Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Robust Solutions of Optimization Problems Affected by Uncertain Probabilities

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen,

To cite this article:
Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen, (2013) Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. Management Science 59(2):341-357. http://dx.doi.org/10.1287/mnsc.1120.1641

MANAGEMENT SCIENCE
Vol. 59, No. 2, February 2013, pp. 341–357
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1641

Robust Solutions of Optimization Problems Affected by Uncertain Probabilities

Aharon Ben-Tal
Department of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel; and CentER, Tilburg University, 5000 LE Tilburg, The Netherlands, abental@ie.technion.ac.il

Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen
Department of Econometrics and Operations Research, Tilburg University, 5000 LE Tilburg, The Netherlands
{d.denhertog@uvt.nl, a.m.b.dewaegenaere@uvt.nl, b.melenberg@uvt.nl, g.rennen@gmail.com}

In this paper we focus on robust linear optimization problems with uncertainty regions defined by ϕ-divergences (for example, chi-squared, Hellinger, Kullback–Leibler). We show how uncertainty regions based on ϕ-divergences arise in a natural way as confidence sets if the uncertain parameters contain elements of a probability vector. Such problems frequently occur in, for example, optimization problems in inventory control or finance that involve terms containing moments of random variables, expected utility, etc. We show that the robust counterpart of a linear optimization problem with ϕ-divergence uncertainty is tractable for most of the choices of ϕ typically considered in the literature. We extend the results to problems that are nonlinear in the optimization variables. Several applications, including an asset pricing example and a numerical multi-item newsvendor example, illustrate the relevance of the proposed approach.

Key words: robust optimization; ϕ-divergence; goodness-of-fit statistics
History: Received April 20, 2011; accepted December 16, 2011, by Gérard P. Cachon, optimization. Published online in *Articles in Advance* November 9, 2012.

1. Introduction

Several papers in the late 1990s (Kouvelis and Yu 1997; Ben-Tal and Nemirovski 1997, 1998; El Ghaoui and Lebret 1997; El Ghaoui et al. 1998) started a revival of robust optimization (RO), both in terms of theoretical aspects, as well as practical applications. For a survey, we refer to Ben-Tal et al. (2009) or Bertsimas et al. (2011). Consider, for example, a linear constraint with uncertain parameters. The idea of robust optimization is to define a so-called *uncertainty region* for the uncertain parameters, and then to require that the constraint should hold for all parameter values in this uncertainty region. The optimization problem modeling this requirement is called the *robust counterpart problem* (*RCP*). Although the RCP typically has an infinite number of constraints, it is still tractable (polynomially solvable) for several optimization problems and several choices of the uncertainty region. In particular, the robust counterpart for a linear programming (LP) problem with polyhedral or ellipsoidal uncertainty regions reduces to a linear programming and a conic quadratic programming (CQP) problem, respectively.

When applying the RO methodology to a practical problem, a major modeling decision concerns the choice of the uncertainty region U. Such a choice should fulfil three basic requirements. First, U should be consistent with whatever data (and information) are available on the uncertain parameters. Second, U should be statistically meaningful. Third, U should be such that the corresponding RCP is tractable. The latter requirement is essential when confronting an optimization problem having a large-scale design dimension and/or large-scale parameter space.

In this paper we are concerned mainly with optimization problems where the uncertain parameters are probabilities. This is the case when the objective function and/or the constraint functions involve terms with expectations (such as moments of random variables, or expected utility, etc.). For such problems, we advocate the use of uncertainty regions that are constructed as confidence sets using ϕ-divergence functionals. Such functionals include the Hellinger distance, the Kullback–Leibler and the chi-squared divergence, the Burg entropy-based divergence, and many others. We choose ϕ-divergences because these play a fundamental role in statistics (see Liese and Vajda 2006, Reid and Williamson 2011). The main contribution of this paper is showing that the choice of U as such uncertainty sets indeed fulfils the above three requirements:

- U is based on empirical probability estimates obtained from historical data.
- U is shown to relate to a statistical confidence region based on asymptotic theory.
- U is such that the corresponding RCP is shown to be tractable: for basically all significant

ϕ-divergence functionals, the resulting robust counterpart problem is polynomially solvable. In fact, in many cases it reduces to a linear, or a conic quadratic problem.

Using (smooth) ϕ-divergences, uncertainty regions can easily be constructed as (approximate) confidence sets, when the probabilities can be estimated from historical data. This follows from applying asymptotic theory. Moreover, ϕ-divergences also allow the construction of confidence sets when the probabilities are calculated using additional information, represented by some underlying statistical model. In this way, smaller confidence sets can be obtained without reducing the confidence level. This is a consequence of the so-called information processing theorem, valid for ϕ-divergences; see Liese and Vajda (2006). The size of the uncertainty region can be controlled by the confidence level of the confidence set. For example, the choice of a 95% confidence level will result in an uncertainty set that is (statistically) significant. Combined with the tractability of the RCP with these uncertainty sets, ϕ-divergences therefore present an appealing approach in robust optimization.

We illustrate the relevance of the proposed approach by applying it first to an investment problem. We show that our approach yields a natural link with standard asset pricing theory. We also present a numerical illustration in terms of a multi-item newsvendor problem. Using our robust optimization approach leads to solutions that are quite robust, while at the same time exhibiting good average optimal performance.

We now discuss related papers. In Chapter 2 of Ben-Tal et al. (2009), probabilistic arguments are used to construct an uncertainty region by using partial a priori knowledge on the underlying distribution of the uncertain parameters. Klabjan et al. (2013) use the well-known chi-squared statistic, which is a special case of a ϕ-divergence statistic, to define uncertainty regions for the unknown demand distribution in an inventory control problem. In their approach, a robust dynamic programming problem has to be solved. Calafiore (2007) studied portfolio selection problems in which the true distribution of asset returns is unknown. He assumed that the true distribution is only known to lie within a certain distance from an estimated one and used the Kullback–Leibler divergence to measure the distance. Our analysis includes the Kullback–Leibler divergence as a special case. Moreover, whereas Calafiore's (2007) approach requires solving a "nested" optimization problem, our approach allows for a tractable reformulation of the robust counterpart. Wang et al. (2009) studied robust optimization for data-driven newsvendor problems, in which the uncertainty set for the unknown distribution is defined as a "likelihood region." Bertsimas and Brown (2009) interpreted robust optimization in terms of coherent risk measures. Ben-Tal et al. (2010) considered the soft robust optimization approach and established for such optimization problems a link with convex risk measures. Related research on robust optimal portfolio choice with uncertainty sets based on confidence sets includes Delage and Ye (2010), Garlappi et al. (2007), and Goldfarb and Iyengar (2003) (for an overview, see Fabozzi et al. 2010). These papers typically use mean or covariance matrix-based confidence sets, whereas we use confidence sets based on ϕ-divergences.

The remainder of this paper is organized as follows. We start with an introduction to robust linear optimization in §2, and to ϕ-divergences in §3. In §3 we also discuss the construction of uncertainty sets as confidence sets using ϕ-divergences. In §4 we study the robust counterparts for problems with ϕ-divergence uncertainty regions. In §5 we show that for different choices of the ϕ-divergence, the robust counterpart can be reformulated as a tractable problem. In §6 we present some applications, including a numerical multi-item newsvendor example, and in §7 we conclude the paper with topics for further research.

2. Introduction to Robust Linear Optimization

In this paper, the main focus is on robust linear optimization. Without loss of generality, we focus on robust counterpart problems of the form

$$\min\{c^T x \mid Ax \le b,\ \forall A \in \hat{U}\},$$

where $x \in \mathbb{R}^n$ is the optimization vector, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^l$ are given (known) parameters, $A \in \mathbb{R}^{l\times n}$ is a matrix with uncertain parameters, and $\hat{U}$ is a given uncertainty region for A. Indeed, as shown by Ben-Tal et al. (2009), for robust linear optimization we can, without loss of generality, assume that the objective and the right-hand side of the constraints are certain.

Moreover, as also shown by Ben-Tal et al. (2009), for robust linear optimization, we can, without loss of generality, assume constraint-wise uncertainty. Hence, we focus on a single constraint, which we assume to be of the form

$$(a + Bp)^T x \le \beta, \quad \forall p \in U, \tag{1}$$

where $x \in \mathbb{R}^n$ is the design vector; $a \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times m}$, and $\beta \in \mathbb{R}$ are given (known) parameters; $p \in \mathbb{R}^m$ is the uncertain parameter; and U the uncertainty region for p.

In Table 1, the tractability results for several standard choices of U are given. For a detailed treatment, see Ben-Tal et al. (2009). The last line in the table is

Table 1 **Robust Linear Optimization for Different Choices for the Uncertainty Region in Terms of $p = (p_1, \ldots, p_m)^T$**

Uncertainty region	U	Robust counterpart	Tractability
Box	$\|p\|_\infty \le 1$	$a^T x + \|B^T x\|_1 \le \beta$	LP
Ball	$\|p\|_2 \le 1$	$a^T x + \|B^T x\|_2 \le \beta$	CQP
Polyhedral	$Cp + d \ge 0$	$\begin{cases} a^T x + d^T y \le \beta, \\ C^T y = -B^T x, \\ y \ge 0 \end{cases}$	LP
Cone	$Cp + d \in K$	$\begin{cases} a^T x + d^T y \le \beta, \\ C^T y = -B^T x, \\ y \in K^* \end{cases}$	Conic opt.
Separable functions	$\begin{cases} \sum_i f_{\ell i}(p_i) \le 0, \\ \ell \in \{1, \ldots, L\} \end{cases}$	$\begin{cases} a^T x + \sum_\ell \sum_i \lambda_\ell f^*_{\ell i}(s_{\ell i}/\lambda_\ell) \le \beta, \\ \sum_\ell s_{\ell i} = b_i^T x, \quad i \in \{1, \ldots, m\}, \\ \lambda \ge 0 \end{cases}$	Convex opt.

Notes. The cone K is closed, convex, and pointed. The functions $f_{\ell i}$ are assumed to be convex, with conjugate $f^*_{\ell i}$, and K^* is the dual cone of K.

a new result and will be proved in this paper. Here, we briefly discuss the derivation of the robust counterparts of the standard choices of U, illustrating the general principles. We shall apply these general principles in our approach as well. To start, the results for the box and ball uncertainty region can easily be obtained by finding the worst-case solution with respect to p, i.e., by solving

$$\max\{p^T B^T x \mid p \in U\}.$$

For the polyhedral and cone uncertainty region, we can use duality. Specifically, under the assumption that the uncertainty region is a cone K that contains a strictly feasible solution (i.e., there exists a $\bar{p}$ such that $C\bar{p} + d \in \operatorname{int} K$), it holds that

$$\begin{aligned} &\max\{p^T B^T x \mid Cp + d \in K\} \\ &\quad = \min\{d^T y \mid C^T y = -B^T x,\ y \in K^*\}, \end{aligned}$$

where K^* denotes the dual cone of K. This means that x satisfies (1) if and only if x satisfies

$$a^T x + \min\{d^T y \mid C^T y = -B^T x,\ y \in K^*\} \le \beta.$$

Hence, we have that x satisfies (1) if and only if (x, y) satisfies

$$\begin{cases} a^T x + d^T y \le \beta, \\ C^T y = -B^T x, \\ y \in K^*. \end{cases}$$

Moreover, Ben-Tal et al. (2009) showed that if U is the intersection of different "tractable cones," then the robust counterpart can also be reformulated as a tractable problem.

In this paper we shall show that if the uncertainty region U is based on a ϕ-divergence, the robust counterpart can also be reformulated as a tractable optimization problem.

3. Introduction to ϕ-Divergence

In this section, we first define the concept of ϕ-divergence and discuss some properties that will be useful in obtaining tractable reformulations of the robust counterpart of problem (1), when the uncertainty region U is defined in terms of a ϕ-divergence. Next, we discuss how to construct uncertainty regions as (approximate) confidence sets based on a ϕ-divergence.

