TÜBİTAK

Turkish Journal of Mathematics

http://journals.tubitak.gov.tr/math/

Research Article

Turk J Math
(2018) 42: 257 – 271

doi:10.3906/mat-1607-20

Worst-case large deviations upper bounds for i.i.d. sequences under ambiguity

Mustafa Çelebi PINAR*
Department of Industrial Engineering, Faculty of Engineering, Bilkent University, Bilkent, Ankara, Turkey

Received: 11.07.2016 • **Accepted/Published Online:** 24.04.2017 • **Final Version:** 22.01.2018

Abstract: An introductory study of large deviations upper bounds from a worst-case perspective under parameter uncertainty (referred to as ambiguity) of the underlying distributions is given. Borrowing ideas from robust optimization, suitable sets of ambiguity are defined for imprecise parameters of underlying distributions. Both univariate and multivariate i.i.d. sequences of random variables are considered. The resulting optimization problems are challenging min–max (or max–min) problems that admit some simplifications and some explicit results, mostly in the case of the normal probability law.

Key words: Large deviations, ambiguity, robust optimization, ellipsoids, Legendre–Fenchel transform, min–max theorem.

1. Introduction

The purpose of this paper is to present an investigation of *large deviations* (see [9, 13] for gentle introductions to large deviations) upper bounds for i.i.d. sequences of random vectors (or random variables) when ambiguity believed to affect parameters of the underlying probability law is taken into account in a pessimistic, i.e. worst-case, fashion for an unwelcome event. It is well accepted that key parameters of commonly used distributions are rarely known with precision in practice. Therefore, addressing this imprecision is of great importance in modeling probabilistic phenomena. The present study is the result of an effort to apply some ideas from robust optimization to large deviations. Robust optimization was initiated by the seminal contributions of Ben-Tal and Nemirosvki [1] and El-Ghaoui and Lebret [10], and it is presently a very active field of investigation; see [3] for a comprehensive review. The spirit of robust optimization can be summarized as follows: faced with an optimization problem (e.g., an engineering design problem) where the data are subject to imprecision (typically, imprecision due to errors of estimation), find the best solution against the worst possible values of imprecise data in a judiciously chosen set of ambiguity. The specification of the set of ambiguity for the imprecise parameters typically reflects the degree to which one wishes to preserve one's design in the face of adversities of nature. In other words, a set of ambiguity that takes into account all possible occurrences of imprecise data may result in a very conservative or expensive solution, which may be impossible to implement. At the other extreme, a set of ambiguity leaving important information out may result in an unstable or fragile solution. Hence, the need to strike a balance in the choice of the ambiguity set. A second issue in the choice of ambiguity set is the geometry of the set, which affects the numerical solvability of the resulting problem. Here, it is important to specify sets

*Correspondence: mustafap@bilkent.edu.tr

Several useful suggestions of anonymous referees and the associate editor are acknowledged. Thanks are also due to Professor Francesco Caravenna from University of Milano-Bicocca for explaining large deviations to the author.

leading to convex and thus numerically solvable robust optimization problems, namely the so-called ellipsoidal, polyhedral, or norm sets; see [3]. On the other hand, the level of conservatism of the optimal robust solution also depends on the specification of the ambiguity set, e.g., a polyhedral ambiguity set based on the infinity norm may ignore dependencies among parameters and result in the worst values of all parameters at once. Ellipsoidal uncertainty sets are preferable in that respect since they mimic the engineering design approach that the value of a random quantity should not exceed a constant times its standard deviation. The reader is referred to the recent book [2] for a comprehensive coverage of robust optimization.

The present paper is not the first to explore worst-case large deviations asymptotics; see, e.g., [12, 14, 16]. The worst-case probability of an event A with respect to a set of probability measures (a capacity) is defined, and a general version of Cramér's theorem is proved in [12]. In [14], univariate i.i.d. processes are considered on a compact metric space with marginal distribution assumed to lie in a so-called *moment class* (a set of distributions with fixed first, and/or second, and/or third moment and so on). Then the worst-case rate function with respect to this moment class is studied in detail with application to queuing and information theory. In [16], large deviations theory is used to study the exponential rate of decrease of error probabilities for a sequence of decisions based on a test statistic sequence whose distribution is a member of a parametric class of distributions. An application to i.i.d. detection is also given. In particular, the set of distributions is specified as the *ϵ-contamination class* around a nominal distribution. This reference also studies the impact of applying convex conjugation to a worst-case cumulant generating function with respect to the set of distributions, instead of finding the convex conjugate function first and then passing to the worst-case estimate. The former operation leads to a lower bound to the tightest exponential rate, which is exact if the cumulant generating function is a closed, proper convex function for each distribution. Our research effort is also linked to a thread of research in mathematical finance referred to as "model uncertainty"; see, e.g., [5], where a set of distributions is given as potentially governing the evolution of a financial variable (e.g., a stock) and worst-case calculations are performed with respect to that set. In a reference related to the present paper [11], *robust* large deviations (among other things) for a coherent version of the entropic risk measure applied to risk pooling in the insurance industry are studied. In contrast to these references that usually deal with function spaces, the present paper focuses on specific distributions with uncertain parameters taking values in a specific set of ambiguity (ellipsoidal in the multivariate case) and explores (explicit) solvability of resulting optimization problems, with the exception of Section 4 where we deal with all discrete probability vectors resulting in a fixed mean for finite alphabets.

Consider the empirical means $\bar{S}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$, for i.i.d. d-dimensional random sequence $\{X_n\}$. Let θ be a vector of parameters controlling the probability law of X_1, and for $n \geq 1$, let $\mu_n^{(\theta)}$ be the law of the empirical mean of the n i.i.d. random variables. The "true" value of θ is assumed to lie in an ambiguity set $\mathcal{U}_\epsilon$ where ϵ controls the level of ambiguity against which one is prepared to protect oneself.

