag{14.20}$$

while $I(\theta, x)$ has elements $I_{ij}(\theta, x)$ defined

$$I_{ij} = E_\theta \left[ \frac{\partial \log f(X; \theta)}{\partial \theta_i} \frac{\partial \log f(X; \theta)}{\partial \theta_j} |_\theta \right] \tag{14.21}$$

In the case that for the response $y$ holds $E(y) = f^T(x)\theta$ then $I^*(\theta, \xi) \equiv I(\xi)$, i.e., it is not a function of $\theta$.

The idea to whether the information matrix is considered to be "small" or "large" is behind the criterion function $\Phi$. The notation $\Phi$ is widely used, Kiefer and Wolfowitz [23], Silvey [38], Wynn [47, 48]. A criterion function is a function from the non-negative defined set of matrices $p \times n$ $NMat(p, n)$ to $R$, i.e.,

$$\Phi : NMat(p, n) \to \mathbb{R} : I(\xi) \to \Phi[I(\xi)] \tag{14.22}$$

If we consider the set of design measures

$$\Xi^* = \left\{ \xi^* \in \Xi \, , \; \xi^* \text{ is } \Phi - \text{optimal} \right\} \tag{14.23}$$

The $\Phi-$ efficiency can be defined as

$$\text{eff}\,(\xi) = \frac{\Phi\,[I\,(\xi)]}{\Phi\,[I\,(\xi^*)]}\ ,\ \xi \in \Xi\ ,\ \xi^* \in \Xi^* \tag{14.24}$$

It was [37] who suggested there is a duality between the D-optimality design measure and the minimum ellipsoid problem, see also Sibson [36], Whittle (1971). So the optimality criteria were limited with the "induced design space", with geometry eventuality, Ford et al. [18].

For the Non-linear case the geometrical interpretation for the alphabetic optimality criteria has been tracked by Kitsos [24].

The duality problem considers the minimal ellipsoid, $\mathcal{E}$ say, defined as

$$\mathcal{E} = \left\{w \in \mathbb{R}^p : w^T \mathcal{M}^{-1}(\xi^*) w = p\right\} \tag{14.25}$$

with $\xi^*$ being the D-optimal design measure. The D-optimal designs remain invariant to any non-degenerate linear transformation. A compact review of optimality criteria is in the next section.

With the minus sign in the optimal criterion, see Sect. 14.3, it is requested to be maximum and there are convex on the set $\mathcal{M}$, while with the plus sign is the dual.

For the defined (average-per-observation) information matrix $I^*(\theta, \xi)$ as in (14.23), the idea of Caratheodory's Theorem, Silvey [38], among others, provides food for thought even in non-linear case, for further development.

It was indeed [42] who proved the following Theorem for the Non-Linear Design Theory:

**Theorem 14.4.1** *For any parameter $\theta$ from the parameter space $\theta \in \Theta \subseteq \mathbb{R}^P$.*

*i* $I^*(\theta, \xi)$ *is symmetric, non-negative definite for any* $\xi \in \Xi$.
*ii* *The set* $\mathcal{M}(\theta) = \{I^*(\theta, \xi),\ \xi \in \Xi\}$ *is convex and compact.*
*iii* *The extreme points of* $\mathcal{M}(\theta)$ *are each of the form* $I^*(\theta, \xi)$ *for some* $\xi$ *from the design space.*
*iv* *For any* $\xi \in \Xi$ *there exists* $\zeta \in \Xi$ *assigning positive weight to at most* $\frac{1}{2}p(p+1)+1$ *points in the design space* $X$ *and such that* $I^*(\theta, \xi) = I^*(\theta, \zeta)$.

For a discuss on the choice on $X$ convex decreasing function of the set of non-negative definite matrices, as well as, the special case of the optimal design, for a subset of parameters and the implications of the "induced design space" in linear case to a more complicated one in Non-linear case [17].