3.1. Definition and Some Characteristics

The ϕ-divergence ("distance") between two vectors[1] $p = (p_1, \ldots, p_m)^T \ge 0$, $q = (q_1, \ldots, q_m)^T \ge 0$ in $\mathbb{R}^m$ is defined as

$$I_\phi(p, q) = \sum_{i=1}^m q_i \phi\left(\frac{p_i}{q_i}\right), \tag{2}$$

where $\phi(t)$ is convex for $t \ge 0$, $\phi(1) = 0$, $0\phi(a/0) := a \lim_{t\to\infty} \phi(t)/t$ for $a > 0$, and $0\phi(0/0) := 0$. We refer to the function ϕ as the ϕ-divergence function. We shall mainly focus on probability vectors p and q that satisfy the additional constraint $p^T e = 1$ and $q^T e = 1$, where e denotes a column vector of ones of the same dimension as p and q. However, many of our results are also valid more generally for $p \ge 0$ and $q \ge 0$.

Different choices for ϕ have been proposed in the literature. For a good overview, see Pardo (2006). Table 2 contains the most known and used choices for ϕ. The power divergence class presented in the bottom row of Table 2 was proposed by Cressie and Read (1984) to be used in case of multinomial data, and since then has been extensively studied; see, for

[1] In case of a vector, we interpret the inequality componentwise, i.e., $p = (p_1, \ldots, p_m)^T \ge 0$ means $p_i \ge 0$ for $i = 1, \ldots, m$. Similarly, $p > 0$ means $p_i > 0$ for $i = 1, \ldots, m$.

Table 2 Some ϕ-Divergence Examples

Divergence	$\phi(t)$	$\phi(t)$, $t \geq 0$	$I_\phi(p, q)$
Kullback–Leibler	$\phi_{kl}(t)$	$t \log t - t + 1$	$\sum p_i \log\left(\frac{p_i}{q_i}\right)$
Burg entropy	$\phi_b(t)$	$-\log t + t - 1$	$\sum q_i \log\left(\frac{q_i}{p_i}\right)$
J-divergence	$\phi_j(t)$	$(t-1)\log t$	$\sum (p_i - q_i)\log\left(\frac{p_i}{q_i}\right)$
χ^2-distance	$\phi_c(t)$	$\frac{1}{t}(t-1)^2$	$\sum \frac{(p_i - q_i)^2}{p_i}$
Modified χ^2-distance	$\phi_{mc}(t)$	$(t-1)^2$	$\sum \frac{(p_i - q_i)^2}{q_i}$
Hellinger distance	$\phi_h(t)$	$(\sqrt{t}-1)^2$	$\sum (\sqrt{p_i} - \sqrt{q_i})^2$
χ-divergence of order $\theta > 1$	$\phi^\theta_{ca}(t)$	$\lvert t-1\rvert^\theta$	$\sum q_i \left\lvert 1 - \frac{p_i}{q_i}\right\rvert^\theta$
Variation distance	$\phi_v(t)$	$\lvert t-1\rvert$	$\sum \lvert p_i - q_i\rvert$
Cressie–Read	$\phi^\theta_{cr}(t)$	$\frac{1-\theta+\theta t - t^\theta}{\theta(1-\theta)}$, $\theta \neq 0, 1$	$\frac{1}{\theta(1-\theta)}(1 - \sum p_i^\theta q_i^{1-\theta})$

Note. $\phi(t) = \infty$, for $t < 0$; $\phi^1_{cr}(t) = \phi_b(t)$; and $\phi^0_{cr}(t) = \phi_{kl}(t)$.

example, Jager and Wellner (2007). The expression for $\theta = 0$ is obtained by taking the limit $\theta \to 0$, and the expression for $\theta = 1$ by taking the limit $\theta \to 1$. Table 3 shows that the Cressie and Read (1984) class contains several well-known ϕ-divergence functions proposed in the literature (up to normalization). Notice that when the ϕ-divergence function corresponding to a ϕ-divergence is differentiable at $t = 1$, the function $\varphi(t) = \phi(t) - \phi'(t)(t-1)$ also yields a ϕ-divergence, satisfying (for probability vectors) $I_\varphi(p, q) = I_\phi(p, q)$, with $\varphi(1) = \varphi'(1) = 0$ and $\varphi(t) \geq 0$; see Pardo (2006).

Given some ϕ-divergence function ϕ with corresponding ϕ-divergence $I_\phi(p, q)$, the so-called adjoint of ϕ is defined for $t \geq 0$ as (see Ben-Tal et al. 1991)

$$\tilde{\phi}(t) := t\phi\left(\frac{1}{t}\right). \tag{3}$$

Table 3 Some Specific Choices for θ for the Cressie and Read (1984) ϕ-Divergence Class

θ	$\phi^\theta_{cr}(t)$	Equivalent with
2	$\frac{1}{2}(t^2 - 2t + 1) = \frac{1}{2}(t-1)^2$	Modified χ^2-distance
1	$t(\log t - 1) + 1$	Kullback–Leibler
$\frac{1}{2}$	$4\left(\frac{1}{2} + \frac{1}{2}t - \sqrt{t}\right) = 2(1-\sqrt{t})^2$	Hellinger distance
-1	$\frac{1}{2}\left(-2 + t + \frac{1}{t}\right) = \frac{1}{2}\left(\sqrt{t} - \frac{1}{\sqrt{t}}\right)^2$	χ^2-distance

Table 4 Some ϕ-Divergence Examples with Their Conjugates and Adjoints

Divergence	$\phi^*(s)$	$\tilde{\phi}(t)$	RCP
Kullback–Leibler	$e^s - 1$	$\phi_b(t)$	S.C.
Burg entropy	$-\log(1-s)$, $s < 1$	$\phi_{kl}(t)$	S.C.
J-divergence	No closed form	$\phi_j(t)$	S.C.
χ^2-distance	$2 - 2\sqrt{1-s}$, $s < 1$	$\phi_{mc}(t)$	CQP
Modified χ^2-distance	$\begin{cases} -1 & s < -2, \\ s + s^2/4 & s \geq -2 \end{cases}$	$\phi_c(t)$	CQP
Hellinger distance	$\frac{s}{1-s}$, $s < 1$	$\phi_h(t)$	CQP
χ-divergence of order $\theta > 1$	$s + (\theta - 1)\left(\frac{\lvert s\rvert}{\theta}\right)^{\theta/(\theta-1)}$	$t^{1-\theta}\phi^\theta_{ca}(t)$	CQP
Variation distance	$\begin{cases} -1 & s \leq -1, \\ s & -1 \leq s \leq 1 \end{cases}$	$\phi_v(t)$	LP
Cressie–Read	$\frac{1}{\theta}(1 - s(1-\theta))^{\theta/(\theta-1)} - \frac{1}{\theta}$, $s < \frac{1}{1-\theta}$	$\phi^{1-\theta}_{cr}(t)$	CQP

Notes. The last column indicates the tractability of (1). S.C., admits self-concordant barrier.

It holds that $\tilde{\phi}$ satisfies the conditions for ϕ-divergence functions, and $I_{\tilde{\phi}}(p, q) = I_\phi(q, p)$. Later in this paper we will also use other properties of $\tilde{\phi}$. For example, it is easy to see that the adjoint of the adjoint function is the function itself, i.e., $\tilde{\tilde{\phi}} = \phi$. Moreover, the function ϕ is called self-adjoint if $\tilde{\phi} = \phi$. Table 4 presents the adjoints (as well as some other characteristics) of the examples presented in Table 2. As can be seen from Table 4, the J-divergence and the variation distance are self-adjoint. For other interesting properties of $\tilde{\phi}$, we refer to Ben-Tal et al. (1991).

We will show in §4 that the robust counterpart of a linear constraint with ϕ-divergence uncertainty can be reformulated in terms of the so-called conjugate of ϕ. The conjugate is a function ϕ^*: $\mathbb{R} \to \mathbb{R} \cup \{\infty\}$, which is defined as follows:

$$\phi^*(s) = \sup_{t \geq 0}\{st - \phi(t)\}. \tag{4}$$

In Table 4 we present the expressions of ϕ^* for the examples of Table 2. We only present the expressions of ϕ^* on their effective domains $\mathrm{dom}(\phi^*)$, i.e., the part of the domain where $\phi^*(s) < \infty$.[2]

In some cases ϕ^* does not exist in a (known) closed form. This is, for example, the case for the J-divergence distance measure (see Table 4). In the sequel we will use the following two propositions

[2] These ϕ^* correspond to ϕ with effective domain $\mathrm{dom}(\phi) = (0, \infty)$. Thus, we set $\phi(t) = \infty$ for $t \leq 0$.

to determine tractable reformulations of the robust counterpart in cases where ϕ^* does not exist in closed form. The first proposition applies when ϕ can be written as the sum of two ϕ-divergence functions, ϕ_1 and ϕ_2. The conditions required in the proposition are fulfilled in case the functions f_1 and f_2 are ϕ-divergence functions.

Proposition 1 (Rockafellar 1970). *Assume that f_1 and f_2 are convex, and the intersection of the relative interiors of the effective domains of f_1 and f_2 is nonempty, i.e.,* $\mathrm{ri}(\mathrm{dom}\, f_1) \cap \mathrm{ri}(\mathrm{dom}\, f_2) \neq \varnothing$. *Then*

$$(f_1 + f_2)^*(s) = \inf_{s_1+s_2=s} (f_1^*(s_1) + f_2^*(s_2)),$$

and the inf is attained for some s_1, s_2.

The following proposition relates the conjugate of the adjoint function to the conjugate of the original function.

Proposition 2 (Gushchin 2008). *For the conjugate of a ϕ-divergence function and the conjugate of its adjoint, we have*

$$\phi^*(s) = \inf\{y \in \mathbb{R}\colon (\tilde{\phi})^*(-y) \leq -s\}.$$

3.2. Construction of Uncertainty Regions

In this subsection we describe how to construct uncertainty regions for probability vectors p as (approximate) confidence sets using ϕ-divergences. We consider settings in which there is a fixed number m of given scenarios for a random variable Z, where the components of the probability vector $p = (p_1, \ldots, p_m)^T$ are given by $p_i \equiv \mathbb{P}(Z \in C_i)$, $i = 1, \ldots, m$ (when $\mathbb{P}$ would be the probability distribution of Z). Here, p_i represents the probability that scenario i will occur, where C_i, $i = 1, \ldots, m$, form a partition (of measurable sets) of the outcome space of Z. As *basic case* we take the case where we only observe $Z \in C_i$, $i = 1, \ldots, m$. In this situation we can assume without loss of generality that $Z \in \{1, \ldots, m\}$, where $Z = i$ in case of scenario i. But we shall also consider cases where Z contains more information than just which of the m scenarios occurs. To capture both the basic case and more general cases, we assume the existence of a (measurable) transformation G, such that $G(Z) = i$ if $Z \in C_i$, $i = 1, \ldots, m$. The basic case then corresponds to the situation where G is a bijection (i.e., a one-to-one correspondence), whereas G is a surjection when Z contains more information.

Denote by $\mathbb{P}_Z$ the true probability distribution of Z. We shall assume that $\mathbb{P}_Z$ belongs to a parameterized set of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$, i.e., there exists some $\theta_0 \in \Theta$, such that $\mathbb{P}_Z = \mathbb{P}_{\theta_0}$. We write $p_\theta = (p_{1,\theta}, \ldots, p_{m,\theta})^T$, with $p_{i,\theta} = \mathbb{P}_\theta(G(Z) = i)$, and we write $p_0 = p_{\theta_0}$. We consider the case where the probability distributions $\mathbb{P}_\theta$ are dominated by a common σ-finite measure μ.[3] The density of $\mathbb{P}_\theta$ with respect to μ is denoted by f_θ, where we shall write $f_0 = f_{\theta_0}$. In the basic case, when we only observe the scenarios, we have $Z \in \{1, \ldots, m\}$, and we can take, for example, $\Theta = \mathbb{R}^{m-1}$, μ the counting measure, and

$$\mathbb{P}_\theta(Z = i) = f_\theta(i) \times 1 = \exp(\theta_i) \Big/ \sum_{j=1}^{m} \exp(\theta_j), \quad i = 1, \ldots, m,$$

for $\theta = (\theta_1, \ldots, \theta_m)^T$, with normalization $\theta_m \equiv 0$. We then have

$$\mathbb{P}_\theta(Z = i) = p_{i,\theta} = f_\theta(i), \quad i = 1, \ldots, m,$$

so that there is a one-to-one correspondence between the sets $\mathcal{P} := \{p \in \mathbb{R}^m \mid p > 0, p^T e = 1\}$ and $\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}$.