The logarithmic moment generating function (a.k.a. cumulant generating function) associated with the probability law $\mu_1^{(\theta)}$ of X_1 is defined as

$$\Lambda(z) = \ln \mathbb{E}[\mathrm{e}^{z^T X_1}]. \tag{1.1}$$

The Legendre–Fenchel transform of $\Lambda(z)$ is

$$\Lambda^*(x) = \sup_{z \in \mathbb{R}^d} \{z^T x - \Lambda(z)\}.$$

For fixed θ, it is well known that (see, e.g., [8], pp. 36–42)

$$\frac{1}{n}\ln \mu_n^{(\theta)}(\mathcal{C}) \leq -\inf_{y\in\mathcal{C}} \Lambda^*(y) \tag{1.2}$$

for every closed set $\mathcal{C}$. In the present paper we shall be dealing with the problem of obtaining upper bounds for the following quantity:

$$\sup_{\theta\in\mathcal{U}_\epsilon} \frac{1}{n}\ln \mu_n^{(\theta)}(\mathcal{C})$$

for every closed set $\mathcal{C}$, i.e. we shall concern ourselves with studying optimization problems of the form

$$\sup_{\theta\in\mathcal{U}_\epsilon} \{-\inf_{y\in\mathcal{C}} \Lambda^*(y)\}$$

since we have immediately using (1.2) the worst-case upper bound:

$$\sup_{\theta\in\mathcal{U}_\epsilon} \frac{1}{n}\ln \mu_n^{(\theta)}(\mathcal{C}) \leq \sup_{\theta\in\mathcal{U}_\epsilon} \{-\inf_{y\in\mathcal{C}} \Lambda^*(y)\}. \tag{1.3}$$

The paper is organized as follows. In Section 2, we shall treat the problem in two cases of univariate random sequences where the controlling parameter(s) are subject to ambiguity. In Section 3, we pass to random vector sequences. We obtain our most explicit worst-case bounds in the Gaussian case. A slightly more general result is obtained for a "shifted" sequence where ambiguity is placed on the shift parameter and no specific assumption on the ambiguity set is made except for closedness and convexity. We also look at a Poisson random vector sequence example from queuing theory. A brief excursion into the Sanov theorem and the method of types is given in Section 4. It is our hope that the present paper will trigger further work on the subject of large deviations estimation under model uncertainty.

2. Univariate examples

In this section as an introduction two cases illustrate the ideas of the paper in the context of unidimensional i.i.d. sequences.

2.1. An exponentially distributed sequence

We begin with the exponential distribution, i.e. we assume the law governing the i.i.d. sequence X_i is the exponential law with mean $1/\lambda$. It is well known that Λ^* is given as:

$$\Lambda^*(x) = \lambda x - \ln \lambda x - 1, \text{ for } x > 0,$$

(it is equal to ∞ otherwise). Specifying the natural ambiguity set $\mathcal{U} = [a, b]$ (we omit ϵ), after straightforward algebraic calculation one obtains for any closed interval $\mathcal{C}$ the following worst-case large deviations principle (LDP) upper bound:

$$\sup_{\lambda\in[a,b]} \frac{1}{n}\ln \mu_n^{(\lambda)}(\mathcal{C}) \leq -\inf_{x\in\mathcal{C}} \phi(x)$$

where

$$\phi(x) = \begin{cases} bx - \ln bx - 1 & x < 1/b \\ ax - \ln ax - 1 & x > 1/a \\ 0 & 1/b \leq x \leq 1/a. \end{cases}$$

Figure 1 exhibits plots of the functions Λ^* and the piecewise function above resulting from the worst-case LDP bound for $a = 1$ and $b = 2$ and $\lambda = 1.8$ for Λ^*. Figure 2 contains the two functions when $\lambda = 1.2$ in Λ^*. In both figures, the dotted curve is the piecewise function of the worst-case LDP bound, while the dashed curve is the Legendre–Fenchel function Λ^*.

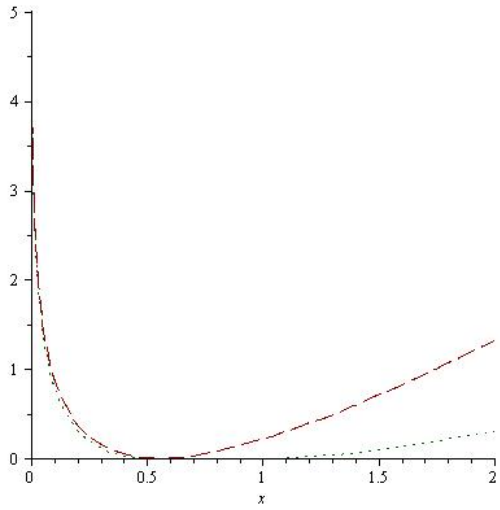

Figure 1. Exponential case: the Legendre–Fenchel function Λ^* for $\lambda = 1.8$ (dashed curve) and the worst-case LDP bound function (dotted curve) for $a = 1$ and $b = 2$.

Note that for "true" λ close to the upper end of the interval the two functions are very close for small values of x and differ for larger values. This observation is reversed when the true λ is closer to the lower end of the interval. We note that the rate function is zeroed out in the ambiguity interval (or an interval induced by the ambiguity interval), an observation also made in [14] (see fig. 2 of [14]).

2.2. A normally distributed sequence under joint (μ, σ)-ambiguity

The final example in this section is for a normally distributed i.i.d. sequence X_i with the Legendre–Fenchel transform of the cumulant generating function given as

$$\Lambda^*(x) = \frac{(x-\mu)^2}{\sigma^2}$$

where μ and σ^2 are the mean and the variance of the normal probability law governing X_1. For ease of notation, we use s for the variance σ^2. We shall consider a joint ambiguity structure on μ, s of the following form:

$$\mathcal{U}_\epsilon = \{(\mu, s) : \sqrt{(\mu-\hat{\mu})^2 + (s-\hat{s})^2} \leq \epsilon\}.$$

One could certainly consider separate/independent ambiguity in μ and σ^2. However, this independent structure again leads to rather predictable extreme behaviour for μ and σ as the reader can easily verify. Furthermore, a joint structure remains tractable in the univariate case as opposed to the multivariate normal case, which is treated in the next section.

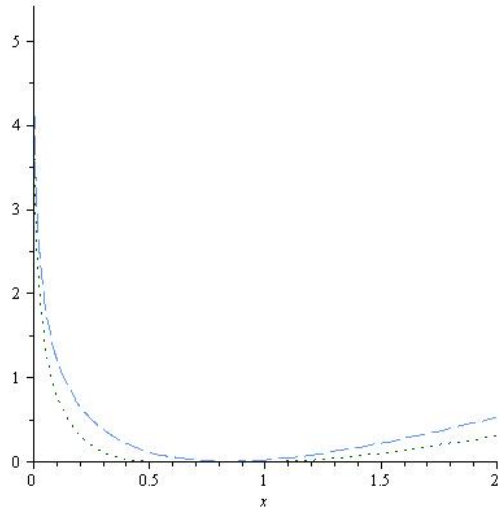

Figure 2. Exponential case: the Legendre–Fenchel function Λ^* for $\lambda = 1.2$ (dashed curve) and the worst-case LDP bound function (dotted curve) for $a = 1$ and $b = 2$.