The geometrical insight, for the convex criterion function, in the linear case has been tackled by the "equivalence theorem" of Kiefer and Wolfwitz [23], Kiefer [22].

The idea behind is that due to the transformation of the design space $X$ by a continuous function $g(X) = X_0$, the family $n$ design measure has been reduced to

$$\Xi_0 = \left\{\xi \in \Xi,\ \xi_0 = \xi\ g^{-1}\right\} \tag{14.26}$$

And thus the measure is preserved. Moreover for the convex decreasing criterion function [36] proves the following theorem

**Theorem 14.4.2** *Let $X_0$ be a compact set which spans $\mathbb{R}^p$. The D-optimal design problem from $X_0$ is the dual of the minimum ellipsoid problems from $X_0$ and the two problems share a common extreme value.*

In practice the kernel of the minimal ellipsoid problem is $I^{-1}(\xi)$ in linear case, $I^{*-1}(\xi, \widehat{\theta})$ in Non-linear case, Kitsos [24]. For the same $X_0$ as above [39] proved for the $D_s$-optimality criterion.

**Theorem 14.4.3** *Let $X_0$ span the leading S-dimensional co-ordinate subspace. Then for $X_0$ the $D_s$-optimal design problem is the dual of the thinnest cylinder problem and the two problems share a common extreme volume.*

Thus with the duality theory, for convex criterion, the optimal experimental design problem was reduced to a geometrical one. In the next section the most popular optimality criteria for the experimental design theory are introduced and discussed.

### *14.4.3 Optimality Criteria*

For the information matrix $I(\xi)$, different criteria have been imposed either in the linear case, Silvey [38], Federov [15], for the non-linear case, Kitsos [26], Ford et al. [17], known as alphabetical criteria. All these criteria are based on the convex criterion function $\Phi$. The defined criteria are reviewed here, with the most popular-one:

$$\Phi_D(I(\xi)) = -\frac{1}{p}\log|I(\xi)|$$

to manipulate the D-optimality. The D-optimality was introduced by Wald (1973). It was [37] who suggested that there is a analogy between D-optimality and the minimum ellipsoid problem, also Sibson [36], Whittle (1971), while Pázman et al. [33] worked for the Experimental Design from the Convex Analysis point of view.

The A-optimality criterion was introduced by Elving [14] and Chernoff [9] as

$$\Phi_A(I(\xi)) = -\frac{1}{p}\mathrm{Tr}I^{-1}(\xi)$$

The E-optimality was introduced by Ehrenfeld (1955), using the minimum eigenvalue of the Information Matrix $I(\xi)$

$$\Phi_E(I(\xi)) = \lambda_{\min}[I(\xi)]$$

For estimating a linear combination of the parameters $\theta$, through a vector $c$, say $c^T\theta$, with $\|c\| = 1$ we consider

$$\Phi_*(I(\xi)) = \lambda_{\min}[I(\xi)]$$

The G-optimality criteria is defined as

$$\Phi_G[I(\xi)] = -\inf\left[f^T(x)I^{-1}(\xi)f(x)\right],\ \widehat{y} = f^T(x)\widehat{\theta}$$

Notice that $\mathrm{Var}(\widehat{y})$ is proportional to $f^T(x)I^{-1}(\xi)f(x)$.

The c-optimality criterion Elving [14] it used to estimate the linear combination of estimates $c^T\theta$, with $\|c\| = 1$ we consider

$$\Phi_c(I(\xi)) = -c^T I^{-1}(\xi)c$$
$$\Phi_c(I(\xi)) = \mathrm{Tr}\left[I^{-1}(\xi)cc^T\right]$$

The geometrical interpretation of c-optimality is really very nice and it has been adopted in applications.