Given this setting, we shall discuss the construction of uncertainty sets as confidence sets, under the assumption that a sample $Z_1, \ldots, Z_N$, randomly drawn from $\mathbb{P}_Z$, is given. The ϕ-divergence between the densities f_θ and f_0 is given by

$$I_\phi(f_\theta, f_0) = \int \phi\left(\frac{f_\theta}{f_0}\right) f_0 \, d\mu.$$

We shall first construct a confidence set in terms of θ, which we then use to construct a confidence set in terms of p_θ. Let $\hat{\theta}$ denote the maximum likelihood estimator of θ, and denote $\hat{f}_0 = f_{\hat{\theta}}$. In the basic case we get $\hat{f}_0 = q_N$, where $q_N = (q_{1,N}, \ldots, q_{m,N})^T$ is the m-dimensional vector containing as components the sample frequencies of the m scenarios based on the random sample $Z_1, \ldots, Z_N$. We shall use $I_\phi(f_\theta, \hat{f}_0)$ as estimator for $I_\phi(f_\theta, f_0)$.[4] Pardo (2006) presents the characteristics of this estimator under the assumption that ϕ is twice continuously differentiable in a neighborhood of 1, with $\phi''(1) > 0$. Most ϕ-divergences reported in Table 2 satisfy this condition. Under the probability distribution $\mathbb{P}_Z = \mathbb{P}_{\theta_0}$ and under appropriate additional regularity conditions, he shows that the normalized estimated ϕ-divergence

$$\frac{2N}{\phi''(1)} I_\phi(f_0, \hat{f}_0) \tag{5}$$

asymptotically (i.e., for $N \to \infty$) follows a χ^2_d-distribution, with degrees of freedom determined by

[3] This means that we can write, for each relevant (measurable) set, A: $\mathbb{P}_\theta(Z \in A) = \int_A f_\theta(z)\, d\mu(z)$, with μ not depending on θ. The measure μ is σ-finite if we can write the outcome space of Z as $A_1 \cup A_2 \cup \cdots$ for some finite of countable sequence of (measurable) sets, satisfying $\mu(A_k) < \infty$, $k = 1, 2, \ldots$ (see, for example, Billingsley 1995 for further details).

[4] It is also possible to avoid the use of the maximum likelihood estimator; see, for example, Broniatowski and Keziou (2009) or Liese and Vajda (2006).

the dimension of the parameter set Θ. In terms of θ we therefore have the following (approximate) $(1-\alpha)$-confidence set around $\hat{\theta}$:

$$\{\theta \in \Theta \mid I_\phi(f_\theta, \hat{f}_0) \le \rho\}, \tag{6}$$

where[5]

$$\rho = \rho_\phi(N, d, \alpha) := \frac{\phi''(1)}{2N}\chi^2_{d,1-\alpha}. \tag{7}$$

Based on the *information* (*or data*) *processing theorem*, see Liese and Vajda (2006) for a new proof, we have

$$I_\phi(p_\theta, \hat{p}_0) \le I_\phi(f_\theta, \hat{f}_0),$$

where $\hat{p}_0 = p_{\hat{\theta}}$ is the estimator of p_0, using (the maximum likelihood estimator) $\hat{\theta}$ as estimator for θ. Thus, we have

$$\{\theta \in \Theta \mid I_\phi(p_\theta, \hat{p}_0) \le \rho\} \supset \{\theta \in \Theta \mid I_\phi(f_\theta, \hat{f}_0) \le \rho\},$$

where the left-hand side is an (approximate) confidence set, with confidence level at least equal to the confidence level of the right-hand set. Transforming the parameters θ into the corresponding probabilities p_θ, the left-hand side can be transformed in a confidence set as subset of $\mathcal{P}_\Theta$, given by

$$\{p_\theta \in \mathcal{P}_\Theta \mid I_\phi(p_\theta, \hat{p}_0) \le \rho\}.$$

This implies that the uncertainty set U given by

$$\begin{aligned} U &= \{p \in \mathbb{R}^m \mid p \ge 0, e^T p = 1, I_\phi(p, \hat{p}_0) \le \rho\} \\ &\supset \{p_\theta \in \mathcal{P}_\Theta \mid I_\phi(p_\theta, \hat{p}_0) \le \rho\}, \end{aligned} \tag{8}$$

is an (approximate) confidence set of confidence level at least $(1-\alpha)$ for p around $\hat{p}_0$. In the basic case, i.e., when we only observe the scenarios, we have that the dimension of Θ equals $m-1$, so that $d = m-1$ in (7). But with additional information we might be able to parameterize f_θ by means of $\Theta \subset \mathbb{R}^d$ with $d < m-1$. Then, using (7) with $d < m-1$, we get a smaller confidence set, but of the same confidence level.

The following example illustrates the difference between the basic case, where we only observe the scenarios, and a case where we exploit that the scenarios are generated by an underlying parameterized distribution.[6]

Example 1. Assume $Z \sim N(\mu_0, \sigma_0^2)$, i.e., Z follows a normal distribution with mean μ_0 and variance σ_0^2. We have $\theta = (\mu, \sigma)^T$. Suppose $C_i = (c_{i-1}, c_i)$, $i = 1, \ldots, m$, with $c_0 = -\infty < c_1 < \cdots < c_{m-1} < c_m = +\infty$. Then $p_{i,\theta} = \Phi((c_i - \mu)/\sigma) - \Phi((c_{i-1} - \mu)/\sigma)$, where Φ denotes the distribution function of the $N(0,1)$-distribution. Suppose we estimate θ_0 by the maximum likelihood estimator $\hat{\theta}$, and we estimate $\hat{p}_0 = p_{\hat{\theta}}$. For ρ we then have (7) with $d = 2$ (denote this value of ρ by ρ_{ml}). If we would not use $Z \sim N(\mu_0, \sigma_0^2)$, but instead assume that we are in the basic case, then we find for ρ the value given in (7) with $d = m-1$ (denote this value of ρ by ρ_{bc}). Let $\alpha = 0.05$. For $m = 10$ we then have $\rho_{bc}/\rho_{ml} \approx 2.8$, whereas for $m = 30$ this increases to $\rho_{bc}/\rho_{ml} \approx 7.1$. Thus, exploiting the information that $Z \sim N(\mu_0, \sigma_0^2)$ limits the size of the uncertainty set.

The confidence set (8) is based on asymptotics $(N \to \infty)$, and therefore only approximately valid. To improve the approximation, several possibilities exist; see Pardo (2006). One possibility in the basic case is to consider the statistic

$$\frac{1}{\sqrt{\delta_\phi}}\left(\frac{2N}{\phi''(1)} I_\phi(p, q_N) - \gamma_\phi\right), \tag{9}$$

instead of (5). The "correction parameters" δ_ϕ and γ_ϕ, satisfying $\delta_\phi \to 1$ and $\gamma_\phi \to 0$ for $N \to \infty$, are defined by Pardo (2006, p. 190). These corrections ensure that the test statistic has the same mean and variance as the limiting χ^2-distribution, up to order $1/N$. We can use (9) to construct an approximate confidence interval, similar to (6), but due to the correction terms the approximation might be better for smaller sample sizes.

In the literature also several so-called (h, ϕ)-divergence statistics have been proposed. Such a (h, ϕ)-divergence between two probability vectors $p \ge 0$ and $q \ge 0$ in $\mathbb{R}^m$ is defined as $h(I_\phi(p, q))$, for some appropriately chosen h. Some examples, taken from Pardo (2006), are given in Table 5. Let h be increasing and continuously differentiable. Then, under $\mathbb{P}_Z$, the statistic

$$\frac{2N}{h'(0)\phi''(1)} h(I_\phi(f_0, \hat{f}_0))$$

follows the same distribution as the statistic in (5). Therefore, the uncertainty region in (8), with

$$\rho = \rho_{(h,\phi)}(N, d, \alpha) := h^{-1}\left(\frac{h'(0)\phi''(1)}{2N}\chi^2_{d,1-\alpha}\right), \tag{10}$$

is an approximate $(1-\alpha)$-confidence interval. Thus, (8) with this choice of ρ yields a (h, ϕ)-divergence-based uncertainty region.

Table 5 Examples of (h, ϕ)-Divergence Statistics

Divergence	$h(t)$	ϕ
Rényi	$\frac{1}{\theta(\theta-1)}\log(\theta(\theta-1)t+1);\quad \theta \ne 0, 1$	$\phi^\theta_{cr};\quad \theta \ne 0, 1$
Sharma–Mittal	$\frac{1}{\nu-1}((1+\theta(\theta-1)x)^{(\nu-1)/(\theta-1)}-1);\quad \theta, \nu \ne 1$	$\phi^\theta_{cr};\quad \theta \ne 0, 1$
Bhattacharyya	$-\log\left(1-\frac{1}{4}t\right)$	$\phi^{1/2}_{cr}$

[5] In this expression, $\chi^2_{m-1,1-\alpha}$ is the $1-\alpha$ percentile of the χ^2_{m-1}-distribution, i.e., $P(X \ge \chi^2_{m-1,1-\alpha}) = \alpha$, with X following a χ^2_{m-1}-distribution.

[6] We thank an anonymous referee for the suggestion to include such an example.

4. Robust Counterpart with ϕ-Divergence Uncertainty

In this section we derive the robust counterpart for (1) with a ϕ-divergence-based uncertainty region. We consider the following robust linear constraint:

$$(a + Bp)^T x \le \beta, \quad \forall p \in U, \tag{11}$$

where $x \in \mathbb{R}^n$ is the optimization vector; $a \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times m}$, and $\beta \in \mathbb{R}$ are given parameters; $p \in \mathbb{R}^m$ is the uncertain parameter; and

$$U = \{p \in \mathbb{R}^m \mid p \ge 0,\ Cp \le d,\ I_\phi(p, q) \le \rho\}, \tag{12}$$

where $q \in \mathbb{R}^m$ (with $q \ge 0$), $\rho > 0$, $d \in \mathbb{R}^k$, and $C \in \mathbb{R}^{k\times m}$ are given. As mentioned before, our main focus is on the case where p is a probability vector. Observe that in the case p is a probability vector, constraint (11) can be interpreted as an expected value constraint. When the uncertainty region is constructed as confidence set (as in (8)), we will have $q = \hat{p}_0$, the empirical or data-based estimate. Moreover, the constraints then include $e^T p \le 1$ and $e^T p \ge 1$. Formulation (12), however, is somewhat more general than (8). For example, if some additional information concerning p is available that can be expressed in terms of linear (in)equalities, these can also be included in the uncertainty region as given by (12), as long as the additional constraints are such that $q \in U$.

We prove the following theorem.