We are thus dealing with this problem:

$$\sup_{(\mu,s)\in\mathcal{U}_\epsilon} -\inf_{x\in\mathcal{C}} \frac{(x-\mu)^2}{s},$$

or equivalently with

$$\sup_{x\in\mathcal{C}} \sup_{(\mu,s)\in\mathcal{U}_\epsilon} -\frac{(x-\mu)^2}{s}.$$

The solution of the inner sup problem boils down to a unidimensional root finding problem for a second-degree polynomial equation.

Proposition 1 *For a normally distributed i.i.d. sequence* $\{X_n\}$ *where the parameters* μ *and* σ^2 *are confined to the ball* $\mathcal{U}_\epsilon = \{(\mu, s) : \sqrt{(\mu-\hat{\mu})^2 + (s-\hat{s})^2} \leq \epsilon\}$ *the following hold:*

1. *For* $x > \hat{\mu} + \epsilon$ *we have*

$$\sup_{(\mu,\sigma^2)\in\mathcal{U}_\epsilon} \frac{1}{n} \ln \mu_n^{(\theta)}(\mathcal{C}) \leq \sup_{x\in\mathcal{C}} -\frac{(x-\mu^*)^2}{s^*},$$

where $\mu^* = \hat{\mu} + \gamma^*$, $s^* = \frac{(x-\hat{\mu}-\gamma^*)\gamma^*}{2\sqrt{\epsilon^2-(\gamma^*)^2}}$, *and* γ^* *is a positive root (in the interval* $(0, \epsilon)$*) of the equation*

$$\gamma^2 + \gamma(x-\hat{\mu}) - 2\hat{s}\sqrt{\epsilon^2-\gamma^2} - 2\epsilon^2 = 0.$$

2. *For* $x < \hat{\mu} - \epsilon$ *we have*

$$\sup_{(\mu,\sigma^2)\in\mathcal{U}_\epsilon} \frac{1}{n} \ln \mu_n^{(\theta)}(\mathcal{C}) \leq \sup_{x\in\mathcal{C}} -\frac{(x-\mu^*)^2}{s^*},$$

where $\mu^* = \hat{\mu} - \gamma^*$, $s^* = \frac{(-x+\hat{\mu}-\gamma^*)\gamma^*}{2\sqrt{\epsilon^2-(\gamma^*)^2}}$, *and* γ^* *is a positive root (in the interval* $(0,\epsilon)$*) of the equation*

$$\gamma^2 + \gamma(\hat{\mu} - x) - 2\hat{s}\sqrt{\epsilon^2 - \gamma^2} - 2\epsilon^2 = 0.$$

3. *For* $x \in [\hat{\mu} - \epsilon, \hat{\mu} + \epsilon]$

$$\sup_{(\mu,\sigma^2)\in\mathcal{U}_\epsilon} \frac{1}{n} \ln \mu_n^{(\theta)}(\mathcal{C}) \leq 0,$$

i.e. $\mu^* = x$, $s^* = \hat{s}$ *(s^* is irrelevant).*

Proof The inner problem $\sup_{(\mu,s)\in\mathcal{U}_\epsilon} -\frac{(x-\mu)^2}{s}$ is a convex optimization problem (the objective function is concave and the set of feasible solutions is convex). Since the set of feasible solutions is compact, we can replace the sup by max. The necessary and sufficient Karush–Kuhn–Tucker (KKT) conditions (with nonnegative multiplier λ) give:

$$-\frac{1}{s}(x-\mu) + \lambda(\mu - \hat{\mu}) = 0, \tag{2.1}$$

$$-\frac{(x-\mu)^2}{s^2} + 2\lambda(s - \hat{s}) = 0, \tag{2.2}$$

$$(\mu - \hat{\mu})^2 + (s - \hat{s})^2 = \epsilon^2, \tag{2.3}$$

$$\lambda\left(\epsilon^2 - (\mu - \hat{\mu})^2 - (s - \hat{s})^2\right). \tag{2.4}$$

We ignore momentarily the requirement that $s > 0$. We make the ansatz $\mu^* = \hat{\mu} + \gamma$ where γ is positive. If we can find μ^*, s^*, λ^* satisfying the KKT optimality conditions (with a positive γ and s^*), the proof is complete. From (2.3) we have $s - \hat{s} = \epsilon^2 - \gamma^2$. Using this in (2.1) we obtain $\lambda = \frac{2\sqrt{\epsilon^2-\gamma^2}}{\gamma^2}$. Since we have two expressions for s^* from (2.2) and (2.3), they should agree, i.e. we have the equation

$$\frac{(x - \hat{\mu} - \gamma)\gamma}{2\sqrt{\epsilon^2 - \gamma^2}} = \hat{s} + \sqrt{\epsilon^2 - \gamma^2},$$

which gives the nonlinear equation

$$\gamma^2 + \gamma(x - \hat{\mu}) - 2\hat{s}\sqrt{\epsilon^2 - \gamma^2} - 2\epsilon^2 = 0.$$

The function on the left of the equation has a negative value at $\gamma = 0$ and a positive value at $\gamma = \epsilon$, which implies by continuity that the equation has a positive root in the interval $(0, \epsilon)$ provided that $x > \hat{\mu} + \epsilon$.

If $x \leq \hat{\mu} - \epsilon$ then we take the ansatz $\mu = \hat{\mu} = \gamma$ for $\gamma > 0$, and we proceed exactly as in the previous part to obtain the nonlinear equation:

$$\gamma^2 + \gamma(-x + \hat{\mu}) - 2\hat{s}\sqrt{\epsilon^2 - \gamma^2} - 2\epsilon^2 = 0,$$

where the function on the left of the equation has a negative root at $\gamma = 0$ and a positive root at $\gamma = \epsilon$ provided $x < \hat{\mu} - \epsilon$.

Finally, for part 3, it is easy to verify that $\mu^* = x$ and $s^* = \hat{s}$ satisfy the optimality conditions with $\lambda = 0$ and the constraint inactive. □

3. The multivariate case

In this section we examine worst-case uniform LDP bounds under model uncertainty for the empirical means $\bar{S}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$, for i.i.d. d-dimensional random sequences. We start with the Gaussian case.