For all these criteria, for each case it holds on an "equivalence theorem" of the form,

**Theorem 14.4.4** *The function $\Phi_i$ $i = D, A, E, G$ is:*

1. *Continuous on the set of information matrix $I(\xi)$.*
2. *Convex on $\mathcal{T}$ and strictly convex on $\mathcal{T}_T$.*
3. *Differentiable whenever is finite. Moreover for*

$$\textit{D-criterion: } \nabla\left[-\log|I(\xi)|\right] = -I^{-1}(\xi).$$
$$\textit{A-criterion: } \nabla\left[tr(I(\xi))\right] = -I^{-2}(\xi).$$

*(Where $\nabla\Phi[I(\xi)]$ is the gradient of the function $\Phi$)*

It was under this consideration that Kiefer and Wolfowitz [23] proved their very well known general equivalence theorem. Moreover it results that for the set of all optimal design measure $\xi^*$, say $\Xi^*$, under (14.23) the set $\Xi^*$ is convex and $I(\xi^*)$ is the same for all $\xi^* \in \Xi^*$.

For Non-linear problems the vector of the unknown parameters is involved in Fisher's information matrix $I(\xi, \theta)$ and Titterington [41], proved that when $\theta$ takes it's "true" value, the linear case can be applied. Certainly there is a need of an initial guess for $\theta$, to work, usually sequentially to estimate $\theta$, Kitsos [25]. Although most of the fundamental results of Optimal Design of Experiments took place some years ago, still theory is mainly developed due to Convex Analysis.

### *14.4.4 Convex Assumptions on Information Theory*

In this section we present a compact consideration of the necessity of convex assumption to statistical divergence measures, Cover and Thomas (1991), in Information

Theory. Given two probability distribution with respective densities $p$ and $q$, say, the Bregman divergence is obtained for a strictly convex generator $G(\cdot)$ as

$$D_B(p \parallel q) = \int \left\{ [G(p(x)) - G(q(x))] - (p(x) - q(x))G^{'}(q(x)) \right\} dx \quad (14.27)$$

The Kullback-Leibler divergence is obtained for the Generator $G(x) = x \log x$, namely

$$D_{KL}(p \parallel q) = \int_{\mathbb{R}^P} p(x) \log \frac{p(x)}{q(x)} dx \quad (14.28)$$

For a strictly convex generator $G(\cdot)$ the Jensen generator is

$$D_{JS}^{(\alpha)}(p \parallel q) = \int [\alpha G(p(x)) + (1-\alpha)G(q(x)) - G(\alpha\ p(x) + (1-\alpha)\ q(x))]\, dx \quad (14.29)$$

It is easy to verify that

$$D_{JS}^{(\alpha)}(p \parallel q) = D_{JS}^{(1-\alpha)}(p \parallel q) \quad (14.30)$$

In limit it can be evaluated that

$$\lim_{\alpha \to 1} D_{JS}^{(\alpha)}(p \parallel q) = \frac{1}{1-\alpha} D_B(p \parallel q) \quad (14.31)$$

$$\lim_{\alpha \to 0} D_{JS}^{(\alpha)}(p \parallel q) = \frac{1}{\alpha} D_B(p \parallel q) \quad (14.32)$$

See also Basseville and Cardoso [6] Amari's $\alpha$-divergence are on the $\alpha$-flat spaces of Information Geometry namely:

$$D_A^{(\alpha)}(p \parallel q) = \begin{cases} \frac{4}{1-\alpha^2}(1 - D_{c_{\frac{1-\alpha}{2}}}(p \parallel q)) & \alpha \neq \pm 1 \\ KL(p \parallel q) & \alpha = -1 \\ KL(q \parallel p) & \alpha = 1 \end{cases} \quad (14.33)$$

with the $D_c^{(\cdot)}(p \parallel q)$ being the Chernoff $\alpha$-coefficient measure of similarity (or divergence) of two distributions $p$ and $q$.

$$C_\alpha(p \parallel q) = -\log \min_{\alpha \in [0,1]} \int p^\alpha(x)\, q^{1-\alpha}(x) dx \quad (14.34)$$

Given the Chernoff $\alpha$-coefficient measure of similarity the Chernoff $\alpha$-diverges is obtained as

$$D_C^\alpha(p \parallel q) = \frac{1}{\alpha(1-\alpha)} (1 - c_\alpha(p \parallel q)) \quad (14.35)$$

see also the pioneering paper of Chernoff [8].