Theorem 1. *A vector $x \in \mathbb{R}^n$ satisfies* (11) *with uncertainty region U given by* (12) *such that $q \in U$ if and only if there exist $\eta \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$ such that (x, λ, η) satisfies*

$$\begin{cases} a^T x + d^T \eta + \rho\lambda + \lambda \sum_i q_i \phi^*\left(\dfrac{b_i^T x - c_i^T \eta}{\lambda}\right) \le \beta, \\ \eta \ge 0,\ \lambda \ge 0, \end{cases} \tag{13}$$

where b_i and c_i are the ith columns of B and C, respectively, and ϕ^* is the conjugate function given by (4)*, with $0\phi^*(s/0) := 0$ if $s \le 0$ and $0\phi^*(s/0) := +\infty$ if $s > 0$.*

Proof of Theorem 1. We have that (11) holds if and only if

$$\begin{aligned} \beta &\ge \max_p \{(a + Bp)^T x \mid p \in U\} \\ &= \max_{p\ge 0} \left\{ (a + Bp)^T x \;\middle|\; Cp \le d,\ \sum_{i=1}^m q_i \phi\left(\frac{p_i}{q_i}\right) \le \rho \right\}. \end{aligned} \tag{14}$$

The Lagrange function for the optimization problem on the right-hand side of (14) is given by

$$L(p, \lambda, \eta) = (a + Bp)^T x + \rho\lambda - \lambda \sum_{i=1}^m q_i \phi(p_i/q_i) + \eta^T (d - Cp),$$

and the dual objective function is

$$g(\lambda, \eta) = \max_{p\ge 0} L(p, \lambda, \eta).$$

Since $q \in U$, it follows that U is regular in the sense that $Cq \le d$ and $I_\phi(q, q) = 0 < \rho$. Because of this regularity of U, strong duality holds. Hence, it follows that x satisfies (11) if and only if $\min_{\lambda,\,\eta\ge 0} g(\lambda, \eta) \le \beta$, where the min is attained for some $\lambda \ge 0$, $\eta \ge 0$. Equivalently, x satisfies (11) if and only if $g(\lambda, \eta) \le \beta$ for some $\lambda \ge 0$ and $\eta \ge 0$. The dual objective function satisfies

$$\begin{aligned} g(\lambda, \eta) &= a^T x + d^T \eta + \rho\lambda \\ &\quad + \max_{p\ge 0} \sum_{i=1}^m \left(p_i (b_i^T x) - p_i (c_i^T \eta) - \lambda q_i \phi(p_i/q_i)\right) \\ &= a^T x + d^T \eta + \rho\lambda \\ &\quad + \sum_{i=1}^m \max_{p_i\ge 0} \left(p_i (b_i^T x - c_i^T \eta) - \lambda q_i \phi(p_i/q_i)\right) \\ &= a^T x + d^T \eta + \rho\lambda \\ &\quad + \sum_{i=1}^m q_i \max_{t\ge 0} \left\{ t (b_i^T x - c_i^T \eta) - \lambda \phi(t) \right\} \\ &= a^T x + d^T \eta + \rho\lambda + \sum_{i=1}^m q_i (\lambda\phi)^*(b_i^T x - c_i^T \eta). \end{aligned} \tag{15}$$

Finally, we have $(\lambda\phi)^*(s) = \lambda\phi^*(s/\lambda)$ for $\lambda \ge 0$, where we define $0\phi^*(s/0) := (0\phi)^*(s)$, which equals 0 if $s \le 0$ and $+\infty$ if $s > 0$. Q.E.D.

In the RCP (13) we need ϕ^*, the conjugate function of ϕ. These conjugates are given in Table 4. However, for the J-divergence, the conjugate function is not available in a closed-form expression. Nevertheless, in the next section, where we discuss the tractability aspects of (13), we also derive a tractable representation of the RCP for this case.

We now present five corollaries. The first corollary specializes the theorem to a probability vector without additional constraints.

Corollary 1. *A vector $x \in \mathbb{R}^n$ satisfies* (11) *with uncertainty region U given by*

$$U = \{p \in \mathbb{R}^m \mid p \ge 0,\ e^T p = 1,\ I_\phi(p, q) \le \rho\},$$

such that $q \in U$ if and only if there exist $\eta \in \mathbb{R}$ and $\lambda \in \mathbb{R}$ such that (x, λ, η) satisfies

$$\begin{cases} a^T x + \eta + \rho\lambda + \lambda \sum_i q_i \phi^*\left(\dfrac{b_i^T x - \eta}{\lambda}\right) \le \beta, \\ \lambda \ge 0. \end{cases} \tag{16}$$

The second corollary considers the following nonlinear constraint in $x \in \mathbb{R}^n$:

$$(a + Bp)^T f(x) \le \beta, \quad \forall p \in U, \tag{17}$$

where $a \in \mathbb{R}^k$, $B \in \mathbb{R}^{k\times m}$, $x \in \mathbb{R}^n$, and $f\colon \mathbb{R}^n \to \mathbb{R}^k$. In the sequel we shall assume that $b_i^T f(\cdot)\colon \mathbb{R}^n \to \mathbb{R}$ is convex for all i (with b_i the ith column of B). We have the following corollary.

Corollary 2. *A vector $x \in \mathbb{R}^n$ satisfies* (17) *with uncertainty region U given by* (12) *such that $q \in U$ if and only if there exist $\eta \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$ such that (x, λ, η) satisfies*

$$\begin{cases} a^T f(x) + d^T \eta + \rho\lambda \\ \quad + \lambda \sum_i q_i \phi^*\left(\dfrac{b_i^T f(x) - c_i^T \eta}{\lambda}\right) \le \beta, \\ \eta \ge 0, \lambda \ge 0. \end{cases} \tag{18}$$

Proof of Corollary 2. The dual objective is given by (15) with x replaced by $f(x)$. Therefore, it follows from (13) that (17) is equivalent to (18). Q.E.D.

Constraints such as (17) may occur if p is a probability vector and if the objective and/or constraints of a nonlinear programming problem depend on moments of a random variable. One example is the class of expected utility maximization problems, which plays a prominent role both in economics and in finance (see also §6). The standard expected utility optimization problem considers a decision maker who faces a problem in which the outcome of the decision is uncertain and depends on which scenario will be realized. Let $x \in \mathbb{R}^n$ denote the decision variable, let $\bar{r}(x, i)$ denote the payoff from decision x if scenario $i = 1, \ldots, m$ occurs, and let $u(r)$ denote the utility that the decision maker attaches to payoff $r \in \mathbb{R}$. Then, the expected utility optimization problem is given by

$$\max_{x \in X} \sum_i p_i \times u(\bar{r}(x, i)), \tag{19}$$

where $X \subset \mathbb{R}^n$ denotes the feasible region for the decision variable x, and where p_i is the probability of scenario i to occur. In case the probability vector $p = (p_1, \ldots, p_m)^T$ is not known, the robust counterpart problem is[7]

$$\max_{x \in X} \min_{p \in U} \sum_i p_i \times u(\bar{r}(x, i)), \tag{20}$$

or, equivalently,

$$\max_{z,\, x \in X,\, p \in U} \left\{ z \;\middle|\; \sum_i p_i \times u(\bar{r}(x, i)) \ge z \right\}.$$

Corollary 3. *The RCP* (20) *with uncertainty region as given in Corollary* 1 *is equivalent to*

$$\max_{x \in X,\, \lambda \ge 0,\, \eta} \left\{ -\eta - \rho\lambda - \lambda \sum_i q_i \phi^*\left(\frac{-u(\bar{r}(x, i)) - \eta}{\lambda}\right) \right\}. \tag{21}$$

This is a concave optimization problem if $u(\bar{r}(x, i))$ is concave in x for all i and the feasible set X is convex.

[7] See Gilboa and Schmeidler (1989) for an axiomatization of this utility.

Proof of Corollary 3. The proof follows from combining Corollaries 1 and 2, with $a = 0$, $B = I^{m \times m}$, and $f_i(x) = u(\bar{r}(x, i))$. Q.E.D.

The fourth corollary considers the possibility to combine several ϕ-divergences, for example, by taking the uncertainty region as an intersection of (a finite number of) ϕ-divergences, given by

$$U = \left\{ p \in \mathbb{R}^m \mid p \ge 0, Cp \le d, I_{\phi_\ell}(p, q) \le \rho_\ell, \ \ell \in \{1, \ldots, L\} \right\}, \tag{22}$$

where ϕ_ℓ are the corresponding ϕ-divergence functions, and $\rho_\ell > 0$ are given. Again, we assume $q \in U$. We have the following corollary.

Corollary 4. *A vector $x \in \mathbb{R}^n$ satisfies* (11) *with uncertainty region U given by* (22) *such that $q \in U$ if and only if there exist $\eta \in \mathbb{R}^k$ and $\lambda = (\lambda_1, \ldots, \lambda_L)^T \in \mathbb{R}^L$ such that (x, λ, η) satisfies*

$$\begin{cases} a^T x + d^T \eta + \sum_\ell \lambda_\ell \rho_\ell + \sum_\ell \lambda_\ell \sum_i q_i \phi_\ell^*\left(\dfrac{s_{\ell i}}{\lambda_\ell}\right) \le \beta, \\ \sum_\ell s_{\ell i} = b_i^T x - c_i^T \eta, \quad i \in \{1, \ldots, m\}, \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

Proof of Corollary 4. In case of (15) we get $\sum_\ell \lambda_\ell \rho_\ell$ instead of $\lambda\rho$ and $(\sum_\ell \lambda_\ell \phi_\ell)^*(b_i^T x - c_i^T \eta)$ instead of $(\lambda\phi)^*(b_i^T x - c_i^T \eta)$. Using Proposition 1, we get

$$\left(\sum_\ell \lambda_\ell \phi_\ell\right)^* (b_i^T x - c_i^T \eta) = \min_{\sum_\ell s_{\ell i} = b_i^T x - c_i^T \eta} \sum_\ell (\lambda_\ell \phi_\ell)^*(s_{\ell i}).$$

Because this expression appears in the "$\le$" inequality in (13), we may ignore the "min." Finally, using $(\lambda_\ell \phi_\ell)^*(s) = \lambda_\ell \phi_\ell^*(s/\lambda_\ell)$, we arrive at the result of the corollary. Q.E.D.

The fifth corollary considers a more general form of the uncertainty set. In the derivation of the RCP we did not exploit the special structure of the ϕ-divergence functions. Therefore, suppose that the uncertainty region in (11) is defined more generally by separable constraint functions:

$$U = \left\{ p \in \mathbb{R}^m \;\middle|\; \sum_i f_{\ell i}(p_i) \le 0, \ \forall\, \ell \in \{1, \ldots, L\} \right\}, \tag{23}$$

where $f_{\ell i}$ are convex functions such that for each i we have $\bigcap_{\ell=1}^L \mathrm{ri}(\mathrm{dom} f_{\ell i}) \ne \varnothing$. Then the following corollary gives a convex reformulation of the RCP. This result extends the classes of uncertainty regions for which finite and explicit RCPs are derived in the literature. See also Table 1.

Corollary 5. *A vector* $x \in \mathbb{R}^n$ *satisfies* (11) *with uncertainty region* U *given by* (23) *such that, for some* $\bar{p} = (\bar{p}_1, \ldots, \bar{p}_m)^T \in U$,

$$\sum_i f_{\ell i}(\bar{p}_i) < 0, \quad \forall \ell \in \{1, \ldots, L\} \tag{24}$$

if and only if there exist $\lambda = (\lambda_1, \ldots, \lambda_L)^T \in \mathbb{R}^L$ *such that* (x, λ) *satisfies*

$$\begin{cases} a^T x + \sum_\ell \sum_i \lambda_\ell f^*_{\ell i}\left(\dfrac{s_{\ell i}}{\lambda_\ell}\right) \le \beta, \\ \sum_\ell s_{\ell i} = b_i^T x, \quad i \in \{1, \ldots, m\}, \\ \lambda \ge 0, \end{cases}$$

where $f^*_{\ell i}$ *denotes the conjugate of* $f_{\ell i}$, $\ell \in \{1, \ldots, L\}$, $i \in \{1, \ldots, m\}$.

Proof of Corollary 5. The proof follows from the proof of Theorem 1 combined with that of Corollary 4, where (24) is Slater's condition guaranteeing that strong duality holds in this case. Q.E.D.