3.1. Gaussian sequences

Let $\bar{S}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ denote the empirical means for i.i.d. d-dimensional Gaussian sequence $\{X_n\}$ with mean m and covariance matrix K assumed invertible. For all $m \in \mathbb{R}^d$ and $n \geq 1$, let $\mu_n^{(m)}$ be the law of the empirical mean of n i.i.d. $\mathcal{N}(m, K)$ random variables. The "true" value of m is assumed to lie in an ellipsoid $\mathcal{U}_\epsilon = \{m | \|K^{-1/2}(m - \bar{m})\| \leq \epsilon\}$ around a nominal mean value $\bar{m}$, where ϵ controls the ambiguity. We define the weighted norm of vector $x \in \mathbb{R}^d$ as $\|x\|_K = \sqrt{x^T K^{-1} x}$. Therefore, the set $\mathcal{U}_\epsilon$ is the closed ϵ-ball centered $\mathcal{U}_\epsilon$ at $\bar{m}$, $\bar{\mathcal{B}}(\bar{m}; \epsilon)$, with respect to that norm.

We give below, for each closed subset $\mathcal{C}$ of $\mathbb{R}^d$, an upper bound for $n^{-1} \ln \mu_n^{(m)}(\mathcal{C})$, uniform in $m \in \mathcal{U}_\epsilon$ (and $n \geq 1$). The proof is a simple exercise in KKT optimality conditions.

Proposition 2 *Under the above hypotheses,*

$$\sup_{m \in \mathcal{U}_\epsilon} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq - \inf_{y \in \mathcal{C}} \left[\mathbb{1}_{y \in \mathcal{U}_\epsilon^c} \frac{1}{2} (\|y - \bar{m}\|_K - \epsilon)^2 \right] y \in \mathcal{U}_\epsilon^c$$

for every closed set C.

Proof For fixed m and K, we have

$$\frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq - \inf_{y \in \mathcal{C}} \Lambda^*(y) = - \inf_{y \in C} \frac{1}{2} \|y - m\|_K^2$$

for every closed set $\mathcal{C}$. Now, consider the worst-case bound:

$$\sup_{m \in \mathcal{U}_\epsilon} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq \sup_{m \in \mathcal{U}_\epsilon} \sup_{y \in \mathcal{C}} -\frac{1}{2} \|y - m\|_K^2 m \in \mathcal{U}_\epsilon.$$

We have

$$\sup_{m \in \mathcal{U}_\epsilon} \sup_{y \in \mathcal{C}} -\frac{1}{2} \|y - m\|_K^2 = \begin{cases} 0 & \text{if } \mathcal{C} \cap \mathcal{U}_\epsilon \neq \emptyset \\ \sup_{y \in \mathcal{C}} -\frac{1}{2} \left(\|y - \bar{m}\| - \epsilon \right)^2 & \text{otherwise.} \end{cases}$$

Notice that this computation of the supremum admits a nice geometric interpretation: it is the problem of computing the projection of y onto $\mathcal{U}_\epsilon$ with respect to the weighted norm $\|.\|_K$. Obviously, when $y \in \mathcal{U}_\epsilon$, the solution is to take $m^* = y$. It is geometrically evident that the point in $\mathcal{U}_\epsilon$ closest to y with respect to the norm $\|.\|_K$ is the point

$$m^* = \bar{m} + \frac{\epsilon}{\|y - \bar{m}\|_K} (y - \bar{m}).$$

This solution can be obtained by direct application of the KKT theorem to the convex optimization problem over m for fixed y:

$$\max_{m \in \mathcal{U}_\epsilon} -\frac{1}{2} (y - m)^T K^{-1} (y - m).$$

One forms the Lagrange function with a nonnegative multiplier λ:

$$L(m,\lambda) = -\frac{1}{2}(y-m)^T K^{-1}(y-m) + \lambda(\epsilon^2 - (y-m)^T K^{-1}(y-m)).$$

The first-order conditions yield $m^* = \frac{y+2\lambda\bar{m}}{2\lambda+1}$. Substituting into the constraint assumed to be active, one gets $\lambda^* = \frac{\sqrt{(y-m)^T K^{-1}(y-m)}}{2\epsilon} - \frac{1}{2}$, from which the result follows after straightforward algebra. □

Remark. We note that the Legendre–Fenchel transform expression of the multivariate Gaussian, given as $(y-m)^T K^{-1}(y-m)$, is equal to (up to a constant) the Mahalanobis distance between two Gaussian distributions with means m and y and common variance-covariance matrix K, which is in turn equal to the differential relative entropy between these two Gaussians; see, e.g., [7] for this connection to machine learning and information theory.

Now, we assume that K is also ambiguous, independently from m. Hence, we consider ambiguity in (μ, K) where $\mu \in \mathcal{U}_\epsilon$ as above and K takes values in the set $\mathcal{K}_\delta = \{K \succeq 0 | \|K - \hat{K}\|_F \leq \delta\}$, where $\hat{K}$ is a symmetric positive definite matrix. Here, $\|X\|_F$ is the Frobenius norm of the matrix X, given as $\mathrm{Tr}(X^T X)$. Recalling the trace inner product of symmetric $n \times n$ matrices X and Y as $\langle X, Y\rangle = \mathrm{Tr}(XY)$, the norm constraint on K is equivalently written as $\sqrt{\langle K - \hat{K}, K - \hat{K}\rangle} \leq \delta$.