Now recall (14.9), (14.10) and let $(\Omega, A, P)$ be a probability space. Moreover let $\Lambda$ be a set of probability measures on $(\Omega, A)$ such that

$$\Lambda = \{q << p : D_{KL}(q \parallel p) \leq -\ln a\} \tag{14.36}$$

with $D_{KL}(. \parallel .)$ as in (14.28). Then the dual definition of the defined already in Sect. 3.1 $\text{EVaR}_{1-\alpha}$ is

$$\text{EVaR}_{1-\alpha}(X) = \sup\left\{\left[E_q(X)\right], q \in \Lambda\right\} \tag{14.37}$$

see also Ahmade-Javid [1].

The Kullaback-Leiber divergence attracts interest in applications and therefore the Kullaback-Leiber divergence of the $\gamma$-order generalized Normal distribution over the $t$-distribution was evaluated, Toulias and Kitsos [43, 44] while a wave theoretical approach to the "distance" between two distributions can be faced through the Hellinger distance, Kitsos and Toulias [30].

## 14.5 Discussion

The target of this paper is to provide a compact consideration of the optimality in various Mathematical areas under the influence of the Convex Analysis. There are different lines of thoughts considering a Statistical problem: from the classical development of optimality based on Caratheodory's theory, Silvey [38] to the development of a canonical form, Ford et al. [18], to a convex approach [33].

It is not only the Convex Analysis working as a link to the Statistical areas we are referred: the provided "information" is already there, through convex considerations, in Optimal Design Theory, in Information Theory, although hidden sometimes through Geometry, Amari and Nagaoka [2]. We provide as an example the entropic value of risk (14.12), through Chernoff bounds and we came across the same idea through the Information Theory (14.37).

Therefore we provided evidence that Convex Analysis it is not only crucial in typical optimality methods, such as Least Squeares and Linear Programming, but to a number of other applications, faced early, Box and Lucas [7]. Simply we need to remain within the area of interest, through the convex combination, and we are looking for the optimum at the "edges" of the area, due to Caratheodory's theorem. The surprising, for the experimentalist, is that some points participate to the "optimum" with "zero measure", i.e., no experimentation at that points. As the computation is so easy in our-days the Convex Analysis can provide any kind of complicated "hill climbing" algorithm to reach the optimum, but the simplicity and beauty always follows Mathematics!

## References


1. Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012.
2. Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.
3. Ashay Ash and Samad Hedayat. An introduction to designh optimality with an overview of the literature. *Communications in Statistics-Theory and Methods*, 7(14):1295–1325, 1978.
4. Nielsen Barndorff. Information and exponential families; in statistical theory. 1978.
5. O. Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 1978.
6. Michele Basseville and J. F. Cardoso. On entropies, divergences, and mean values. In *Proceedings of 1995 IEEE international symposium on information theory*, page 330. IEEE, 1995.
7. George E. P. Box and Henry L. Lucas. Design of experiments in non-linear situations. *Biometrika*, 46(1/2):77–90, 1959.
8. Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.
9. Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pages 586–602, 1953.
10. Herman Chernoff. A note on an inequality involving the normal distribution. *The Annals of Probability*, pages 533–535, 1981.
11. Ding-Zhu Du, Panos M. Pardalos, and Weili Wu. *Mathematical theory of optimization*, volume 56. Springer Science & Business Media, 2013.
12. Edward J. Dudewicz. Introduction to statistics and probability. *Holt, Rinehart and Winston*, 1976.
13. Sylvain Ehrenfeld. Complete class theorems in experimental design. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 57–67. University of California Press, 1956.
14. Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, pages 255–262, 1952.
15. Valerii V. Fedorov. *Theory of Optimal Experiments*. Elsevier, 2013.
16. Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press, 1968.
17. Ian Ford, Michael D. Titterington, and Christos P. Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31(1):49–60x, 1989.
18. Ian Ford, Bernard Torsney, and C. F. Jeff Wu. The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 54(2):569–583, 1992.
19. Anders Hald. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14(2):214–222, 1999.
20. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.
21. S. Karlin and H. M. Taylor. A first course in stochastic processes, new york: Academic. *Cited on*, page 64, 1975.
22. Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, pages 849–879, 1974.
23. Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.
24. Christos P. Kitsos. *Design and inference in nonlinear problems*. University of Glasgow (United Kingdom), 1986.
25. Christos P. Kitsos. On the optimal continuous experimental design problem. *Discussiones Mathematicae Probability and Statistics*, 31(1–2):59–70, 2011.