Finally, we briefly consider the case of possibly misclassified data, due to, for example, measurement or observation errors. Potential solutions are discussed by Pardo (2006). For example, one could use a double sampling scheme, where one sampling scheme is without errors but expensive, and the other sampling scheme is inexpensive but (likely) not error free. Alternatively, one could add a robustness layer with respect to q, by replacing the constraint in (13) by

$$a^T x + d^T \eta + \rho\lambda + \lambda \sum_i q_i \phi^*\left(\frac{b_i^T x - c_i^T \eta}{\lambda}\right) \le \beta, \quad \forall q \in \bar{U}, \tag{25}$$

where $\bar{U}$ is the uncertainty region for q. Note that the left-hand side of (25) is an affine function in q, so this constraint is a special case of (17). Therefore, if $\bar{U}$ is again a ϕ-divergence-based uncertainty region, we can use the results obtained in this paper to determine tractable reformulations, and if U is polyhedral or ellipsoidal, we can use the results of Ben-Tal et al. (2009). Suppose, for example, that

$$\bar{U} = \left\{ q \ge 0 \;\middle|\; \sum_i |q_i - \bar{q}_i| \le \varrho, \sum_i q_i = 1 \right\},$$

for some $\varrho > 0$. Then it can easily be proven that (x, η, λ) satisfies (25) if and only if (x, η, λ, μ) satisfies

$$\begin{cases} a^T x + d^T \eta + \rho\lambda + \mu + \varrho \|\gamma'\|_\infty + \bar{q}^T \gamma' \le \beta, \\ \gamma'_i = \lambda \phi^*\left(\dfrac{b_i^T x - c_i^T \eta}{\lambda}\right) \le \mu, \quad i \in \{1, \ldots, m\}, \\ \eta \ge 0, \lambda \ge 0. \end{cases} \tag{26}$$

5. Tractability of the Robust Counterpart

In this section we investigate a number of questions related to the RCP (13), and by answering these questions we illustrate tractable reformulations for a selection of ϕ-divergence functions, including the Kullback–Leibler divergence and the J-divergence, and the Burg entropy-based divergence. We present the tractability results of the other ϕ-divergences, which can be treated in a similar way, in the appendix. The last column of Table 4 summarizes the tractability results.

Questions that need to be addressed to derive tractable RCPs (13) are as follows:

1. What do we do if ϕ^* is not differentiable?
2. What do we do if ϕ^* does not exist in a closed form?
3. What is the convexity status of the first constraint function in (13)?
4. Does the constraint set (13) admit a self-concordant barrier?

The first question is relevant because some ϕ^* functions presented in Table 4 are not differentiable. However, for all these cases we can reformulate the problem as a differentiable problem by adding extra variables and constraints.

Question 2 will be addressed below, when we discuss uncertainty regions based on the J-divergence.

To answer question 3, concerning the convexity issue, observe that for a ϕ-divergence function ϕ its conjugate ϕ^* is also convex. Moreover, because

$$\lambda\phi^*\left(\frac{b_i^T x - c_i^T \eta}{\lambda}\right) = \sup_{t \ge 0}\{(b_i^T x - c_i^T \eta)t - \lambda\phi(t)\} \tag{27}$$

and the supremum over linear functions is convex, we obtain that the left-hand side of (27) is jointly convex in λ, x, and η, which means that the constraint function in (13) is convex. In a similar way we find that the constraint function in (18) is also convex, because we assume that $b_i^T f(\cdot)$ is convex for all i.

An affirmative answer to question 4 is very desirable because it implies the possibility to use polynomial-time interior point algorithms (see Nesterov and Nemirovski 1993). We shall address this question for the Kullback–Leibler- and the Burg entropy-based divergences. As we shall see later, after answering question 2 in case of the J-divergence, we will also be able to answer question 4 for the J-divergence.[8]

[8] In case of the other ϕ-divergences presented in Tables 2 and 4 we find that the RCP can even be reformulated as a CQP or LP problem. See the appendix.

To investigate question 4, we reformulate the constraint set (13) as follows:

$$\begin{cases} a^T x + d^T \eta + \rho\lambda + q^T z \le \beta, \\ \lambda\phi^*\left(\dfrac{b_i^T x - c_i^T \eta}{\lambda}\right) \le z_i, \quad \forall i, \\ \eta \ge 0,\ \lambda \ge 0, \end{cases} \tag{28}$$

with $z = (z_1, \ldots, z_m)^T$. In case of the Burg entropy-based divergence (like many others), we have $\mathrm{dom}(\phi^*) = (-\infty, u)$, for $u < \infty$. As a consequence, the middle inequalities of (28) can be reformulated as

$$\lambda f\left(\frac{s_i}{\lambda}\right) \le z_i, \quad s_i = \lambda u - (b_i^T x - c_i^T \eta) \ge 0, \quad \forall i, \tag{29}$$

with $f(s) := \phi^*(u - s)$. In the sequel we shall use this reformulation to answer question 4 for the Burg entropy-based divergence.

Reformulation (29) cannot be used in case of the Kullback–Leibler divergence, because the effective domain of the conjugate of its ϕ-divergence function equals the real line. In this case we apply Proposition 2 and obtain that the RCP (13) is equivalent to

$$\begin{cases} a^T x + d^T \eta + \rho\lambda \\ \quad + \lambda \sum_i q_i \inf\left[y \in \mathbb{R}\colon (\tilde{\phi})^*(-y) \le \dfrac{-b_i^T x + c_i^T \eta}{\lambda} \right] \le \beta, \\ \eta \ge 0,\ \lambda \ge 0, \end{cases} \tag{30}$$

which, in turn, is equivalent to

$$\begin{cases} a^T x + d^T \eta + \rho\lambda + q^T z \le \beta, \\ \lambda(\tilde{\phi})^*\left(\dfrac{-z_i}{\lambda}\right) \le -b_i^T x + c_i^T \eta, \quad \forall i, \\ \eta \ge 0,\ \lambda \ge 0, \end{cases} \tag{31}$$

with, again, $z = (z_1, \ldots, z_m)^T$. In case of the Kullback–Leibler divergence (like many others), we have $\mathrm{dom}((\tilde{\phi})^*) = (-\infty, \tilde{u})$, for $\tilde{u} < \infty$. As a consequence, the middle inequalities of (31) can be reformulated as

$$\lambda f\left(\frac{s_i}{\lambda}\right) \le -b_i^T x + c_i^T \eta, \quad s_i = \lambda\tilde{u} + z_i \ge 0, \quad \forall i, \tag{32}$$

now with $f(s) := (\tilde{\phi})^*(\tilde{u} - s)$. We shall use this reformulation to answer question 4 for the Kullback–Leibler divergence.

Our aim is now to establish self-concordance for the logarithmic barrier function for the constraint set (28) combined with (29) (in case of the Burg entropy-based divergence) and for the constraint set (31) combined with (32) (in case of the Kullback–Leibler divergence). We first recall the definition of a self-concordant function.

Definition 1. Let $F \subset \mathbb{R}^n$ be an open and convex set. A function $\varphi\colon F \to \mathbb{R}$ is called *κ-self-concordant on F*, with $\kappa \ge 0$, if φ is $C^3(F)$, and $\forall\, y \in F$ and $\forall\, h \in \mathbb{R}^n$ the following inequality holds:

$$|\nabla^3\varphi(y)[h, h, h]| \le 2\kappa(h^T\nabla^2\varphi(y)h)^{3/2},$$

where $\nabla^3\varphi(y)[h, h, h]$ denotes the third-order differential of φ at y and h.

We shall use the next theorem.

Theorem 2. *If for a convex function $f\colon \mathbb{R}^+ \to \mathbb{R}$ it holds that*

$$|f'''(s)| \le \kappa f''(s)/s, \quad \textit{for some } \kappa > 0, \tag{33}$$

then the logarithmic barrier function $\varphi_B(s, y, z)$ for

$$\{yf(s/y) \le z,\ s \ge 0,\ y \ge 0\}, \tag{34}$$

given by

$$\varphi_B(s, y, z) = -\ln(z - yf(s/y)) - \ln s - \ln y, \tag{35}$$

is $(2 + (\sqrt{2}/3)\kappa)$-self-concordant.

Proof of Theorem 2. Define $g(s, y) = yf(s/y)$. According to Lemma A.2 of den Hertog (1994) it holds that if there exists a β such that

$$|\nabla^3 g(s, y)[h, h, h]| \le \beta h^T\nabla^2 g(s, y)h\sqrt{\frac{h_1^2}{s^2} + \frac{h_2^2}{y^2}}, \quad \forall\, h \in \mathbb{R}^2, \tag{36}$$

in which $\nabla^3 g(s, y)[h, h, h]$ is the third-order differential, then the logarithmic barrier function for (34), given by (35), is $(1 + (1/3)\beta)$-self-concordant. We now prove that (36) holds for $\beta = 3 + \kappa\sqrt{2}$. It can easily be verified that for the second-order differential we have

$$\begin{aligned} \nabla^2 g(s, y)[h, h] &= h^T\nabla^2 g(s, y)h \\ &= f''(s/y)\left(\frac{h_1^2}{y} - \frac{2sh_1h_2}{y^2} + \frac{s^2h_2^2}{y^3}\right). \end{aligned} \tag{37}$$

Moreover, for the third-order differential we have

$$\begin{aligned} &\nabla^3 g(s, y)[h, h, h] \\ &\quad = f''(s/y)\left(-\frac{3h_1^2h_2}{y^2} + \frac{6sh_1h_2^2}{y^3} - \frac{3s^2h_2^3}{y^4}\right) \\ &\qquad + f'''(s/y)\left(\frac{h_1^3}{y^2} - \frac{3sh_1^2h_2}{y^3} + \frac{3s^2h_1h_2^2}{y^4} - \frac{s^3h_x^3}{y^5}\right). \end{aligned}$$

Using $|f'''(s)| \le \kappa f''(s)/s$, for some $\kappa > 0$, we have

$$\begin{aligned} &|\nabla^3 g(s, y)[h, h, h]| \\ &\quad \le f''(s/y)\left|-\frac{3h_1^2h_2}{y^2} + \frac{6sh_1h_2^2}{y^3} - \frac{3s^2h_2^3}{y^4}\right| \end{aligned}$$

$$
\begin{aligned}
&+\kappa\frac{y}{s}f''(s/y)\left|\frac{h_1^3}{y^2}-\frac{3sh_1^2h_2}{y^3}+\frac{3s^2h_1h_2^2}{y^4}-\frac{s^3h_x^3}{y^5}\right|\\
&\le 3f''(s/y)\left(\frac{h_1^2}{y}-\frac{2sh_1h_2}{y^2}+\frac{s^2h_2^2}{y^3}\right)\frac{|h_2|}{y}\\
&\quad+\kappa f''(s/y)\left(\frac{h_1^2}{y}-\frac{2sh_1h_2}{y^2}+\frac{s^2h_2^2}{y^3}\right)\left(\frac{|h_1|}{s}+\frac{|h_2|}{y}\right)\\
&\le(3+\kappa\sqrt{2})h^T\nabla^2 g(s,y)h\sqrt{\frac{h_1^2}{s^2}+\frac{h_2^2}{y^2}}.
\end{aligned}
$$

This proves that (36) holds for $\beta=3+\kappa\sqrt{2}$, and hence that the corresponding logarithmic barrier function is $(2+(\sqrt{2}/3)\kappa)$-self-concordant. Q.E.D.

To apply this theorem, notice that we reformulated the relevant parts of both the constraint set (28) combined with (29) (in case of the Burg entropy-based divergence) and the constraint (31) combined with (32) (in case of the Kullback–Leibler divergence) as (34). Thus, if f in (29) or in (32) satisfies condition (33), then the theorem implies that the logarithmic barrier function for the corresponding constraint set is self-concordant. In case of the *Burg entropy-based divergence* we have $\mathrm{dom}(\phi^*)=(-\infty,1)$, resulting in $f(s)=-\log(s)$. This function f satisfies condition (33) with $\kappa=2$. Therefore, it follows that in case of the Burg entropy-based divergence, (28) combined with (29) is tractable. As an immediate consequence, we also have that (31) combined with (32) is tractable in case of the *Kullback–Leibler divergence*.

Finally, we return to question 2: What do we do if ϕ^* does not exist in a closed form? For these cases Propositions 1 and 2 may be of help.