Now, we consider the problem

$$\sup_{m\in\mathcal{U}_\epsilon, K\in\mathcal{K}_\delta} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq \underbrace{\sup_{m\in\mathcal{U}_\epsilon, K\in\mathcal{K}_\delta} \left\{ -\inf_{y\in\mathcal{C}} \Lambda^*(y) \right\}}_{RHS}.$$

Proposition 3 *For i.i.d. d-dimensional Gaussian random sequence $\{X_n\}$ with mean m and covariance matrix K taking values in $\mathcal{U}_\epsilon$ and $\mathcal{K}_\delta$, respectively, we have*

$$\sup_{m\in\mathcal{U}_\epsilon, K\in\mathcal{K}_\delta} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq \sup_{y\in\mathcal{C}} \inf_{\lambda\in\mathbb{R}^d} F(\lambda),$$

where

$$F(\lambda) = \frac{1}{2}\lambda^T \hat{K} \lambda + \delta \|\lambda\lambda^T\|_F + \epsilon \sqrt{\lambda^T \hat{K} \lambda + \delta \|\lambda\lambda^T\|_F} + \lambda^T(\bar{m} - y).$$

Proof Here we shall deviate from the proof of the previous result since the Legendre–Fenchel transform of the cumulant generating function depends on K^{-1}, whereas we wish to work directly on K when K is ambiguous. We proceed as follows. Rewrite the RHS:

$$\sup_{m\in\mathcal{U}_\epsilon, K\in\mathcal{K}_\delta} \left\{ -\inf_{y\in\mathcal{C}} \Lambda^*(y) \right\} = \sup_{K\in\mathcal{K}_\delta} \sup_{m\in\mathcal{U}_\epsilon} \left\{ \sup_{y\in\mathcal{C}} -\Lambda^*(y) \right\}.$$

Using the definition of Λ^* we have

$$\sup_{K\in\mathcal{K}_\delta} \sup_{m\in\mathcal{U}_\epsilon} \sup_{y\in\mathcal{C}} \left\{ -\sup_{\lambda\in\mathbb{R}^d} \{\lambda^T y - \Lambda(\lambda)\} \right\}.$$

Since the sequence $\{X_n\}$ is Gaussian, we have

$$\mathbb{E}[\mathrm{e}^{\lambda^T X_1}] = \mathrm{e}^{\lambda^T m + \frac{1}{2}\lambda^T K \lambda},$$

and therefore, after exchanging the order of the suprema, we can rewrite the RHS as

$$\sup_{y\in\mathcal{C}} \sup_{K\in\mathcal{K}_\delta} \sup_{m\in\mathcal{U}_\epsilon} \left\{ - \sup_{\lambda\in\mathbb{R}^d} [\lambda^T y - \lambda^T m - \frac{1}{2}\lambda^T K \lambda] \right\},$$

or as

$$\sup_{y\in\mathcal{C}} \sup_{K\in\mathcal{K}_\delta} \sup_{m\in\mathcal{U}_\epsilon} \left\{ \inf_{\lambda\in\mathbb{R}^d} [-\lambda^T y + \lambda^T m + \frac{1}{2}\lambda^T K \lambda] \right\}.$$

Now, using an appropriate min–max theorem for exchanging the order of the third sup and the inf (see, e.g., [15], Cor. 37.3.2), since the function is concave (linear) in m and (strictly) convex in λ, and $\mathcal{U}_\epsilon$ is compact, the above is equal to

$$\sup_{y\in\mathcal{C}} \sup_{K\in\mathcal{K}_\delta} \left\{ \inf_{\lambda\in\mathbb{R}^d} \sup_{m\in\mathcal{U}_\epsilon} [-\lambda^T y + \lambda^T m + \frac{1}{2}\lambda^T K \lambda] \right\}.$$

We can calculate the inner supremum

$$\sup_{m\in\mathcal{U}_\epsilon} [-\lambda^T y + \lambda^T m + \frac{1}{2}\lambda^T K \lambda]$$

in closed-form as

$$-\lambda^T y + \lambda^T \bar{m} + \frac{1}{2}\lambda^T K \lambda + \epsilon\sqrt{\lambda^T K \lambda}$$

since the function to be maximized is linear, and the set $\mathcal{U}_\epsilon$ is a convex, compact (and conic) set. This follows easily from KKT optimality conditions. Thus, the RHS has been transformed into

$$\sup_{y\in\mathcal{C}} \sup_{K\in\mathcal{K}_\delta} \inf_{\lambda\in\mathbb{R}^d} -\lambda^T y + \lambda^T \bar{m} + \frac{1}{2}\lambda^T K \lambda + \epsilon\sqrt{\lambda^T K \lambda}.$$

Now, invoking the min–max theorem one more time, we can equivalently rewrite the above as

$$\sup_{y\in\mathcal{C}} \inf_{\lambda\in\mathbb{R}^d} \sup_{K\in\mathcal{K}_\delta} -\lambda^T y + \lambda^T \bar{m} + \frac{1}{2}\lambda^T K \lambda + \epsilon\sqrt{\lambda^T K \lambda}$$

and concentrate on the problem:

$$\sup_{K\in\mathcal{K}_\delta} \frac{1}{2}\lambda^T K \lambda + \epsilon\sqrt{\lambda^T K \lambda}.$$

One can further rewrite the objective function as

$$\frac{1}{2}\langle C, K\rangle + \epsilon\sqrt{\langle C, K\rangle},$$

where $C \equiv \lambda\lambda^T$, or as

$$\frac{1}{2}\langle C, X + \hat{K}\rangle + \epsilon\sqrt{\langle C, X + \hat{K}\rangle},$$

and treat the problem over the symmetric matrix variable $X \equiv K - \hat{K}$. Now, one writes the Lagrange function

$$L(X,\gamma) = \frac{1}{2}\langle C, X + \hat{K}\rangle + \epsilon\sqrt{\langle C, X + \hat{K}\rangle} + \gamma(\delta^2 - \langle X, X\rangle)$$

with a positive multiplier γ. First-order conditions give

$$X = \frac{1}{4\gamma}(1 + \frac{\epsilon}{\sigma})C$$

where $\sigma \equiv \sqrt{\langle C, X + \hat{K}\rangle}$. Using the definition of σ and supposing that the constraint is active we have two equations in two unknowns σ, γ:

$$\frac{1}{4\gamma}(1 + \frac{\epsilon}{\sigma})B + A = \sigma^2,$$

$$\frac{1}{16\gamma^2}(1 + \frac{\epsilon}{\sigma})^2 B = \delta^2,$$

where $B \equiv \|C\|_F^2$ and $A \equiv \langle C, \hat{K}\rangle$. The solutions are obtained as $\sigma = \sqrt{A + \delta\sqrt{B}}$ and $\gamma = \frac{1}{4}\frac{(\sqrt{A+\delta\sqrt{B}}+\epsilon)\sqrt{B}}{\sqrt{A+\delta\sqrt{B}}}$, which results in $X^* = \frac{\delta}{\|C\|_F}C$ after evident simplification, thus giving $K^* = \hat{K} + \delta\frac{C}{\|C\|_F}$, a positive definite matrix. □

Note that $G(y) \equiv \inf_{\lambda\in\mathbb{R}^d} F(\lambda)$ is a concave function of y since it is the infimum of a collection of affine functions.