26. Christos P. Kitsos. *Optimal experimental design for non-linear models: theory and applications*. Springer Science & Business Media, 2014.
27. Christos P. Kitsos. Risk analysis in practice and theory. *Statistical Modelling and Risk Analysis, by C.P Kitsos, T. Oliveira, F. Pierri, and M. Restaino*, 430:107–118, 2022.
28. Christos P. Kitsos and Stavros Fatouros. Geometry in quantitative methods and applications. In *Analysis, Geometry, Nonlinear Optimization and Applications*, pages 535–563. World Scientific Pub Co., 2023.
29. Christos P. Kitsos, Teresa Oliveira, and Amilcar Oliveira. Geometrical and optimization aspects for approximationg the response surface in industry. 2022.
30. Christos P. Kitsos and Thomas L. Toulias. Hellinger distance between generalized normal distributions. *Br. J. Math. Comput. Sci*, 21(2):1–16, 2017.
31. P. Lévy. *Processus stochastiques et mouvement brownien*. Gauthier-Villars, Paris, 1948.
32. James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.
33. Andrej Pázman, Markus Hainy, and Werner G Müller. A convex approach to optimum design of experiments with correlated observations. *Electronic Journal of Statistics*, 16(2):5659–5691, 2022.
34. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1969.
35. Horst Schubert. *Topology*. Allyn and Bacon series in advanced mathematics. Allyn and Bacon, 1968.
36. Robin Sibson. Contribution to discussion of results in the theory and construction of d-optimum experimental designs' by hp wynn. *J. Roy. Statist. Soc., Ser. B*, 34:181–183, 1972.
37. Samuel D. Silvey. Discussion of paper by H.P. Wynn. *Journal of the Royal Statistical Society: Series B*, 34:174–175, 1972.
38. Samuel D. Silvey. Optimal design. *Chapman and Hall*, 1980.
39. Samuel D. Silvey and Michael Titterington. A geometric approach to optimal design theory. 1973.
40. Joy A. Thomas and Thomas M. Cover. *Elements of information theory*. Wiley-Interscience, 2006.
41. Michael D. Titterington. Optimal design: some geometrical aspects of d-optimality. *Biometrika*, 62(2):313–320, 1975.
42. Michael D. Titterington. Geometric approaches to design of experiment. *Series Statistics*, 11(1):151–163, 1980.
43. Thomas L. Toulias and Christos P. Kitsos. Kullback-leibler divergence of the $\gamma$-ordered normal over t-distribution. *British Journal of Mathematics & Computer Science*, 2(4):198–212, 2012.
44. Thomas L. Toulias and Christos P. Kitsos. Information divergence and the generalized normal distribution: A study on symmetricity. *Communications in Mathematics and Statistics*, 9(4):439–465, 2019.
45. Abraham Wald. On the efficient design of statistical investigations. *The annals of mathematical statistics*, 14(2):134–140, 1943.
46. Peter Whittle. Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(1):123–130, 1973.
47. Henry P. Wynn. The sequential generation of $d$-optimum experimental designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664, 1970.
48. Henry P. Wynn. Jack kiefer's contributions to experimental design. *The Annals of Statistics*, 12(2):416–423, 1984.