First, consider the case where ϕ^* is not available in closed-form expression, but there exist ϕ-divergences ϕ_1 and ϕ_2 such that $\phi=\phi_1+\phi_2$, and ϕ_1^* and ϕ_2^* are available in closed form. Then, applying Proposition 1, we obtain that the RCP (13) is equivalent to

$$
\begin{cases}
a^Tx+d^T\eta+\rho\lambda\\
\quad+\lambda\sum_i q_i\min_{s_{1i}+s_{2i}=b_i^Tx-c_i^T\eta}[\phi_1^*(s_{1i}/\lambda)+\phi_2^*(s_{2i}/\lambda)]\le\beta,\\
\eta\ge 0,\ \lambda\ge 0.
\end{cases}
$$

Because the first inequality is a "$\le$" one, we may delete the "min" and get the following system of inequalities in (x,η,λ,s_1,s_2) to represent the RCP (13):

$$
\begin{cases}
a^Tx+d^T\eta+\lambda\rho\\
\quad+\lambda\sum_i q_i[\phi_1^*(s_{1i}/\lambda)+\phi_2^*(s_{2i}/\lambda)]\le\beta,\\
s_{1i}+s_{2i}=b_i^Tx-c_i^T\eta,\quad\forall i,\\
\eta\ge 0,\ \lambda\ge 0.
\end{cases}
\tag{38}
$$

This is a tractable problem if, loosely speaking, the corresponding problems for ϕ_1 and ϕ_2 are tractable.

We can apply this approach to the *J-divergence* for which there is no closed-form expression available for ϕ^*. The crucial observation is that in case of the J-divergence we have $\phi_j(t)=(t-1)\log t=t\log t-\log t=\phi_{kl}(t)+\phi_b(t)$, where $\phi_{kl}(t)$ is the Kullback–Leibler ϕ-divergence function and $\phi_b(t)$ the ϕ-divergence function based on the Burg entropy.

To complete our analysis, we give an example of a ϕ-divergence function, for which a closed-form expression for its conjugate is not available, but for which still a tractable RCP can be derived by using Proposition 2. Suppose that

$$\phi(t)=|t-1|^\theta t^{1-\theta}.$$

It can be verified that this is a ϕ-divergence function corresponding to a well-defined ϕ-divergence. However, ϕ^* is not available in a closed-form expression, and hence (13) cannot be used directly. To overcome this problem we observe that $\tilde{\phi}=\phi_{ca}^\theta$, i.e., the adjoint of ϕ is the χ-divergence function of order θ, for which a closed-form expression for its conjugate is available (see Table 4). Therefore, one can obtain a tractable RCP for this choice of ϕ by using (31), based on an application of Proposition 2.

6. Implementation Issues and Examples

In this section we first discuss some implementation issues. Next, we present two applications of robust expected utility maximization, namely, an investment problem and a newsvendor problem. Finally, to illustrate the performance of ϕ-divergence-based robust optimization, we present as numerical example a multi-item newsvendor optimization problem.

6.1. Implementation Issues

In this subsection we first discuss three important issues concerning the choice of a suitable ϕ-divergence. We then discuss computational issues related to the number of scenarios.

The first issue concerning the choice of ϕ is the tractability of the final optimization problem. In Table 4, the tractability of the resulting optimization problem is given for many choices of ϕ. In general, LP is preferred to CQP, and CQP is preferred to more general convex problems that still allow a self-concordant barrier (indicated by "S.C." in Table 4).

The second important issue concerning the choice of ϕ is the "statistical" behavior. In this paper we use first-order asymptotics, yielding as result that the test statistic (5) asymptotically ($N\to\infty$) follows a χ_d^2-distribution. Based on first-order asymptotics, there is no difference between the various ϕ-divergences. Using higher-order asymptotics might result in differences between the ϕ-divergences as

higher-order terms of the test statistic (5) might be smaller (closer to zero) for one ϕ-divergence than for another ϕ-divergence. However, these higher-order terms are rather complicated (and also partly unknown) (see Pardo 2006).

The third important issue is related to the fact that, for finite N, the uncertainty regions for different choices of ϕ may differ. The simulation study presented in §6.4 shows that the choice of ϕ indeed might matter. However, it is a priori not possible to judge which uncertainty region (i.e., which ϕ) is the "best." We advocate the pragmatic approach to perform the robust optimization for different choices of ϕ, and then select the one that leads to the best optimal objective value. The family of Cressie and Read (1984) divergences, which is parameterized by the parameter $\theta \in \mathbb{R}$ (see Table 2), is rich enough to contain most of the ϕ-divergence functionals (see Table 3). Therefore, the task of choosing the best ϕ-divergence from this large class can be reduced essentially to the choice of the single parameter θ. Consider, for example, the case of expected utility maximization considered in Corollary 3. Let $\Psi(\phi)$ be the optimum value function of problem (21) for a given ϕ. For the Cressie–Read divergence ϕ^{θ}_{cr}, let $\tilde{\Psi}(\theta) := \Psi(\phi^{\theta}_{cr})$. The best θ can then be obtained by solving the single variable optimization problem $\max_{\theta} \tilde{\Psi}(\theta)$.

We conclude this section by discussing computational issues related to the number of scenarios. Note that the robust counterpart (13) is also tractable with respect to the number of scenarios. In fact, the number of variables and constraints in (13) does not depend on the number of scenarios. However, if the number of scenarios is huge, then still computational issues can arise, because the summation in constraint (13) is over all scenarios. For such cases one can try to reduce the number of scenarios by using a factor model, which is typically done in portfolio analysis. This reduces the number of stochastic parameters, and therefore the number of scenarios. One can also try to group scenarios to reduce the number of scenarios. If the number of scenarios is really large, one can try to approximate the discrete probability distribution by a continuous distribution. In a future paper we will extend the approach of this paper to continuous probability distributions.

6.2. Investment Example

As a special case of the expected utility maximization problem considered in (19), we consider an investment problem. Let $R_i \in \mathbb{R}^n$ be an n-dimensional vector of gross returns in case of scenario i. Investors can choose portfolios represented by a vector of weights x belonging to the set $X \equiv \{x \in \mathbb{R}^n \mid x^T e = 1\}$. If scenario i occurs, the portfolio with weights x yields as gross return $\bar{r}(x, i) = x^T R_i$. Let $R_i = (R_{1i}, \tilde{R}_i^T)^T$, with $\tilde{R}_i$ the $(n-1)$-dimensional subvector of R_i, containing the gross returns of assets 2 to n. Similarly, let $x = (x_1, \tilde{x}^T)^T$ and $e = (1, \tilde{e}^T)^T$. Then optimization problem (19) becomes

$$\max_{\tilde{x}} \sum_i p_i \times u(R_{1i} + \tilde{x}^T \tilde{R}_i^e), \tag{39}$$

with $\tilde{R}_i^e = \tilde{R}_i - R_{1i}\tilde{e}$. Similarly, the RCP becomes (20), with $\bar{r}(x, i) = R_{1i} + \tilde{x}^T \tilde{R}^e$ and U as given in Corollary 1. We shall assume that u is differentiable and that its derivative satisfies $u'(\cdot) > 0$.

The first-order optimality conditions for problem (39) are given by

$$\sum_i p_i \times u'(R_{1i} + \tilde{x}^T \tilde{R}_i^e) \times (R_{ji} - R_{1i}) = 0, \quad j = 2, \ldots, n. \tag{40}$$

This equation is a special case of the "basic equation of asset pricing" (Campbell 2000, p. 1517).[9] It is an equilibrium condition, stating that the weighted average of the excess return of any asset j in excess of a reference asset (in our case, asset 1) equals zero. The positive random variable realizing these weights $u'(R_{1i} + \tilde{x}^T \tilde{R}_i^e)$ is a so-called *stochastic discount factor* (SDF). In the risk neutral case, i.e., when u is linear, the SDF would be constant, and the equilibrium condition (40) becomes the condition that all expected returns are equal. As its name suggests, the "basic equation of asset pricing" is heavily used in finance, particularly in equilibrium pricing. But also when estimating and testing a particular asset pricing model, one typically makes use of the implied SDF.

A natural question is whether the first-order conditions of the RCP (20) or its equivalence (21) is also a special case of the "basic equation of asset pricing." To obtain the first-order conditions of the RCP, we shall assume that ϕ^* is differentiable, with $(\phi^*)'(\cdot) > 0$,[10] and we shall assume that $\lambda > 0$. Then we find, as first-order conditions,

$$\text{(w.r.t. } \eta\text{:)} \quad -1 + \sum_i \tilde{q}_i = 0,$$

$$\text{(w.r.t. } \lambda\text{:)} \quad -\rho - \sum_i q_i \times \phi^*\left(\frac{-u(R_{1i} + \tilde{x}^T \tilde{R}_i^e) - \eta}{\lambda}\right) - \frac{1}{\lambda} \sum_i \tilde{q}_i \times (R_{1i} + \tilde{x}^T \tilde{R}_i^e + \eta) = 0,$$

[9] We assume rational expectations; i.e., the probabilities p_i represent the "true" probabilities. The derived SDF is up to normalization, because Equation (40) is in terms of excess returns.

[10] It follows from the assumptions that $\mathrm{dom}\phi = \mathbb{R}^+$, and hence $(\phi^*)'(\cdot) \geq 0$. However, from Tables 2 and 4 we see that for some choices of ϕ, like the modified χ^2-distance or the variation distance, $(\phi^*)'(\cdot)$ may be zero. Such choices of ϕ are excluded in the sequel, as they do not result in a strictly positive SDF, required in the "basic equation of asset pricing."

$$\text{(w.r.t. } \tilde{x}\text{:)} \quad \sum_i \tilde{q}_i \times u'(R_{1i} + \tilde{x}^T \tilde{R}_i^e) \times (R_{ji} - R_{1i}) = 0, \quad j = 2, \ldots, n,$$

where

$$\tilde{q}_i \equiv q_i \times (\phi^*)'\left(\frac{-u(R_{1i} + \tilde{x}^T \tilde{R}_i^e) - \eta}{\lambda}\right), \quad i = 1, \ldots, m.$$

If we combine the first-order conditions with respect to η and $\tilde{x}$, we see that we have to solve the same system of equations as in the case of (40). The difference is that the probabilities p_i are replaced by $\tilde{q}_i$, $i = 1, \ldots, m$.[11] To become a special case of the "basic equation of asset pricing," we consider the equations as expectations with respect to $q_{i,N}$, the empirical counterparts of p_i. We then find as the SDF[12]

$$(\phi^*)'\left(\frac{-u(R_{1i} + \tilde{x}^T \tilde{R}_i^e) - \eta}{\lambda}\right) \times u'(R_{1i'} + \tilde{x}^T \tilde{R}_i^e).$$

Thus, our reformulation (21), specialized to the investment problem, allows a straightforward way to retrieve the SDF in case of the robust optimization problem. This makes reformulation (21) also relevant from the point of view of equilibrium pricing and empirical finance.

A special case is obtained in case of risk neutrality, i.e., when $u(r) = r$. Then the RCP can be reformulated as

$$\min_{\tilde{x}} \max_{p \in U} \sum_i p_i \times (-(R_{1i} + \tilde{x}^T \tilde{R}_i^e)).$$

The inner maximization represents a coherent risk measure (see Artzner et al. 1999). Thus, in this special case of the RCP the portfolio weights are determined by minimizing a coherent risk measure. Pelsser (2010) provides an application.

6.3. Newsvendor Example

In this subsection, we consider as application of utility maximization the single-item *newsvendor* problem. The newsvendor's problem is how many units of a product (item) to order, taking into account that the demand for the product is stochastic. Because of uncertainty, the newsvendor can face both unsold items or unmet demand. The unsold items will return a loss because their salvage value is lower than the purchase price. In the case of unmet demand the newsvendor incurs a cost of lost sales, which may include a penalty for the lost customer goodwill.

Let $u(r)$ denote the newsvendor's utility from net profit $r \in \mathbb{R}$. His objective is to choose the order quantity $x = Q$ to maximize the expected utility (19) of his net profit

$$\tilde{r}(Q, i) = v \min(d_i, Q) + s(Q - d_i)^+ - l(d_i - Q)^+ - cQ,$$

where $d_i \geq 0$ is the uncertain demand in scenario i, v is the unit selling price, s is the salvage value per unsold item, l is the shortage cost per unit of unsatisfied demand, and c is the purchasing price per unit. A standard assumption for this problem is $v + l \geq s$.