As a variation on the theme of Proposition 2, consider the mean ambiguity set defined as a box around a nominal value $\bar{m}$:

$$\mathcal{U}_\infty = \{m | \|m - \bar{m}\|_\infty \leq \epsilon\}.$$

We assume K known with certainty. We obtain the following result, which is less explicit than our Proposition 2 above.

Proposition 4 *Under the hypotheses of Proposition 2,*

$$\sup_{m\in\mathcal{U}_\infty} \frac{1}{n}\ln \mu_n^{(m)}(\mathcal{C}) \leq \sup_{y\in\mathcal{C}}\left[(\bar{m} - y)^T\lambda^* + \frac{1}{2}(\lambda^*)^T K\lambda^* + \epsilon\|\lambda^*\|_1\right]$$

for every closed set C, where λ^* is any d-vector satisfying the inclusion

$$0 \in (\bar{m} - y) + K\lambda + \epsilon\{g \in \mathbb{R}^d : \|g\|_\infty \leq 1, g^T\lambda = \|\lambda\|_1\}.$$

Proof We proceed as in the proof of the previous proposition to arrive at the right-hand side

$$RHS \equiv \sup_{y\in\mathcal{C}}\left\{\inf_{\lambda\in\mathbb{R}^d}\sup_{m\in\mathcal{U}_\infty}[-\lambda^T y + \lambda^T m + \frac{1}{2}\lambda^T K\lambda]\right\}.$$

Now, taking the inner supremum over m yields

$$RHS = \sup_{y\in\mathcal{C}}\left\{\inf_{\lambda\in\mathbb{R}^d}[-\lambda^T y + \lambda^T\bar{m} + \epsilon\|\lambda\|_1 + \frac{1}{2}\lambda^T K\lambda]\right\}.$$

Now, since the function in the expression above is convex in λ, but not everywhere differentiable, we use the subdifferential characterization of the minimizer [15], and the proof is complete. □

The above proposition serves to appreciate the virtues of the specific ellipsoidal ambiguity set used in the present paper (defined via the covariance matrix K), which allows closed-form expressions for multivariate Gaussian random sequences, essentially the only case in multivariate analysis where we were able to obtain explicit bounds. Another case allowing to make progress towards explicit bounds is discussed next.

3.2. A shifted sequence

Consider a sequence of d-dimensional random vectors $X_1, X_2, \ldots$ where $X_n = m + Y_n$ with m a deterministic but ambiguous vector (the shift) and Y_n a random d-dimensional vector sequence. No specific assumption about the probability law governing Y is made. However, we shall assume the shift vector m takes values in the closed, convex set $\mathcal{U}$. After straightforward algebra, we have that the cumulant generating function $\Lambda(z)$ of X_1 is given as

$$\Lambda(z) = z^T m + \lambda(z)$$

where $\lambda(z)$ is the cumulant generating function corresponding to Y_1. Let $\mu_n^{(m)}$ denote the probability law of $\bar{S}_n = \sum_{i=1}^n X_i$ as usual. Then, from the worst-case Cramér bound, we have that for every closed set $\mathcal{C}$

$$\sup_{m\in\mathcal{U}} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq \sup_{x\in\mathcal{C}} \inf_{z\in\mathbb{R}^d} \{\lambda(z) - z^T x + \sup_{m\in\mathcal{U}} z^T m\}$$

using the definition of the Legendre–Fenchel transform and the usual infimum/supremum manipulations (we use again the min–max theorem for exchanging the order of the third sup and the inf [15], Cor. 37.3.2). Now, the term $\sup_{m\in\mathcal{U}} z^T m$ is actually the *support function* $S_U(z)$ (evaluated at z) of the closed convex set U from convex analysis [15]. Hence, the right-hand side of the inequality above becomes

$$\sup_{x\in\mathcal{C}} \inf_{z\in\mathbb{R}^d} \{g(z) - z^T x\},$$

where $g(z) \equiv \lambda(z) + S_U(z)$. Therefore, we have proved:

Proposition 5 *For a sequence of d-dimensional random vectors $X_1, X_2, \ldots$ where $X_n = m + Y_n$ with m (the shift) taking values in the closed, convex set $\mathcal{U}$, and Y_n is a random d-dimensional vector sequence, we have*

$$\sup_{m\in\mathcal{U}} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq - \inf_{x\in\mathcal{C}} g^*(x)$$

for every closed set C, where g^* is the Legendre–Fenchel transform of g defined as $\lambda(z) + S_U(z)$.

The above result furnishes a way to incorporate different probability laws and ambiguity sets into large deviations.

Now, as an application, consider the case where $\{Y_n\}$ is a d-dimensional normally distributed random sequence with mean 0 and variance-covariance K (we do not need mean equal to zero here, it is only for convenience). Furthermore, we revert to ellipsoidal ambiguity set $\mathcal{U}_\epsilon = \{m | \|K^{-1/2}(m - \bar{m})\| \leq \epsilon\}$ instead of the unspecified closed, convex set U. The term $\sup_{m\in\mathcal{U}_\epsilon} z^T m$ is equal to $z^T\bar{m} + \epsilon\sqrt{z^T K z}$ (notice that the

term $z^T\bar{m} + \epsilon\sqrt{z^TKz}$ can be interpreted to reflect the engineering design methodology that random variable z^Tm with mean $z^T\bar{m}$ most likely lies within ϵ standard deviation, i.e. $\epsilon\sqrt{z^TKz}$, of its mean.) For the multivariate Gaussian we have that $\lambda(z) = \frac{1}{2}z^TKz$. Now, we can evaluate $g(z)$ and its Legendre–Fenchel transform explicitly. We solve the inner inf problem, which is a quadratic-norm problem

$$z^T(\bar{m} - x) + \frac{1}{2}z^TKz + \epsilon\|z\|_K,$$

(there is a quadratic and a weighted norm term: $\|z\|_K = \sqrt{z^TKz}$) in closed-form. From the first-order conditions (they are sufficient as the function is convex), one obtains:

$$\bar{m} - x + K\lambda + \frac{\epsilon}{\sqrt{z^TKz}}Kz = 0,$$

which gives

$$z^* = \frac{\sigma}{\sigma + \epsilon}K^{-1}(x - \bar{m})$$

where we have defined $\sigma = \sqrt{z^TKz}$. Substituting the expression for z^* into the definition of σ one obtains the quadratic equation in σ as