In case the probability distribution of the demand is unknown, the RCP is given by

$$\max_Q \min_{p \in U} \left\{ \sum_i p_i \times u\big(v \min(d_i, Q) + s(Q - d_i)^+ - l(d_i - Q)^+ - cQ\big) \right\}.$$

It follows immediately from Corollary 3 that with a ϕ-divergence uncertainty region U as given by Corollary 1 this problem is equivalent to

$$\min_{Q,\, \lambda \geq 0,\, \eta} \left\{ \eta + \rho\lambda + \lambda \sum_i q_i \phi^*\big[(-u(v \min(d_i, Q) + s(Q - d_i)^+ - l(d_i - Q)^+ - cQ) - \eta) \cdot (\lambda)^{-1}\big] \right\}.$$

With a concave utility function $u(\cdot)$, the assumption $v + l \geq s$ ensures that

$$-u(v \min(d_i, Q) + s(Q - d_i)^+ - l(d_i - Q)^+ - cQ)$$

is convex in Q for all i.

Several papers study risk aversion in the newsvendor model by using as objective function expected utility (Eeckhoudt et al. 2009), mean–variance (Chen and Federgruen 2000), or conditional value at risk (Chen et al. 2009). Still, all of these papers assume that the entire demand distribution is known. Our approach can be used to add risk aversion with respect to the unknown demand distribution in cases where only some historical data are given.

Note that our approach can also be applied to regret approaches for the newsvendor model. Perakis and Roels (2008) studied regret in newsvendor models in which only partial information is given, for example, mean, variance, symmetry, or unimodality. Our result here can be used to minimize robustly the regret when only some historical demand data are available.

[11] Moreover, the expectation of the optimal RCP portfolio return with respect to these probabilities $\tilde{q}_i$ has to equal the maximum value of the objective function (21). This latter requirement follows from a reformulation of the first-order conditions with respect to λ:

$$\sum_i \tilde{q}_i (R_{1i} + \tilde{x}^T \tilde{R}_i^e) = -\eta - \rho\lambda - \lambda \sum_i q_i \phi^*\left(\frac{-u(R_{1i} + \tilde{x}^T \tilde{R}_i^e) - \eta}{\lambda}\right).$$

[12] The SDF is again up to normalization; see Footnote 9. Moreover, this SDF is the relevant one from an empirical point of view in case $q = q_N$, which is consistent for the true probability vector.

6.4. Numerical Illustration: Multi-item Newsvendor Example

As numerical illustration, we consider a multi-item newsvendor problem (see, for example, Erlebacher 2000, Moon and Silver 2000). This problem deals with optimizing the inventory of several items that can only be sold in one period. Because of the uncertain demand, this newsvendor can face both unsold items or unmet demand. As in the single-item case, the unsold items will return a loss, and unmet demand generates a cost of lost sales. For each item j, we define the purchase cost c_j, the selling price v_j, the salvage value of unsold items s_j, and the cost of lost sales l_j. Furthermore, we denote γ for the budget that is available for the purchase of the items.

We assume that demand for item j is a random variable that can take on m values, denoted as d_i, $i=1,\dots,m$ (i.e., for simplicity, the same possible outcomes for all items j). We denote by $p_i^{(j)}$ the unknown probability that the demand for item j equals d_i, and we let the uncertainty region for $p^{(j)}=(p_1^{(j)},\dots,p_m^{(j)})^T$ be given by

$$U^{(j)} := \{p^{(j)}\in\mathbb{R}^m \mid p^{(j)}\geq 0,\ (p^{(j)})^T e=1,\ I_\phi(p^{(j)},q_N^{(j)})\leq\rho\}, \quad (41)$$

where $q_N^{(j)}$ represents the sample-based estimated probability distribution for item j, with N denoting sample size.

Denote by Q_j the order quantity for item j. We consider two types of multi-item newsvendor problems. The first is to maximize the sum of the expected profits,

$$\max_Q \sum_j \sum_i p_i^{(j)} \bar{r}_j(Q_j,i),$$

and the second is to maximize the minimal expected profit,

$$\max_Q \min_j \sum_i p_i^{(j)} \bar{r}_j(Q_j,i).$$

The robust versions of these problems can be stated as

$$\begin{aligned}\max\ & \|z\|\\ \text{s.t.}\ & -c_jQ_j+\sum_i p_i^{(j)} f_{i,j}(Q_j)\geq z_j, \quad \forall j,\ \forall p^{(j)}\in U^{(j)},\\ & \sum_j c_jQ_j\leq\gamma\end{aligned}$$

for the case where the norm in the objective is either the 1-norm or the ∞-norm, respectively, and with

$$f_{i,j}(Q_j)=v_j\min\{d_i,Q_j\}+s_j(Q_j-d_i)^+-l_j(d_i-Q_j)^+.$$

It follows from (18) that this problem can be reformulated as

$$\begin{aligned}\max\ & \|z\|\\ \text{s.t.}\ & -c_jQ_j-\eta_j-\lambda_j\rho\\ & \quad -\lambda_j\sum_i q_{i,N}^{(j)}\phi^*\left(\frac{-f_{i,j}(Q_j)-\eta_j}{\lambda_j}\right)\geq z_j, \quad \forall j,\\ & \sum_j c_jQ_j\leq\gamma,\\ & \lambda\geq 0.\end{aligned}$$

Our numerical results apply to the case with $n=12$ different items, and $m=3$ scenarios for the demand for each item: low demand (4), medium demand (8), and high demand (10), denoted as $d_1=4$, $d_2=8$, and $d_3=10$, respectively. The parameter values of the revenue functions, as well as the values of $q_{i,N}^{(j)}$, are as given in Table 6. Furthermore, the budget is set at $\gamma=1{,}000$.

We solve the RCP for the Burg entropy-based divergence (or the Kullback–Leibler divergence in terms of $\tilde{\phi}$) and for the Cressie and Read (1984) ϕ-divergence function with $\theta=0.5$. For both ϕ-divergence functions, we consider the confidence set based on the test statistic (5), with corresponding ρ denoted by ρ_ϕ^a, and the confidence set based on the corrected test statistic (9), with the corresponding ρ denoted by ρ_ϕ^c. In each case, the confidence level is set at $\alpha=0.05$, and we determine the robust optimal solutions for different sample sizes $N=10,20,\dots,1{,}000$.

Using the solutions of the RCP problems and the solution of the nonrobust problem (i.e., assuming that q_N is the true probability vector), we make several comparisons. First, we compare the performance of the robust versus the nonrobust solutions for the different values of the sample size N (which in turn yields different values for ρ_ϕ^a and ρ_ϕ^c). Second, we compare the results of the two ϕ-divergence functions. Third, for each ϕ-divergence function, we look

Table 6 Parameter Values for the Multi-Item Newsvendor Example

Item (j)	1	2	3	4	5	6	7	8	9	10	11	12
c	4	5	6	4	5	6	4	5	6	4	5	6
v	6	8	9	5	9	8	6	8	9	6.5	7	8
s	2	2.5	1.5	1.5	2.5	2	2.5	1.5	2	2	1.5	1
l	4	3	5	4	3.5	4.5	3.5	3	5	3.5	3	5
$q_{1,N}^{(j)}$	0.375	0.250	0.375	0.127	0.958	0.158	0.485	0.142	0.679	0.392	0.171	0.046
$q_{2,N}^{(j)}$	0.375	0.250	0.250	0.786	0.007	0.813	0.472	0.658	0.079	0.351	0.484	0.231
$q_{3,N}^{(j)}$	0.250	0.500	0.375	0.087	0.035	0.029	0.043	0.200	0.242	0.257	0.345	0.723

at the effect of using the corrected test statistic instead of the approximate test statistic, i.e., using $\rho = \rho^c_\phi$ instead of $\rho = \rho^a_\phi$.

To make comparisons, we proceed as follows. First, we sample 10,000 hypothetically true p-vectors. Next, for each sampled probability vector p, we calculate the value of the objective function for the nonrobust as well as for the robust optimal solutions. We then compare the performance of the different solutions by determining the mean and the range (i.e., the minimum and the maximum values) of the objective values corresponding to the sampled p-vectors.

The p-vectors are sampled such that approximately 95% of the sample satisfies $I_{\phi_{mc}}(p, q_N) \le \bar\rho := \rho^a_{\phi_{mc}}$, where ϕ_{mc} denotes the modified χ^2-divergence. Specifically, we sample p_i, for $i = 1, \ldots, m-1$, from a normal distribution $N(q_{i,N}, \sigma_i)$, and set $p_m = 1 - \sum_{i=1}^{m-1} p_i$. If this sampling returns a probability vector (i.e., $p_i \ge 0$ for $i = 1, \ldots, m$), we accept the vector. Otherwise, we repeat the sampling until a valid p-vector is found. To satisfy $I_{\phi_{mc}}(p, q_N) \le \bar\rho$ for approximately 95% of the sampled p-vectors, we determine the value of σ_i of the normal distribution as follows. We know that the condition $I_{\phi_{mc}}(p, q_N) \le \bar\rho$ is satisfied if (but not only if) the following holds:

$$\frac{(p_i - q_{i,N})^2}{q_{i,N}} \le \frac{\bar\rho}{m} \quad \Leftrightarrow \quad q_{i,N} - \sqrt{\frac{\bar\rho}{m} q_{i,N}} \le p_i \le q_{i,N} + \sqrt{\frac{\bar\rho}{m} q_{i,N}}.$$

For the normal distribution, approximately 95% of the values are within two standard deviations from the mean. Therefore, we take $\sigma_i = (1/2)\sqrt{(\bar\rho/m) q_{i,N}}$. Because $\bar\rho$ can be relatively large for small values of N and to avoid too many invalid samples, we put an upper bound of $(1/2) q_{i,N}$ on σ_i. Figures 1 and 2 display the range and the mean of the objective values corresponding to the sampled p-vectors for the Cressie and Read (1984) divergence function

Figure 1 **Cressie–Read for $\theta = 0.5$, and ρ^c_ϕ, and the 1-Norm**

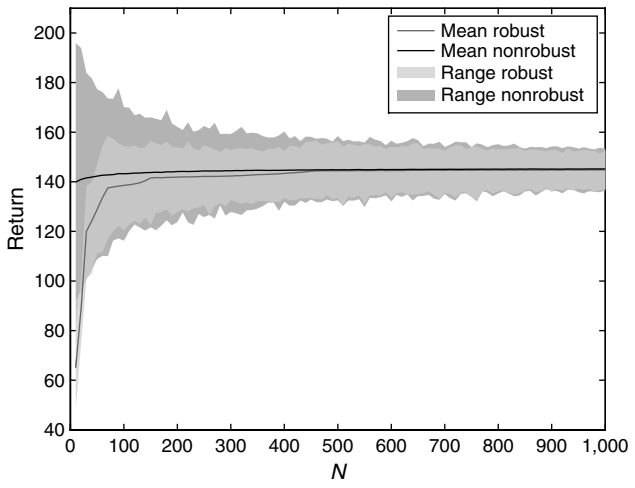
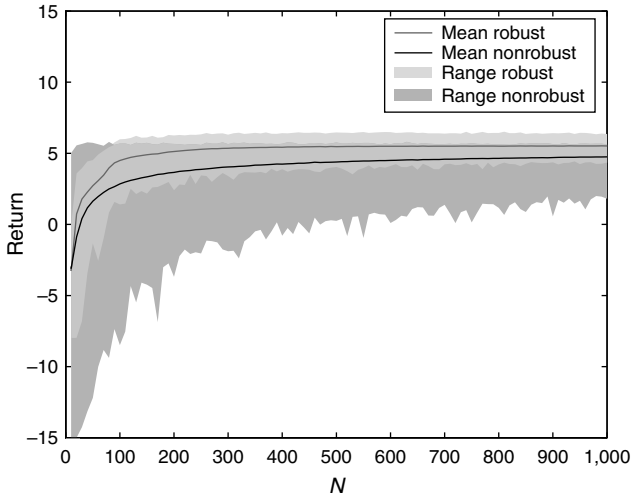


Figure 2 **Cressie–Read for $\theta = 0.5$, and ρ^c_ϕ, and the ∞-Norm**



with $\rho = \rho^c_\phi$, for the 1-norm and the ∞-norm, respectively. The results for the Burg entropy-based divergence function are qualitatively similar.