$$\sigma^2 + 2\epsilon\sigma + \epsilon^2 - H^2 = 0,$$

where $H = \sqrt{(x - \bar{m})^TK^{-1}(x - \bar{m})}$. The positive root of the equation is given by $H - \epsilon$, for $H \geq \epsilon$. The result, which is identical to the result of Proposition 2, follows by substituting the solution

$$z^* = \frac{H - \epsilon}{H}K^{-1}(x - \bar{m})$$

into the function. When $H < \epsilon$, one simply takes $z^* = 0$. Therefore, we have

$$\sup_{m\in\mathcal{U}_\epsilon}\frac{1}{n}\ln\mu_n^{(m)}(\mathcal{C}) \leq -\inf_{x\in\mathcal{C}}\left[\mathbb{1}_{x\in\mathcal{U}^c}\frac{1}{2}(\|x - \bar{m}\|_K - \epsilon)^2\right]$$

for every closed set C.

3.3. A multivariate Poisson sequence

Now, we consider an example from queuing theory [17]. Suppose y_{ij} are i.i.d. random variables following a Poisson law with rate λ_j. Define the vectors

$$x_i = \sum_{j=1}^{J} y_{ij}e_j,$$

where $e_j \in \mathbb{R}^d$ are given vectors for $j = 1, \ldots, J$. We shall be interested in a worst-case LDP upper bound estimate for the average $\frac{x_1+\ldots+x_n}{n}$ as in the previous paragraphs. For $n \geq 1$, let $\mu_n^{(\mathbf{\Lambda})}$ be the law of the empirical mean of the n i.i.d. random variables, where $\mathbf{\Lambda}$ is the J-vector with components λ_j. We shall confine ambiguity in the rates λ_j to the ambiguity set

$$\mathcal{L} = \{\mathbf{\Lambda} \in \mathbb{R}^J : \|\mathbf{\Lambda} - \hat{\mathbf{\Lambda}}\|_2 \leq \epsilon\}.$$

We are interested in the bound:

$$\sup_{\mathbf{\Lambda}\in\mathcal{L}} \frac{1}{n} \ln \mu_n^{(m)}(\mathcal{C}) \leq \sup_{\mathbf{\Lambda}\in\mathcal{L}} \{-\inf_{x\in\mathcal{C}} \ell(x)\},$$

where $\ell(x)$ is given as

$$\ell(x) = \sup_{\theta} \{\theta^T x - g(\theta)\}$$

with the cumulant generating function

$$g(\theta) = \sum_{j=1}^{J} \lambda_j (\mathrm{e}^{\theta^T e_j} - 1).$$

Going through the usual motions we have the right-hand side of the inequality as

$$\sup_{x\in\mathcal{C}} \inf_{\theta\in\mathbb{R}^d} \sup_{\mathbf{\Lambda}\in\mathcal{L}} \{-\theta^T x + \sum_{j=1}^{J} \lambda_j (\mathrm{e}^{\theta^T e_j} - 1)\}.$$

For ease of notation denote by $\xi_j(\theta)$ the quantity $\mathrm{e}^{\theta^T e_j} - 1$, and hence by $\xi(\theta)$ the J-vector with components $\xi_j(\theta)$. Now, evaluation of the innermost supremum gives the right-hand side:

$$\sup_{x\in\mathcal{C}} \inf_{\theta\in\mathbb{R}^d} \{-\theta^T x + \xi(\theta)^T \hat{\mathbf{\Lambda}} + \epsilon \|\xi(\theta)\|_2\}.$$

We note that the function $H(x)$, defined as

$$H(x) \equiv \inf_{\theta\in\mathbb{R}^d} \{-\theta^T x + \xi(\theta)^T \hat{\mathbf{\Lambda}} + \epsilon \|\xi(\theta)\|_2\},$$

is a concave function since it is the pointwise infimum of a collection of affine functions. However, an explicit expression for H is not possible. Hence, calculations involving H have to be done numerically. For illustration, we consider $d = 2 = J$ with $e_1 = (1\ 0)^T$ and $e_2 = (0\ 1)^T$, the unit vectors, $\hat{\mathbf{\Lambda}} = (10\ 10)^T$. For $x_1 \geq 100$ and $x_2 \geq 100$, the function H attains its maximum at $(100, 100)$. Figure 3 shows the behavior of $H(100, 100)$ as ϵ increases. It is almost a linear curve.

4. Sanov's theorem under ambiguity

In this section, we shall briefly explore worst-case bounds within the method of types and Sanov's theorem, which can be viewed as an application of large deviations theory (more precisely, of the Gärtner–Ellis theorem; see, e.g., [4]). Sanov's theorem is also heavily used in information theory; see [6]. This section is related to the work reported in [14] where the worst-case rate function is characterized using a variational formula involving the solution of a semiinfinite linear optimization problem.

Our desktop reference for Sanov's theorem is [8]. We denote by Σ the finite alphabet $\{a_1, a_2, \ldots, a_N\}$ (we also use the N-vector a to denote the vector with components $(a_1, a_2, \ldots, a_N)$). Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of random variables that are i.i.d. according to the law $\mu \in M_1(\Sigma)$ where $M_1(\Sigma)$ denotes the space of all probability laws on Σ. The type $\mathbf{L}_n^{\mathbf{y}}$ of a finite sequence $y = (y_1, \ldots, y_n) \in \Sigma^n$ is the empirical measure

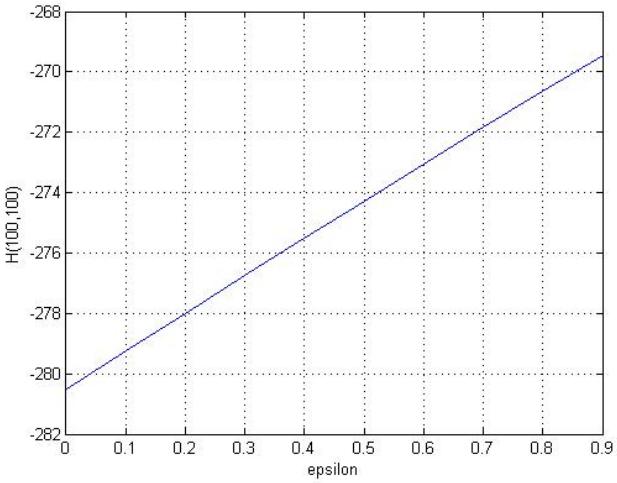

Figure 3. A plot of $H(100,100)$ versus ϵ with $\hat{\mathbf{\Lambda}} = (10\ 10)^T$.

induced by that sequence, i.e. $\mathbf{L}_n^{\mathbf{y}}(a_i)$ is the fraction of occurrences of a_i in the sequence $y = (y_1, \ldots, y_n)$. The relative entropy of a probability vector ν with respect to another probability vector μ is

$$H(\nu|\mu) = \sum_{i=1}^{|\Sigma|} \nu(a_i) \ln \frac{\nu(a_i)}{\mu(a_i)}.$$

Let $\mathcal{P}$ denote the set of probability measures of which μ is a member. The following estimate follows immediately from Sanov's theorem (see Th. 2.1.10 [8]).

Proposition 6 *For every set Γ of probability vectors in $M_1(\Sigma)$, we have*

$$\limsup_{n\to\infty} \sup_{\mu\in\mathcal{P}} \frac{1}{n} \ln P(\mathbf{L}_n^{\mathbf{y}} \in \Gamma) \leq \sup_{\mu\in\mathcal{P}} \{ - \inf_{\nu\in\Gamma} H(\nu|\mu) \}.$$

Proposition 6 notes that when one would like to generalize Sanov's theorem to a case where the actual measure is known to come from a given set of measures, the LDP rate for the empirical measure is exactly the relative entropy distance between two sets of measures. Computing such distances is a topic currently studied in computer science; reference [7] cited above in the remark after the proof of Proposition 2 is an example. Thus, Proposition 6 provides a connection between these two problems and research areas.

In general, it is extremely difficult to obtain explicit expressions for the right-hand side in the above bound. However, considering the case $\mathcal{P}_m = \{\mu : \mathbf{1}^T\mu = 1, \mu \geq 0, a^T\mu = \alpha\}$ (we assume now that the alphabet has numeric values), i.e. the set of probability vectors resulting in a mean value equal to α, we were able to show a (somewhat limited) result. Assuming that $a_1 < a_2 < a_3$, for every set Γ of probability vectors in $M_1(\Sigma)$, we have for $N = 3$ and $\alpha = a_2$:

$$\limsup_{n\to\infty} \sup_{\mu\in\mathcal{P}_m} \frac{1}{n} \ln P(\mathbf{L}_n^{\mathbf{y}} \in \Gamma) \leq - \inf_{\nu\in\Gamma} H(\nu|\mu^*),$$

where $\mu_1^* = \frac{a_1(a_2-a_3)(\nu_1+\nu_3)}{a_1-a_3}$, $\mu_2^* = \nu_2$, $\mu_3^* = \frac{(a_1-a_2)(\nu_1+\nu_3)}{a_1-a_3}$. Admittedly, the specification $\alpha = a_2$ is restrictive. However, an explicit result for general α was not possible. In general, one has to solve Nth degree polynomial equations to find the solution of the inner problem. Hence, one must resort to numerical methods. As a result, our efforts to extend the above result to general N, different α, and other specifications of $\mathcal{P}$ (e.g., $\mathcal{P} = \{p \in M_1(\Sigma) : \text{dist}(P, \bar{P}) \leq \varepsilon\}$ for a nominal probability vector $\bar{P}$ and a suitable distance measure) have so far borne no fruit. This is the subject of future investigations.

5. Concluding remarks

We investigated the impact of ambiguity in parameters for common distributions on large deviations upper bounds in a worst-case sense inspired by the last decade of development in robust optimization. In particular, we adopted the ellipsoid specification of ambiguity for multivariate random sequences since ellipsoids help mimic the engineering design approach that a random variable affecting the design will most likely not exceed a constant times its standard deviation, and leads to tractable (at least in some cases) optimization problems and explicit worst-case bounds. Much remains to be explored: some examples are hypothesis testing under ambiguity and large deviations for Markov chains under ambiguity, among others.

References

[1] Ben-Tal A, Nemirovski A. Robust solutions of uncertain linear programs. Oper Res Lett 1999; 25: 1-13.

[2] Ben-Tal A, El Ghaoui L, Nemirovski A. Robust Optimization. Princeton, NJ, USA: Princeton University Press, 2009.

[3] Bertsimas D, Brown DB, Caramanis C. Theory and applications of robust optimization. SIAM Rev 2011; 53: 464-501.

[4] Bucklew JA. Large Deviation Techniques in Decision, Simulation and Estimation. New York, NY, USA: Wiley, 1990.

[5] Cont R. Model uncertainty and its impact on the pricing of derivative instruments. Math Financ 2006; 16: 519-547.

[6] Cover TJ, Thomas JA. Elements of Information Theory. New York, NY, USA: Wiley, 1991.

[7] Davis JV, Dhillon I. Differential entropic clustering of multivariate Gaussians. In: Scholkopf B, Platt J, Hoffman T, editors. Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2006, pp. 337-344.

[8] Dembo A, Zeitouni O. Large Deviations Techniques and Applications. 2nd ed. New York, NY, USA: Springer, 1998.

[9] den Hollander F. Large Deviations. Fields Institute Monographs. Providence, RI, USA: American Mathematical Society, 2008.

[10] El Ghaoui L, Lebret H. Robust solutions to least squares problems with uncertain data. SIAM J Matrix Anal A 1997; 18: 1035-1064.

[11] Föllmer H, Knispel T. Entropic risk measures: coherence vs. convexity, model ambiguity, and robust large deviations. Stoch Dynam 2011; 11: 333-351.

[12] Hu F. On Cramer's theorem for capacities. CR Acad Sci I-Math 2010; 348: 1009-1013.

[13] Lewis JT, Russell R. An Introduction to Large Deviations for Teletraffic Engineers. Dublin, Ireland: Dublin Institute for Advanced Studies, 1997.

[14] Pandit C, Meyn S. Worst-case large-deviation asymptotics with application to queueing and information theory. Stoch Proc Appl 2006; 116: 724-756.

[15] Rockafellar TR. Convex Analysis. Princeton, NJ, USA: Princeton University Press, 1970.

[16] Sadowsky JS. Robust large deviations performance analysis for large sample detectors. IEEE T Inform Theory 1989; 35: 917-920.

[17] Shwartz A, Weiss A. Large Deviations for Performance Analysis. London, UK: Chapman & Hall, 1994.