Concerning the value of the objectives of the robust optimizations, there is a significant difference between using the 1-norm and the ∞-norm. For the 1-norm (Figure 1), it holds that for small values of N, the mean of the objective values for the robust solution is lower than the mean of the objective values for the nonrobust solution, but as N grows the two methods have practically the same mean profit. In contrast, for the ∞-norm (Figure 2), the mean of the objective values for the robust solution is higher than the mean for the nonrobust solution. Moreover, the range of objective values for the robust solution is significantly smaller than the range of objective values for the nonrobust solution for the ∞-norm. In particular, the robust solution avoids substantial losses.

Concerning the effect of N, the effect of using ρ^c_ϕ versus ρ^a_ϕ, and the differences between the two ϕ-divergence functions, we observe the following

Effect of N. Because 95% of the sampled p-vectors needs to satisfy $I_{\phi_{mc}}(p, q_N) \le \bar\rho$, and because $\bar\rho$ is decreasing in N, the range of the expected returns becomes smaller as N increases. However, because 5% of the sampled p-vectors does not need to satisfy $I_{\phi_{mc}}(p, q_N) \le \bar\rho$, the range does not converge to a single value.

Effect of ρ^c_ϕ vs. ρ^a_ϕ. With regard to the differences between the robust solutions in case where ρ^a_ϕ is used (i.e., the uncertainty region is based on the approximate test statistic) and when ρ^c_ϕ is used (i.e., the uncertainty region is based on the corrected test statistic), we observe that there are significant differences only for relatively small values for N. This occurs of course because the effect of the correction becomes smaller as N increases.

Comparison of Different ϕ-Divergence Functions. Although the qualitative effects of robust optimization are similar for different ϕ-divergence functions, the optimal expected profit can differ significantly. For example, for the ∞-norm, we considered the Cressie–Read divergence function with $\theta = -1, 1/2, 1$. The average of the expected profit as well as the range of the expected profits over the sampled p-vectors and over the values of N depend on θ. We found that $\theta = 1/2$ yields an average expected profit that is 6.4% higher than the average for $\theta = 1$, and 7.6% higher than the average for $\theta = -1$. For the 1-norm, we found that $\theta = 1/2$ yields an average expected profit that is 1.4% higher than the average for $\theta = 1$, but 1% lower than the average for $\theta = -1$. The difference between $\theta = 1/2$ and $\theta = 1$ over the sampled p-vectors and over the values of N ranges from 0% to 10.1%, and the difference between $\theta = 1/2$ and $\theta = -1$ ranges from -2.3% to 3.7%. The approach proposed in §6.1 may be used to select the "best" θ.

7. Concluding Remarks

In this paper we have shown that the robust counterpart of linear and nonlinear optimization problems with uncertainty regions defined by ϕ-divergence distance measures can be reformulated as tractable optimization problems. Thus, these uncertainty regions are useful alternatives to uncertainty regions considered in the existing literature, particularly so when the uncertainty is associated with probabilities. In this latter case, we have shown that uncertainty regions based on ϕ-divergence test statistics have a natural interpretation in terms of statistical confidence sets. This allows for an approach that is fully data driven.

Our approach also has other applications. For example, ϕ-divergence distances can be directly used as the distance in the so-called *globalized robust counterpart* methodology (see Ben-Tal et al. 2009, Chap. 3).

Let us now mention some directions for further research. First, different choices of ϕ have been proposed in the literature (Pardo 2006), each of them with different statistical properties. It could be interesting to study the differences in performance of optimal solutions of robust counterpart problems such as (13) for different choices of ϕ.

Next, in the classical statistical literature, many goodness-of-fit statistics are considered that do not belong to the ϕ-divergence class. It is an interesting topic for further research to analyze whether the corresponding robust counterparts are tractable.

In this paper we consider the setting where there is only a finite number of scenarios. It is possible to extend the results to general continuous distributions, which will be the subject of a future paper.

In terms of practical applications, it may be useful to extensively study the applicability of the proposed approach to, for example, asset liability management problems and other inventory control problems. In particular, with respect to inventory control problems it may be interesting to extend the work of Wagner (2010). In that paper, the Wagner–Whitin model with backlogged demand and period-dependent costs is analyzed in settings in which the demand distribution is not known. Our analysis can most likely be used to extend the analysis to the more practical case where only some historical demand data are given.

Finally, several commonly used risk measures in finance (for example, mean–variance, expected shortfall) are nonlinear in probabilities. It is a challenging question whether the proposed approach can be extended to problems in which the unknown probability vector p appears nonlinearly. In Ben-Tal et al. (2009), techniques are described to deal with certain types of nonlinear uncertainty, and maybe similar techniques can be used in this case.

Acknowledgments

The authors thank the department editor, an associate editor, and two anonymous referees for useful comments that helped to improve this paper. This paper was written when the first author was visiting Centrum Wiskunde and Informatica in Amsterdam, the Netherlands, as a CWI Distinguished Scientist.

Appendix. Tractable Reformulations

In this appendix we give the final tractable reformulations for (13) for different choices of ϕ. The tractable reformulations for the Kullback–Leibler divergence and J-divergence and the Burg entropy-based divergence are already derived in §5.

χ^2-distance (CQP):

$$\begin{cases} a^T x + d^T \eta + \lambda\rho + 2\lambda e^T q - 2q^T y \le \beta; \\ \sqrt{y_i^2 + \frac{1}{4}(b_i^T x - c_i^T \eta)^2} \le \frac{1}{2}(2\lambda - b_i^T x + c_i^T \eta), \quad \forall i; \\ b_i^T x - c_i^T \eta \le \lambda, \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

Modified χ^2 distance (CQP):

$$\begin{cases} a^T x + d^T \eta + \lambda(\rho - e^T q) + \frac{1}{4} q^T y \le \beta; \\ \sqrt{z_i^2 + \frac{1}{4}(\lambda - \mu_i)^2} \le \frac{1}{2}(\lambda + \mu_i), \quad \forall i; \\ z_i \ge 0, \quad \forall i; \\ z_i \ge b_i^T x - c_i^T \eta + 2\lambda, \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

Hellinger distance (CQP):

$$\begin{cases} a^T x + d^T \eta + \lambda\rho - \lambda e^T q + q^T y \le \beta; \\ \sqrt{\lambda^2 + \frac{1}{4}(y_i - \lambda + b_i^T x - c_i^T \eta)^2} \le \frac{1}{2}(y_i + \lambda - b_i^T x + c_i^T \eta), \quad \forall i; \\ b_i^T x - c_i^T \eta \le \lambda, \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

χ divergence of order θ (CQP):

$$\begin{cases} a^T x + d^T \eta + \lambda\rho + \sum_i q_i (b_i^T x - c_i^T \eta) \\ \quad + \lambda(\theta - 1) \sum_i q_i (z_i/(\theta\lambda))^{\theta/(\theta-1)} \le \beta; \\ z_i \ge -b_i^T x + c_i^T \eta, \quad \forall i; \\ z_i \ge b_i^T x - c_i^T \eta, \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

Variation distance (LP):

$$\begin{cases} a^T x + d^T \eta + \lambda\rho + q^T y \le \beta; \\ y_i \ge -\lambda, \quad \forall i; \\ y_i \ge b_i^T x - c_i^T \eta, \quad \forall i; \\ b_i^T x - c_i^T \eta \ge \lambda, \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

Cressie–Read (CQP):

$$\begin{cases} a^T x + d^T \eta + \lambda\rho + \dfrac{\lambda}{\theta} \sum q_i \left((y_i/\lambda)^{\theta/(\theta-1)} - 1 \right) \le \beta; \\ y_i = \lambda - (1 - \theta)(b_i^T x - c_i^T \eta), \quad \forall i; \\ \eta \ge 0, \lambda \ge 0. \end{cases}$$

References

Artzner P, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. *Math. Finance* 9(3):203–228.

Ben-Tal A, Nemirovski A (1997) Robust truss topology design via semidefinite programming. *SIAM J. Optim.* 7(4):991–1016.

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Math. Oper. Res.* 23(4):769–805.

Ben-Tal A, Ben-Israel A, Teboulle M (1991) Certainty equivalents and information measures: Duality and extremal principles. *J. Math. Anal. Appl.* 157(1):211–236.

Ben-Tal A, Bertsimas D, Brown D (2010) A soft robust model for optimizing under ambiguity. *Oper. Res.* 58(4):1220–1234.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Bertsimas D, Brown D (2009) Constructing uncertainty sets for robust linear optimization. *Oper. Res.* 57(6):1483–1495.

Bertsimas D, Brown D, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53(3):464–501.

Billingsley P (1995) *Probability and Measure*, 3rd ed. (John Wiley & Sons, New York).

Broniatowski M, Keziou A (2009) Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.* 100(1):16–36.

Calafiore GC (2007) Ambiguous risk measures and optimal robust portfolios. *SIAM J. Optim.* 18(3):853–877.

Campbell JY (2000) Asset pricing at the millennium. *J. Finance* 55(4):1515–1567.

Chen Y, Federgruen A (2000) Mean-variance analysis of basic inventory models. Working paper, Columbia University, New York.

Chen Y, Xu M, Zhang ZG (2009) A risk-averse newsvendor model under the cvar criterion. *Oper. Res.* 57(4):1040–1044.

Cressie N, Read TR (1984) Multinomial goodness-of-fit tests. *J. Royal Statist. Soc., Series B* 46(3):440–464.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

den Hertog D (1994) *Interior Point Approach to Linear, Quadratic and Convex Programming* (Kluwer Academic, Dordrecht, The Netherlands).

Eeckhoudt L, Gollier C, Schlesinger H (2009) Risk averse (and prudent) newsboy. *Management Sci.* 41(5):786–794.

El Ghaoui L, Lebret H (1997) Robust solution to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.* 18(4):1035–1064.

El Ghaoui L, Oustry F, Lebret H (1998) Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.* 9(1):33–52.

Erlebacher SJ (2000) Optimal and heuristic solutions for the multi-item newsvendor problem with a single capacity constraint. *Production Oper. Management* 9(3):303–318.

Fabozzi F, Huang D, Zhou G (2010) Robust portfolios: Contributions from operations research and finance. *Ann. Oper. Res.* 176(1):191–220.

Garlappi L, Uppal R, Wang T (2007) Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Rev. Financial Stud.* 20(1):41–81.

Gilboa I, Schmeidler D (1989) Maxmin expected utility with non-unique prior. *J. Math. Econom.* 18(2):141–153.

Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Math. Oper. Res.* 28(1):1–38.

Gushchin AA (2008) On an extension of the notion of f-divergence. *Theory. Probab. Appl.* 52(3):439–455.

Jager L, Wellner JA (2007) Goodness-of-fit tests via phi-divergences. *Ann. Statist.* 35(5):2018–2053.

Klabjan D, Simchi-Levi D, Song M (2013) Robust stochastic lot-sizing by means of histograms. *Production Oper. Management* Forthcoming.

Kouvelis P, Yu G (1997) *Robust Discrete Optimization and Its Application* (Kluwer Academic, London).

Liese F, Vajda I (2006) On divergences and information in statistics and information theory. *IEEE Trans. Inform. Theory* 52(10):4394–4412.

Moon I, Silver EA (2000) The multi-item newsvendor problem with a budget constraint and fixed ordering costs. *J. Oper. Res. Soc.* 51(5):602–608.

Nesterov Y, Nemirovski A (1993) *Interior-Point Polynomial Algorithms in Convex Programming* (Society for Industrial and Applied Mathematics, Philadelphia).

Pardo L (2006) *Statistical Inference Based on Divergence Measures* (Chapman & Hall/CRC, Boca Raton, FL).

Pelsser A (2010) Robustness, model uncertainty and pricing. Working paper, Maastricht University, Maastricht, The Netherlands.

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Oper. Res.* 56(1):188–203.

Reid M, Williamson R (2011) Information, divergence and risk for binary experiments. *J. Machine Learn.* 12:731–817.

Rockafellar RT (1970) *Convex Analysis* (Princeton University Press, Princeton, NJ).

Wagner MR (2010) Fully distribution-free profit maximization: The inventory management case. *Math. Oper. Res.* 35(4):728–741.

Wang Z, Glynn PW, Ye Y (2009) Likelihood robust optimization for data-driven newsvendor problems. Working paper, Stanford University, Stanford, CA.